Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 4, pp. 847 ~ 857

A Study on Error Detection Algorithm of COD Measurement Machine[1)]

Hyun Seok Choi[2)] · Gyu Moon Song[3)] · Tae Yoon Kim[4)]

Abstract

This paper provides a statistical algorithm which detects COD (chemical oxygen demand) measurement machine error on real-time. For this we propose to use regression model fitting and check its validity against the current observations. The main idea is that the normal regression relation between COD measurement and other parameters inside the machine will be violated when the machine is out of order.

Keywords : COD, Machine Error Detection, Regression Model

1. 서론

COD(Chemical Oxygen Demand, 화학적 산소요구량)는 수중의 유기물질 함유량을 측정하기 위한 간접지표로서 산화제를 이용하여 단시간 내에 유기물질을 산화시키는 방법이며, 수중의 유기물이 산화제에 의해 분해될 때 발생하는 산소량으로 측정된다. 단시간 내에 측정 가능하다는 장점 때문에 BOD(Biochemical Oxygen Demend: 생물학적 산소요구량)량을 예측하거나 생물학적 난분해성 유기물질 혹은 미생물활동에 영향을 주는 독성물질이 함유된 경우 유기물 질량을 측정하는데 주로 사용된다 ([1],[2]). 특히 COD는 수질환경기준에 흔히 사용되는데 그 기준은 [표 1]과 같다. 이와 같은 다양한 장점들로 인해 COD는 실시간 수질 오염여부 확인을 위해 필수적으로 사용해야 할 지표로 간주되고 있다.

---

1) This research was supported by the Program for the Training of Graduate Students in Regional Innovation which was conducted by the Ministry of Commerce Industry and Energy of the Korean Government

2) Full time instructor, Department of Statistics, Keimyung University, Daegu, 704-701, Korea
E-mail: chsuk1@kmu.ac.kr

3) Professor, Department of Statistics, Keimyung University, Daegu, 704-701, Korea
E-mail: kms252@kmu.ac.kr

4) Professor, Department of Statistics, Keimyung University, Daegu, 704-701, Korea
E-mail: tykim@kmu.ac.kr

[표 1] 수질 환경기준

등급	I	II	III	IV	V
적용대상	상수원수1급	상수원수2급 수산용수1급	상수원수3급 수산용수2급 공업용수1급	공업용수2급 농업용수	공업용수3급
COD(*mg*/*l*)	1이하	3이하	6이하	8이하	10이하

하천의 수질오염문제는 인간의 건강뿐만 아니라 자연생태계의 보전차원에서도 매우 중요한 과제이다. 최근 낙동강, 한강, 영산강 등을 중심으로 국내 여러 하천에 수질 감시를 위한 자동 측정기가 설치되고 있는데 이 경우 COD 자동 측정기는 반드시 포함되고 있다. 본 연구에서는 실시간 COD 자동 측정기의 기계적 오류가 발생할 경우 이를 탐지해낼 수 있는 알고리즘을 구축한다. 이와 같은 알고리즘이 중요한 이유는 하천의 특정 위치에 설치된 COD 측정기의 COD값이 높게 나타날 경우 그 원인이 오염물질 유입 등에 있을 수도 있지만 자동 측정기 내부 오류로 인해 높게 나타날 수도 있기 때문이다. 이와 같이 기계적 오류인지를 올바르게 신속하게 판단하는 문제는 자동 측정기를 제작하거나 수입하여 판매하는 회사들뿐만 아니라 이를 유지 관리하는 관련 감시기관들에게도 매우 중요한 과제가 되고 있다.

본 논문은 다음과 같이 구성되어 있다. 2절에서는 알고리즘 구축에 대하여 기술하고, 3절에서는 실제 데이터의 분석을 통한 알고리즘 적용에 대하여 기술하며 4절에서 결론을 맺는다.

2. 알고리즘

Y_t를 t시점의 COD 값이며 $X_{t1}, X_{t2}, \cdots, X_{tp}$를 Y_t 와 동시에 관찰 가능한 기계적 내부값이라 하자. 여기서 기계적 내부값이란 자동 측정기내에서 COD 값을 측정하여 계산하는데 필요한 값이거나 COD 측정기와 동시에 측정되는 관련 변수들이다. COD 측정기의 기계적 측정 오차를 탐지하는 절차는 기본적으로 $(X_{t1}, X_{t2}, \cdots, X_{tp})$와 Y_t와의 관계를 분석한 후 실제 관찰값이 기존의 분석된 관계에서 크게 벗어날 경우 그 값을 기계적 측정 오류에 의한 것으로 판단한다. 이러한 절차를 구체적으로 기술하면 다음과 같다.

먼저 $(X_{11}, X_{12}, \cdots, X_{1p}, Y_1), \cdots, (X_{n1}, X_{n2}, \cdots, X_{np}, Y_n)$ 등 n개의 데이터가 관찰되었으며

$$Y_t = g(X_{t1}, X_{t2}, \cdots, X_{tp}) + \nu_t, \qquad t = 1, \cdots, n \tag{1}$$

이라는 회귀함수 모형(regression model)이 성립한다고 하자. 여기서 g는 회귀함수, ν_t는 자기회귀 오차로서 다음과 같이 정의된다.

$$\nu_t = \epsilon_t - \phi_1 \nu_{t-1} - \cdots - \phi_m \nu_{t-m}$$

단 ϵ_t는 동일하고 독립적인 분포를 갖는 오차 (iid 오차)이다.

모형 (1)에서 iid 오차 대신 상관된 오차 모형을 사용하는 이유는 $(X_{11}, X_{12}, \cdots, X_{1p}, Y_1), \cdots, (X_{n1}, X_{n2}, \cdots, X_{np}, Y_n)$ 들이 순차적으로 관찰되므로 상관될 가능성이 높기 때문이다. 모형 (1)을 이용하여 기계적 오류 측정값을 탐지하는 절차는 다음과 같다.

(i) 주어진 데이터를 이용하여 회귀 함수 $\hat{g}$ 및 $\hat{\nu}$를 추정한다.

(ii) $t = 1, \cdots, n$에 대해 잔차 $e_t = Y_t - \hat{g}(X_{t1}, X_{t2}, \cdots, X_{tp}) - \widehat{\nu_t}$ 를 구하여 그 분포 F_e를 $\widehat{F_e}$로 추정한다.

(iii) T시점의 새로운 관찰치 $(X_{T1}, \cdots, X_{Tp}, Y_T)$에 대해
$e_T = Y_T - \hat{g}(X_{T1}, \cdots, X_{Tp}) - \widehat{\nu_T}$ 및 $\widehat{F_e}$ 를 이용하여 p-value를 계산한다.
여기서 p-value는 $P(e > e_T | F_e)$이다. 만일 p-value가 사전에 주어진 유의수준보다 작으면 기계적 오류라고 판단한다.

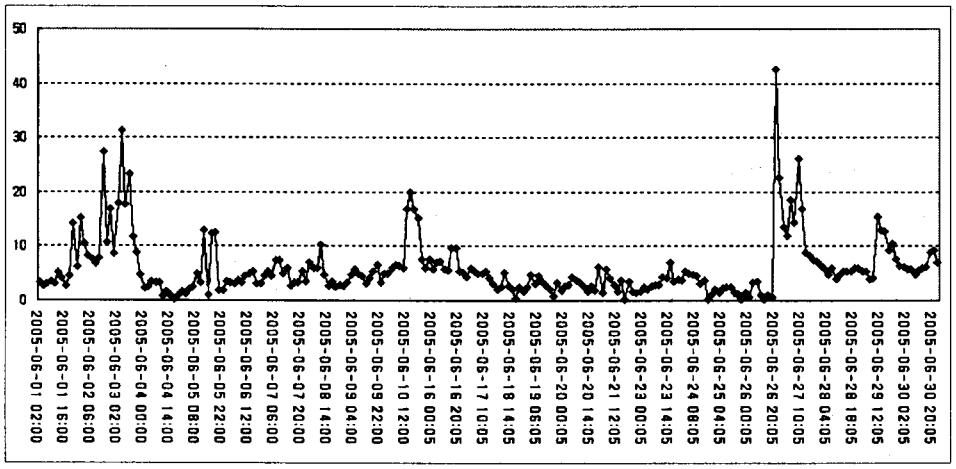
위의 알고리즘에서 $\hat{g}$를 구하는 과정에서 모수적 혹은 비모수적 회귀분석 방법을 모두 사용할 수 있으나 모수적 모형을 선호하는데 그 이유는 비모수 회귀분석을 사용하면 관찰값이 없는 지역에서의 적합이 어렵기 때문에 새로운 관찰치가 그러한 지역에서 발생할 경우 기계적 오류로 잘못 판단할 가능성이 높기 때문이다. 이는 비모수 기법을 사용하여 밀도 추정을 하면 작은 $\alpha > 0$에 대해 α백분위수와 $(1-\alpha)$백분위수를 효율적으로 추정하기 어렵다는 사실과 연관되어 있다 ([7]). 또한 반드시 요구되는 사항은 아니지만 적합 후 잔차 $e_1, \cdots, e_n$이 가능하면 정규분포 iid가 되도록 적합한다. 이는 적합이 제대로 이루어졌다는 것을 뜻할 뿐만 아니라 그 경우 p-value를 구하기가 용이하다.

3. 알고리즘의 실제 COD 데이터 적용

본 절에서 사용된 자료는 서울 근교의 하천 A의 특정 지역 B에 설치된 자동 측정기로부터 관찰된 자료로써 동시 측정변수로는 NH4 (암모늄이온), NO2 (이산화질소), NO3 (질산염), PO4 (인산), TN (총질소), TP(총인) 등이 포함되어 있다.

관측기간은 2005년 6월1일에서 6월30일까지이며, 관찰 주기는 2시간이다. 이 기간의 COD 자료에 대한 평상적인 움직임을 모형화하기 위해 COD (Y_t)의 시도표([그림 1]) 를 살펴보면 이상치들이 다수 존재하며 정상 시계열(stationary time series)로 보는데 다소 무리가 있는 것으로 판단된다 ([3],[4]). 여기서 [그림 1]의 시도표는 총 측정자료 251개 중 결측치, 중복된 값, 음수 값을 제외한 240개에 대한 시도표이다. 물론 2시간 측정간격일 때 원 자료의 수가 30일×12=360개가 되어야 하나 109개의 자료가 원 자료로부터 기계 작동 중지 등에 의해 유실된 상태이다. 본 논문에는 포함되지 않았으나

유실된 데이터의 복구를 위해 보간법 등을 시도하여 본 결과, 복구 이전이나 복구 후에 시계열 분석결과가 크게 다르지 않았다. 이는 데이터 복구가 유실된 데이터의 시계열 움직임이 기존의 데이터와 크게 다르지 않다는 전제하에 이루어진 결과 때문인 듯하며 이와 같은 이유로 데이터 복구 절차를 생략하였다. 참고로 유실된 데이터의 시계열 움직임이 기존의 데이터와 크게 다르다고 가정하는 경우 추가적인 정보가 없는 한 데이터 복구는 불가능한 것으로 판단된다.



[그림 1] COD (Y_t)의 시도표

COD기기 오작동 탐지 알고리즘을 구축하기 전에 하천 A의 주어진 지점의 COD값들의 움직임에 대한 시계열 특성 및 그 분포 특성을 살펴보았다. 먼저 적절한 시계열 모형 선택에 앞서 하천 A의 B지점에서 COD 값에 주야 효과(day and night effect)가 있을 수 있다는 판단에 의해 주야 효과를 제거하기 위한 절차를 수행하였다. 이 효과는 해당 지점에서 오염 물질의 투기가 주로 밤에 이루어진다는 사실에 근거한 것이다. 그 결과는 [표 2]에 주어져 있다.

[표 2] 밤과 낮의 COD 평균과 표준편차

	전체	밤	낮
평균	5.82	6.25	5.21
표준편차	5.44	5.95	4.62

$$Z_t = Y_t - 6.25 I_t,\ t \in \text{밤}$$

$$Z_t = Y_t - 5.21 I_t,\ t \in \text{낮}$$

Z_t에 대한 표본 자기상관계수(sacf) 및 편자기 상관 계수(pacf)를 살펴보면 ([그림

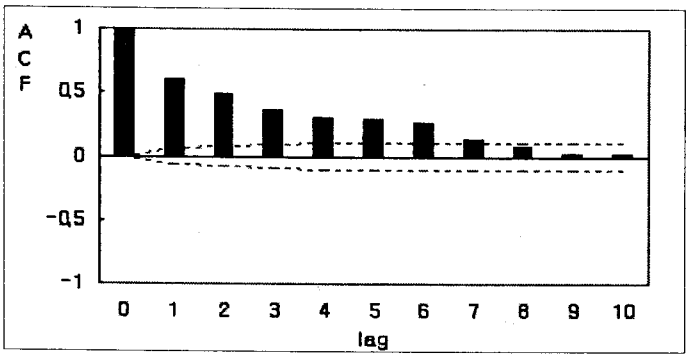
2]와 [그림 3]) ACF는 빠르게 감소하는 반면, PACF는 3시점이후 절단점을 갖게 되므로 AR(2)모형을 선택하였다 ([3],[4]). 적합한 결과 ([표 3])는 아래와 같다.

$$\hat{Z}_t = 0.47618 Z_{t-1} + 0.21003 Z_{t-2} \tag{2}$$

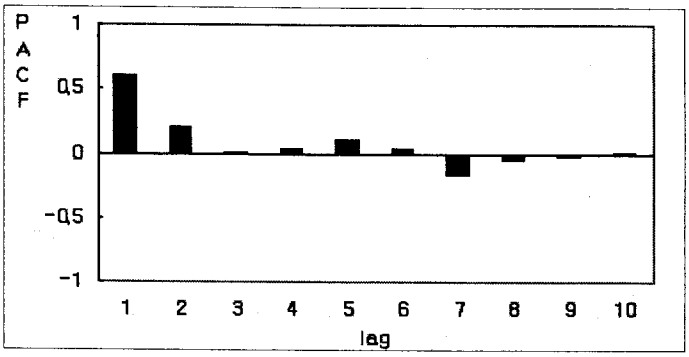
[표 3]과 [그림 4]를 통해 (2)의 적합이 적절히 이루어 졌음을 확인할 수 있다. 이제 식(2)를 이용하면 $\hat{Y}_t$는 다음과 같이 추정된다.

$$\hat{Y}_t = \hat{Z}_t + 6.25 I_t, \quad t \in \text{밤}$$

$$\hat{Y}_t = \hat{Z}_t + 5.21 I_t, \quad t \in \text{낮}$$



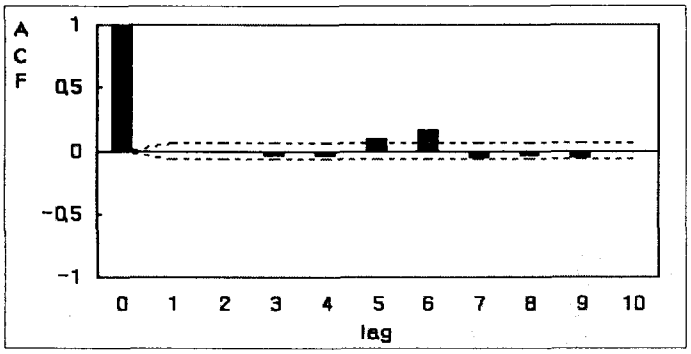
[그림 2] Z_t의 ACF



[그림 3] Z_t의 PACF

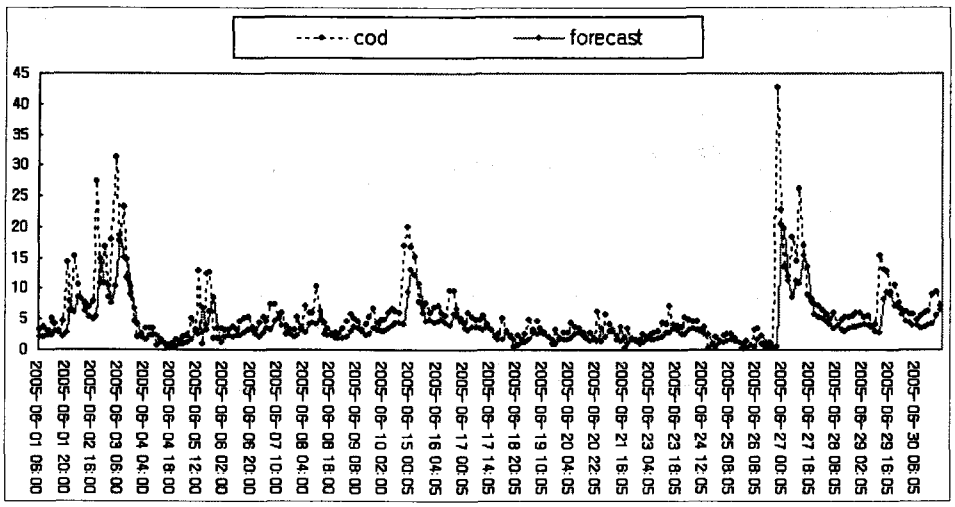
[표 3] 모형식 추정

Parameter	Estimate	Standard error	t-value	p-value
μ	-0.07268	0.85196	-0.09	0.9321
ϕ_1	0.47618	0.06352	7.50	0.0000
ϕ_2	0.21003	0.06356	3.30	0.0011

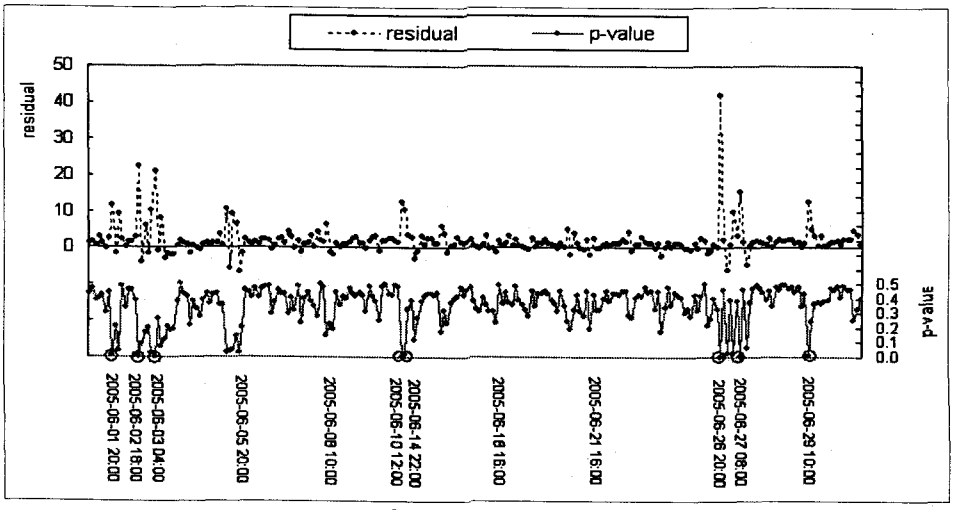


[그림 4] $\hat{Z}_t - Z_t$의 ACF

[그림 5]와 [그림 6]에 식(2)를 사용한 예측값 $\hat{Y}_t$와 실제값 Y_t 및 해당 잔차 ($\hat{Y}_t - Y_t$)들의 p-value가 주어져 있다. p-value가 0.01이하 시점이 [표 5]의 첫 번째 열 (시계열 모형)에 나타나 있다.



[그림 5] Y_t(cod값)과 $\hat{Y}_t$(시계열 모형의 예측값)



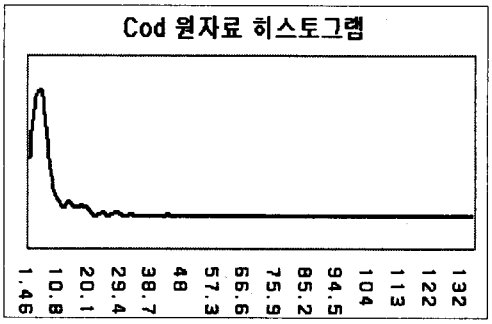
[그림 6] $Y_t - \hat{Y}_t$ (residual값) 과 해당 p-value

하천 A의 B지점에서 COD(Y_t)의 분포적 특성을 살펴보기 위해 다음과 같은 절차를 사용하였다. 먼저 Y_t의 히스토그램을 그려보면 ([그림 7]) 다수의 이상점 존재하므로 Y_t의 분포를 꼬리가 두터운 모수 m과 $b>0$인 로지스틱 확률분포 $f_{m,b}$로 추정하는 것이 타당한 것으로 판단된다([5],[6]). 여기서

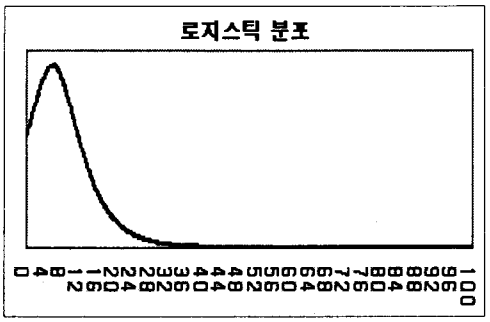
$$f_{m,b}(y) = \frac{e^{-[y-m]/b}}{b[1+e^{-[y-m]/b}]^2}$$

이며 추정된 $\hat{m}$=5.82159 이고 $\hat{b}$=5.4273 이다.

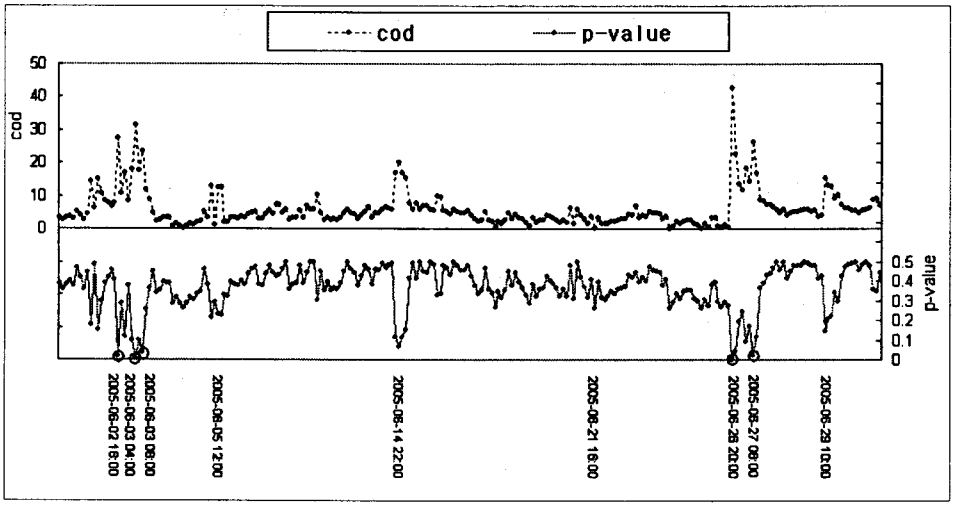
[그림 8]에 추정된 로지스틱 밀도함수의 그림이 주어져 있다. $\hat{f}_{m,b}$에 의한 p-value는 [그림 9]에 주어져 있으며 특히 p-value가 0.01이하인 시점들은 [표 5]의 세 번째 열 (밀도함수 모형)에 나타나 있다.



[그림 7] COD (Y) 히스토그램



[그림 8] Y (COD)의 추정된 로지스틱 밀도함수



[그림 9] Y_t (cod 값) 과 Logistic 분포에 따른 p-value

최종 목표인 COD 기계 오류 탐지 알고리즘 구축을 위해 COD를 종속변수로 하는 회귀모형을 구축한다. 설명변수 선택을 위해 동시에 관찰된 변수들인 NH4, NO2, NO3, PO4, TN, TP 및 COD 계산에 사용된 변수들 BlueLED, RedLED, Heat_Temp, Sample들을 사용하여 설명변수 선택 절차를 밟은 결과 BlueLED, RedLED, Heat_Temp, Sample 등이 의미 있는 변수들로 선택되었다. 여기서 BlueLED, RedLED, Heat_Temp, Sample 등은 자동측정기를 사용하여 COD 값을 측정할 때 필요한 변수들로써 BlueLED 와 RedLed는 시료의 오염 농도, Heat_Temp는 시료의 온도, Sample은 강물의 발색값을 뜻한다. 참고로 본 연구의 대상이 된 COD 측정기는 망간법을 이용하여 COD를 측정하고 있다. 모형 (1)을 사용한 회귀 분석 결과는 [표 4]에 주어져 있으며 추정된 모형식은 다음과 같다.

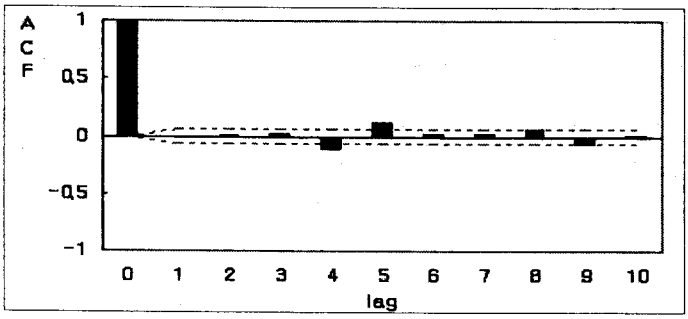
$$\widehat{Y_{t,r}} = 54.548 - 0.021 BlueLED - 0.005 RedLED - 0.037 Heat\_Temp - 0.002 Sample + 1.41163 - 0.31648 e_{t-1} - 0.31302 e_{t-2}$$

위 식의 적합 잔차에 대한 분포를 살펴본 결과 정규 분포를 따른다는 것을 확인할 수 있으며 ([그림 11]) 그에 따른 p-value의 값을 계산할 수 있다. 잔차 및 p-value는 [그림 12]에 주어져 있으며 이상점 탐지 결과는 [표 5] 가운데 열에 주어져 있다.

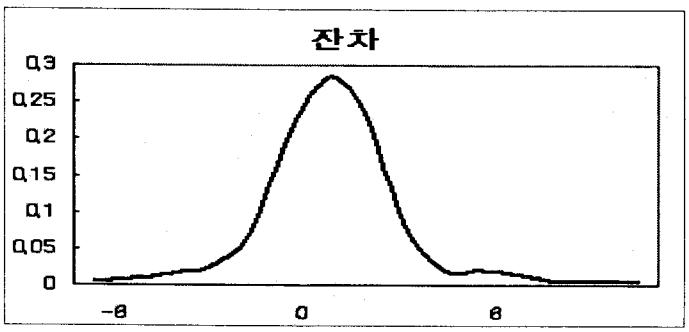
[표 4] 회귀분석 결과

모형요약			
R	R^2	수정된 R^2	추정값의 표준오차
0.817	0.668	0.662	3.15

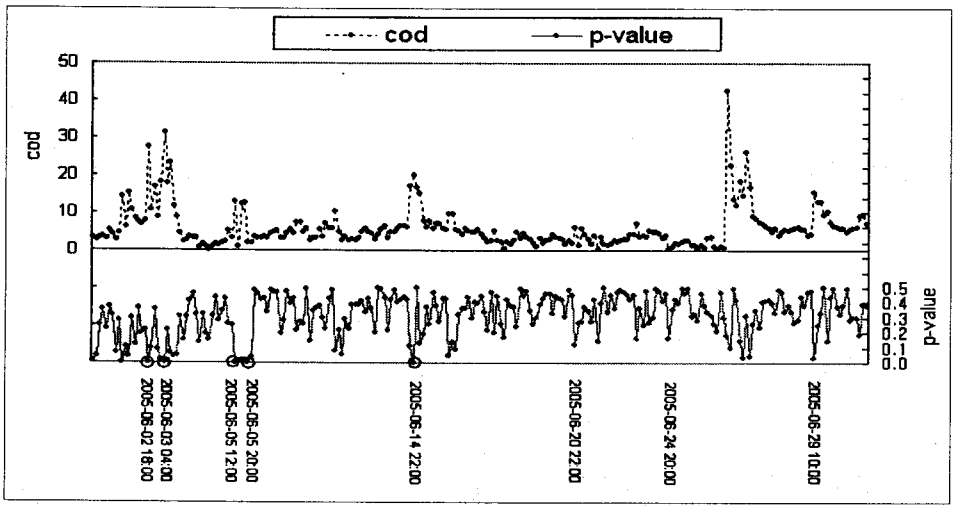
모형	비표준화계수		표준화계수	t	유의확률
	B	표준오차	베타		
(상수)	54.458	12.540		4.343	.000
BlueLED	-.021	.002	-.481	-10.173	.000
RedLED	-.005	.001	-.351	-7.604	.000
Heat_Temp	-.037	.088	-.016	-.419	.675
Sample	-.002	.000	-.204	-5.196	.000



[그림 10] $COD - \widehat{COD_t}$의 ACF



[그림 11] 잔차($Y_t - \widehat{Y_{t,r}}$)에 대한 분포곡선



[그림 12] $\hat{Y}_{t,r}$ 값과 p-value

[표 5] 이상점 탐지 결과

시계열 모형	회귀함수 모형	밀도함수 모형
2005-06-26 20:00	**2005-06-03 04:00**	**2005-06-02 18:00**
2005-06-02 18:00	**2005-06-02 18:00**	**2005-06-03 4:00**
2005-06-03 04:00	2005-06-05 12:00	2005-06-26 20:00
2005-06-27 08:00	2005-06-14 22:00	2005-06-27 8:00
2005-06-10 12:00	2005-06-05 20:00	2005-06-03 08:00
2005-06-29 10:00		
2005-06-01 20:00		
2005-06-14 22:00		

[표 5]를 보면 1단계 시계열 및 밀도함수를 이용한 이상점 탐지 결과가 기록되어 있는데 쉽게 알 수 있는 것은 시계열에 의한 이상점 탐지가 상대적으로 많이(8회) 이루어졌다는 사실이다. 이는 A하천의 B지점에 COD의 시간에 걸친 움직임을 나타내주는 시계열 모형이 이상점에 상당히 민감한 반면 밀도함수는 단순히 임계치에 벗어난 값들만을 나타내 주고 있기 때문인 것으로 보인다. 회귀함수를 이용한 COD 측정기 오류 탐지 결과와 비교해 보면 기존의 두 모형에 의해 탐지된 이상점들 중 2005-06-02 18:00와 2005-06-03 04:00 시점에 측정 오류 가능성이 높다는 사실을 알 수 있다. 이러한 사실은 특히 시계열이나 밀도함수 모형을 이용하여 오염 경보를 발령하고자 할 때 유용한데 예를 들어 2005-06-02 18:00인 경우 시계열과 밀도함수에 의한 이상점이므로 경보 발령이 필요하나 회귀모형에 의하면 측정 오류일 가능성이 높으므로 경보 발령 이전에 기계 오작동 여부를 확인하는 단계가 필요함을 나타내 주고 있다.

4. 결론

과학과 기술이 발전함에 따라 주어진 현상에 대해 좀 더 정밀하고 과학적인 분석을 가능하게 하는 실시간 데이터 들이 관찰되고 있으며 본 논문의 COD 자동 측정 데이터도 그 중에 하나이다. 본 연구에서는 통계적 모형을 사용하여 실시간으로 기기측정 오류를 탐지해 낼 수 있는 알고리즘을 제안하여 구축하였다. 이를 위해 COD 값과 기계 내부 변수간의 관계를 회귀 모형을 통해 분석하고 새로운 데이터에 대해 모형의 타당성을 판단하는 절차를 제시하였다.

참고문헌

1. 이홍근,송준상,이지윤,방기웅,안령미,어수미,이영신,이준호,백도현 (2005). 수질오 염관리, 신광출판사 ,서울.
2. 환경부 (2006). 수질측정망 운영계획.
3. 이상열 (2001). 시계열분석의 원리, 자유아카데미, 서울.
4. Box, G., Jenkins, G. and Reinsel, G. (1994). *Time Series Analysis: Forecasting and Control*, Prentice-Hall, New York.
5. Feller W. (1971), *An Introduction to Probability Theory and Its Applications*, Vol 2, 2nd ed., John Wiley & Sons, Inc., New York.
6 Lehmann, E .L. (1986). *Testing Statistical Hypotheses*, 2nd ed., John Wiley & Sons, Inc., New York.
7. Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

[2007년 6월 접수, 2007년 9월 채택]