Forecasting influent flow rate and composition with occasional data for supervisory management system by time series model

J.R. Kim*, J.H. Ko, J.H. Im*, S.H. Lee*, S.H. Kim*, C.W. Kim* and T.J. Park***

*Dept of Environmental Engineering, Pusan National University, Busan, 609-735, Korea
(E-mail: *jong93@pusan.ac.kr*; *hoonyijh@pusan.ac.kr*; *reallife2@pusan.ac.kr*; *imsangh@pusan.ac.kr*; *cwkim@pusan.ac.kr*; *taejoo@pusan.ac.kr*)

**R&D Centre POSCO E&C, Kyunggi-Do, 445-810, Korea (E-mail: *joohko@poscoenc.com*)

Water Science & Technology **Vol 53 No 4–5 pp 185–192** © IWA Publishing 2006

Abstract The information on the incoming load to wastewater treatment plants is not often available to apply modelling for evaluating the effect of control actions on a full-scale plant. In this paper, a time series model was developed to forecast flow rate, COD, NH_4^+-N and PO_4^{3-}-P in influent by using 250 days data of field plant operation data. The data for 150 days and 100 days were used for model development and model validation, respectively. The missing data were interpolated by the spline method and the time series model. Three different methods were proposed for model development: one model and one-step to seven-step ahead forecasting (Method 1); seven models and one-step-ahead forecasting (Method 2); and one model and one-step-ahead forecasting (Method 3). Method 3 featured only one-step-ahead forecasting that could avoid the accumulated error and give simple estimation of coefficients. Therefore, Method 3 was the reliable approach to developing the time series model for the purpose of this research.

Keywords ARIMA; influent forecasting; supervisory management system; time series model

Introduction

Information about influent and effluent are necessary economically to manage wastewater treatment plants (WWTPs) while maintaining stable effluent quality. In reality, operators usually manage WWTPs by expert knowledge from previous years' experience in real fields, but even well-trained operators cannot deal with a rapid increase in the number of WWTPs. In order to deal with this problem, research into instrumentation, control and automation (ICA) for wastewater treatment has been performed for the last few decades. Various control strategies have been suggested for WWTPs and some of them have been applied to real-scale treatment plants. Good examples are SMAC (smart control of wastewater systems; Krüger, 2004) and TELEMAC (telemonitoring and advanced telecontrol of high yield wastewater treatment plants; Bernard *et al.*, 2004). Information on the incoming load to WWTPs is not often available to apply to modelling for evaluating the effect of control actions on the full-scale plant (Krüger, 2004).

Once influent concentrations are provided, the mathematical models can be used as a tool for estimating optimal management strategy, design of the automatic controller and several feeding strategies (Bernard *et al.*, 2004). A time series model was suggested to forecast influent data as mentioned above, as, generally, a variation in wastewater flow rate and concentrations of components occurs daily, monthly and yearly, in certain patterns that can possibly be described by a time series model. Hiraoka and Tsumura (1989) developed a model for control of MLSS, effluent organics and suspended solids using a multivariate autoregressive model. Naghdy and Helliwell (1989) used a time series model showing excellent prediction performance for influent flow rate and NH_4^+-N hourly loading rate. Even though there was a time delay between predicted and measured

doi: 10.2166/wst.2006.123

values for abnormal data patterns, it was proved that daily variation of influent flow rate and NH_4^+-N loading rate in the target wastewater plant represented a constant pattern. Novotny *et al.* (1991) presented a time series model for MLSS derived partly from causal relationships, with influent BOD and suspended solids. Van Dongen and Geuens (1998) also used influent filtered COD, suspended solid and food-to-microorganisms ratio to control nitrate concentration and aerating costs in an anoxic reactor.

The purpose of this paper is to propose a time series model that has a simple structure with low prediction errors, capable of forecasting influent flow rate, COD, NH_4^+-N, PO_4^{3-}-P and temperature for 7 days in advance. Equal interval data are required for the time series model. This research interpolated missing data to generate daily data based on 2–4 day interval data from WWTPs, which led to the development of the time series model.

Materials and methods

Approaches for time series model development

Three feasible methods were proposed for developing a simple forecasting model according to the number of models and forecasting steps.

Method 1:One model and one-step to seven-step-ahead forecasting. Each time series model was constructed for influent flow rate, COD, NH_4^+-N, PO_4^{3-}-P and temperature for one-day to seven-day-ahead forecasting. Model identification, estimation and diagnosis were performed following the procedure suggested by Box and Jenkins (1976) and SPSS V.11 for numeric calculation.

Method 2: Seven models and one-step-ahead forecasting. In the case of Method 1, because previous forecasted values were used to predict next day values after two-day-ahead forecasting, the error generally accumulated with the increase of step numbers. In order to overcome this barrier, the “seven different models with one-step-ahead forecasting” approach was employed. These models used one-day previous values for one-day-ahead forecasting and seven-day previous values for 7-day-ahead forecasting. Seven models had the identical model structure (linear regressive model) and different parameter values that were estimated for objective function of minimising root mean square errors (RMSE).

Method 3: One model and one-step-ahead forecasting. Method 2 was the opposite approach for our research purpose, “development of the simplest influent prediction model”. Therefore, a time series model having one model and minimum error, as in Method 3, was used. It used a similar forecasting approach to Method 2. One model was proposed using the same parameter values as those of the seven models in Method 2. Method 3 also used the objective function of minimising RMSE for parameter estimation.

Procedure

Field plant data. Enough data for model development were obtained from D-city field wastewater plant which operated as a five-stage step-feed EBPR (*fs* EBPR) process for 250 days from July 2002 to March 2003 (Lee *et al.*, 2005). Table 1 shows the influent characteristics of the D-city wastewater plant. Data for 150 days were utilised for model development and the remainder were used for model validation.

Interpolation by spline method. The influent data were not measured daily. The data were measured at intervals of 2–4 days on average and 10 days at a maximum. Missing

Table 1 Influent characteristics of D-city wastewater plant

	Range	Average
Flow rate (m^3/day)	10,573–24,872	18,288
COD (mg/L)	60.2–428.0	230.7
NH_4^+-N (mg/L)	7.6–31.7	23.8
PO_4^{3-}-P (mg/L)	0.6–3.3	2.1
Temperature (°C)	12.5–26.7	19.9

values were interpolated with the spline method for application to the time series analysis.

Re-interpolation with the time series model. The spline method is useful for interpolation. However, interpolated values often cannot follow a trend of original time series data when the data have a lot of missing values. ARIMA analysis (Box and Jenkins, 1976) was conducted to generate re-interpolated values of flow rate, COD, NH_4^+-N, PO_4^{3-}-P and temperature in influent. The time series models of Methods 2 and 3 used re-interpolated data.

Development of linear regressive model. The autocorrelation function (ACF) and partial autocorrelation function (PACF) of the influent flow rate, COD, NH_4^+-N, and PO_4^{3-}-P were used to identify a linear regressive model.

Validation. Data obtained during 151–250 days were utilized for model validation. Figure 1 shows the overall procedure for development of the time series model.

Results and discussion

Time series model identification

An ARIMA analysis was performed for the time series model identification of flow rate, COD, NH_4^+-N, PO_4^{3-}-P and temperature in influent. Figure 2 shows ACF and PACF for

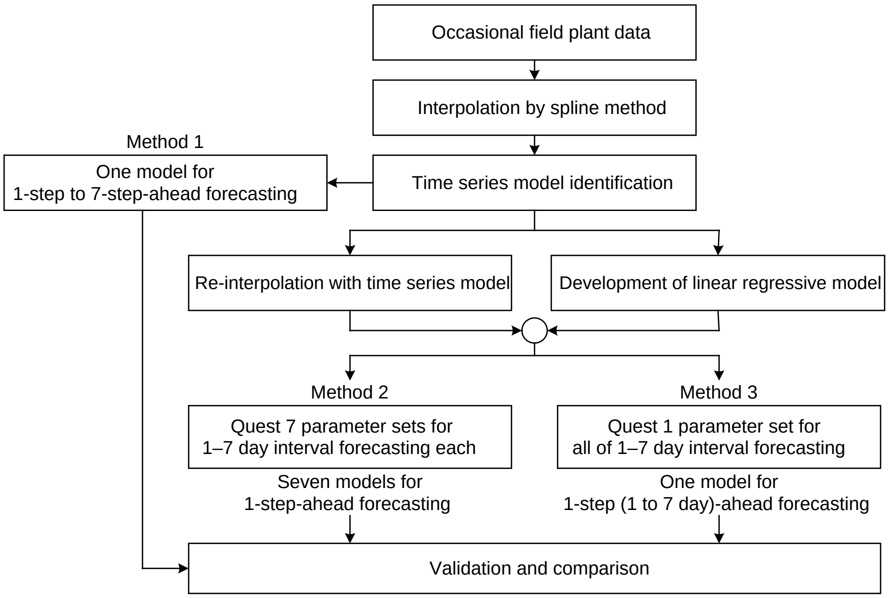
Occasional field plant data
Interpolation by spline method
Method 1
One model for 1-step to 7-step-ahead forecasting
Time series model identification
Re-interpolation with time series model
Development of linear regressive model
Method 2
Quest 7 parameter sets for 1–7 day interval forecasting each
Seven models for 1-step-ahead forecasting
Method 3
Quest 1 parameter set for all of 1–7 day interval forecasting
One model for 1-step (1 to 7 day)-ahead forecasting
Validation and comparison

Figure 1 Overall procedure for development of time series model

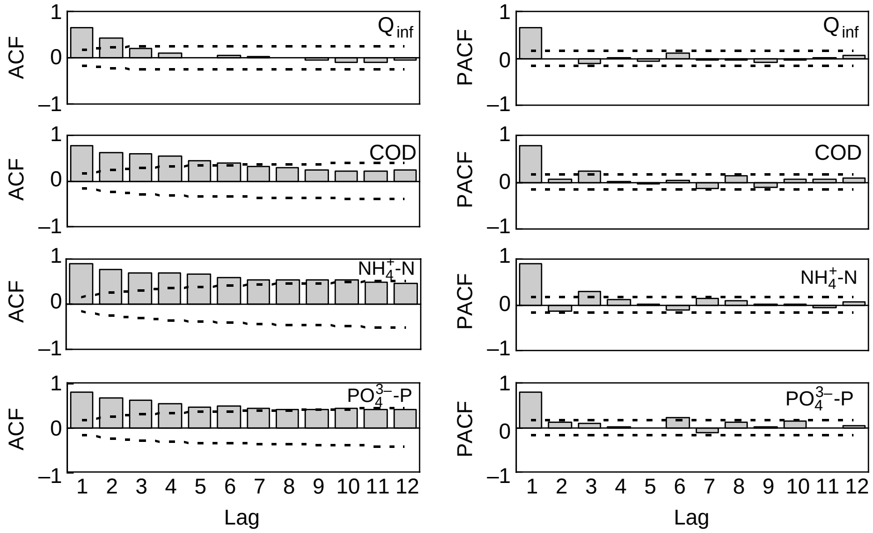

Figure 2 ACF and PACF for flow rate, COD, NH_4^+-N and PO_4^{3-}-P of influent

each component. Table 2 presents the univariate time series model developed by ARIMA analysis.

Interpolation and re-interpolation of missing data

Flow rate, COD, NH_4^+-N, PO_4^{3-}-P and temperature were interpolated using the spline method. They were then re-interpolated using the time series model shown in Table 2. Figure 3 shows the interpolated data of flow rate, COD, NH_4^+-N and PO_4^{3-}-P by the spline method and the re-interpolated data by the time series model. The re-interpolated data provided a better trend of measured concentrations compared with the interpolated data.

Model development and validation

Figures 4 and 5 show the results of one-day and seven-day-ahead forecasting for influent flow rate and concentrations with the three different methods.

Method 1: One model and seven-step-ahead forecasting. There was no significant difference in the three methods in the results of one-day-ahead forecasting, all yielding a good prediction. However, seven-day-ahead values of flow rate, COD and PO_4^{3-}-P showed a relatively constant value whereas seven-day-ahead values of NH_4^+-N oscillated extremely. These phenomena depended on characteristics of model structure. The time

Table 2 The univariate time series model developed through ARIMA analysis

	Model	Equation
Flow rate	ARIMA(1,0,0)	$Z_t = 6332 + 0.6438\,Z_{t-1} + a_t$
COD	ARIMA(1,0,0)	$Z_t = 48.7 + 0.7776\,Z_{t-1} + a_t$
NH_4^+-N	ARIMA(3,1,0)	$Z_t = Z_{t-1} - 0.2996\,Z_{t-2} + 0.1281\,Z_{t-3} + 0.1715\,Z_{t-4} + a_t$
PO_4^{3-}-P	ARIMA(1,0,0)	$Z_t = 0.4 + 0.7926\,Z_{t-1} + a_t$
Temperature	ARIMA(0,1,0)	$Z_t = Z_{t-1} + a_t$

*a_t = white noise series with a zero mean and a constant variance σ_a^2

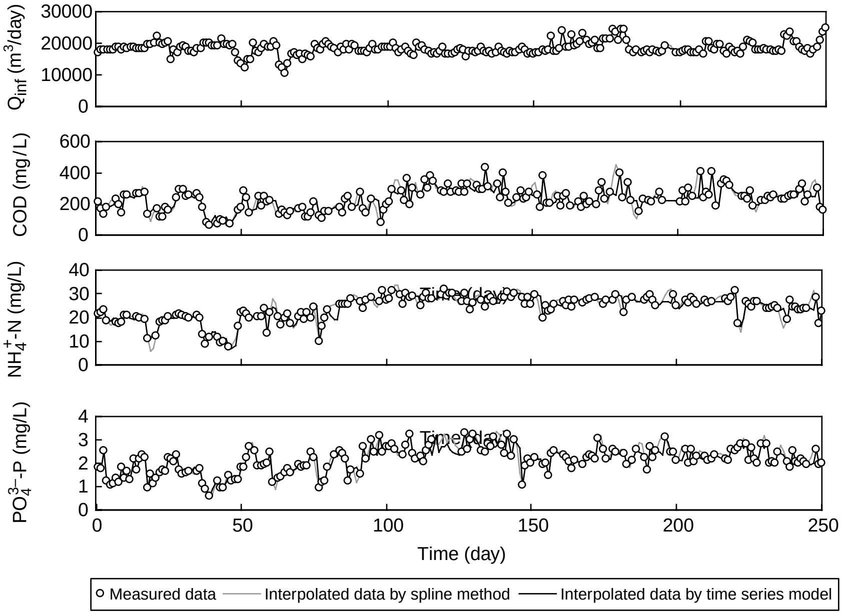

Figure 3 Interpolation results by spline method and by time series model

series models for flow rate, COD, and PO_4^{3-}-P were equally identified to ARIMA(1,0,0), which was $Z_{t+1} = C + \phi_1 Z_t$. An equation for n-step-ahead forecasting is:

$$Z_{t+n} = C + \phi_1 Z_{t+n-1} = C + \phi_1(C + \phi_1 Z_{t+n-2}) = C(1+\phi_1) + \phi_1^2 Z_{t+n-2} = \cdots$$
$$= C(1 + \phi_1 \cdots + \phi_1^{n-1}) + \phi_1^n Z_t \qquad (1)$$

If the number of steps (n) increased, $\phi_1^n Z_{t+n}$ reached 0 due to $|\phi_1| < 1$. Therefore, seven-day-ahead forecasting values converged on a constant.

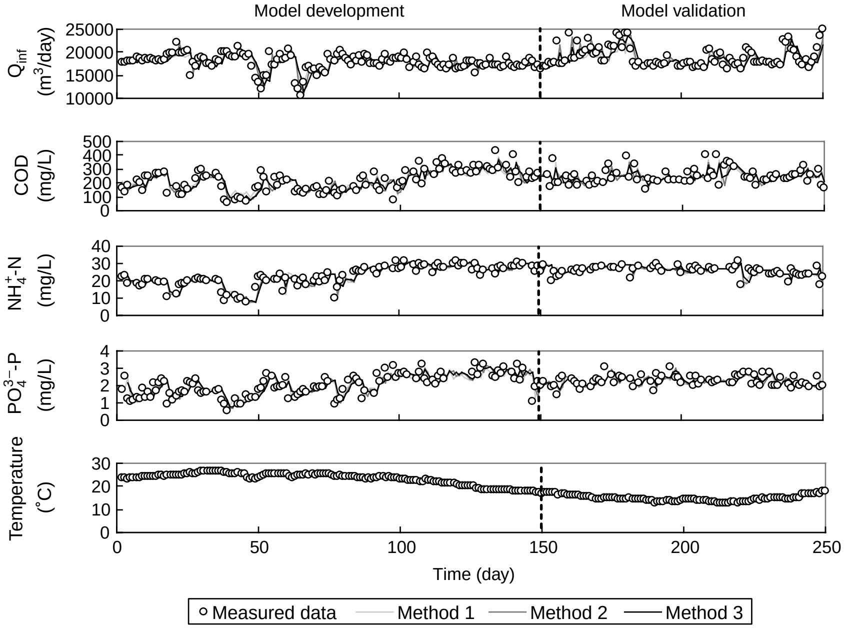

Figure 4 Comparison of one-day-ahead forecasting results by each method

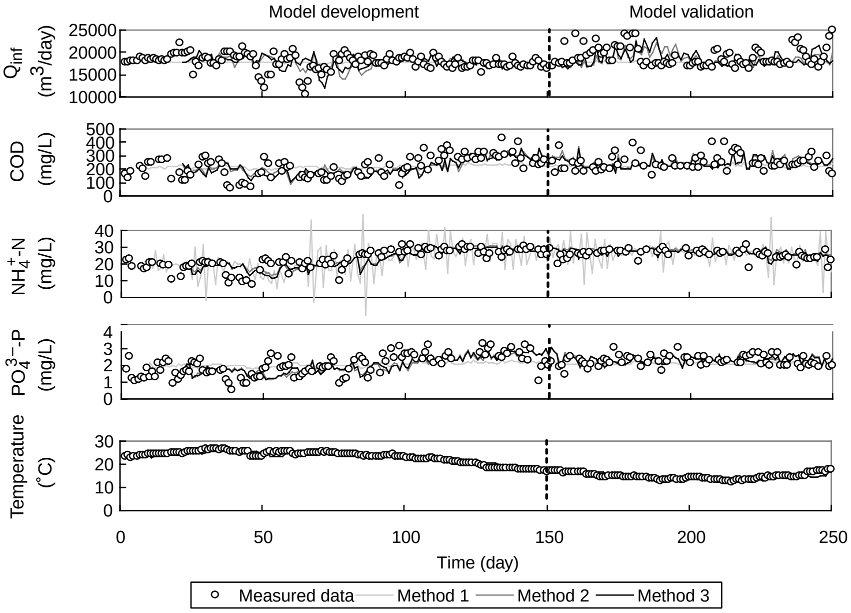

Figure 5 Comparison of seven-day-ahead forecasting results by each method

Since the mean of NH_4^+-N was non-stationary, difference data were used for development of the model. The predicted data oscillated greatly for seven-day-ahead forecasting, indicating that four predicted values (three-, four-, five-, six-day-ahead forecasting values) were used for seven-day-ahead forecasting, which greatly accumulated the prediction error. ARIMA(0,1,0) was identified as prediction model for temperature. There was no significant variation of data for the prediction horizon within 7 days. This model did not include a coefficient, implying that there were identical prediction values from one-day to seven-day-ahead forecasting.

Method 2: Seven models and one-step-ahead forecasting. Daily time series data were reformed to the time series data set for 1–7 days interval. Difference data were used for developing the time series model for NH_4^+-N in Method 1 because the average of NH_4^+-N data was non-stationary. In applying Method 2, the difference term was not necessary because the average of the three days interval data became stationary. The model structure was derived from the ACF and PACF of COD, NH_4^+-N and PO_4^{3-}-P. PACF showed significant values at lag 1 and lag 3 for influent flow rate and COD, and at lags 1, 2 and 3 for NH_4^+-N and PO_4^{3-}-P (Figure 2). This implied that every component was influenced by data within three days. Therefore, seven linear regressive models, one-step-ahead forecasting, were made for one-day to seven-day-ahead forecasting for each component (Eq. 2). The parameter sets were estimated for each model with the objective function of minimising RMSE.

$$Z_{t+1} = \phi_0 Z_t + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} \tag{2}$$

Figure 4 shows the one-day-ahead forecasting results for each method. There were some fluctuations of flow rate in the early validation period. In this period, flow rate increased very clearly. The model did not follow the variation of flow rate during this period because the model was developed with the data of non-fluctuated flow rate. The one-day-ahead forecasting gave good prediction results for NH_4^+-N and PO_4^{3-}-P. Overall prediction accuracy was better than for Method 1. The linear regressive model was therefore determined to be a reasonable structure.

Method 3: One model and one-step-ahead forecasting. The model structure in Method 3 was the same as in Method 2 and included one set of coefficients for every model. The prediction results of Method 3 were very similar to those of Method 2. Nothing was significantly different for any component during the validation period. It was concluded that only *one model and one-step-ahead forecasting* was enough to forecast the seven-day-ahead values for influent flow rate and concentrations. The method of one-step-ahead forecasting did not result in any accumulated error, and the coefficients estimation was simple and easy as a result of using only one model.

Further discussion for field application

A number of statistical modelling approaches, including time series modelling, can forecast more accurate values if there is sufficient data obtained from various conditions. In this research, the data for 150 days from a field plant were used for model development. The quantity of data was not considered sufficient because it did not include seasonal and annual variation patterns of influent characteristics. Therefore, it was essential to acquire enough data to develop a reliable prediction model. The purpose of this research was to suggest the simplest model to forecast seven-day-ahead data for the influent flow rate and concentration of each component. The model was successfully developed and applied. When more data were accumulated after operation, the model could be modified for better prediction.

Conclusions

The influent information was necessary to compare and evaluate various control strategies for a wastewater treatment process. The time series model presented the possibility to forecast seven-day-ahead influent information. The data used for model development were influent flow rate, COD, NH_4^+-N and PO_4^{3-}-P. These data did not include seasonal and annual variation patterns, so the developed model did not represent these characteristics. However, the model produced good results for one-day to seven-day-ahead forecasting. Since the model in Method 3 featured one-step-ahead forecasting only without accumulated errors and with simple coefficient estimation, it was considered to have excellent field applicability.

Acknowledgements

This study was supported financially by the Basic Research program of the Korea Science and Engineering Foundation (project No. R01-2003-000-10714-0) and Pusan National University (Post-Doc. program 2004). We wish to acknowledge the great contribution to this work made by Dr Haejin Woo, who passed away due to cancer, during her carrying out of this project.

References

Asano, T., Maeda, M. and Takaki, M. (1996). Wastewater reclamation and reuse in Japan: overview and implementation examples. *Wat. Sci. Tech.*, **34**(11), 219–226.

Bernard, O., Dantec, B.L., Chahuat, B., Steyer, J.-P., Lardon, L., Lambert, S., Ratini, P., Lema, J., Ruiz, G., Rodriguez, J., Vanrolleghem, P., Zaher, U. *et al*. (2005). An integrated system to remote monitor and control anaerobic wastewater treatment plants through the internet. *Wat. Sci. Tech.*, **52**(1–2), 457–464.

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

Hiraoka, M. and Tsumura, K. (1989). System identification and control of the activated sludge process by use of a statistical model. *Wat. Sci. Tech.*, **21**(8–9), 1161–1172.

Krüger (2004). SMAC SMArt Control of Wastewater Systems Deliverable No. 18. Summary of Implementations. Report EVK1-CT-2000-00056, Krüger, Denmark.

Lee, S.H., Ko, J.H., Im, J.H., Park, J.B., Kim, C.W. and Woo, H.J. (2005). Practice of ASM3 and EAWAG bio-P module for simulating five-stage step-feed EBPR process comparing with ASM2d. Submitted to ICA2005, 29 May–2 June, Busan, Korea.

Naghdy, G. and Helliwell, P. (1989). Process improvement by computer-aided load smoothing in activated sludge treatment. *Wat. Sci. Tech.*, **21**(8–9), 1225–1237.

Novotny, V., Jones, H., Feng, X. and Capodaglio, A. (1991). Time series analysis models of activated sludge plants. *Wat. Sci. Tech.*, **23**(5–6), 1107–1116.

Van Dongen, G. and Geuens, L. (1998). Multivariate time series analysis for design and operation of a biological wastewater treatment plant. *Wat. Res.*, **32**(3), 691–700.