Environmental Engineering Research

Environ. Eng. Res. 2014 March,19(1) : 31-36

Research Paper

http://dx.doi.org/10.4491/eer.2014.19.1.031
pISSN 1226-1025 eISSN 2005-968X

# A Study on the Influence of a Sewage Treatment Plant's Operational Parameters using the Multiple Regression Analysis Model

**Seung-Pil Lee, Sang-Yun Min, Jin-Sik Kim, Jong-Un Park, Man-Soo Kim†**

*Environmental Technology Institute, Samchully Enbio Co., Ltd., Seoul 150-972, Korea*

## Abstract

In this study, the influence of the control and operational parameters within a sewage treatment plant were reviewed by performing multiple regression analysis on the effluent quality of the sewage treatment. The data used for this review are based on the actual data from a sewage treatment plant using the media process within the year 2012. The prediction models of chemical oxygen demand ($COD_{Mn}$) and total nitrogen (T-N) within the effluent of the 2nd settling tank based on the multiple regression analysis yielded the prediction accuracy measurements of 0.93 and 0.84, respectively; and it was concluded that the model was accurately predicting the variances of the actual observed values. If the data on the energy spent on each operating condition can be collected, then the operating parameter that conserves energy without violating the effluent quality standards of COD and T-N can be determined using the regression model and the standardized regression coefficients. These results can provide appropriate operation guidelines to conserve energy to the operators at sewage treatment plants that consume a lot of energy.

**Keywords:** Control, Modeling, Prediction, Regression analysis, Regression model

## 1. Introduction

The first sewage treatment plant of Korea, Jungrang Sewage Treatment plant, was introduced in 1976, which was late compared to the installments of other city infrastructures. However, according to the data in 2009, there were 438 sewage treatment plants treating over 500 $m^3$ per day with the total facility capacity of 24,735,000 $m^3$ per day. Also, there were 2,332 sewage treatment plants treating below 500 $m^3$ per day with the total facility capacity of 171,000 $m^3$ per day; and consequently, the percent of population being supplied with sewage was 89.4%, which was not lacking compared to other developed countries in the world [1]. Also, a vast improvement has been made in terms of sewage treatment processes compared to the days when there was only the activated sludge process, which targeted organic matters; now, there are advanced sewage treatment processes that target various nutrients and poisonous substances. However, there is still huge room for improvement in terms of the operation and management technology of sewage treatment plants.

Operating a sewage treatment plant involves biological processes with complex reactions, which make it difficult to understand the behavior of the processes. Therefore, it requires a lot of analysis in order to fully understand the processing conditions and behaviors. Consequently, almost all of the operations within the sewage treatment plants rely heavily on experience. This is not desirable in terms of the automation of sewage treatment processes and reduction of personnel expenses. If a sewage treatment plant is operated based on the operator's experience, it is very difficult to conserve the energy spent within the sewage treatment plant.

In 1987, the International Association on Water Pollution Research and Control (IAWPRC) introduced the ASM1 model based on the analysis of the sewage treatment process using mathematical models. The ASM1 model was followed up by ASM2, ASM2d, and ASM3. These models are rational models with excellent emulation abilities. There are commercial programs applying these models that are being developed and used. But the analysis for substance classification and the measurement of the parameters using these models is difficult. Therefore, it is difficult to predict and control the sewage treatment process using these models in real-time. To solve these problems, there are various researches going on, such as the Benchmark Simulation Model, ARIMA Model, Neural Network Model, Multi Objective Model, ASMs simplifications, etc. [2-10].

Therefore, a multiple regression model that predicts and controls the sewage treatment process was developed utilizing the



**Received** Jun 03, 2013 **Accepted** October 23, 2013

†Corresponding Author
E-mail: SE080008@samchully.co.kr
Tel: +82-2-6309-7700 Fax: +82-2-6309-7704

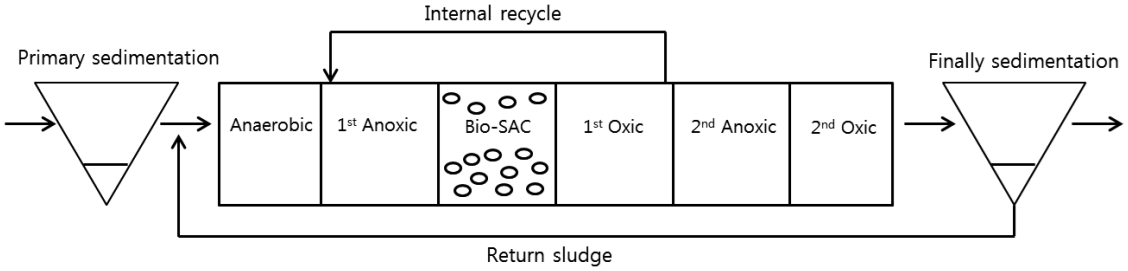


**Fig. 1.** A schematic of the Bio-SAC process.

**Table 1.** Variables of multiple regression analysis

| Category | Variable | |
|---|---|---|
| Inflow rate | INF | Independent |
| 1st settling tank influent | | |
| Temperature | I_TEMP | Independent |
| pH | I_PH | Independent |
| BOD | I_BOD | Independent |
| $COD_{Mn}$ | I_COD | Independent |
| SS | I_SS | Independent |
| T-N | I_TN | Independent |
| T-P | I_TP | Independent |
| Salt | I_SALT | Independent |
| Bioreactor influent | | |
| Temperature | RI_TEMP | Independent |
| pH | RI_PH | Independent |
| BOD | RI_BOD | Independent |
| COD | RI_COD | Independent |
| SS | RI_SS | Independent |
| T-N | RI_TN | Independent |
| T-P | RI_TP | Independent |
| Salt | RI_SALT | Independent |
| 1st Oxic reactor | | |
| Temperature | AE_TEMP | Independent |
| pH | AE_PH | Independent |
| DO | AE_DO | Independent |
| MLSS | AE_MLSS | Independent |
| SVI | AE_SVI | Independent |
| SRT | SRT | Independent |
| ASRT | ASRT | Independent |
| F/M ratio | FM | Independent |
| Return sludge amount | IRSF | Independent |
| Excess sludge amount | SUR | Independent |
| C/N ratio | CN | Independent |
| C/P ratio | CP | Independent |
| 2nd settling tank effluent | OUTF | Independent |
| $COD_{Mn}$ | RO_COD | Dependent |
| T-N | RO_TN | Dependent |

BOD: biochemical oxygen demand, COD: chemical oxygen demand, SS: suspended solids, T-N: total nitrogen, T-P: total phosphorus, DO: dissolved oxygen, MLSS: mixed liquor suspended solids, SVI: sludge volume index, SRT: solids retention time, ASRT: aerobic solids retention time, F/M ratio: food-to-microorganism ratio, C/N ratio: carbon-to-nitrogen ratio, C/P ratio: carbon-to-phosphorus ratio.

actual operational data being analyzed on a daily basis at the sewage treatment plants in this study. Also, the operational parameter prioritization methods with respect to effluent quality standards were developed by analyzing the influence of each operational parameter based on the results from the model above.

## 2. Materials and Methods

### 2.1. Selected Sewage Treatment Plant

The sewage treatment plant within this study, the P sewage treatment plant (232,000 ton/day) uses the media process (Fig. 1). The bioreactor process was chosen for the model development process and we reviewed the process starting with the intake of the influent and ending up with the release of the effluent. All of the data used in this research are within the year 2012.

### 2.2. Multiple Regression Analysis

Multiple regression analysis is a statistical method that analyzes the relationship between the dependent variable (Y), and independent variables (X1, X2, X3, ...). Any variables leading to multicollinearity during the multiple regression analysis were eliminated from the model, and any outliers and influential observations were considered through residual analysis in the model.

Multiple regression analysis used the SAS 9.2 program (SAS Institute Inc., Cary, NC, USA) and the stepwise method was implemented as the variable selection method. The significance level for the model verification and variable selection through the variable selection method was set at 0.05.

There were a total of 32 variables for the multiple regression analysis. RO_COD (2nd settling tank effluent $COD_{Mn}$) and RO_TN (2nd settling tank effluent T-N) were set as dependent variables and the remaining 30 variables were set as independent variables (Table 1).

### 2.3. Analysis Method

In the multiple regression analysis, the model was developed using the data from January 2012 to October 2012 and the data from November 2012 to December 2012 were used to verify the accuracy of the model.

The studentized residual suggested by Belsley et al. [11] and the DFFITS were used to verify the outliers and influential observations in this study. The following conditions had to be estab-

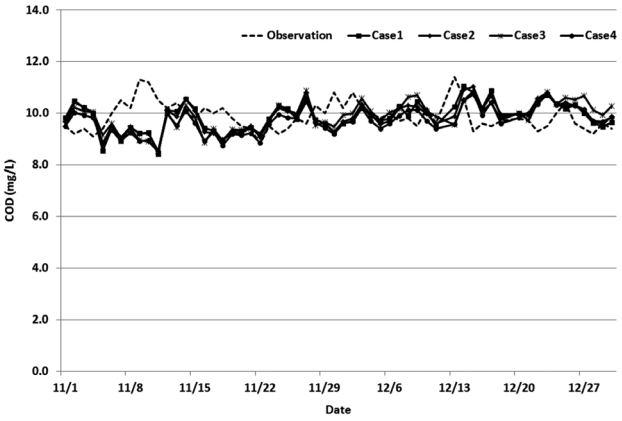


**Fig. 2.** Predictions of RO_COD (2nd settling tank effluent $COD_{Mn}$) cases 1 to 4. COD: chemical oxygen demand.

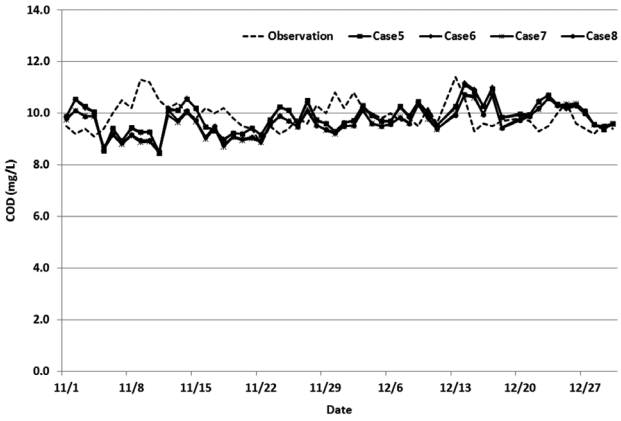


**Fig. 3.** Predictions of RO_COD (2nd settling tank effluent $COD_{Mn}$) cases 5 to 8. COD: chemical oxygen demand.

lished between the studentized residual and the DFFITS in order to determine whether they were outliers or influential observations [11-14].

Outliers criterion using the studentized residual:

$$\left|r_i^*\right| \geq t(n-k-2; a/2).$$

Influential observations criterion using the DFFITS:

$$\left|DFFITS(i)\right| \geq 2\left[\frac{k+1}{n}\right]^{1/2}.$$

In this case, $k$ is the number of independent variables and $n$ is the number of samples.

For checking multicollinearity, the stepwise method was used to set a variable and the variable's variance inflation factor (VIF) was calculated. If there was a variable greater than the VIF by 10 or more, it would be eliminated from the model development process after review and another variable would be selected by the stepwise method.

For verifying the accuracy, the following equation was used.

$$\text{Accuracy of prediction} = 1-\left(\sum_{i=1}^{i}\frac{\left|y_i-\hat{y}_i\right|}{y_i}\right)/i$$

In this case, $y_i$ is the $i$-th observed value, $\hat{y}_i$ is the $i$-th predicted value, and $l$ is the total number of values in the prediction data.

Multiple regression analysis was carried out through 8 cases depending on the variable selection method, elimination of outliers and influential observations, and log conversions (Table 2). In cases 1 to 4, the stepwise method was implemented as the variable selection method for all variables, and from cases 4 to 8, variables that are capable of being operated and controlled at the sewage treatment plant (aerobic reactor DO, aerobic reactor MLSS, return sludge amount, excess sludge amount) were selected and they were all included in the model development, if the stepwise method was applied. Standardized regression coefficients were considered in order to compare the effects of each parameter through multiple regression analysis for each case.

## 3. Results and Discussion

After reviewing the multiple regression model based on two variables RO_COD (2nd settling tank effluent $COD_{Mn}$) and RO_TN (2nd settling tank effluent T-N), all values resulting from the F-test on each case were below 0.05, which is statistically significant.

### 3.1. RO_COD (2nd Settling Tank Effluent $COD_{Mn}$)

After performing multiple regression analysis through 8 different cases, it was discovered that the variance patterns of the actual values and the predicted values for every case were similar. Also after reviewing factors that could explain the actual values resulting from the multiple regression model, such as the coefficient of determination, accuracy of prediction, and root mean squared error, all 8 cases displayed similar values (Table 3). The graphs comparing the observed values and predicted values for each case are show in Figs. 2 and 3.

**Table 2.** Regression analysis method for each case

| Case | Variable selection method | Elimination of outliers and influential observations | Log conversion |
|---|---|---|---|
| 1 | No included 4 variables | x | x |
| 2 | No included 4 variables | x | o |
| 3 | No included 4 variables | o | x |
| 4 | No included 4 variables | o | o |
| 5 | Included 4 variables | x | x |
| 6 | Included 4 variables | x | o |
| 7 | Included 4 variables | o | x |
| 8 | Included 4 variables | o | o |

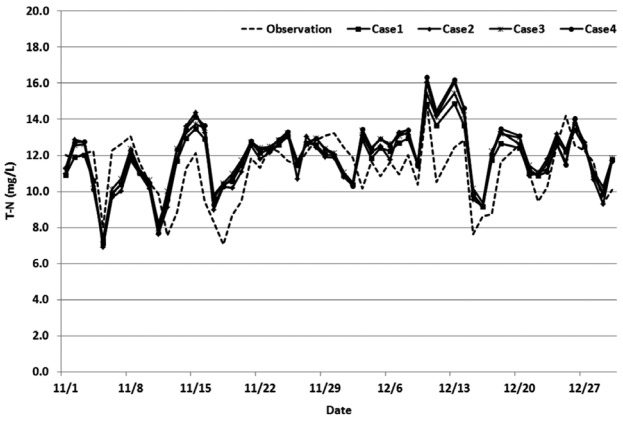


**Fig. 4.** Predictions of RO_TN (2nd settling tank effluent T-N) cases 1 to 4. T-N: total nitrogen.

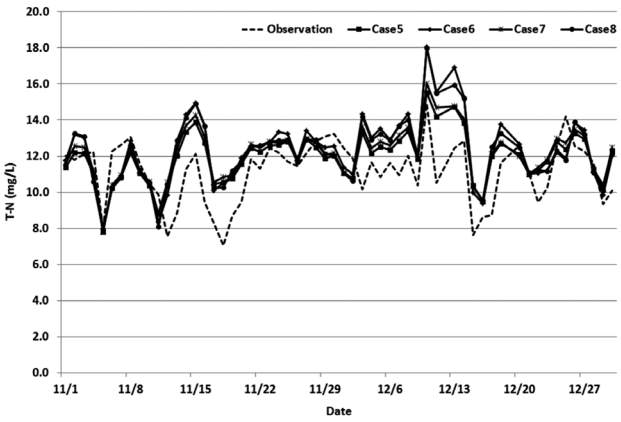


**Fig. 5.** Predictions of RO_TN (2nd settling tank effluent T-N) cases 5 to 8. T-N: total nitrogen.

### 3.2. RO_TN (2nd Settling Tank Effluent T-N)

After performing multiple regression analysis through 8 different cases, it was discovered that the variance patterns of the actual values and the predicted values for every case were similar. Also, after reviewing factors which could explain the actual values resulting from the multiple regression model, such as the coefficient of determination, accuracy of prediction, and root mean squared error, all 8 cases displayed similar values (Table 4). The graphs comparing the actual values and predicted values for each case are shown in Figs. 4 and 5.

### 3.3. Selection of Multiple Regression Model

The regression model for each dependent variable was selected after reviewing the $R^2$ (coefficient of determination) value, accuracy of prediction, root mean squared error, and variance pattern's predictabilities.

In the case of dependent variable RO_COD (2nd settling tank effluent $COD_{Mn}$), case 4, which scored the highest in terms of $R^2$ (coefficient of determination) value and accuracy of prediction, was selected and a regression model could be constructed based on the regression coefficients of each independent variable

**Table 3.** RO_COD (2nd settling tank effluent $COD_{Mn}$) regression analysis results

| Case | Number of selected variables | Coefficient of determination ($R^2$) | Accuracy of prediction | Root mean squared error |
|---|---|---|---|---|
| 1 | 9 | 0.5003 | 0.932523 | 0.854084 |
| 2 | 12 | 0.5273 | 0.932060 | 0.860667 |
| 3 | 10 | 0.5581 | 0.930142 | 0.908087 |
| 4 | 11 | 0.5911 | 0.934084 | 0.876982 |
| 5 | 11 | 0.5064 | 0.931929 | 0.856496 |
| 6 | 11 | 0.5152 | 0.932474 | 0.855621 |
| 7 | 10 | 0.5621 | 0.932987 | 0.879436 |
| 8 | 10 | 0.5890 | 0.934277 | 0.863299 |

COD: chemical oxygen demand.

**Table 4.** RO_TN (2nd settling tank effluent T-N) regression analysis results

| Case | Number of selected variables | Coefficient of determination ($R^2$) | Accuracy of prediction | Root mean squared error |
|---|---|---|---|---|
| 1 | 9 | 0.6686 | 0.879202 | 1.537822 |
| 2 | 10 | 0.6780 | 0.864789 | 1.718906 |
| 3 | 8 | 0.7229 | 0.861227 | 1.748826 |
| 4 | 8 | 0.6969 | 0.859342 | 1.811602 |
| 5 | 11 | 0.6678 | 0.868458 | 1.646804 |
| 6 | 13 | 0.6718 | 0.844277 | 2.040300 |
| 7 | 11 | 0.7185 | 0.857723 | 1.777473 |
| 8 | 13 | 0.7044 | 0.841517 | 2.033755 |

T-N: total nitrogen.

(Table 5).

In the case of dependent variable RO_TN (2nd settling tank effluent T-N), case 1, which scored the highest in terms of root mean squared error and accuracy of prediction, was selected and a regression model could be constructed based on the regression coefficients of each independent variable (Table 6).

A multiple regression equation for each dependent variable could then be formulated using the regression coefficients in Tables 5 and 6 above, and a prediction model for water quality predictions could be developed for each dependent variable depending on the changes made to independent variables, using the multiple regression equation.

### 3.4. Reviewing the Influence of Operation Parameters

Standardized regression coefficients were considered in order to review the influence of the operation parameters based on the model selected for each dependent variable. As the absolute value of the standardized regression coefficient grew larger, the influence on the model became greater.

In case 4, which selected from the dependent variable of RO_COD (2nd settling tank effluent $COD_{Mn}$), the order of the highest standardized regression coefficient among 11 independent variables was OUTF, I_COD, AE_MLSS (Table 7).

In case 1, which selected from the dependent variable RO_TN (2nd settling tank effluent $COD_{Mn}$), the order of highest standardized regression coefficient among the 8 independent variables was AE_TEMP, I_TN, FM (Table 8).

Based on the results of the standardized regression coefficients for each dependent variable, an independent variable with the biggest influence on dependent variables could be selected. After reviewing the influence on dependent variables mentioned above, it was concluded that dependent variables COD and T-N could be controlled by controlling the operation parameter AE_MLSS (aerobic reactor MLSS) for COD and operation parameter FM (F/M ratio) for T-N at the sewage treatment plant from which data was collected to formulate the multiple regression equations.

### 3.5. Reviewing the Economic Operation Condition of the Bioreactor

The method of process control utilizing the selected multiple regression model can provide the economic operation condition that saves energy without violating the effluent quality standards of $COD_{Mn}$ and TN in terms of energy savings. In order to provide the economic operation condition, first, the power consumption of the equipment (blowers, internal return sludge pumps,

**Table 5.** RO_COD regression coefficients (case 4)

| Variable | Coefficient |
|---|---|
| Intercept | 3.329518 |
| I_PH | 0.265276 |
| I_COD | 6.03E-05 |
| I_TN | -5.85E-05 |
| AE_TEMP | 0.003917 |
| AE_PH | 0.100492 |
| AE_DO | 0.021162 |
| AE_MLSS | 8.30E-05 |
| AE_SVI | 0.00106 |
| FM | 1.135763 |
| CP | -0.00393 |
| OUTF | 5.26E-06 |

All abbreviations used in this table are listed in Table 1.

**Table 6.** RO_TN regression coefficients (case 1)

| Variable | Coefficient |
|---|---|
| Intercept | -4991.35 |
| INF | -0.00529 |
| I_TN | 0.443086 |
| RI_PH | 524.3343 |
| RI_TP | 1.122865 |
| RI_SALT | 0.000652 |
| AE_TEMP | -39.3536 |
| AE_PH | 416.47 |
| FM | 4,175.656 |

All abbreviations used in this table are listed in Table 1.

**Table 7.** RO_COD standardized regression coefficient (case 4)

| Variable | Coefficient |
|---|---|
| OUTF | 0.66678 |
| I_COD | 0.38608 |
| I_PH | 0.23773 |
| I_TN | -0.20973 |
| AE_MLSS | 0.30001 |
| AE_SVI | 0.26458 |
| AE_TEMP | 0.22015 |
| AE_DO | 0.20773 |
| AE_PH | 0.11453 |
| FM | 0.21111 |
| CP | -0.14350 |

All abbreviations used in this table are listed in Table 1.

**Table 8.** RO_TN standardized regression coefficients (case 1)

| Variable | Coefficient |
|---|---|
| AE_TEMP | -0.49246 |
| AE_PH | 0.10546 |
| I_TN | 0.35140 |
| FM | 0.17037 |
| RI_SALT | 0.16888 |
| RI_TP | 0.15465 |
| RI_PH | 0.11428 |
| INF | -0.12166 |

All abbreviations used in this table are listed in Table 1.

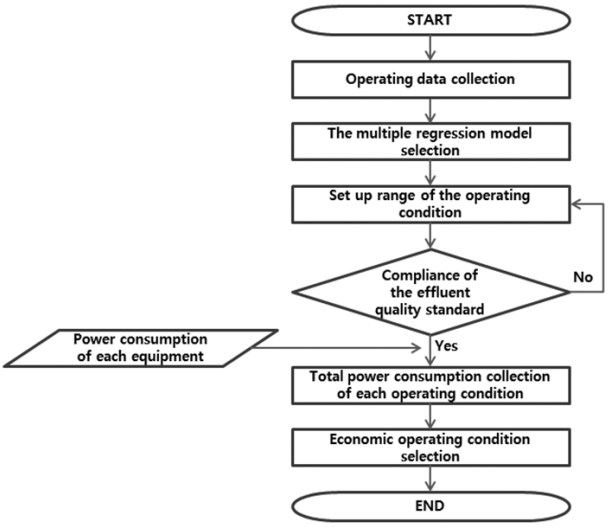


**Fig. 6.** The economic operation condition mimetic diagram.

excess sludge pumps, etc.) for operating the bioreactor should be reviewed. The economic operation condition can be provided by considering the effluent quality standards and the amount of power consumption (Fig. 6).

## 4. Conclusions

Using the prediction models for the 2nd settling tank effluent $COD_{Mn}$ and T-N based on multiple regression analysis, it was shown that the accuracy of prediction came out to be above 0.93 and 0.84, respectively. Also, through this model, the variance pattern of the actual values was proven to have been predicted fairly well.

In the case of dependent variable RO_COD (2nd settling tank effluent $COD_{Mn}$), case 4 was selected, because it scored the highest in terms of $R^2$ (coefficient of determination) value and accuracy of prediction; and in the case of dependent variable RO_TN (2nd settling tank effluent T-N), case 1, which scored the highest in terms of root mean squared error and accuracy of prediction, was selected.

Based on the results after reviewing the effectiveness of the operation parameters, the most effective operation parameter for controlling $COD_{Mn}$ was AE_MLSS and F/M for controlling T-N.

The effectiveness of the operational parameters on the 2nd settling tank effluent $COD_{Mn}$ and T-N were verified by reviewing the standardized regression coefficients of the multiple regression models constructed in this study. Predicting the concentration of $COD_{Mn}$ and T-N in the 2nd settling tank effluent was possible by following the operation conditions using the selected multiple regression model. If the data on the energy spent on each operation parameter can be collected, then the operation parameter that conserves energy without violating the effluent quality standards of $COD_{Mn}$ and TN can be determined using the regression model and the standardized regression coefficients. These results can provide appropriate operation guidelines to conserve energy to the operator at a sewage treatment plant that consumes a lot of energy.

## Acknowledgments

This research was supported by the Korea Ministry of the Environment as a “Global Top Project” (No. GT-11-B-02-014-2).

## References

1. Korea Ministry of Environment. Sewage statistics. Gwacheon: Ministry of Environment; 2010.
2. Henze M, Grady CP, Gujer V, Marais GV, Matsuo T. Activated sludge model No. 1. London: International Association on Water Pollution Research and Control; 1987.
3. Woo DJ. Model based predictive control algorism development and application in A2/O process [master’s thesis]. Busan: Pusan National University; 2011.
4. Choi SY. Fault diagnosis of a biological wastewater treatment plant by multivariate statistical approaches and development of a simplified activated sludge model [master’s thesis]. Daegu: Kyungpook National University; 2011.
5. Woo DJ, Kim H, Kim YJ, et al. Development and evaluation of model-based predictive control algorithm for effluent NH4-N in A2/O process. *J. Korean Soc. Environ. Eng.* 2011;33:25-31.
6. Min SY, Lee SP, Kim JS, Park JU, Kim MS. Development and validation of multiple regression models for the prediction of effluent concentration in a sewage treatment process. *J. Korean Soc. Environ. Eng.* 2012;34:312-315.
7. Benedetti L, De Baets B, Nopens I, Vanrolleghem PA. Multicriteria analysis of wastewater treatment plant design and control scenarios under uncertainty. *Environ. Model. Softw.* 2010;25:616-621.
8. Dellana SA, West D. Predictive modeling for wastewater applications: linear and nonlinear approaches. *Environ. Model. Softw.* 2009;24;96-106.
9. Hakanen J, Sahlstedt K, Miettinen K. Wastewater treatment plant design and operation under multiple conflicting objective functions. *Environ. Model. Softw.* 2013;46:240-249.
10. Fu G, Butler D, Khu ST. Multiple objective optimal control of integrated urban wastewater systems. *Environ. Model. Softw.* 2008;23:225-234.
11. Belsley DA, Kuh E, Welsch RE. Regression diagnostics: identifying influential data and sources of collinearity. New York: John Wiley & Sons; 1980.
12. Kim JD. Linear regression analysis using SAS. Seoul: Free Academy; 2002.
13. Park BJ. Theory and application of modern statistics. Seoul: Sigma Press; 2006.
14. Jung KM, Kim MG. Multivariate analysis. Seoul: Kyo Woo Sa; 2007.