Center for Quality and Productivity Improvement
UNIVERSITY OF WISCONSIN
610 Walnut Street
Madison, Wisconsin 53705

(608) 263–2520
(608) 263–1425 FAX

Report No. 139

Time Series Models for Forecasting Wastewater Treatment Plant Performance

P.M. Berthouex and George E. Box

February 1996

C Q P I
UNIVERSITY OF WISCONSIN

Report No. 139

Report Series
in Quality and Productivity

Time Series Models for Forecasting Wastewater Treatment Plant Performance

P.M. Berthouex

Professor of Civil and Environmental Engineering

University of Wisconsin
Madison, Wisconsin

George E. Box

Center for Quality and Productivity Improvement

Univeristy of Wisconsin
Madison, Wisconsin

ABSTRACT

This paper describes a time series modeling procedure that can be useful for calculating predictions, with confidence intervals, of effluent quality one to five days ahead, and it explains how these predictions can serve as an early warning of process upsets that will sometimes enable an operator to take preventive action. The time series model has the form of an exponentially weighted moving average (EWMA). The interpretation of the model is that response of the system can be predicted by deviations from the EWMA smoothed values of the predictor variables.

KEYWORDS: ARIMA models, IMA models , EWMA, wastewater treatment, time series models, moving average, forecasts, process upsets

CQPI Report No. 139, February 1996

Time Series Models for Forecasting Wastewater Treatment Plant Performance

P.M. Berthouex and George E. Box

This paper describes a time series modeling procedure that can be useful for calculating predictions, with confidence intervals, of effluent quality one to five days ahead, and it explains how these predictions can serve as an early warning of process upsets that will sometimes enable an operator to take preventive action. The time series model has the form of an exponentially weighted moving average (EWMA). The interpretation of the model is that response of the system can be predicted by deviations from the EWMA smoothed values of the predictor variables.

INTRODUCTION

Anticipatory control is desirable whenever (1) the results of key measurements on the final output are delivered too slowly to be useful and (2) when an upset, once established, cannot be quickly rectified (for example in a process that has a long residence time or great inertia). In environmental systems having these characteristics the operator needs to be able to foretell whether a particular set of operating conditions may result in a good product or a process upset. The term upset is used to describe a condition that is outside the normal pattern of operation. It may or may not be a violation of a legal constraint. It may involve an internal process stream for which no legal constraints exist, but which somehow influence the quality of the plant output.

The time series transfer function model to be described produces a numerical estimate of effluent quality one or more days ahead and a confidence interval for the forecasted value. This forecast can suggest a process upset in two ways: when the effluent BOD forecast is high and when the forecasted current value is unusually low relative to the observed value (as judged using the forecast confidence interval). This second warning results from a sudden increase in the process level that the model cannot predict, such as upsets caused a sudden equipment failure or a toxic spill). This information may help the operator decide whether some preventive action should be taken. Prescribing a suitable preventive action is a separate problem; the method presented does not provide a prescription.

TIMES SERIES MODEL BUILDING

The optimal forecasts of future values of a time series are determined by the nature of the stochastic model which describes that series. An important principle (the principle of parsimony) is that the model should adequately represent the data using as few parameters as possible. The main effort is directed to obtaining a suitable stochastic model for forecasting future values of the series. The stochastic models employed can be interpreted as descriptions of physical phenomena having the right general character, but they do not represent exact physical reality and are fitted to data empirically. This section is a brief review of the class of time series models, widely known as Box-Jenkins models (Box and Jenkins, 1976; Box, Jenkins and Reinsel, 1994).

Suppose that n consecutive observations $y_1, y_2, y_3, \ldots$ from a series are available and we wish to determine a suitable model. We denote values of a times series at equispaced times $t, t-1, t-2, \ldots$ by $y_t, y_{t-1}, \ldots$. Let B be the backward shift operator so that

$$By_t = y_{t-1}$$

and

$$(1-B)y_t = y_t - y_{t-1} = \nabla y_t \qquad (1)$$

A useful model to represent a nonstationary time series, such as occurs in many environmental applications, is the autoregressive integrated moving average model (ARIMA(p,d,q) model:

$$\phi_p(B)\nabla^d y_t = \theta_q(B)a_t \qquad (2)$$

The polynomial $\phi_p(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$ is the autoregressive operator and the polynomial $\theta_q(B) = 1 - \theta_1 B - \ldots - \theta_q B^q$ is the moving average operator. The reason for introducing both a finite moving average operator $\theta_q(B)$ and a finite autoregressive operator $\phi_p(B)$ is that a finite moving average is equivalent to an infinite autoregressive series, and vice versa, so that including both types of terms makes for parsimony. The difference operator

∇^d is introduced to allow for what may be called homogeneous non-stationarity. The a_t, a_{t-1}, a_{t-2} etc. are a sequence of random "shocks" which are random variables identically, independent, and approximately normally distributed.

Suppose the model for a single time series has been tentatively identified as of some specific form within the family $\phi(B)\nabla^d y_t = \theta(B)a_t$. Then, if for a sequence of observations y_t and for given $\phi(B)$ and $\theta(B)$, we can compute

$$a_t = \frac{\phi(B)}{\theta(B)}(1-B)^d y_t \qquad t = 1, 2 \ldots, n \quad (3)$$

Then the log likelihood for specific ϕ and θ is closely approximated by a linear function of the sum of the squares

$$S(\phi, \theta) = \sum_{i=1}^{n} a_t^2 \quad (4)$$

In practice the a_t resulting from a trial choice of the ϕ's and θ's are conveniently calculated recursively and approximate maximum likelihood estimates are obtained by minimizing $S(\phi, \theta)$ (Liu and Hudak 1992).

Two procedures for checking the tentative fitted model are (a) examination of residual a_t's and (b) overfitting. If the model is adequate and the number of fitted observations is not too small, then the estimated values $(\hat{\phi}, \hat{\theta})$ obtained from the fitted model will be sufficiently close to the values (ϕ, θ) and that the residuals $a_t(\hat{\phi}, \hat{\theta})$ will be uncorrelated deviations. When there is a particular elaboration of the tentatively identified model to be checked, a more sensitive check is provided by comparing the fits of the more elaborate and the less elaborate model (i.e. overfitting).

A considerable widening of the range of useful application of the model is achieved if the possibility of transformation is allowed. Thus $y_t^{(\lambda)}$ is substituted for y_t where $y_t^{(\lambda)}$ is some non-linear transformation of y_t involving one or more transformation parameters λ. A suitable transformation may be suggested by the physical situation or in some cases be estimated from the data. For example, if y_t were increasing at a rapid rate and the *percentage* fluctuation showed stability rather than the absolute fluctuation, it would be sensible to analyze the logarithm of y_t. When the transformation is to be estimated from the data one way to proceed is given Box and Cox (1964).

FORECASTING

Suppose a model of the form $\phi(B)\nabla^d y_t = \theta(B)a_t$ has been fitted. Then it may be used to make a minimum mean square error forecast $\hat{y}_t(l)$ of some future value $y_{t+l}(l \geq 1)$, which has the origin t and lead time l It is a linear function of current and previous observations $y_t, y_{t-1}, \ldots$ and can also be written as a linear function of current and previous shocks $a_t, a_{t-1}, \ldots$. In practice the forecasts are most easily calculated directly from the fitted stochastic difference equation model. Writing

$$\phi(B) = \phi(B)(1-B)^d = 1 - \phi B - \phi_2 B^2 - \ldots - \phi_{p+d} B^{p+d}$$

we can use the fitted model to express the observation to be forecast in terms of previous y's and a's as follows

$$y_{t+l} = \phi_1 y_{t+l-1} + \ldots + \phi_{p+d} y_{t+1-p-d} - \theta_1 a_{t+l-1} - \ldots - \theta_q a_{t+l-q} \quad (5)$$

The forecast at origin t for lead time l is found by taking conditional expectations at time t on both sides of the equation. The y_{t-j} (j = 0, 1, 2, ...), which have already happened at origin t, are left unchanged. The y_{t+j} (j = 1, 2, ...), which have not yet happened, are replaced by their forecasts $\hat{y}_t(l)$ at origin t. The a_{t-j} (j = 0, 1, 2,...), which have happened, are available from $y_{t-j} - \hat{y}_{t-j-1}$. The a_{t+j} (j = 1, 2, ...), which have not happened, are replaced by zeros. This provides the forecast $\hat{y}_t(l)$ entirely in terms of previous forecasts and known values of the series (Box and Jenkins 1976).

In practice we do not have the exact values of the model parameters but only estimates which we substitute instead. Experience shows that the forecast is rather robust to moderate changes in the parameter values and that the approximation is very good whenever the number of observations used to estimate the coefficients is reasonably large.

Forecasts of the kind so far discussed might be called *autoforecasts* since they use only their own past to predict the future. For example, a series of past values of effluent suspended solids might be used to identify, fit and check a time series ARIMA model in the manner described above and use it to forecast the future suspended solids values. Using this approach, the only variable in the model is effluent suspended solids and only the past values of suspended solids would be used to forecast future values. The autoforecast model does not entirely ignore the effect of the other variables on effluent suspended solids because they are partially taken account of by past values of the suspended solids themselves. Nevertheless, a better result frequently

can be obtained from a transfer function model which takes explicit account of other concomitant variables (such as raw sewage BOD_5, sewage flow rate, etc.). Another way to say this is that the forecast errors of the autoforecasts may be correlated with past and present values of the concomitant variables. In this case, the forecast errors contain information that was not fully extracted by the autoforecast model and improved forecasts should be possible if some function of the correlated variables is added to the autoforecast model. Adding these new variates creates a multivariate transfer function model in which the variables are considered in terms of inputs and outputs.

A particularly interesting situation relates to forecasting effluent BOD_5 values. Unfortunately, there is a necessary delay of five days in knowing a value for this indicator of the effluent quality. An interesting question therefore, is how well can a transfer function model, which uses only information available today, forecast BOD_5 values which will in fact not be available for five further days, and how well can such a transfer function model predict into the future.

TRANSFER FUNCTION MODELS

Transfer function models incorporate concomitant variables via linear dynamic models expressed as linear difference equations of the same general form as the stochastic models already discussed. Suppose that values are available on predictor variables $x_{1t}, x_{2t}, \ldots$, (e.g. flow, suspended solids) and an output variable y_t (effluent BOD_5). Then the dynamic characteristics of such systems can often be approximated parsimoniously by linear difference equations of the form

$$\phi(B)y_t = \theta(B)x_t$$

or

$$y_t = \frac{\theta(B)}{\phi(B)}x_t \qquad (6)$$

where $\theta(B) = \left(\theta_0 - \theta_1 B - \ldots - \theta_p B^p\right)B^b$ and $\phi(B) = \phi_0 - \phi_1 B - \ldots - \phi_q B^q$. If there are several inputs and we allow for a general ARIMA model, we can write

$$y_t = \sum_{i=1}^{k} \frac{\theta_i(B)}{\phi_i(B)}x_{it} + \frac{\theta(B)}{\phi(B)}\alpha_t \qquad (7)$$

where α_t are independent and identically distributed normal series which are not necessarily the same as the shock series, a_t, of the univariate model. In general $\theta_i(B) \neq \theta(B)$ and $\phi_i(B) \neq \phi(B)$. Methods for identifying, fitting, and checking models of this kind closely parallel those of the univariate models and are fully described in Box and Jenkins (1976).

ACTIVATED SLUDGE TREATMENT PROCESS

Models were identified for three treatment plants using data obtained in an earlier study (Berthouex and Fan 1986). These plants are identified as A, B and C. The variables available in the model building are listed below. Input or process variables for the activated sludge treatment process are denoted by x and output variables by y. Effluent suspended solids can be either an output (y_2) or a process variable (x_{14}), depending upon the model under consideration. The notation used is:

y_1	=	effluent BOD_5 (mg/L)
y_2	=	effluent suspended solids (mg/L) when used as a process output
x_1	=	raw sewage flow rate (MGD)
x_2	=	raw sewage temperature
x_3	=	raw sewage BOD_5 (mg/L)
x_4	=	raw sewage suspended solids (mg/L)
x_5	=	sludge age (day)
x_6	=	food to microorganism ratio (day^{-1})
x_7	=	MLSS (mg/L)
x_8	=	dissolved oxygen (DO) concentration, (mg/L)
x_9	=	settleability (%)
x_{10}	=	waste sludge (1000 gal)
x_{11}	=	return sludge ratio (%)

x_{12} = sludge volume index (mL/g)

x_{13} = dimensionless integer value indicating DO level (x_8) as follows:

$x_{13} = 0$ if $x_8 < 0.75$
and $x_{13} = 1$ if $x_8 \geq 0.75$

x_{14} = effluent suspended solids (mg/L) when used as a process variable

The variable x_{13} was constructed from variable x_8 to represent DO on a binary rather than a continuous scale, reasoning that low DO was harmful but having DO above some critical level probably provided no additional benefit (the critical level was arbitrarily selected to be 0.75 mg/L DO). Effluent SS (x_{13}) is denoted as y_2 when it is regarded as a dependent variable in a transfer function model.

Frequently the linear time series and transfer function models are best fitted to transformed data values. In this project, a log transformation was applied to both the input and output variables. Upper case letters are used to denote the logged values, such that $X_i = ln\,(x_i)$. In support of this transformation it should be noted that for many of the variables a proportional rather than a linear relationship is expected between outputs and inputs. Furthermore, such proportional relations are implied by a predisposition toward using ratios. Thus the F/M ratio and recirculation ratio are sometimes regarded as natural process variables. On a log scale the ratios simplify to linear functions of the log-transformed variables. For example, the F/M ratio, x_{6t}, is proportional to $x_{1t}x_{5t}/x_{7t}$ which can be written as $\ln(x_{6t}) = \ln(x_{1t}) + \ln(x_{3t}) - \ln(x_{7t})$, or as $X_{6t} = x_{1t} + x_{3t} - x_{7t}$.

The transfer function model was found to be a function of differences between log-transformed values on consecutive days, which is, correspondingly, the ratio of consecutive days on the original scale of measurement. An advantage of using logarithms is that this ratio is dimensionless, so units of variables are unimportant and the user does not need to worry about consistency or conversion of units. Also, errors in the output variables (e.g., BOD_5 and suspended solids) tend to be proportional to the magnitude of the variable, rather than being constant at all levels, so that the logged variables have constant variance, which is a condition for obtaining optimal estimates when using the method of least squares. Since the standard deviation of the logarithm is roughly proportional to the coefficient of variation, $SD[ln(y)] \alpha\, \sigma / \mu$, a given reduction in the standard deviation of some variable $Y = ln\,(y)$ allows the approximate percentage reduction in the variation of y to be easily calculated.

UNIVARIATE MODELS

Some of the effluent quality data are shown by Figure 1. Considerable univariate model identification work on the effluent BOD_5 and SS series indicated that a simple IMA(0,1,1) model of the form

$$(1-B)Y_t = (1-\theta B)a_t \qquad (8)$$

provided the best fit for all three treatment plants (Plant A, Plant B, and Plant C). This model is consistent with the form of transfer function models that will be presented in a later section. Table 1 summarizes the univariate models for the three plants. The standard deviations of the estimated parameters are small relative to the parameter values. The forecasting effectiveness of the models will be presented later.

For convenience, the notation for forecast functions are developed here. The IMA(0,1,1) model for time $t+1$ may be written as

$$Y_{t+1} = \tilde{Y}_t(\theta) + a_t \qquad (9)$$

where $\tilde{Y}_t(\theta)$ is an *exponentially weighted moving average* (EWMA) calculated at time t from the current and previous values of Y and has often been used as a forecast of Y_{t+1}. Thus

$$\tilde{Y}_t(\theta) = \sum_{j=1}^{\infty} (1-\theta)\theta^{j-1}Y_{t-j} \qquad (10)$$

where the weights $1-\theta, (1-\theta)\theta, (1-\theta)(\theta)^2, \ldots$ decrease exponentially as the data go back in time from t, t–1, t–2, and so on. For example, the IMA(0,1,1) model

$(1-B)Y_{t+1} = (1-0.5B)a_t$ leads to

$$\tilde{Y}_t(\theta) = 0.5Y_t + 0.25Y_{t-1} + 0.125Y_{t-2} + \ldots .$$

IMPORTANCE OF ACCOUNTING FOR SERIAL DEPENDENCE OF DATA AND SYSTEM INERTIA

A particular kind of multivariate modeling uses regression analysis. This usually consists of fitting a linear equation

$$Y_t = b_0 + b_1X_{1t} + b_2X_{2t} + \ldots + b_kX_{kt} + e_t.$$

For purpose of modeling wastewater treatment plant data such a model may be totally invalid and also misleading. Problems arise because of (1) autocorrelated data, (2) dependence between input (predictor) variables, (3) dynamic relations between input variables and out variables, (4) and parsimonious use of variables.

Ordinary least squares used in regression analysis is only appropriate for uncorrelated data. For data from a waste treatment system, proper allowance must be made for serial dependence that is likely to occur between successive observations. This is done by the use of time series models.

Dependence between potential predictor variables is especially common when the data are obtained from the routine operation of a process, unlike data from a well-designed experiment in which the input variables can be manipulated according to some balanced experimental design. As an example of dependence between input variables, Draper and Smith (1981) mention that ".... the operating policy, for example 'if X_1 goes high, reduce X_2 to compensate' ,often causes large correlation between [variables]. This make it impossible to see if changes in Y are associated with X_1, or X_2, or both." The consequence is that a variable which is known to be highly important may appear insignificant because the operator, knowing its importance, has carefully controlled and confined it within a narrow range so it has no chance to affect the process output. As a consequence, a set of data may be adequately fitted by several alternate transfer function models that involve different variables because of correlation between the *X* variables.

Ordinary regression analysis does not take account of system inertia such as occurs in treatment plants. Thus a change in level of an input may not be transmitted immediately to the output, but is gradually felt. Dynamic models can allow for this behavior.

Parsimony is used to mean having the fewest number of parameters necessary to provide an adequate model. It might at first be thought that a larger number of parameters would provide greater accuracy. In fact, the addition of unnecessary parameters results in less predictive accuracy. For example, it can be shown that in the case of linear models fitted by least squares, the average variance of the predicted value $\hat{y}_i$ (i = 1, 2, 3, ... n) at each of the observed conditions is $V(\hat{y}_i) = p\sigma^2/n$, where p is the number of parameters and n is the number of observations. Thus if two parameters are used when one would suffice, on average the predictive variance is doubled.

All these difficulties can arise when working with historical (happenstance) data of the kind that is available from wastewater treatment plants. Even for the models considered here, which do take account of dynamic relations and dependence between data, it is important to establish what one can expect to achieve. Happenstance data, such as those analyzed here, have specific limitations. Two different purposes for which the model may be used are: (a) to forecast future values and (b) to establish a causal relationship. It is known that although happenstance data can produce models which can forecast satisfactorily, they cannot be relied upon to unequivocally show causal relationships, although results may be helpful in suggesting possible relations or in helping to verify expectations.

TRANSFER FUNCTION MODELS FOR BOD AND SS

For the present data, all three treatment plants showed strong correlations between some of the potential predictor variables. For example, raw sewage flow, raw sewage SS, and raw sewage BOD provide much the same information. Thus one variable in a model can often do as well as a combination of several variables. It also makes it possible to construct a collection of models that differ in appearance but are nearly identical in forecasting accuracy. This can cause considerable confusion if one tries to interpret the model as a cause-and-effect description of the process. In this study, our main interest is only to obtain an accurate forecast.

The model building process was started assuming that the BOD values were known at the same time as the other variables. Later, other models were developed that did not use influent BOD as a predictor variable. These models with and without BOD will be compared to show that the models with BOD were only marginally better than those that omitted BOD. These comparisons provide some

measure of the benefit that would obtain if a quickly-measured BOD surrogate, like COD or TOC, were available.

Model Building and Simplification. Considerable effort went into model building to select useful independent variables and to develop the basic model structure. A key step was discovering that all the time series of all variables could be modeled as first-order moving averages. A rough idea of the model building process can be gained by considering this model for effluent BOD_5 at Plant B, which was obtained by fitting data for one full year (n = 365):

$$\underset{(0.03)}{(1-0.91B)}Y_{t+1} = \underset{(0.06)}{(0.19-0.07B)}X_{3t} + \underset{(0.04)}{(0.06-0.09B)}X_{4t} + \underset{(0.05)}{(0.37-0.34B)}X_{6t} - \underset{(0.07)}{(0.08-0.13B)}X_{9t} - \underset{(0.03)}{(0.16-0.15B)}Y_{2t} + \underset{(0.06)}{(1-0.63B)}\alpha_{t+1} \quad (11)$$

with $\hat{\sigma}_\alpha = 0.272$ ($\hat{\sigma}_\alpha$ is on the natural log scale). The small numbers in parentheses are the estimated standard deviations of the fitted parameters.

This model is already more simple than many were investigated, but it still contains 12 parameters and further simplification would be advantageous. Note that the term $(1 - 0.91\text{B})Y_{1t}$ is close to being $(1-\text{B})Y_{1t}$, which suggests that it might be worthwhile to fit a model having $(1-\text{B})Y_{1t}$ instead of $(1-\theta\text{B})Y_{1t}$. This simplification was supported by the refitting. The result is as model expressed in terms of simple differences of subsequent Y values.

Two other characteristics of this model allowed further simplification. Each term on the right hand side has the general form $(a_i - b_i\,\text{B})X_i$, and the values of a_i and b_i in each term have nearly the same value, in fact they are not significantly different. Taking the term $(0.06 - 0.09\text{B})Y_{4t}$ as a specific example, the confidence intervals of the coefficients 0.06 and 0.09 overlap and support an assumption that their values might be equal. Applying the assumption that $b_i = a_i$ simplifies $(a_i - b_i\,\text{B})X_i$ to $b_i(1-\text{B})\,X_{it}$ and the number of parameters in the model is reduced by roughly half.

These simplifications led to a model of the simple form

$$(1-\text{B})\text{Y}_{t+1} = \sum_{i=1}^{k} \text{b}_i(1-\text{B})\text{X}_{it} + (1-\theta\text{B})\alpha_{t+1} \quad (12)$$

where the Y_{t+1} can be either BOD or SS. An example of the refitted model of this form for BOD at Plant B is:

$$\underset{(0.06)}{(1-B)}Y_{t+1} = \underset{(0.04)}{0.18}(1-B)X_{3t} + \underset{(0.05)}{0.07}(1-B)X_{4t} + \underset{(0.06)}{0.38}(1-B)X_{6t} - \underset{(0.03)}{0.13}(1-B)X_{9t} - \underset{(0.04)}{0.15}(1-B)X_{14t} + \underset{(0.04)}{(1-0.75B)}\alpha_{t+1} \quad (13)$$

which has $\hat{\sigma}_a$=0.277, essentially the same as the $\hat{\sigma}_a$=0.272 obtained from fitting the more complicated Equation 11.

We shall see later that a model of this form has the great advantage of simplicity and ease of interpretation, and it will give forecasts which are equivalent, in practical terms, to more complicated models. The BOD and SS models need not use the same X variables, and of course the parameter values will be different.

This model contains no seasonal terms, even though one might expect a weekly cycle in the raw sewage flow and strength. Any 7-day seasonality that was identified in the influent was weak and it seemed to be blended out in the treatment process because the effluent quality did not display this seasonality. Putting a 7-day seasonal difference operator (1-B^7) into the model produced slightly worse results than the nonseasonal model.

Summary of the Fitted Models. The results of fitting the effluent BOD transfer function models to the three treatment plants are summarized in Table 2. Influent BOD and variables computed from it (F/M ratio) were not allowed to enter the model. The standard deviations of the parameters are given in parentheses under the estimated values. The estimated parameter values are $\hat{b}_i$. Estimates are given only for the variables that should be included in the model, based on the standard deviations of the estimates. For example $\hat{b}_1 = 0.52$ is the coefficient for the term $0.52(1 - \text{B})X_{1t}$. The standard deviation of this parameter estimate is 0.15, and the approximate confidence interval would be $b_1 = 0.52 \pm 2(0.15)$. $\hat{\sigma}_a$ is the estimated standard deviation of the residuals; this is also the standard deviation of the one-step ahead forecasts of the model.

Table 3 lists the estimated parameter values for the effluent suspended solids transfer function models. Again, influent BOD and F/M were not used as predictors in these models.

INTERPRETATION OF THE TRANSFER FUNCTION MODEL

In this section we show that the deviation of outputs from a smoothed (EWMA) series can be interpreted and might be modeled as a function of the deviation of inputs from their smoothed (EWMA) series. This kind of a model has an intuitive appeal for modeling a treatment plant if one imagines that the propulsive forces for change in the system's effluent quality are changes in the inputs *relative to the current input levels.* That is to say, whatever the current levels of the inputs and outputs, the change relative to the levels is more important than the levels themselves.

The transfer function model may be written in the form

$$\frac{(1-B)}{(1-\theta B)} Y_{t+1} = \sum_{i=1}^{k} b_i \frac{(1-B)}{(1-\theta B)} X_{it} + \alpha_t \qquad (14)$$

Now

$$\frac{(1-B)}{(1-\theta B)} Y_{t+1} = Y_{t+1} - \tilde{Y}_t = e_{t+1}$$

and

$$\frac{(1-B)}{(1-\theta B)} X_{it} = X_{it} - \tilde{X}_{i,t-1} = e_{it} \qquad (15)$$

Thus the quantity $Y_{t+1} - \tilde{Y}_t = e_{t+1}$ is the deviation of Y_{t+1} from its EWMA (smoothed) level, $\tilde{Y}_t$. Likewise, each term of the form $X_{it} - \tilde{X}_{i,t-1} = e_{it}$ represents the deviation of ${}_tX_{it}$ from its EWMA reference level, $\tilde{X}_{i,t-1}$, and the model may be written in its equivalent form:

$$Y_{t+1} - \tilde{Y}_t = \sum_{i=1}^{k} b_i (X_{it} - \tilde{X}_{it-1}) + \alpha_{t+1} \qquad (16)$$

or

$$Y_{t+1} = \tilde{Y}_t + \sum_{i=1}^{k} b_i e_{it} + \alpha_{t+1} \qquad (17)$$

The forecast $\hat{Y}_t(1)$ one-step ahead from origin t may now be obtained by taking the conditional expectation at time t of Y_{t+1} in equation 17.

$$\hat{Y}_t(1) = \tilde{Y}_t + \sum_{i=1}^{k} b_i e_{it} \qquad (18)$$

This form of the model shows that the deviations of outputs from an EWMA-smoothed series are modeled as a function of the deviations of inputs from the EWMA-smoothed series of each X variable.

When $i > 1$, then the minimum mean square error of the forecast $\hat{Y}_t(1)$ is also the conditional expectation from origin t of Y_{t+i}. The error of the one-step ahead forecast is obtained by subtracting Eq. 18 from Eq. 17, thus

$$Y_{t+1} - \hat{Y}_t(1) = \alpha_{t+1} = e_{t+1} - \sum_{i=1}^{k} b_i e_{it} \qquad (19)$$

This can be thought of as an augmented version of the autoforecast model $Y_{t+1} = \tilde{Y}_t + e_{t+1}$. A linear combination

$$\sum_{i=1}^{k} b_i e_{it}$$

of the deviations $e_{it} = X_{it} - \tilde{X}_{i,t-1}$ that occurred in the concomitant variables from the smoothed input variables (i.e., the $\tilde{X}_{i,t-1}$) is used to partially forecast the error e_{t+1} of the autoforecast model:

$$e_{t+1} = \sum_{i=1}^{k} b_i e_{it} + \alpha_{t+1} \qquad (20)$$

This allows Eq. 17 to be written as

$$\hat{Y}_t(1) = \tilde{Y}_t + \sum_{i=1}^{k} b_i \left(X_{it} - \tilde{X}_{i,t-1} \right) \qquad (21)$$

This shows that the forecast $\hat{Y}_t(1)$ is a function of the deviations $e_{it} = X_{it} - \tilde{X}_{i,t-1}$ of the inputs X_{it} of the smoothed values (trend line). This kind of a model has an intuitive appeal for modeling a process if one imagines that the propulsive forces for change in the system's effluent quality are changes in the inputs *relative to the current input levels.* That is to say, the deviation from the value that was expected in the output of any past values must be modified by the deviations of the input values from where they would have been expected to be.

This idea is shown graphically in Figure 2. Plot (a) shows the exponentially decaying weights of an EWMA. Plot (b) shows system outputs; the dots are the observed value and the line following the trend is the EWMA smoothed value computed using the weights shown in (a). Plot (c) shows the observations of one of the inputs and the EWMA smoothed series of a predictor variable *X*. Plots (d) and (e) show the deviations.

The forecast $\tilde{Y}_t$ contains information that is also contained in the input variables taken up to time t, but in addition it even contains information about analytical and sampling errors in the inputs.

IMPLEMENTING THE TRANSFER FUNCTION MODEL FORECAST

The forecast model, equation 21, can be used directly for effluent suspend solids forecasts. To forecast effluent BOD_5 it is modified to

$$\hat{Y}(1) = \tilde{Y}_{t-5} + \sum_{i=1}^{k} b_i \left(X_{it} - \tilde{X}_{i,t-1} \right) \quad (22)$$

to account for the fact that the most recent BOD measurement is five days old.

Equation 20 may be written as

$$\hat{Y}_t = \tilde{Y}_t + Z_t - \tilde{Z}_{t-1} \quad (23)$$

where $Z_t = \sum_{i=1}^{k} b_i X_{it}$ and

$$\tilde{Z}_{t-1} = \sum_{i=1}^{k} b_i \tilde{X}_{i,t-1}$$

The forecast is accomplished, bearing in mind that $\tilde{Y}_t$ and $\tilde{Z}_t$ may be readily updated, from

$$\tilde{Z}_t = \theta \tilde{Z}_{t-1} + (1-\theta) Z_{t-1}$$

and

$$\tilde{Y}_t = \theta \tilde{Y}_{t-1} + (1-\theta) Y_{t-1} \quad (24)$$

An alternative expression of the forecasting function is Eq. 18,

$$\hat{Y}_t(1) = \tilde{Y}_t + \sum_{i=1}^{k} b_i e_{it} \quad (25)$$

and operating on both sides of the equation with $(1-\theta B)$ which gives

$$(1-\theta B)\hat{Y}_t(1) = (1-\theta B)\tilde{Y}_t + \sum_{i=1}^{k} b_i (1-\theta B) e_{it} \quad (26)$$

Now, $(1-\theta B)\tilde{Y}_t = (1-\theta) Y_t$ and

$$(1-\theta B) e_{it} = X_{it} - X_{i,t-1}, \text{ and so}$$

$$\hat{Y}_t = \theta \hat{Y}_{t-1} + (1-\theta) Y_t + \sum b_i \left(X_{it-Xi,t-1} \right) \quad (27)$$

This model can be used for suspended solids, but it needs to be modified slightly for BOD forecasting to account for the fact that the most recent BOD measurement is five days old. In that case we need

$$\hat{Y}_t = \theta \hat{Y}_{t-1} + (1-\theta) Y_{t-5} + \sum b_i \left(X_{it} - X_{i,t-1} \right) \quad (28)$$

We will construct an example from the model for Plant A. An equivalent form that may be more convenient for actual forecasting is

$$\hat{Y}_t(1) = 0.42 \hat{Y}_{t-1} + (1-0.42) Y_{t-5} + 0.61(1-B) X_{1t} - 0.06(1-B) X_{8t} + 0.38(1-B) X_{14t} \quad (29)$$

where $\hat{Y}_{t-1}$ is the forecast of the previous value and Y_{t-5} is the latest available measured value.

The implementation of Equation 22 for effluent BOD in Plant A becomes,

$$\hat{Y}(1) = \tilde{Y}_{t-5} + 0.61 \left(X_{1,t} - \tilde{X}_{1,t-1} \right) - 0.06 \left(X_{8,t} - \tilde{X}_{8,t-1} \right) + 0.38 \left(X_{14,t} - \tilde{X}_{14,t-1} \right) \quad (30)$$

The EWMA, $\tilde{Y}_t$, calculated using $\theta = 0.42$, which gives the weights as shown below

$$\tilde{Y}_{t-5} = 0.58 Y_{t-5} + 0.2436 Y_{t-6} + 0.1023 Y_{t-7} + 0.043 Y_{t-8} + 0.018 Y_{t-9} \quad (31)$$

Terms beyond Y_{t-9} are negligible because the weights have decayed to almost zero.

FORECASTING ACCURACY

It is to be expected, however, that the transfer function which uses data on the process inputs would give more accurate forecasts than a univariate model. However, to the extent that the process output is a function of process inputs, the output actually does incorporate information about the inputs even when these inputs are explicitly shown in the model, and this information is indirectly incorporated in the univariate model forecast. The accuracy of the univariate model (UVM) and transfer function models (TFM) was compared by forecasting a series of data that was not used to construct the models.

We will use the BOD models for Plant A as a specific example. The IMA (0,1,1) univariate model for Plant A is $(1-B)Y_{t+1} = (1-0.34B)a_{t+1}$, or $Y_t = \tilde{Y}_{t-1} + a_t$, where $\tilde{Y}_{t-1}$ is computed using $\theta = 0.34$ for BOD. To use this we must assume that BOD values, or some equivalent measurement are available. The corresponding forecast function is $\hat{Y}_{t+1} = \tilde{Y}_t$, where the EWMA $\tilde{Y}_t$ is computed $\theta = 0.34$.

Figure 3 shows the one-step ahead forecasts of the UVM and the TFM for plant A. Similar results were obtained for all three plants. The univariate model does rather well considering its simplicity, but it is only useful if such a model could be constructed using a surrogate for BOD. The more complicated transfer function should do better, and it does, even though it does not use influent BOD as a predictor,

and it uses only the available effluent BOD values (those from 5 days earlier). Incorporating the other predictor variable makes up the apparent weakness due to lack of current BOD values.

The forecasting accuracy of the two models may be compared quantitatively using the value of $\hat{\sigma}_a$ (on the natural log scale). The values of $\hat{\sigma}_a$ of the TFM were about 45% less than $\hat{\sigma}_a$ of the univariate model (from Tables 1 and 2, $\hat{\sigma}_a$ = 0.24 for the TFM compared with $\hat{\sigma}_a$ = 0.32 for the UVM). The amount of information is proportional to the inverse of the variance. Thus, the percentage improvement supplied by the TFM is

$$\left(\left(.032^2/0.24^2\right)-1\right)100 = 80\%.$$

Another criterion for comparing the forecasts is the mean square error (MSE), which is an estimate of the variance of the l-step ahead forecast. A model with a small MSE gives more accurate forecasts than a model with large MSE. An equivalent but more convenient measure, since it is measured on the same scale as the variable being predicted, is the square root of the mean square error, $\sqrt{\text{MSE}}$, which is an estimate of the standard deviation of the forecast; that is $\hat{\sigma}_a = \sqrt{\text{MSE}}$ for the fitted models. Since the models were developed in terms of natural logarithms of the dependent and independent variables, the $\sqrt{\text{MSE}}$ is also on the natural logarithmic scale.

The MSE is defined as the average square of residuals of the forecasts, where the residual is the difference between forecasted and the observed values.

$$\text{MSE} = \frac{1}{n}\sum_{i=t_0}^{t_0+n}\left(Y_i - \hat{Y}_i\right)^2 \qquad (32)$$

The MSE is computed over a period of $n+1$ forecasts, running from day t_0 to day t_0+n. To compare the forecasting accuracy of the models, the models were developed using data up to day 335 (days measured with t=1 as January 1), and the fitted model was used to make one-step to five-step ahead forecasts for the 30 days from day 336 to day 365. For this period, 29 one-step ahead forecasts can be made, 28 two-step ahead forecasts, and 25 five-step ahead forecasts can be made. The $\sqrt{\text{MSE}}$ values for the forecasts from days 336 to 365 for each of the three forecasting models at each treatment plant are given in Table 4. Note that the values of $\sqrt{\text{MSE}}$ in Table 4 is computed by comparing the forecasts and observed values for the forecasting period after day 335, whereas the values of $\hat{\sigma}_a$ in the other tables are for the model fitted to data prior to day 335. Nevertheless, if the model is adequate, the values should be similar, and they are in reasonable agreement.

The approximate $(1-\alpha)$ confidence interval of the forecasts is $\hat{Y}_t \pm t_{v,\alpha/2}\sqrt{\text{MSE}}$, where $t_{\upsilon,\alpha/2}$ is the t statistic with υ degrees of freedom. For the approximate 95% confidence interval, $t_{\upsilon,\alpha/2} = 2 = 2$. The confidence interval increases as the length of the forecasting period increases, but the increase is small as the forecasting period goes from one day to three days.

The confidence intervals are symmetric on the logarithmic scale, but not on the scale of the original measurements. If we predicted $\hat{Y}_t$ = 2.5, then $\hat{y}_t$ = exp(2.5) = 12.1. Suppose that the 95% confidence interval is $\hat{Y}_t \pm 2\sqrt{\text{MSE}}$ = 2.5 ± 2(0.17) = 2.5 ± 0.34, giving upper and lower confidence limits of 2.84 and 2.16. When converted to the original measurement scale the confidence limits are exp(2.84) = 17.1 and exp(2.16) = 8.7. Instead of a 95% confidence it may be more meaningful to an operator to work with the 50% confidence interval, which would be computed using $t_{\upsilon,\alpha/2}$ = 0.67.

COULD WE FORECAST BETTER IF INFLUENT BOD WERE KNOWN?

To those of us accustomed to measuring performance in terms of BOD it might seem that having knowledge of the BOD would improve our ability to forecast. Out of curiosity about how the value of BOD as a predictor variable, transfer functions were also developed that treated BOD as though it were known at the same time as the other variables. If this model were greatly superior to the TFM it would indicate that measuring a surrogate of BOD, and developing models the incorporate this surrogate, would be worthwhile. The results showed that putting BOD into the model was of little value, presumably because influent BOD was so strongly related with other variables, like flow and SS, that these variables already serve as surrogates for BOD. Likewise, the effluent suspended solids seemed to serve as a useful surrogate for effluent BOD. This result will not surprise operators who after all, do operate their treatment plants without contemporaneous data on BOD.

CONCLUSIONS

The objective was to develop a model for effluent quality that could be used to forecast BOD and SS values beyond those known from direct measurements. Some progress has been made toward

this goal. The transfer function model presented here cannot be expected to fit data from all treatment plants, but the general structure of the model should provide a way to begin when one is faced with modeling a new data set.

The model is based on first differences of the dependent and independent log-transformed variables. The log-transformation helps to achieve uniform variance and the first-order difference accounts for nonstationarity in the inputs and outputs. The model amounts to smoothing the input and output series as exponentially weighted moving averages (EWMAs) and then fitting the deviations from these averages. The smoothing parameters of the moving averages are determined by nonlinear least squares. The model interpretation is that changes in the output relative to the current level of the moving average, and this change is a linear function of differences between the current levels and the moving averages of the independent variables.

ACKNOWLEDGMENTS

This research represents a portion of work done under NSF grant BCS-9113124. Thanks are due to Chang-I Wu and K. T. Wu for their considerable contributions.

P. M. Berthouex is Professor of civil and Environmental Engineering, University of Wisconsin-Madison. George Box is Research Director, Center for Quality and Productivity Improvement, University of Wisconsin-Madison.

REFERENCES

Berthouex, P. M. and R. Fan (1986). "Evaluation of Treatment Plant Performance: Cause, Frequency and Duration of Upsets," *J. Water Poll. Cont. Fed.*, 58, 368-375

Box, G. E. P. and D. R. Cox (1964). "An analysis of transformations," *J. Royal Stat. Soc.*, Series B, 26, no. 2, 211-252.

Box, G. E. P., W. G. Hunter and J. S. Hunter (1978). *Statistics for Experimenters: an introduction to design, data analysis, and model building.* Wiley, New York.

Box, G. E. P., and G. M. Jenkins (1976). *Time Series Analysis - Forecasting and Control,* Holden Day, San Francisco, CA.

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel (1994). *Time Series Analysis: Forecasting and Control*, Englewood Cliffs, NJ, Prentice-Hall.

Draper, N. R. and H. Smith (1964). *Applied Regression Analysis*, John Wiley & Sons, NY

Liu, L. M. and G. B. Hudak (1992). *Forecasting and Time Series Analysis Using the SCA Statistical System,* DeKalb, IL, Scientific Computing Associates.