



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

工學碩士 學位論文

다변량 회귀분석에 의한 일차침전지
유출수 수질 예측



2009年 2月

釜慶大學校 産業大學院

環 境 工 學 科

金 泰 炫

工學碩士 學位論文

다변량 회귀분석에 의한 일차침전지
유출수 수질 예측

指導教授 李 柄 憲

이 論文을 工學碩士 學位論文으로 提出함



2009年 2月

釜慶大學校 産業大學院

環 境 工 學 科

金 泰 炫

金泰炫의 工學碩士 學位論文을
認准함

2008 年 12 月



主 審 工學博士 姜 壬 錫 ㉠

委 員 工學博士 李 太 允 ㉠

委 員 工學博士 李 柄 憲 ㉠

목 차

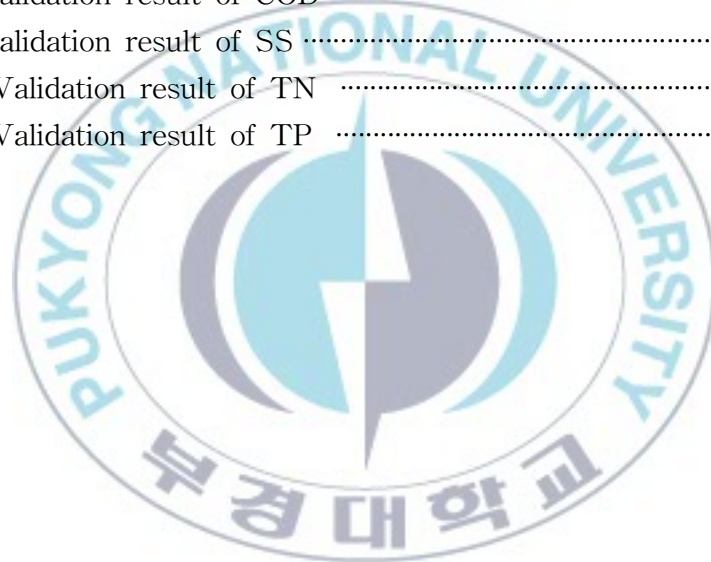
제 1 장 서론	1
제 2 장 이론적 배경	3
2.1 회귀분석(Regression)	3
2.1.1 회귀모형의 종류	4
2.1.2 단순선형회귀모형	4
2.1.3 중회귀모형	6
2.1.4 잔차의 성질	8
2.1.5 회귀방정식의 신뢰성	9
2.1.6 자료의 진단	9
2.1.7 모형의 진단	10
제 3 장 연구방법	12
3.1 회귀분석을 이용한 수질예측	12
3.2 일차침전지 모형 구축	14
3.2.1 반응표면분석을 이용한 회귀계수 산정	14
3.2.2 모형의 검정	16
3.2.3 모형의 적합성 평가	17
제 4 장 실험결과 및 고찰	18
4.1 회귀분석 모형의 추정	18
4.1.1 회귀계수의 산정	18
4.1.2 추정한 회귀분석 모형과 실측값 비교	25
4.2 회귀분석 모형의 검정	29
제 5 장 결론	33
제 6 장 참고문헌	34

List of Table

Table 2.1 Classify the regression model	4
Table 3.1 Ability of multi-regression	12
Table 3.2 Water quality of S-WWTP influent in 2006	15
Table 3.3 Water quality of S-WWTP primary settling effluent in 2006	15
Table 3.4 Water quality of S-WWTP influent in 2007	16
Table 3.5 Water quality of S-WWTP primary settling effluent in 2007	17
Table 4.1 Analysis dispersion table of BOD	20
Table 4.2 A regression coefficient of BOD	20
Table 4.3 Analysis dispersion table of COD	21
Table 4.4 A regression coefficient of COD	21
Table 4.5 Analysis dispersion table of SS	22
Table 4.6 A regression coefficient of SS	22
Table 4.7 Analysis dispersion table of TN	23
Table 4.8 A regression coefficient of TN	23
Table 4.9 Analysis dispersion table of TP	24
Table 4.10 A regression coefficient of TP	24
Table 4.11 Estimation result of RMSE between the measured value and predictive value	25
Tabel 4.12 Validation result of RMSE between the measured value and predictive value	29

List of Figure

Fig. 4.1 Biplot between the influent and the primary settling effluent	18
Fig. 4.2 Estimation result of BOD	26
Fig. 4.3 Estimation result of COD	26
Fig. 4.4 Estimation result of SS	27
Fig. 4.5 Estimation result of TN	27
Fig. 4.6 Estimation result of TP	28
Fig. 4.7 Validation result of BOD	30
Fig. 4.8 Validation result of COD	30
Fig. 4.9 Validation result of SS	31
Fig. 4.10 Validation result of TN	31
Fig. 4.11 Validation result of TP	31



Prediction of effluent quality from primary sedimentation tank using multi-variable regression method

Tae-Hyun Kim

*Department of Environmental Engineering, Graduate school,
Pukyong National University*

Abstract

The release of polluted water by an inadequate sewage treatment cause a serious environmental problems and disease in animals and plants. Recently, strict environmental controls and Water Quality Standards have been adopted and consequently ICA(Instrumentation, Control and Automation) techniques have been developed rapidly in order to manage and control the sewage treatment process. As a result, various predictive models have been proposed by using simulation and mathematical models.

Many models which are developed for the sewage treatment process are focused on a predicting technique which is to detect an irregular situation of the process and prevent a problem from occurring in the process. Recently, various Neural Networks-based models which might be efficient in pattern recognition and prediction problems have been used in previous researches. However, there some unsolved issues in neural network models. The input vectors in neural networks should be normalized between 0 and 1 if the activation function used is the standard sigmoid.

Another problem is the noise of input vectors could effect to the predictive accuracy. Because of these problems, it would not be appropriate to use previous popular models for the process design and the process control optimization of the sewage treatment.

In this research, the Primary Sedimentation Tank Effluent Estimation model, which can be substitute for the ASMs proposed by IWA task group and adaptable to an integrated model of the sewage treatment, has been proposed in order to overcome the problems of previous studies. The main results in this study ara as follows;

1. A multiple regression model in which independent variables are inflow, BOD, SS, TN, TP, and first sludge has been developed to predict the primary sedimentation tank effluent.
2. The RMSE(root mean square error) of BOD, COD, SS, TN and TP are 3.81, 2.74, 4.7, 4, 0.2 respectively.
3. The proposed regression model, which has been developed with data of year 2006, has been tested on the data of year 2007 and produced a good performance in terms of predictive accuracy.

제1장 서론

하수처리장의 부적절한 운전에 의해 유출된 방류수는 심각한 환경 문제를 발생시키고, 동·식물에 다양한 질병을 유발한다. 최근 환경 규제 강화에 의해 하수처리장 방류수 수질기준이 엄격해짐에 따라 하수처리장의 관리 및 운전을 위한 ICA(Instrumentation, Control and Automation) 기술들이 빠르게 발전해 왔으며, 하수처리장 공정 모델링에 대한 연구 또한 활발히 진행되고 있다(Kim, 2006). 그 결과 하수처리공정에 대한 다양한 예측 기법이 제시되었고, 실규모 하수처리장에 실제 적용되는 경우도 있었다. 제안된 제어 방법의 실제 적용에 앞서 모델을 사용하여 적용 가능성을 평가하는 것이 필요하며, 이 과정에서 시뮬레이션의 모형구축에 수학적 모델을 이용할 수 있다(김 등, 2007).

현재 하수처리 공정을 대상으로 개발된 모델들은 공정의 이상 상태를 사전에 감지하여 공정에서 발생할 수 있는 문제를 사전에 방지하기 위한 예측 기법에 중점을 두고 있다. 그에 따라 실측 데이터를 기반으로 입력 변수와 목표 변수간의 패턴 인식 및 예측에 효과적이고, 보정하기 쉬운 인공신경망을 이용한 다양한 예측 모델이 개발되었다(문 등, 2008; 최 등, 2007; 김 등, 2007). 하지만 인공신경망의 경우 입출력 값이 0 과 1 사이로 제한되어 있어 추가적인 전처리가 필요한 단점이 있다. 또한 데이터의 질이 모델의 핵심이므로 실측값의 오류나, 데이터 조작시 발생한 오류는 추후 분석에 있어 문제점을 야기할 수도 있다. 이러한 이유로 하수처리 공정의 효율적인 설계 및 운전을 위한 최적화 및 제어의 측면에 있어서 제한이 있다(최 등, 2007). 하수처리 시뮬레이션 중에서도 유체해석 프로그램을 이용하

는 침전지 모델의 경우 그 연구가 진행되고 있으나(Yong_Na Zhao, 2008), 미비한 실정이다.

따라서 본 연구에서는 기존 연구의 한계점을 보완하고, IWA taskgroup에서 개발한 활성슬러지(ASMs) 모델과 하수처리장 통합 모델에 적용가능한 일차침전지 모형을 구축하고자 한다. 비선형이 강한 하수처리 공정을 선형화 시킬 경우, 모델의 정확성이 떨어질 가능성이 높기 때문에(이 등, 2007) 반응표면분석을 이용하여 다변량 회귀분석을 구하고, 일차 침전지 유출수의 BOD, COD, SS, T-N, T-P 예측 모델을 개발한다.



제2장 이론적 배경

2.1 회귀분석(Regression)

회귀분석은 데이터 분석에서 가장 널리 쓰이는 통계적 분석방법의 하나로서, 모형화의 관점에서 특히 중요시되는 데이터 분석 방법이다. 회귀분석이란 반응변수와 독립 변수들간의 관계를 밝히는 분석방법으로 선형모형을 통해서 모형을 적합하고, 적합화된 모형을 통해 관련된 현상을 연구하는 방법이다. 따라서 이 방법은 사회과학 분야, 자연과학 분야 뿐만 아니라 다른 어느 분야에 있어서도 모형화와 관련된 현상을 연구하는데 있어서 필수 불가결한 방법이라고 할 수 있다.

2.1.1 회귀모형의 종류

회귀모형(regression model)은 회귀분석에 사용되는 통계적 모형으로 Table 2.1과 같은 모형이 있다. 여기에서 $\beta_0, \beta_1, \beta_2$, 등은 회귀계수이고, ε 은 오차항이다.

Table 2.1 Classify the regression model

종류	모형	설명
단순회귀	$Y = \beta_0 + \beta_1 X + \varepsilon$	독립변수가 하나이고 종속변수가 Y와의 관계가 직선
중회귀	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$	독립변수가 2개 이상($X_1, X_2, \dots, X_k$)이고 Y와의 관계가 1차선형 함수
곡선회귀	$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_k X^k + \varepsilon$	독립변수가 하나이고 Y와의 관계가 2차곡선 이상($k \geq 2$)
다항회귀	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X^k + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \dots + \beta_{kk} X_k^2 + \beta_{12} X_1 X_2 + \dots + \varepsilon$	독립변수가 K개이고 2차 다항회귀모형인 경우
비선형회귀	$Y = \beta_0 \log(\beta_1 X_1) + \beta_2 \sin(\beta_3 X_2) + \varepsilon$	회귀계수들이 선형으로 표현되지 않은 모형. 이에 비하여 위의 네가지는 모두 선형회귀 모형

2.1.2 단순선형회귀 모형

회귀분석은 한 변수가 다른 변수의 값에 영향을 미친다고 보고 그 관계의 형태를 규명하고자 하는 시도라고 볼 수 있다. 영향을 미치는 변수를 독립변수(independent variable)라고 하고 영향을 받는 변수를 종속변수 또는 반응변수(dependent or response variable)라고 한다.

단순선형회귀 모형은 다음과 같다.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad (2.1)$$

$$i = 1, 2, \dots, n$$

Y_i : 종속변수의 i 번째 측정치

X_i : 독립변수의 i 번째 측정치

β_0, β_1 : unknown regression parameters (회귀모수)

ε_i : error term (오류항)

가) 모형의 기본적 가정

i. 독립변수와 종속변수 사이에 독립변수가 종속변수를 설명할 수 있는 관계가 존재한다.

ii. 관계는 선형이다.

일반적으로 독립변수와 종속변수의 구분이 명확한 경우가 많으나 변수간의 관계의 존재 여부는 아무도 확인할 수 없는 순수한 가정인 경우가 대부분이며, 설사 관계가 존재한다고 하더라도 그 관계의 형태에 대해서는 알지 못한다. 다만 관계가 존재하고 그 형태는 선형이라는 것은 우리의 가정일 뿐이다.

나) 모형의 기술적 가정

단순선형회귀 모형을 보면 두 부분으로 구성되어 있음을 알 수 있다. 그 첫 번째는 독립변수 X_i 에 기울기라고 할 수 있는 상수 β_1 을 곱한 다음 상수 β_0 을 더해서 얻어지는 부분이며 두 번째는 오차항 ε_i 이다. 따라서 종속

변수 Y_i 의 값은 기본 값 β_0 과 독립변수 X_i 의 값에 의해 자동적으로 결정되는 $\beta_1 X_i$ 에 Random mechanism에 의해 결정되는 오차항 ε_i 가 추가되어 정해지는 것으로 볼 수 있다. 이 오차항은 상호 독립적이므로 같은 독립변수 값 X_i 에 대해서도 관측시 마다 다를 수 있어서 결과적으로 종속변수의 값이 다양해지도록 만든다.

2.1.3 중회귀 모형

종속변수의 변화를 설명하기 위하여 두 개 이상의 독립변수가 사용되는 선형회귀모형을 중선형 회귀(multiple linear regression model)라 부르며, 간단히 중회귀모형(multiple regression model)이라고도 한다. 중회귀분석의 기본적인 개념은 단순회귀분석과 같으나, 독립변수가 두 개 이상이 사용된다는 점이 다르다.

독립변수의 수가 k 개인 중회귀 모형은 일반적으로 다음과 같이 표현된다.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \quad (2.2)$$

$$i = 1, 2, \dots, n$$

$\beta_0, \beta_1, \dots, \beta_k$: 회귀계수

$\varepsilon_i \sim N(0, \sigma^2)$ 이고 서로 독립

여기서 $\beta_0, \beta_1, \dots, \beta_k$ 는 모집단의 회귀계수이고, ε_i 는 반응변수 Y_i 를 측정할 때 발생하는 오차이다.

중회귀 모형에서 독립변수가 2개인 경우에 대해서 먼저 살펴보도록 하자. 식(2.2)에서 $k=2$ 이면 중회귀 모형은

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i \quad (2.3)$$

$$i = 1, 2, \dots, n$$

이고, n 개의 오차 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 은 서로 독립이고 각각 $N(0, \sigma^2)$ 의 분포를 따른다고 가정한다. 식(2.3)에서 I 번째 측정치에 대해서는

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i \\ &= (1, X_1, X_2) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \varepsilon_i \end{aligned} \quad (2.4)$$

와 같이 벡터표현이 가능하다. 여기에서

$$Y_{n \times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad X_{n \times 3} = \begin{pmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{pmatrix}, \quad \beta_{3 \times 1} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \quad \varepsilon_{n \times 1} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (2.5)$$

이라 하고, n 개의 데이터에 대해 행렬을 사용하여 표현하면 다음과 같다.

$$Y = X\beta + \varepsilon, \quad N(0, I\sigma^2) \quad (2.6)$$

식(2.5)에서

$$E(\varepsilon) = 0 \quad (2.7)$$

이고,

$$\begin{aligned}
 Var(\varepsilon) &= E[(\varepsilon - E(\varepsilon))(\varepsilon - E(\varepsilon))'] \\
 &= E[\varepsilon \varepsilon'] \\
 &= \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = I\sigma^2
 \end{aligned} \tag{2.8}$$

이므로,

$$E(Y) = X\beta \tag{2.9}$$

가 되고, 또한 Y의 분산-공분산 행렬은 ε 와 마찬가지로

$$Var(Y) = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = I\sigma^2 \tag{2.10}$$

가 된다.

2.1.4 잔차의 성질

가) 잔차들의 합은 0이다.

$$\sum e_i = 0 \tag{2.11}$$

나) 잔차들의 X_{ij} 에 대한 가중합은 0이다.

$$\sum_{i=1}^n X_{ij}e_i=0, \quad (j=1,2,\cdots,k) \quad (2.12)$$

다) 잔차들의 $\hat{Y}$ 에 대한 가중합도 0이 된다. 즉,

$$\sum \hat{Y}_i e_i = 0 \quad (2.13)$$

라) 중회귀 모형 $Y=X\beta+\varepsilon$ 에서 오차항 ε 는 $N(0, I\sigma^2)$ 의 분포를 하며, ε_i 들 간에는 서로 상관관계가 없이 서로 독립적이나, 잔차 e_i 들간에는 상관관계가 일반적으로 존재한다.

2.1.5 회귀방정식의 신뢰성

추정된 회귀방정식이 어느 정도 믿을만한 것이며, 주어진 자료들을 어느 정도 잘 설명하여 주고 있고, 또한 추정된 회귀방정식을 이용하여 X 의 값에서 $E(Y)$ 를 예측할 때 어느 정도 정확한 값을 검토하는 측도가 된다.

2.1.6 자료의 진단

회귀식을 구하고 추정 및 검정을 하는 과정을 총괄분석(aggregate analysis)이라고 한다. 이것의 목적은 전체자료를 결합하여 적합모형(fitted model)이라고 하는 하나의 요약된 결과를 구하는 것이다. 이러한 총괄분석은 설정한 회귀모형과 가정들이 정확하다는 전제하에 이루어진 것이다. 그러나 어떠한 문제에서도 그렇듯이 제시된 회귀모형 및 가정은 언제든지 틀릴수가 있다. 총괄분석에 이어 회귀분석의 두 번째 중요한 과정은 사용된

회귀모형 및 가정이 과연 타당한지, 그리고 각각의 관측값이 모형 및 가정에 어떠한 영향을 미치는지를 진단하는 것이며, 우리는 이러한 과정을 회귀진단(regression diagnostics)이라고 부른다. 총괄분석과 비교하여 이 과정을 개별분석(case analysis)이라고도 한다.

여기에서는 추정된 모형이 과연 적절하고 유용한지 여부를 판단하게 된다. 그리고 각각의 관측값에 대해 좀 더 세밀한 분석이 행해진다. 개별분석의 결과에서 추정된 모형을 고칠 필요가 있게 되면 자료나 가정을 수정하여 첫 번째 과정인 총괄분석부터 다시 시작하게 된다.

회귀진단은 모형진단(model diagnostics)과 자료진단(data diagnostics)의 두 가지로 구분된다. 모형진단에서는 모형이나 가정에 어떤 문제점이 있는지를 알아보고 자료진단에서는 자료에 조그만 변화가 모형의 추정에 어떠한 영향을 미치는가를 알아본다. 모형진단이 필요한 이유는 회귀모형에 대해 세운 가정이 실제 문제에서는 적당치 않는 경우가 종종 발견되기 때문이며 자료진단이 필요한 이유는 우리가 관측한 자료에서 모든 관측값의 특징이 같을 수 없는 경우가 있기 때문이다. 모형진단이 가정의 타당성을 의심하는 반면에 자료진단은 기본적으로 가정이 적절하다는 전제 하에 이루어진다. 모형진단에서 기본적인고도 핵심적으로 사용되는 통계량은 잔차이다. 잔차의 형태로부터 추정모형이 관측된 자료를 닮은 정도와 모형에 대해 세운 가정이 바람직한지 알 수 있다. 자료진단에서 기본적인 개념은 각각의 관측값이 추정모형에 어떠한 영향을 미치고 있는가를 파악하는 것이다.

2.1.7. 모형의 진단

모형진단이 필요한 이유는 우리가 사용하는 회귀모형이나 회귀모형에 세운 가정이 실제 문제에서는 적당하지 않는 경우가 종종 발견되기 때문이다

다. 여기서도 기본적으로 핵심적으로 사용되는 것은 잔차이다. 잔차의 형태로부터 모형과 가정이 적절한지 여부를 알 수 있다(김, 2001).



제3장 연구방법

본 연구에는 부산의 S-하수처리장의 2006~2007년의 수질 자료를 사용하였다. 2006년의 일일 자료를 토대로 미니탭 통계프로그램을 이용하여 각각의 회귀식을 구하고, 구해진 식을 통하여 수질예측을 하여 2007년 자료와 비교하였다.

3.1. 회귀분석을 이용한 수질 예측

회귀분석은 독립변수와 종속변수의 관계를 요약하여 제시하는 기술의 도구로서 사용될 수 있으며, 또한 표본에서 발견된 변수들 간의 관계를 모집단의 수준으로 일반화시키는 통계적 추론의 도구도 된다. 물론 이 두 가지 용도는 상호 밀접한 연관을 가지나, 개념적으로는 분리하여 생각할 필요가 있다.

Table 3.1 Ability of multi-regression

-
- | |
|---------------------------------|
| 1. 기술 도구 |
| a. 회귀식 추정 및 추정된 회귀식의 정확도 평가 |
| b. 특정 독립변수의 종속변수에 대한 독립적 영향력 측정 |
| 2. 추리 도구 |
| a. 모수 추정 및 신뢰구간 측정 |
| b. 모집단에 관한 가설검증 |
-

Table 3.1에 제시된 바와 같이, 기술도구로서 회귀분석의 가장 중요한 기

능은 회귀모형에 포함된 변수들로 구성되는 가장 적합한 회귀식을 찾아내고 추정된 회귀식의 정확도를 평가하는 것이다. 또한 회귀분석에서는 다른 독립변수들이 고정된 상태에서 특정 독립변수가 종속변수에 독립적으로 미치는 영향력의 측정이 가능하다. 즉, 복합적으로 얽혀 있는 인과관계의 매커니즘 속에서 관련된 다른 변수의 영향력을 배제하고 특정 변수의 중요성을 파악하게 해준다. 그러므로 회귀분석은 변수들간의 복잡한 인과관계 매커니즘을 파악하는 도구가 된다.

한편, 수질의 계측값 또는 분석값은 수질 그 자체의 어떤 변화와 더불어, 계측 과정에 몇가지 오차가 가미되어 계측값이 일반적으로 정해지지 않을 때가 많다. 따라서 수질 계측에 있어서는 여러 개의 반복된 계측값으로부터 중요한 법칙성을 알아내고 그 위에 연역할 수 있는 몇 개의 특성을 합쳐 추정할 수 있으면 더욱 좋을 것이다. 구체적으로는 반복된 수질측정 결과에 통계적 수법을 도입해서 조금이라도 많은 특성을 기술하는 것이 당연한 목적이며 그 기술된 특성에 의해 조금이라도 새로운 지식을 추가할 수 있다면 통계적 해석의 의의는 이미 확인된 것이라 할 수 있다. 따라서 수질 데이터의 통계적 해석의 기본적인고도 구체적인 방법으로는 어떤 변화하는 수질 데이터의 분포 특성을 조사하고, 그 분포를 수식으로 표시할 수 있는 경우에는 분포형에 관련된 수학적 지식을 적극적으로 도입해서 조금이라도 많은 법칙성을 통계적으로, 바꾸어 말하면 수학적으로 추정하려는 것이다(왕 등, 2000).

3.2. 일차침전지 모형 구축

본 연구에서는 하수처리장의 대표적인 물리화학적 공정인 일차침전지 모형을 구축하기 위해, minitab의 반응표면분석을 실행시켜 생성된 회귀계수로 회귀식을 추정하였으며, 검정을 통해 모델의 예측 능력을 평가하였다. 그 후, 추정과 검정 단계에서 RMSE(root mean square error)를 통해 모델의 적합성을 평가했다.

3.2.1 반응표면분석을 이용한 회귀계수 산정

회귀계수의 산정에는 자료 분석을 연구하는 모든 분야에서 널리 사용되고 있는 기초 통계 프로그램인 minitab을 이용하여 반응표면 분석을 실행하였다. 반응표면 분석은 하나 또는 그 이상의 반응 변수와 정량적인 실험변수 또는 인자 집합과의 관계를 구명하는데 사용하는 실험 설계 방법으로 반응 변수를 최적화 할 수 있는 인자의 조건을 찾는 데 목적을 둔다.

생물학적 프로세스의 전단계인 일차침전지 유출수(이하 초침 월류수)를 예측하기 위해 유입수의 유량, BOD, SS, TN, TP, 1차슬러지량, 하루전 SS로 회귀식을 추정한 후, 초침 월류수의 BOD(biological oxygen demand), COD(chemical oxygen demand), SS(suspended solid), TN(total nitrogen), TP(total phosphorous)의 실제치와 예측치를 비교하였다. 초침 월류수의 경우 공휴일에는 수질분석이 이루어지지 않아, 모델 구축에 실제 사용된 데이터는 공휴일을 제외한 188개이다.

Table 3.2와 Table 3.3에 회귀계수를 산정하는데 사용된 부산의 S-하수처리장의 2006년 유입수와 초침 월류수 수질자료가 나타나 있다.

Table 3.2 Water quality of S-WWTP influent in 2006

	유입량	수온	pH	DO	BOD	COD	SS	TN	TP	1차슬러지량	대장균
	m ³ /d	℃		mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	m ³ /d	개/mL
평균	249963	18.3	7.4	1.6	123.6	63.3	135.6	36.4	3.8	1022	138106
표준편차	46094	5.6	0.1	0.2	11.3	6.2	16.7	4.8	0.5	449	67398
최소값	186400	7.9	6.9	1.3	79.2	42.7	86.0	24.2	2.2	0	59000
중위수	237150	18.3	7.4	1.6	124.3	62.7	136.0	37.2	4.0	1146	97000
최대값	409300	29.1	7.8	2.5	158.9	78.0	168.0	46.8	5.0	1963	290000
왜도	1.7	0.0	0.2	0.4	-1.2	-0.2	-0.5	-0.4	-0.9	-0.4	0.7
첨도	2.6	-1.2	0.9	-0.4	3.0	0.6	0.3	-0.5	1.2	-0.7	-1.0

Table 3.3 Water quality of S-WWTP primary settling effluent in 2006

	BOD	COD	SS	TN	TP
	mg/L	mg/L	mg/L	mg/L	mg/L
평균	66.7	36.2	51.4	28.2	2.9
표준편차	8.2	5.7	5.6	3.9	0.4
최소값	34.6	23.8	32.0	16.1	1.7
중위수	68.1	37.4	52.0	28.7	3.0
최대값	83.8	47.8	66.0	37.3	3.7
왜도	-0.8	-0.4	-0.5	-0.6	-0.9
첨도	1.2	-1.0	1.0	0.1	0.6

3.2.2 모델의 검정

추정된 회귀식의 성립여부를 실제 데이터로 판단하여 통계적인 결정을 내리는 것을 검정이라 한다. 앞서 추정한 회귀식이 제대로 예측을 하는지 알아보는 과정으로 모델 구축에 사용하지 않은 새로운 데이터로 모델의 예측 능력을 평가한다.

모델의 검정에는 부산의 S-하수처리장의 2007년 수질자료 177개를 사용하였다. 다음 Table 3.4와 Table 3.5에 각각 유입수와 초침 월류수의 수질 특성이 나타나 있다.

Table 3.4 Water quality of S-WWTP influent in 2007

item	유입량	수온	pH	DO	BOD	COD	SS	TN	TP	1차슬러지량	대장균
	m ³ /d	℃		mg/L	mg/L	mg/L	mg/L	mg/L	mg/L	m ³ /d	개/mL
평균	233031	19.0	7.5	1.6	125.0	74.0	130.2	36.8	4.3	1257	166534
표준편차	36756	5.3	0.1	0.2	21.5	13.6	29.9	6.6	0.6	243	57401
최소값	175900	8.2	7.0	1.3	69.2	35.1	58.0	20.4	2.4	78	56000
중위수	222850	19.7	7.4	1.4	130.7	74.4	131.0	37.1	4.3	1251	184000
최대값	393000	28.2	7.8	2.0	174.1	107.8	212.0	50.2	6.1	1994	320000
왜도	2.0	0.0	0.3	0.6	-0.6	-0.3	-0.1	-0.2	-0.5	-0.68	-0.7
첨도	4.9	-1.3	0.1	-1.3	-0.3	0.3	-0.2	-0.9	0.6	3.8	-0.5

Table 3.5 Water quality of S-WWTP primary settling effluent in 2007

	BOD	COD	SS	TN	TP
	mg/L	mg/L	mg/L	mg/L	mg/L
평균	76.5	39.1	45.8	30.6	3.3
표준편차	16.0	5.8	9.9	4.9	0.6
최소값	34.4	20.1	23.0	15.1	1.8
중위수	79.7	40.4	46.0	30.9	3.3
최대값	113.0	50.9	77.0	40.2	4.4
왜도	-0.4	-0.7	0.1	-0.7	-0.5
첨도	-0.4	0.3	-0.4	0.7	-0.2

3.2.3 모델의 적합성 평가

추정된 회귀식이 실측치를 얼마나 잘 설명하고 있으며, 회귀식을 이용하여 예측할 경우 어느 정도 신뢰성이 있는지를 검토하는 척도로서 RMSE(root mean square error, 제곱근-평균-제곱오차)를 구했다.

RMSE는 각각의 오차값들이 갖는 오차의 평균량을 평가하는 방법으로, 계산된 RMSE 값이 크면 실제값과 예측값의 차, 즉 오차가 더 크다는 것을 의미한다.

$$RMSE = \sqrt{\frac{\sum_{n=1}^N (Y_M - Y_P)^2}{N}}$$

Y_M : 실측치

Y_P : 예측치

N : 데이터 수

제4장 실험결과 및 고찰

4.1 회귀분석모델의 추정

4.1.1 회귀계수 산정

회귀계수를 산정함에 있어 입력변수를 결정하기 위해 하수처리장의 유입수와 초침 월류수의 행렬도를 Fig 4.1에 나타내었다. 행렬도를 통해 인자들 사이의 연관성을 파악할 수 있으며, 일정할수록 영향을 많이 주는 인자라 할 수 있겠다. 이를 바탕으로 유입수의 유량, BOD, SS, TN, TP, 1차슬러지량, 하루전 SS를 입력 변수로 사용하였다.

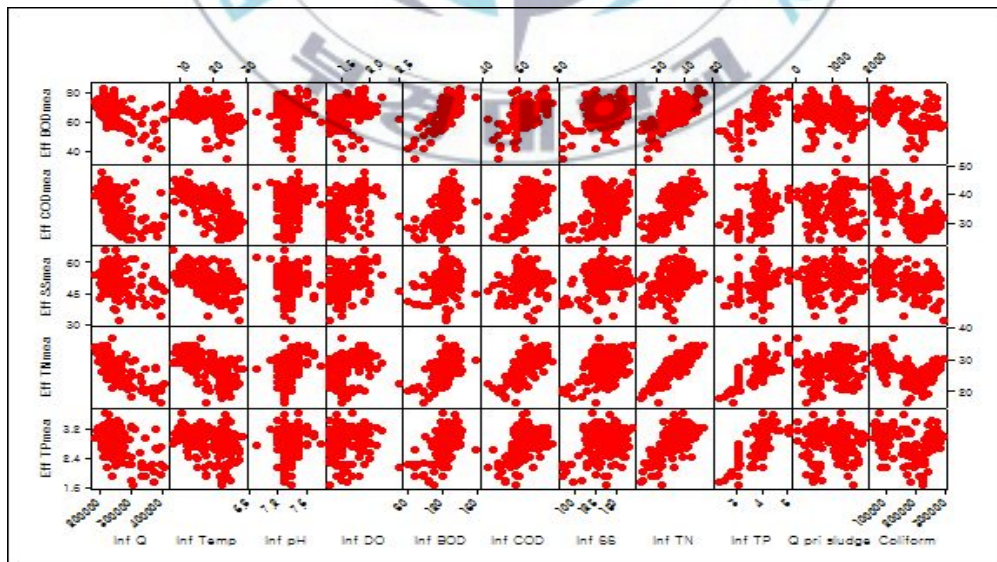


Fig. 4.1 Biplot between the influent and the primary settling effluent

2006년 유입수 자료를 바탕으로 생성된 회귀계수와 분산분석표를 Table 4.1~4.10에 나타내었다. 회귀식은 항과 계수의 곱을 더해서 산정하였으며, 일차침전지 모델 초침 월류수의 BOD, COD, SS, TN, TP의 회귀식이 각각 구해졌다.



Table 4.1. Analysis dispersion table of BOD

출처	DF	SS	adj SS	MS	F	P
회귀분석	35	9861.3	9861.31	281.752	11.21	0
선형	7	8494.2	2420.61	345.801	13.75	0
제곱	7	449.6	618.48	88.355	3.51	0.002
상호작용	21	917.5	917.54	43.692	1.74	0.031
잔차오차	150	3771.3	3771.27	25.142		
전체	185	13632.6				

Table 4.2. A regression coefficient of BOD

항	계수	항	계수
상수	90.641	유입유량*유입BOD	1.67E-06
유입유량	-3.42E-05	유입유량*하루전SS	-4.16E-06
유입BOD	-0.892262	유입유량*유입SS	1.64E-06
하루전SS	2.47781	유입유량*유입TN	-4.18E-07
유입SS	-0.69978	유입유량*유입TP	-2.82E-05
유입TN	-3.76997	유입유량*1차슬러지량	-2.45E-08
유입TP	24.2898	유입BOD*하루전SS	0.00601121
1차슬러지량	0.0215952	유입BOD*유입SS	-0.00189443
유입유량*유입유량	-1.08E-10	유입BOD*유입TN	0.013817
유입BOD*유입BOD	0.0037237	유입BOD*유입TP	-0.116453
하루전SS*하루전SS	-0.00799712	유입BOD*1차슬러지량	-1.82E-04
유입SS*유입SS	-0.00117282	하루전SS*유입SS	-0.00113425
유입TN*유입TN	-0.0577666	하루전SS*유입TN	0.00481162
유입TP*유입TP	-1.61957	하루전SS*유입TP	-0.246208
1차슬러지량*1차슬러지량	-1.82E-06	하루전SS*1차슬러지량	-3.13E-04
		유입SS*유입TN	0.0198508
		유입SS*유입TP	-0.0178446
		유입SS*1차슬러지량	0.000145259
		유입TN*유입TP	0.873581
		유입TN*1차슬러지량	0.000389268
		유입TP*1차슬러지량	-0.00236004

Table 4.3 Analysis dispersion table of COD

출처	DF	SS	adj SS	MS	F	P
회귀분석	35	4499.54	4499.54	128.558	9.15	0
선형	7	3812.48	762	108.857	7.75	0
제곱	7	44.66	44.48	6.355	0.45	0.867
상호작용	21	642.4	642.4	30.591	2.18	0.004
잔차오차	150	2106.41	2106.41	14.043		
전체	185	6605.95				

Table 4.4 A regression coefficient of COD

항	계수	항	계수
상수	49.949	유입유량*유입BOD	-2.66E-07
유입유량	-2.64E-04	유입유량*하루전SS	1.03E-06
유입BOD	0.829775	유입유량*유입SS	1.79E-06
하루전SS	0.443718	유입유량*유입TN	-4.84E-06
유입SS	-0.781829	유입유량*유입TP	-5.92E-06
유입TN	3.10033	유입유량*1차슬러지량	4.70E-08
유입TP	-17.556	유입BOD*하루전SS	-0.00687712
1차슬러지량	-0.0315667	유입BOD*유입SS	0.00422892
유입유량*유입유량	2.12E-10	유입BOD*유입TN	0.00284467
유입BOD*유입BOD	-0.00151686	유입BOD*유입TP	-0.192013
하루전SS*하루전SS	-0.00657721	유입BOD*1차슬러지량	1.80E-05
유입SS*유입SS	9.35E-05	하루전SS*유입SS	-0.0078004
유입TN*유입TN	0.00589006	하루전SS*유입TN	0.0064792
유입TP*유입TP	4.43507	하루전SS*유입TP	0.447369
1차슬러지량*1차슬러지량	3.89E-07	하루전SS*1차슬러지량	-1.38E-05
		유입SS*유입TN	-0.00102133
		유입SS*유입TP	0.0876092
		유입SS*1차슬러지량	-1.41E-04
		유입TN*유입TP	-0.781761
		유입TN*1차슬러지량	0.000656075
		유입TP*1차슬러지량	0.00281817

Table 4.5 Analysis dispersion table of SS

출처	DF	SS	adj SS	MS	F	P
회귀분석	35	3251.3	3251.3	92.8942	4.28	0
선형	7	1982.56	305.28	43.6117	2.01	0.057
제곱	7	262.62	193.39	27.6267	1.27	0.267
상호작용	21	1006.11	1006.11	47.9101	2.21	0.003
잔차오차	150	3254.9	3254.9	21.6993		
전체	185	6506.2				

Table 4.6 A regression coefficient of SS

항	계수	항	계수
상수	-208.57	유입유량*유입BOD	-2.28E-06
유입유량	0.000456261	유입유량*하루전SS	-4.38E-06
유입BOD	2.42321	유입유량*유입SS	-5.87E-07
하루전SS	0.654179	유입유량*유입TN	9.69E-06
유입SS	0.687503	유입유량*유입TP	-5.78E-05
유입TN	-0.230759	유입유량*1차슬러지량	4.08E-08
유입TP	-9.12819	유입BOD*하루전SS	-2.92E-04
1차슬러지량	-0.0100262	유입BOD*유입SS	0.00381879
유입유량*유입유량	-5.55E-11	유입BOD*유입TN	-0.00914837
유입BOD*유입BOD	-0.00575157	유입BOD*유입TP	-0.129228
하루전SS*하루전SS	-8.72E-04	유입BOD*1차슬러지량	-1.53E-04
유입SS*유입SS	0.000196222	하루전SS*유입SS	-0.0110845
유입TN*유입TN	0.092722	하루전SS*유입TN	-0.0206066
유입TP*유입TP	4.51178	하루전SS*유입TP	0.745373
1차슬러지량*1차슬러지량	-2.58E-06	하루전SS*1차슬러지량	0.000239296
		유입SS*유입TN	-0.0196854
		유입SS*유입TP	0.051013
		유입SS*1차슬러지량	7.08E-05
		유입TN*유입TP	-1.0231
		유입TN*1차슬러지량	0.000199079
		유입TP*1차슬러지량	-0.00189991

Table 4.7 Analysis dispersion table of TN

출처	DF	SS	adj SS	MS	F	P
회귀분석	35	2201.74	2201.737	62.9068	18.97	0
선형	7	2043.73	491.218	70.174	21.16	0
제곱	7	85.04	26.881	3.8401	1.16	0.331
상호작용	21	72.97	72.971	3.4748	1.05	0.411
잔차오차	150	497.52	497.522	3.3168		
전체	185	2699.26				

Table 4.8 A regression coefficient of TN

항	계수	항	계수
상수	51.8289	유입유량*유입BOD	4.82E-07
유입유량	-1.64E-04	유입유량*하루전SS	8.80E-07
유입BOD	-0.480227	유입유량*유입SS	2.16E-07
하루전SS	-0.219435	유입유량*유입TN	-2.68E-06
유입SS	0.156143	유입유량*유입TP	1.85E-05
유입TN	-0.626497	유입유량*1차슬러지량	2.29E-08
유입TP	17.796	유입BOD*하루전SS	0.00364242
1차슬러지량	-0.0158432	유입BOD*유입SS	-1.04E-04
유입유량*유입유량	2.35E-11	유입BOD*유입TN	0.0082946
유입BOD*유입BOD	0.00255475	유입BOD*유입TP	-0.211125
하루전SS*하루전SS	0.00195133	유입BOD*1차슬러지량	7.91E-05
유입SS*유입SS	-0.00146356	하루전SS*유입SS	-5.67E-04
유입TN*유입TN	-0.0321651	하루전SS*유입TN	0.00519951
유입TP*유입TP	1.68159	하루전SS*유입TP	-0.166383
1차슬러지량*1차슬러지량	9.15E-07	하루전SS*1차슬러지량	-1.55E-04
		유입SS*유입TN	0.0112029
		유입SS*유입TP	-0.0453257
		유입SS*1차슬러지량	-3.10E-05
		유입TN*유입TP	0.220323
		유입TN*1차슬러지량	0.000429018
		유입TP*1차슬러지량	-0.00112256

Table 4.9 Analysis dispersion table of TP

출처	DF	SS	adj SS	MS	F	P
회귀분석	35	20.7844	20.78439	0.59384	9.59	0
선형	7	18.583	5.61314	0.801877	12.95	0
제곱	7	1.1547	0.22081	0.031544	0.51	0.826
상호작용	21	1.0467	1.04672	0.049844	0.81	0.71
잔차오차	150	9.2858	9.2858	0.061905		
전체	185	30.0702				

Table 4.10 A regression coefficient of TP

항	계수	항	계수
상수	-1.19933	유입 유량*유입BOD	-6.81E-08
유입 유량	5.12E-06	유입 유량*하루 전SS	1.26E-07
유입BOD	0.0410394	유입 유량*유입SS	-1.80E-08
하루 전SS	-0.133686	유입 유량*유입TN	-7.96E-08
유입SS	0.0179249	유입 유량*유입TP	1.82E-06
유입TN	0.196618	유입 유량*1차슬러지량	9.42E-10
유입TP	-0.966065	유입BOD*하루 전SS	0.000358507
1차슬러지량	0.000196734	유입BOD*유입SS	2.20E-06
유입 유량*유입 유량	-1.11E-11	유입BOD*유입TN	0.000575033
유입BOD*유입BOD	-6.02E-05	유입BOD*유입TP	-0.0103976
하루 전SS*하루 전SS	0.000610608	유입BOD*1차슬러지량	-5.13E-06
유입SS*유입SS	-4.39E-05	하루 전SS*유입SS	6.65E-05
유입TN*유입TN	0.00266642	하루 전SS*유입TN	-0.00206941
유입TP*유입TP	0.200224	하루 전SS*유입TP	0.0176932
1차슬러지량*1차슬러지량	2.39E-08	하루 전SS*1차슬러지량	-1.91E-06
		유입SS*유입TN	-0.00111229
		유입SS*유입TP	0.00906337
		유입SS*1차슬러지량	-2.12E-06
		유입TN*유입TP	-0.0428277
		유입TN*1차슬러지량	-2.47E-06
		유입TP*1차슬러지량	0.000128115

4.1.2 추정된 회귀분석 모델과 실측값 비교

추정된 회귀분석이 실제값을 얼마나 잘 표현하고 있는지는 검토하기 위해 RMSE를 구하였다. 각 수질항목의 RMSE를 Table 4.11 나타내었다.

Table 4.11 Estimation result of RMSE between the measured value and predictive value

	BOD	COD	SS	TN	TP
RMSE	3.81	2.74	4.7	4.0	0.2

RMSE가 0일 경우, 회귀식과 실제값이 100% 일치하며, RMSE가 작을수록 회귀식이 실제값을 잘 표현한다고 본다. Table 4.11을 보면 TP의 RMSE가 0.2로 가장 작은 값을 나타내 회귀식이 실제값과 가장 가깝게 표현되었으며, SS의 경우 4.7로 비교적 높은 값을 나타냈다.

Fig 4.2~4.6에 회귀식과 실제값을 비교한 결과를 그래프로 나타내었다.

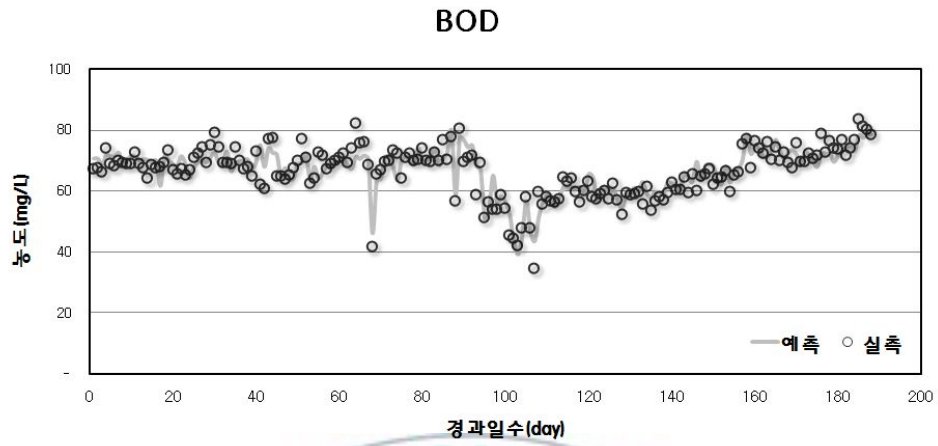


Fig. 4.2 Estimation result of BOD

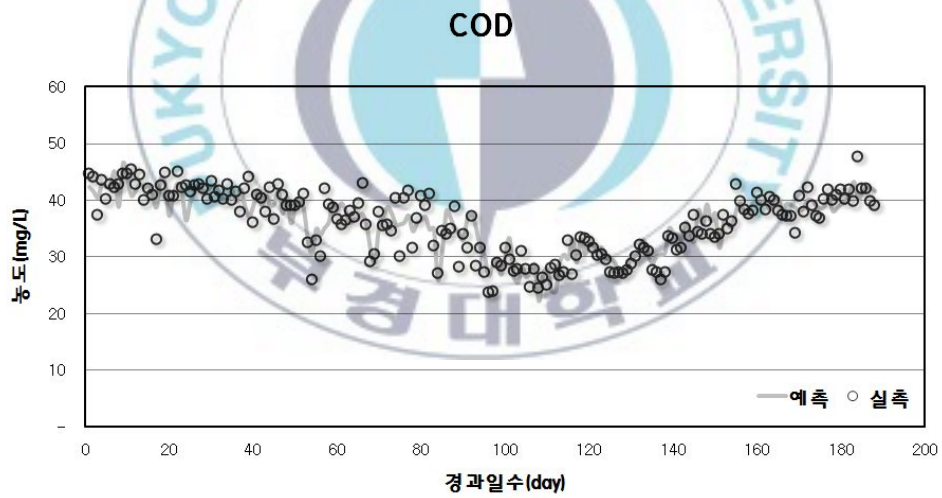


Fig. 4.3 Estimation result of COD

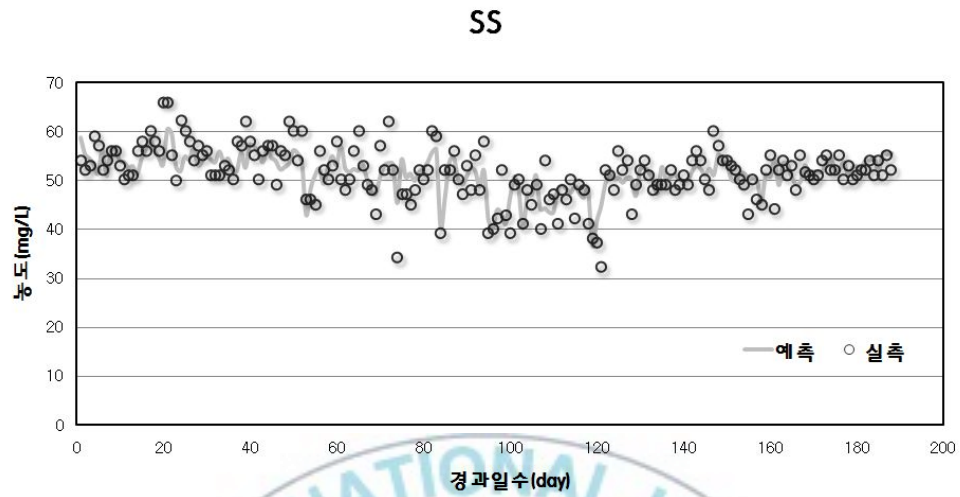


Fig. 4.4 Estimation result of SS

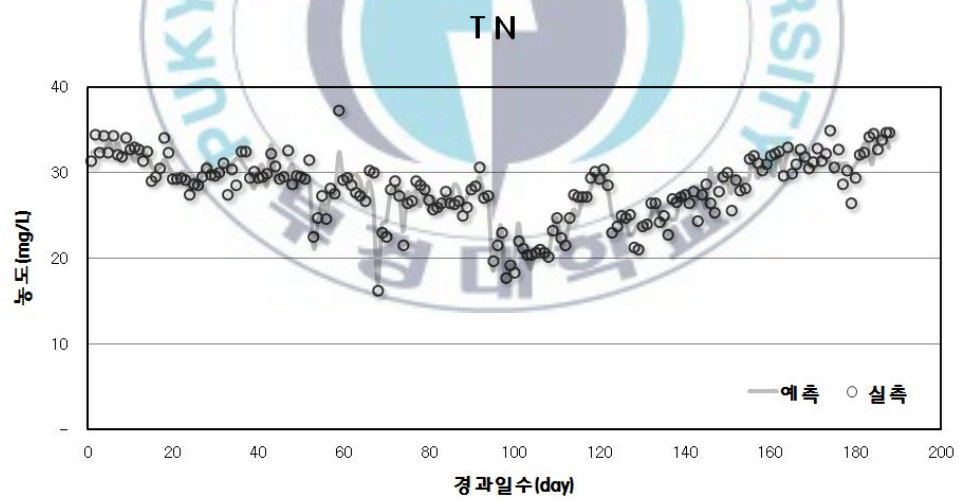


Fig. 4.5 Estimation result of TN

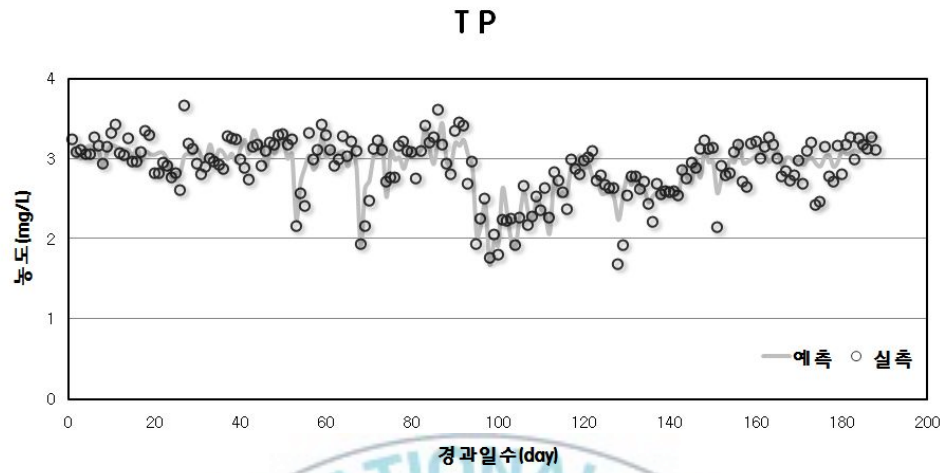


Fig. 4.6 Estimation result of TP



4.2 회귀분석모델의 검증

앞서 추정한 회귀식의 검정을 위해 실제 부산의 S-하수처리장의 2007년 초침 월류수 수질 자료와 회귀식을 비교하였다. 모델의 신뢰성을 판단하기 위해 Table 4.12에 회귀식과 2007년 초침 월류수와의 RMSE를 나타내었다.

Table. 4.12 Validation result of RMSE between the measured value and predictive value

	BOD	COD	SS	TN	TP
RMSE	13.24	6.47	9.92	4.70	0.46

Table 4.12을 보면 앞서 추정할 때와 마찬가지로 TP의 RMSE가 가장 낮은 값을 나타내어 검증에 있어서도 TP의 회귀식이 초침 월류수를 가장 잘 표현함을 알 수 있었다. 반면에 BOD 경우 가장 높은 13.24로 나타났다.

Fig 4.7~4.11에 회귀식과 실제값을 비교한 결과를 그래프로 나타내었다.

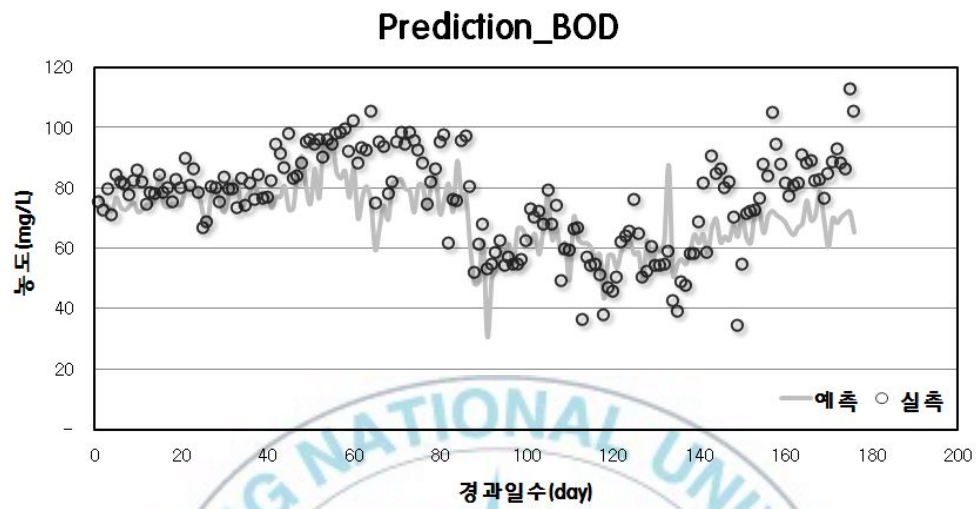


Fig 4.7 Validation result of BOD

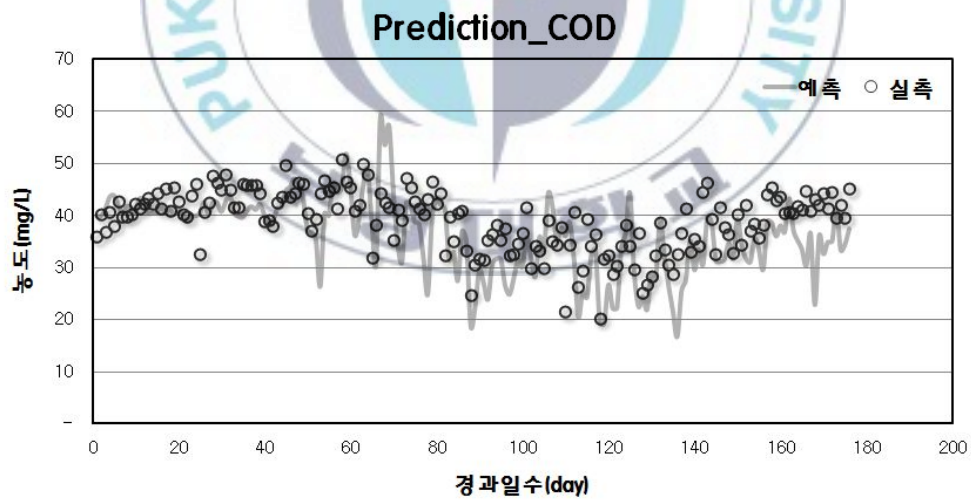


Fig 4.8 Validation result of COD

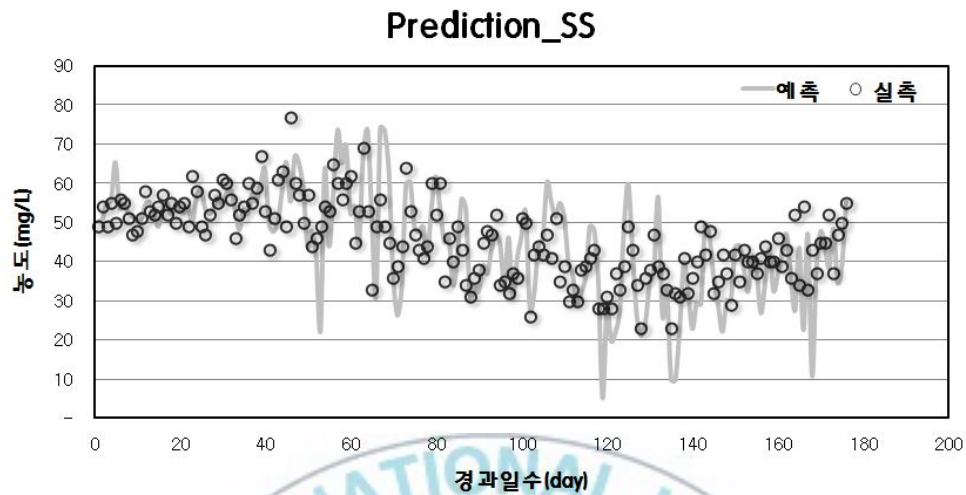


Fig 4.9 Validation result of SS

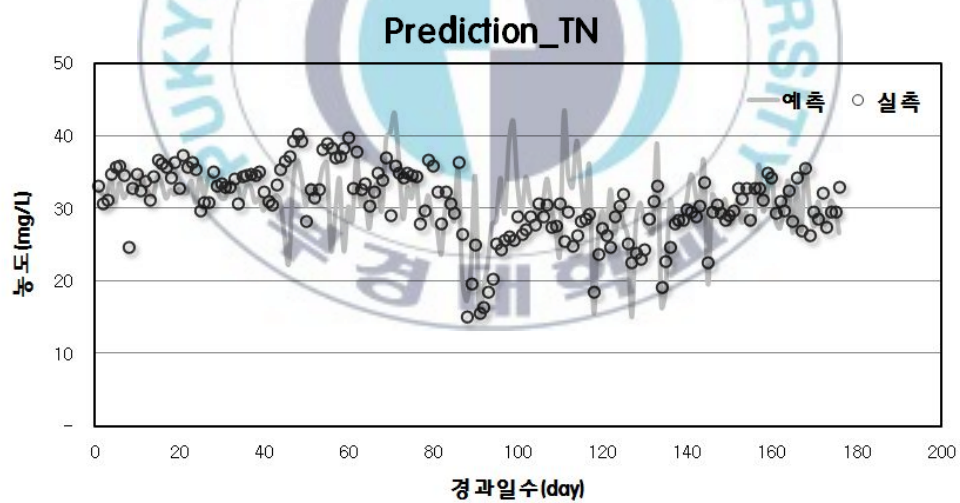


Fig 4.10 Validation result of TN

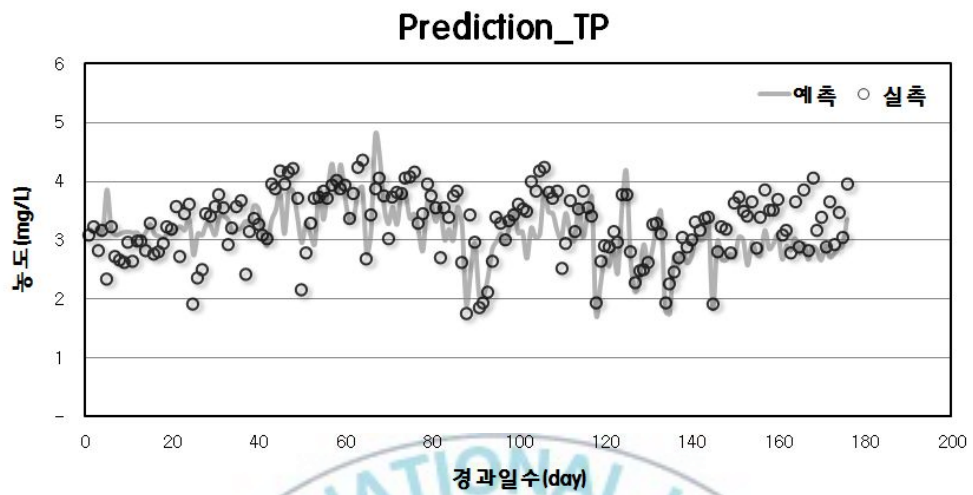
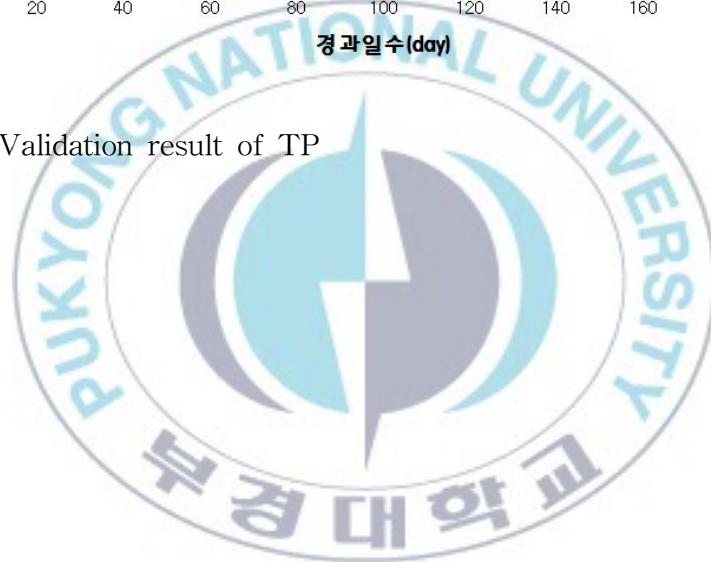


Fig 4.11 Validation result of TP



제5장 결론

다변량 회귀분석에 의한 일차침전지 유출수 수질예측 시뮬레이션에 대한 결론은 다음과 같다.

1. 표면반응 분석을 이용한 다변량 회귀분석으로 유입수의 유량, BOD, SS, TN, TP, 1차슬러지량, 하루전 SS를 이용한 일차침전지 유출수수질을 재현할 수 있는 회귀 모델식을 만들었다.
2. 유출수 BOD, COD, SS, TN, TP의 RMSE(root mean square error)는 각각 3.81, 2.74, 4.7, 4, 0.2 이었다.
3. 2006년 자료로 추정된 회귀모델은 2007년 자료로 검정한 결과 실측값을 잘 재현할 수 있었다.

제6장 참고문헌

Yong_Na Zhao,(2008). Prediction of Wastewater Treatment Plant Performanced using Artificial Neural Network Model. Department of Environmental Engineering. M. A. these, Pukyong National University, Korea

Kim,. Y. J.(2006). Development of inference models and diagnosis algorithms using intelligent methods for sequencing batch reactor operation. Ph. D. these, Pusan National University, Korea

김주아(2001). 선형회귀모형에 관한 연구, 연세대 교육대학원 : 석사논문

최재훈, 문태섭, 김예진, 김성희, 김효수, 김창원(2007). Black-box model을 이용한 하수처리장 유입 유량 및 성상의 예측에 관한 연구, 대한환경공학회 추계학술연구발표회 논문집, Vol.2007, No.2, pp.624~631

문태섭, 최재훈, 김성희, 차재환, 염훈식, 김창원(2008). 인공 신경망(ANN)에 의한 하수처리장의 유입 유량 및 유입 성분 농도의 예측, Journal of Korean Society on Water Quality, Vol.24, No.1, pp.91~98.

최재훈, 문태섭, 김예진, 김성희, 김효수, 김창원(2007). 인공신경망을 이용한 하수처리장 유입수 및 공정성능 예측모델 개발, 대한상하수도학회 공동

추계 학술발표회 논문집, Vol.2007, No.0, pp.116~126.

이중원, 서창원, 홍윤석, 신항식(2007). Nonlinear/non-Gaussian Bayesian state estimation for wastewater treatment plant, 대한환경공학회 학술연구발표회, Vol.2007, No.2, pp.639~646

김두섭·강남준, {회귀분석 기초와 응용}, 나남출판, 2000

왕창근·서정운·신재경·구자용·성낙창·이영형 공역, {수질데이터의 통계적 해석}, 동아출판, 2000

전홍석·이재준·고광백 공역, {환경통계학}, 동아출판, 1997

박준선, {새 MINITAB 실무완성}, (주)이레테크, 2005

