ISSN(Print): 1225-7672 / ISSN(Online): 2287-822X
DOI http://dx.doi.org/10.11001/jksww.2015.29.4.469

수질오염 감시체계 구축을 위한 수질 데이터의 통계적 예측 가능성 검토

A Study on the Statistical Predictability of Drinking Water Qualities for Contamination Warning System

박노석[1]·이영주[2]·채선하[2]·윤석민[1*]
No-Suk Park[1]·Young-Joo Lee[2]·Seonha Chae[2]·Sukmin Yoon[1*]

[1]경상대학교 토목공학과 및 공학연구원, [2]K-water 연구원
[1]*Department of Civil Engineering and Engineering Research Institute, Gyeongsang National University*
[2]*K-water Institute*

ABSTRACT

This study have been conducted to analyze the feasibility of establishing Contamination Warning System(CWS) that is capable of monitoring early natural or intentional water quality accidents, and providing active and quick responses for domestic C_water supply system. In order to evaluate the water quality data set, pH, turbidity and free residual chlorine concentration data were collected and each statistical value(mean, variation, range) was calculated, then the seasonal variability of those were analyzed using the independent t-test. From the results of analyzing the distribution of outliers in the measurement data using a high-pass filter, it could be confirmed that a lot of lower outliers appeared due to data missing. In addition, linear filter model based on autoregressive model(AR(1) and AR(2)) was applied for the state estimation of each water quality data set. From the results of analyzing the variability of the autocorrelation coefficient structure according to the change of window size(6hours~48hours), at least the window size longer than 12hours should be necessary for estimating the state of water quality data satisfactorily.

Key words: contamination warning system, outlier, high-pass filter, linear filter model

주제어: 수질오염 감시체계, 이상치, 고주파통과필터, 선형필터 모델

1. 서 론

배·급수 관망을 포함하는 상수도 공급 시스템은 구성 요소(배수지, 관로, 탱크, 밸브 및 펌프 등)가 많고 복잡한 관망으로 이루어져 있어 수질 사고에 취약하다 (Hart and Murray, 2010; Tangena et al., 2011). 특히 외부로부터 오염원이 상수도 공급 시스템 내부로 유입되는 의도적 또는 자연적 오염 사고의 감지는 공간상의 긴 관로, 시간에 따라 변화하는 물소비 경향과 시스템 내의 수리 조건 및 루프(loop)형태의 관상으로 인한 국부 혼화 등으로 인해 아주 어렵다. 미국 환경청(U.S. EPA; Environmental Protection Agency)에서는 2005년부터 2015년 현재까지 약 10년 동안 3단계로 구성된 “Water Security Initiative(물안보 확립 계획)”을 추진 중에 있다. 이 프로그램의 궁극적인 목적은 수돗물을 생산하여 공급하는 전 과정에 대하여 발생할 수 있는 의도적인(테러 등) 또는 자연적인(자연 재해 등) 수질 사고를 조기에 감시하고 혹시 발생하는 경우 가장 빠른 대응을 할 수 있는 총체적인 시스템인

Received 29 May 2015; Revised 3 August 2015; Accepted 5 August 2015
*Corresponding author: Sukmin Yoon (E-mail: gnuysm@gmail.com)

"Contamination Warning System(CWS, 수질오염 감시체계)"을 구축하고 성공적으로 운영하는 것이다(U.S. EPA, 2005; Janke et al., 2006).

CWS는 5개의 구성요소를 포함하는데 첫째는 온라인 수질모니터링 스테이션, 둘째 샘플링과 데이터 분석, 셋째 보안 시스템, 넷째 소비자 불만 감시 및 마지막으로 다섯째는 공중 보건 감시로 구성되어 있다. 이러한 CWS의 구성요소 중에서 가장 기술적으로 개발이 시급한 요소 중에 하나가 바로 온라인 수질 모니터링 스테이션이다. 온라인 수질 모니터링 스테이션은 수질 계측장비, 수질사고 감지를 위한 "Event Detection System(EDS)", 수질사고 분석을 위한 "Agent 및 Plant Library", 경보 시스템, 수질 데이터 전송 장치 및 사고 시 샘플링 장치 등으로 구성되는 것으로 계획하였다. 이중 EDS는 "Brains of the CWS"로 명명될 만큼 그 중요성이 강조되는데 미국 환경청에서는 2007년부터 2014년까지 미국 신시내티 내 고립 배·급수 관망 6곳을 대상으로 "EDS Challenge"라는 기술 개발 프로젝트에 5개의 수질 센서 전문기업인 Sandia National Laboratories(SNL) & U.S. EPA, OptiWater (Elad Salomons), s::can, WhiteWater Security 및 HACH Company를 참여시켜 각기 개발된 기술을 평가한 바 있다(U.S. EPAa, 2014). 하지만 상기 기간의 평가 도중 지속적으로 발생하는 문제점들의 보완과 EDS의 업그레이드 등을 위해서 연속적인 성능 평가가 어려웠으며, EDS의 정확도 또한 70%이하의 성능을 보여 기술개발의 어려움을 여실히 보여주었다.

상기 언급한 "EDS Challenge" 개발 프로젝트 수행기관 중 미국 환경청과 SNL이 공동 개발하고자 한 "CANARY"의 경우 수질을 예측하고 사고 확률을 계산하는 알고리즘을 구성하여 시제품을 만들었는데, 수질예측을 위한 상태 추정모델(state estimation model)로서 시계열 증가모델(time series increments model), 선형필터 모델(linear filter model) 및 다변량 최근접 이웃 알고리즘(multivariate nearest neighbor algorithm)를 적용하였다. 또한 구축된 사고 확률 계산 모델은 이항 사고 선별 모델(BED; Binominal Event Discriminator model)을 적용하였으며, 미국 환경청에서는 기존의 일정한 임계치를 이용하여 수질을 감시하는 'Set Point' 방법에 비해 보다 효율적임을 강조하였다(U.S. EPAb, 2014).

이에 본 연구에서는 국내에서도 잠재되어 있는 상수도 공급 시스템에서의 수질 사고를 조기에 감지하고 대응할 수 있는 CWS의 구축 타당성을 조사하기 위해 국내 C_정수장의 수질 데이터를 대상으로 예측 가능성을 검토하고자 하였다. 과거의 데이터를 이용하여 공급 전 수질 데이터를 예측하고 예측한 데이터와 실제 데이터와의 잔차(residual)를 계산하여 오염여부를 빠른 시간에 결정하는 것이 EDS의 주요한 기능임을 인식할 때 국내에서 실시간으로 저장하고 있는 수질 데이터의 품질을 검토하고 이미 선진국에서 사용하고 있는 통계기법을 이용하여 예측 가능성을 판단하는 것이 본 연구의 궁극적인 목적이라 할 수 있다.

2. 이론적 배경

본 연구에서는 국내 C_정수장을 대상으로 pH, 탁도 및 잔류염소 데이터를 수집하였다. 그리고 수집된 수질 데이터들의 시계열 분포의 변동성 및 수질 데이터 내에 존재하는 이상치들의 분포와 같은 수질 데이터들의 품질 특성을 분석하기 위해 독립표본 t-검정과 고주파 통과필터(high-pass filter)를 이용하였다. 또한 수질예측을 위한 상태 추정모델로서는 미국 환경청과 SNL의 공동연구를 통해 "CANARY" 시스템에 적용된바 있는 선형필터 모델(linear filter model)을 이용하였다.

2.1 독립표본 t-검정

독립표본 $t-$검정은 두 집단 a, b의 평균차이를 비교하기 위해 귀무가설(null hypothesis, H_0)과 대립가설(alternative hypothesis, H_1) 설정하고 통계적 유의수준(significant level) 하에서 어느 쪽이 보다 실현 가능성이 높은지를 주어진 데이터에 의거해서 확률적으로 판단하는 통계적 기법이다. 독립된 두 집단 a, b의 모평균을 μ_a, μ_b라 할 때 독립 t-검정을 위한 가설(hypothesis)은 다음과 같이 설정된다.

- **귀무가설(null hypothesis)**

H_0: $\mu_a = \mu_b$, 두 집단 a와 b의 모평균은 동일하다.

- **대립가설(alternative hypothesis)**

H_1: $\mu_a \neq \mu_b$, 두 집단 a와 b의 모평균은 상이하다.

독립된 두 집단의 표본크기를 n_a, n_b 그리고 표본분산을 S_a^2, S_b^2라 할 때 두 집단의 공통분산(common

variance) S_p^2은 Eq.1와 같이 나타낼 수 있다.

$$S_p^2 = \frac{(n_a-1)S_a^2+(n_b-1)S_b^2}{n_a+n_b-2} \quad \text{(Eq.1)}$$

그리고 독립된 두 집단의 표본평균 $\overline{x_A}$, $\overline{x_b}$의 차이를 검정하기 위한 t-검정 통계량은 Eq.2와 같다.

$$t = \frac{\overline{x_a}-\overline{x_b}}{\sqrt{S_p^2(\frac{1}{n_a}+\frac{1}{n_b})}} \quad \text{(Eq.2)}$$

여기서, P-Value는 귀무가설 H_0가 참(true)일 때 검정통계량에 해당하는 귀무가설 H_0를 기각하는 방향을 형성되는 모든 값들의 확률로서 정의 할 수 있으며, 독립 t-검정에서 P-Value는 주어진 자유도를 갖는 검정 통계량 $|t|$값을 초과할 확률로서 정의할 수 있다.

통계적 유의수준을 α할 때 P-Value는 유의수준과 비교하여 귀무가설의 채택 여부를 결정하는 지표로서 사용되며 양측검정이든 단측검정이든 α와 P-Value의 비교에 P-Value가 α보다 크면 귀무가설을 채택하고, 그 반대이면 귀무가설을 기각하면 된다. 이 사실을 식으로 표시하면 다음 Eq.3과 같다.

$$\begin{cases} P-Value > \alpha, \text{귀무가설 } H_0 \text{ 채택} \\ P-Value < \alpha, \text{귀무가설 } H_0 \text{ 기각} \end{cases} \quad \text{(Eq.3)}$$

2.2 고주파통과필터

일반적으로 수질측정과 같이 현장관측 통해 수집된 원시자료(Raw Data)들은 측정방법, 측정기구 및 측정상태의 정도에 따라 이상치(Outlier)와 같은 다양한 잡음(Noise)들을 포함하게 된다. 이처럼 측정원시자료 내에 발생되는 다양한 잡음들을 참값으로부터 분리하기 위해 개발된 방법론을 데이터 필터(Date Filter)기법이라 한다(Kim, 2010). 본 연구에서는 수질 데이터 내에 분포하는 측정 잡음들을 참값으로부터 분리하고 분리된 잡음들로부터 이상치들의 경향과 분포 특성을 분석하여 수집된 수질 데이터들의 품질을 평가하기 위해 주파수 이론에 기초한 고주파통과필터(high-pass filter)를 적용하였다.

고주파통과필터는 입력되는 신호에서 장기변동을 나타내는 저주파 성분들을 걸러내고 측정된 신호의 잡음들과 같은 고주파 성분만 통과시키게 되며 "Low-cut filter" 혹은 "Bass-cut filter" 라고도 한다(Watkinson, 1998). 고주파통과필터로의 입력신호를 $Y(s)$, 출력신호를 $U(s)$라 할 때 고주파통과필터함수 $G(s)$는 Eq.4와 같이 나타낼 수 있다.

$$G(s) = \frac{Y(s)}{U(s)} \quad \text{(Eq.4)}$$

또한 고주파통과필터함수 $G(s)$는 신호 s, 임의의 양의 상수 d를 이용하여 Eq.5와 같이 간단한 전달함수로서 나타낼 수 있다.

$$G(s) = \frac{s}{s+d} = \frac{s/d}{s/d+1} = \frac{\tau s}{\tau s+1} \quad \text{(Eq.5)}$$

여기서, $\tau \equiv \frac{1}{d}$이다.

Eq.5를 Eq.4의 좌변에 대입하여 Eq.6과 같이 정리한 후 라플라스 역변환을 취하면 Eq.7과 같이 나타낼 수 있다.

$$(\tau s+1)\,Y(s) = \tau s\,U(s) \quad \text{(Eq.6)}$$

$$\tau \dot{y}(t) + y(t) = \tau \dot{u}(t) \quad \text{(Eq.7)}$$

이산시간 k에서 고주파통과필터의 출력값 y_k를 구하기 위해 차분을 이용하여 Eq.7를 이산화한 후 재정리하면 출력값인 입력 신호의 측정 잡음 y_k는 Eq.8과 같다.

$$y_k = \frac{\tau}{\tau+\triangle t} y_{k-1} + \frac{\tau}{\tau+\triangle t}(u_k - u_{k-1}) \quad \text{(Eq.8)}$$

2.3 선형필터 모델

미국 환경청과 SNL는 공동연구를 통해 상수관망의 수질예측을 위한 상태추정 모델로서 시계열증가 모델(time series increments model), 선형필터 모델(linear filter model) 및 다변량 최근접 이웃 알고리즘(multi-variate nearest neighbor algorithm)를 제안한바 있다. 이중 선형필터 모델은 "Linear predictor" 개념을 이용하여 과거 데이터의 가중 합(weighted sum)을 근간으로 시계열상 현재 값을 예측하는 방법론으로서 이러한 접근법은 "Autoregressive Model(AR)"로도 알려져 있다(U.S. EPA, 2010). 임의의 시간 t에서의 관측값을 $z(t)$

라 할 때 다음 시간간격 $(t+1)$에서의 예측값 $\hat{z}(t+1)$에 대한 $AR(P)$ 모델은 Eq.9와 같이 나타낼 수 있다.

$$\hat{z}(t+1) = a_1 z(t) + a_2 z(t-1) + \ldots + a_P z(t-P+1) + \delta(t+1) \quad \text{(Eq.9)}$$

여기서, $a = \{a_1, a_2, \ldots, a_P\}$는 추정계수(estimation coefficient), P는 시차(lag time), $\delta(t+1)$는 추정에러(estimation error) 또는 잔차(residual)로서 Eq.10과 같이 정의할 수 있다.

$$\delta(t+1) = z(t+1) - \hat{z}(t+1) \quad \text{(Eq.10)}$$

Eq.9에서 정의된 $AR(P)$ 모델의 추정계수 a를 구하기 위해 Eq.11에서 나타낸 선형식을 사용하고 이를 행렬로서 나타내면 Eq.12과 같다.

$$Za \approx b \quad \text{(Eq.11)}$$

$$Z = \begin{bmatrix} z(t) & 0 & \cdots & 0 \\ z(t-1) & z(t) & \ddots & \vdots \\ \vdots & z(t-1) & \ddots & 0 \\ z(t-P+1) & \vdots & \ddots & z(t) \\ 0 & z(t-P+1) & \ddots & z(t-1) \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & z(t-P+1) \end{bmatrix},$$

$$a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}, \; b = \begin{bmatrix} z(t+1) \\ z(t) \\ \vdots \\ z(t-P+2) \end{bmatrix} \quad \text{(Eq.12)}$$

온라인 수질측정에 있어서 추정계수 a가 수질 변화에 적응하도록 하기 위해 가장 최근의 시차 P만의 측정값을 이용하게 되며 추정오차를 최소화하기 위해 최소자승법을 이용해 Eq.11를 나타내면 Eq. 13과 같다.

$$a = (Z^T Z)^{-1} Z^T b \quad \text{(Eq.13)}$$

여기서, Z^T는 Z의 전치행렬이다.

각 시차에 대응되는 자기상관계수(autocorrelation coefficient) ρ와 추정계수 a의 관계를 Yule-Walker 방정식을 이용해 나타내면 Eq.14와 같고 이를 행렬로 나타내면 Eq.15와 같다.

$$\begin{aligned} k=1&: \rho_1 = a_1 + \rho_1 a_2 + \cdots + \rho_{p-1} a_p \\ k=2&: \rho_2 = \rho_1 a_1 + a_2 + \cdots + \rho_{p-2} a_p \\ &\vdots \\ k=p&: \rho_p = \rho_{p-1} a_1 + \rho_{p-2} a_2 + \cdots + a_p \end{aligned} \quad \text{(Eq.14)}$$

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{P-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{P-2} \\ \rho_1 & \rho_1 & 1 & \cdots & \rho_{P-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{P-1} & \rho_{P-2} & \rho_{P-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \vdots \\ \rho_P \end{bmatrix} \quad \text{(Eq.15)}$$

상기 Eq.15에서 아래 첨자는 각 시간단계를 나타내는데 이 경우에는 “lag spacing”을 고려한 값을 대입하며 이 시간을 기준으로 수집된 수질자료들의 표본자기상관계수(sample autocorrelation coefficient)이용하면 추정계수 a를 구할 수 있다. 따라서 시계열 수질 측정값에 대한 자기상관 구조를 살펴보면 각기 다른 수질 측정값 데이터 신호와 모니터링 장소에 따라 추정계수 a가 어떻게 변화할지 예측할 수 있게 된다.

3. 연구 방법

본 연구에서는 상수도 공급 시스템에서 잠재적으로 발생될 수 있는 자연적 또는 인위적인 수질 사고를 Online으로 감지하고 조기에 대응할 수 있는 CWS의 국내 구축 타당성을 조사하기 위해 국내 C_정수장을 대상으로 연구를 수행하였다. 현재 국내 C_정수장은 일일 263,800㎥(생활용: 154,800㎥/day, 공업용109,000㎥/day)의 상수를 생산하고 있으며 목표수질은 pH 5.8~8.5, 탁도 0.5NTU 이하, 잔류염소 0.1~4.0mg/L로서 1분 간격의 실시간 Online 측정을 통해 상기 항목들에 대한 측정 데이터들을 생산하고 있다.

앞서 서론에서 언급한 바와 같이 CWS의 구성요소 중에서 가장 기술적으로 개발이 시급한 요소 중에 하나가 바로 온라인 수질 모니터링 스테이션이이며 이중 핵심기술인 EDS는 상수시스템 내에서 실시간으로 측정되는 수질 데이터들로부터 구축되게 된다. 따라서 CWS의 국내 구축 타당성을 검토하기 위해서는 현재 국내 상수시스템 내에서 실시간을 측정되고 있는 수질 데이터들에 대한 품질검토가 우선이라 할 수 있다.

이에 본 연구에서는 현재 국내 C_정수장에서 실시간으로 측정되고 있는 pH, 탁도 및 잔류염소 데이터들에 대한 품질 특성을 분석하기 위해 2014/01/01 ~

2014/12/31 기간 동안의 분단위의 측정 데이터를 수집하였다. 그리고 수집된 수질 데이터들을 월단위로 재구성한 후 독립 t-검정을 이용해 계절별 수질 데이터들의 평균값의 변동성을 고찰하였다. 또한 수집된 수질 데이터 내에 분포하고 있는 측정 잡음들을 고주파통과 필터를 이용해 추출한 후 이상치의 분포 및 발생특성을 분석하였다. 그리고 미국 환경청 및 SNL의 공동연구에서 수질 데이터들의 상태추정 모델로서 적용된바 있는 선형필터 모델의 적용성을 고찰하기 위해 상기의 수질 측정항목들에 대한 자기공분산(autocovariance) 구조를 분석하고 각 측정 항목별 선형필터 모델을 구축하였다.

4. 연구 결과 및 토의

4.1 수질 데이터 품질

Table 1은 자료수집기간 국내 C_정수장으로부터 수집된 pH, 탁도 및 잔류염소 데이터들에 대한 기본통계량을 나타낸 것이다.

일반적으로 평균 μ, 분산 S를 갖는 자료의 95%의 신뢰구간(Confidence interval)은 $\mu \pm 2\frac{s}{\sqrt{n}}$로 나타낼 수 있다(Alfreodo and Wilson, 2006). 따라서 각 수질 데이터들의 95% 신뢰구간을 산정하게 되면 pH 7.24 ~ 7.25, 탁도 0.04 ~ 0.05NTU, 잔류염소 0.78 ~ 0.79mg/L로서 각 수질 데이터들의 연평균 관측값들은 C_정수장의 수질 목표수준 및 국내 상수도 수질기준을 만족하는 것으로 판단할 수 있다.

pH, 탁도 및 잔류염소 데이터들의 년간 변동성을 분석하기 위해 수집된 수질 데이터들을 월단위로 재구성한 후 월별 관측값의 평균과 표준오차를 산정한 후 그 결과를 Fig. 1에 도시하였다.

Fig. 1(a)에 도시한 것과 같이 pH의 월평균 관측값은 6.89~7.50의 범위에서 분포하였으며, 1월~9월 기간 지속적으로 감소하는 경향을 나타내다 9월 이후에는 다시 증가하는 경향을 나타내었다. Fig. 1(b) 및 Fig. 1(c)는 탁도 및 잔류염소의 월별 평균값의 변화를 도시한 것으로 탁도의 월평균 관측값은 1월~6월 기간 0.03~0.04NTU, 7월~8월 기간에는 다소 증가하여 0.05~0.06NTU 범위에서 분포하였으며, 잔류염소의 월

Table 1. Statistical characteristic of water quality data set

	Range		Mean	Variation	Confidence interval*	
	min	max			Lower	Upper
pH	0.00	7.65	7.25	0.091	7.24	7.25
Turbidity	0.00	1.00	0.05	0.004	0.04	0.05
Cl**	0.00	2.10	0.78	0.006	0.78	0.79

*significant level, α=0.05, **Cl=Free residual chlorine

Table 2. The results of t-test for water quality data set

	pH		Turbidity		Cl*	
	t-value	P-value	t-value	P-value	t-value	P-value
Jan-Feb	32.0	0.00	-2.8	0.00	24.3	0.00
Feb-Mar	6.1	0.00	3.7	0.00	11.8	0.00
Mar-Apr	10.8	0.00	-5.6	0.00	24.5	0.00
Apr-May	26.7	0.00	44.0	0.00	-55.5	0.00
May-Jun	45.2	0.00	85.6	0.00	-122.9	0.00
Jun-Jul	85.0	0.00	-137.6	0.00	67.9	0.00
Jul-Aug	75.4	0.00	10.3	0.00	30.4	0.00
Aug-Sep	169.8	0.00	-70.3	0.00	-6.4	0.00
Sep-Oct	-158.5	0.00	-105.7	0.00	-326.9	0.00
Oct-Nov	-42.7	0.00	398.7	0.00	-9.2	0.00
Nov-Dec	-46.5	0.00	121.8	0.00	318.1	0.00

*Cl=Free residual chlorine

평균 관측값은 10월(0.88mg/L), 11월(0.89mg/L)을 제외한 나머지 기간 동안 0.75~0.81mg/L의 범위 내에서 상대적으로 일정한 분포를 나타났다.

Fig. 1에 도시된 각 수질 데이터의 월평균 관측값의 변동성을 통계적으로 검정하기 위해 유의수준 5%($\alpha = 0.05$)하에서 연속하는 2개월 사이의 월평균 관측값들에 대한 독립 t-검정을 자료 전기간에 걸쳐 수행하였다. Table 3는 각 수질 데이터들의 월평균 관측값들에 대한 독립 t-검정의 검정통계량과 검정통계량을 만족하는 P-value의 산정 결과를 나타낸 것이다.

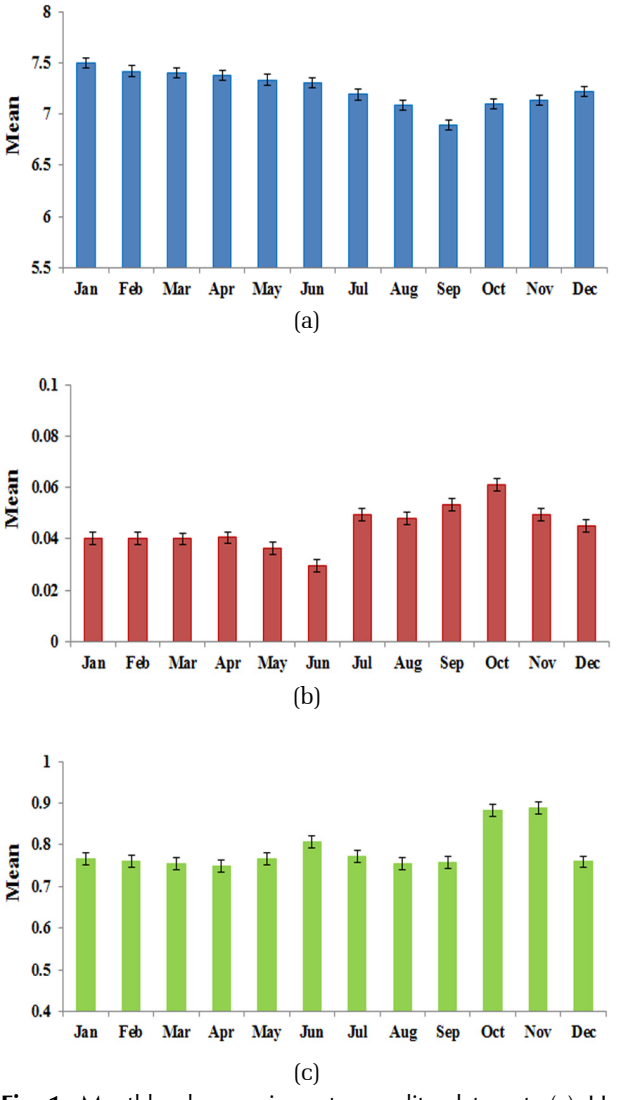

Fig. 1. Monthly changes in water quality data set. (a)pH, (b)Turbidity, and (c)Free residual chlorine. Bar plot indicate mean and error bar indicate standard error.

Table 3에 나타낸 것과 같이 pH, 탁도 및 잔류염소의 월평균 관측값들에 대한 모든 P-value들은 독립 t-검정을 위해 설정된 유의수준($\alpha = 0.05$) 보다 낮은 값을 나타내었다. 이는 각 수질 데이터의 월평균 관측값 사이의 통계적 차이가 자료수집 전기간에 걸쳐 지속적인 발생하는 것을 의미한다. 따라서 Fig. 1에 도시된 각 수질자료의 월별 평균 관측값의 변동은 통계적으로 유의하다고 판단할 수 있다.

본 연구에서는 수집된 수질 데이터 내의 이상치를 포함한 측정 잡음들의 분포와 특성을 분석하기 위해 고주파통과필터를 이용해 각 수질 데이터들로 부터 잡음들을 분리하였고, 그 결과를 Fig. 2에 도시하였다.

Fig. 2(a)는 pH 데이터의 시계열 분포 및 고주파통과필터를 이용해 분리된 데이터 내의 잡음들의 분포를 함께 도시한 것이다. 우선 Fig. 2(a.1)에 도시한 pH 데이터들의 시계열 분포를 살펴보면 pH 데이터들은 연평균 관측값(pH=7.25)을 중심으로 pH=7~8사이의 범위 내에서 비교적 일정하게 분포하였으나 '0'값을 나타내는 하한 이상치가 자료 전기간에 걸쳐 분포하는 것을 알 수 있다. 그리고 Fig. 2(a.2)는 고주파통과필터를 이용해 분리된 pH 데이터 내의 잡음들을 도시한 것으로서 하한 이상치가 발생한 부분들을 제외한 나머지 시간에서의 잡음들은 '0'을 중심으로 비교적 안정적인 분포를 나타내었다. 시계열 과정에 있어서 오차들의 분포가 평균 '0'이고 분산이 일정한 값을 나타내는 경우 이를 백색잡음 과정(White-noise process)이라하며, 정상적인 시계열 자료 내에는 항상 일정한 백색잡음들이 내제 되어있다(Kim, 1990). 따라서 자료수집 전기간에 분포된 pH 데이터의 잡음들은 하한 이상치를 제외하면 pH 측정 프로세스에 내제된 고유적인 측정 오차로서 판단할 수 있다.

Fig. 2(b)는 탁도 데이터의 시계열 분포 및 잡음들의 분포를 함께 도시한 것이다. Fig. 2(b.1)에 도시한 것과 같이 탁도 데이터들의 경우 '0'값을 나타내는 하한 이상치와 더불어 목표 수질인 0.5NTU를 초과하는 다수의 상한 이상치들 또한 자료 전기간에 걸쳐 분포하는 것을 알 수 있다. 그리고 Fig. 2(b.2)는 상·하한 이상치를 포함한 탁도 데이터 내의 잡음들의 분포를 도시한 것으로서 상·하한 이상치가 발생한 기간 이외에서는 측정 잡음들은 백색잡음의 분포를 나타내었다. 하지만 탁도 데이터 내에 분포하는 상한 이상치들은 국내 상수도 수질의 허용범위를 상회는 것으로 이에 대해

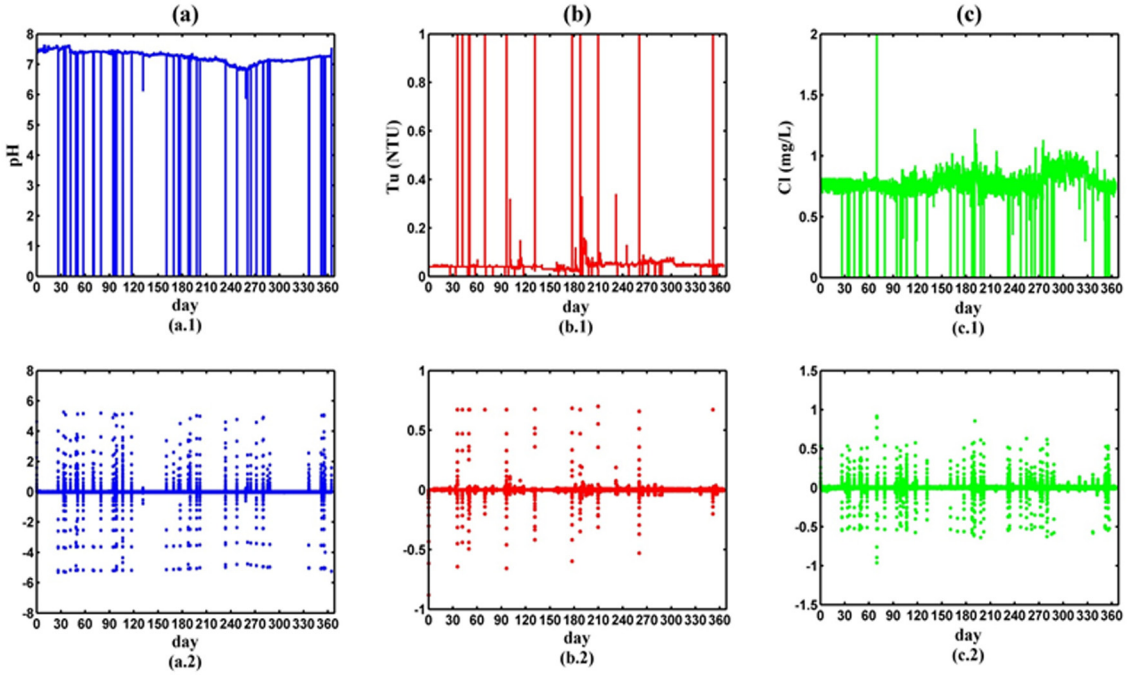

Fig. 2. Plotting of time series for water quality data set(solid line) and noises(point). (a)pH, (b)Turbidity, and (c)Free residual Chlorine.

서는 국내 C_정수장의 운영기록에 대한 추가적인 분석이 필요할 것으로 판단된다.

Fig. 2(c)는 잔류염소 데이터의 시계열 분포 및 측정 잡음들의 분포를 함께 도시한 것이다. Fig. 2(c.1)에 도시한 것과 같이 잔류염소 데이터의 경우 pH 및 탁도 데이터들에 비해 측정값들의 산포도가 상대적으로 높게 나타나는 것을 알 수 있다. 또한 자료 전기간에 걸쳐 분포하는 '0'값의 하한 이상치와 더불어 지협적인 자료의 산포 내부에 작은 빈도의 상한 이상치들이 다수 분포하는 것으로 나타났다. 그리고 Fig. 2(c.2)는 잔류염소 데이터 내의 분포하는 잡음들을 도시한 것으로 pH 및 탁도 데이터에서 관측된 측정 잡음들에 비해 잡음들의 밀도와 빈도가 강하게 나타나는 것을 알 수 있으며, 이는 잔류염소에 대한 측정이 pH 및 탁도의 측정에 비해 해당 정수장의 운영조건 및 측정 프로세스에 상대적으로 민감하게 반응한다는 것을 의미한다.

상기의 분석에서 언급한 바와 같이 pH, 탁도 및 잔류염소 데이터 내에는 '0'값 으로 기록되는 하한 이상치들이 자료수집 전기간에 걸쳐 분포하였다. 일반적으로 국내 상수도 시스템 내에서 pH, 탁도 및 잔류염소와 같은 수질 데이트들은 측정 및 자료전송의 과정에서 자동 프로세스를 기반으로 하는 Online 시스템(e.g. Supervisory Control and Data Acquisition, SCADA)을 이용하게 되며 수질관측 기구의 고장 또는 통신오류 등에 의해 발생되는 결측값들은 수질 데이터 저장 시스템 내에서 '0'값으로 기록되게 된다. 따라서 본 연구에서는 각 수질 데이터 내에서 분포하고 있는 '0'값의 하한 이상치들을 잠재적인 결측치로 분류하였다. 이러한 수질 데이터들의 결측치에 관한 년간 빈도를 살펴보면 pH 782회(13hr), 탁도 2,785회(46.5hr) 및 잔류염소 1,365회(22.8hr)로서 월별 발생 빈도는 Fig. 3에 나타내었다.

Fig. 3에 도시한 것과 같이 pH, 탁도 및 잔류염소에 대한 결측치들은 11월 제외한 전기간에 발생되었으며, 특히 3가지 측정항목 모두 동시에 결측되는 빈도가 매우 높은 것을 알 수 있다. 기존의 통계적 연구들을 살펴보면 다변량 측정에 있어서 자료간의 상관관계 및 자료간의 통계적 거리(e.g. Euclidean distance and Mahalanobis distance)등을 이용하는 자료보간 기법들은 다변량 자료내에 발생되는 결측치들을 효과적으로 보간 할 수 있는 것으로 알려져 있으며(Kim et al., 2010; Kim, 2011; Lee, 2014), 이러한 통계적 자료보간 기법을 적용하기 위해서는 다변량 자료들의 연속성과 더불어 최소한 한 가지 이상의 자료는 반드시 주어져야만 한다.

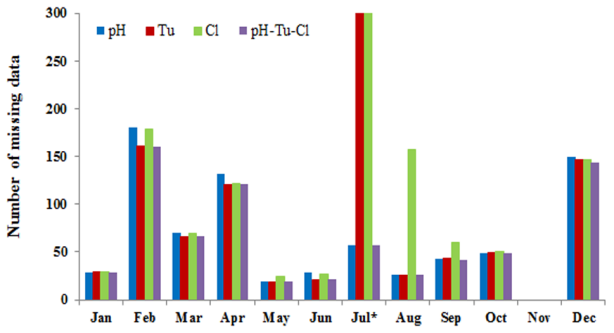

Fig. 3. The number of missing data for water quality data set(Jul*: Turbidity: 2104th/month, Free residual chlorine: 499th/month).

하지만 현재 정수장, 배수지 및 배·급수관망을 포함하는 국내 정수시스템에서는 운영의 편의 및 비용의 절감을 위해 실시간 측정되는 다항목의 수질 데이터들을 단일 통신망에 의존해 송신 및 저장하는 시스템을 운영하고 있으며, 통신장비의 오류 등으로 인한 정상적인 데이터의 송신과 저장이 이루어지지 않는 경우 측정되는 수질데이터 전항목에 대한 결측은 필수적일 수밖에 없다. 따라서 현재 단일 통신망에 의존해 데이터의 전송과 저장을 수행하는 국내 수질측정 프로세스에 내제된 근본적인 문제점과 이로 인한 수질데이터의 결측 문제는 추후 지속적인 논의와 연구가 필요할 것으로 판단된다.

4.2 수질상태 추정모델

본 연구에서는 $AR(P)$ 모형에 기초한 선형필터 모델을 국내 CWS 구축을 위한 수질상태 추정모형으로 선정하였고, 국내 C_정수장의 2014년 자료 중 결측치가 전무하였던 11월01일~11월30일 기간을 대상으로 모델의 검정을 수행하였다.

시계열 모형에 있어서 변량에 대해 일정 기간 동안 변화한 값들을 기록한 것을 그 변량에 대한 시계열 데이터 시퀀스(data sequence)라 부른다. 그리고 $AR(P)$ 모형과 같은 통계적 시계열 모형들은 변량들의 자기공분산 관계를 이용해 미래값에 대한 예측을 수행하기 때문에 변량들의 자기공분산 구조를 분석하기 위한 고정된 시계열 길이를 갖는 서브시퀀스(subsequence)가 필요하며, 이러한 시계열 데이터들의 서브시퀀스들을 시계열분석 윈도우(Window)라 한다(Lim et al., 2006). 변량들의 자기공분산 구조를 분석하기 위한 시계열분석 윈도우 크기는 데이터의 성격에 따라 그 규모가 가변적이나 일반적으로 윈도우 크기가 증가하는 경우 시계열 모형은 강건(Robust)해지는 반면 과도한 분석시간이 요구될 수 있다. 이러한 관점에서 미국 환경청과 SNL의 공동연구에서도 선형필터 모델을 구축하기 위해서는 최적 윈도우 크기를 설정하는 것이 우선임을 언급하였으며, 관망을 대상으로 한 선형필터 모형의 구축을 위해서는 최소 40시간 이상의 크기를 갖는 시계열분석 윈도우가 요구됨을 밝힌바 있다(U.S. EPA, 2010). 이에 본 연구에서는 pH, 탁도 및 잔류염소 데이터들의 선형필터 모형 구축에 앞서 시계열분석 윈도우 크기 변화에 따른 각 수질 데이터들의 자기공분산 구조의 변동성을 분석하기 위해 6시간~48시간의 범위를 갖는 시계열 분석 윈도우를 대상으로 각 시차별(1min~60min) 자기상관계수를 산정하였다.

Fig. 4는 pH 데이터의 자기상관계수(autocorrelation coefficient) 변화를 도시한 것으로서 pH 데이터의 자기상관계수는 시계열분석 윈도우 6시간에서는 시차의 증가에 따라 급격히 감소하는 경향을 나타내었으나, 시계열분석 윈도우 12시간에서 부터는 자기상관계수의 감소폭이 다소 감소하여 시차 24시간이후에는 자기상관계수의 전체적인 구조가 일정한 형태를 나타내었다.

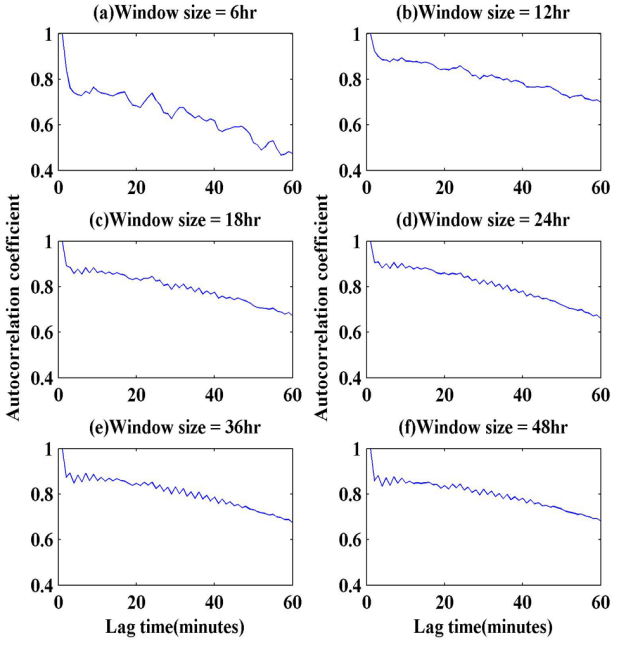

Fig. 4. Variation of autocorrelation coefficient for pH data set.

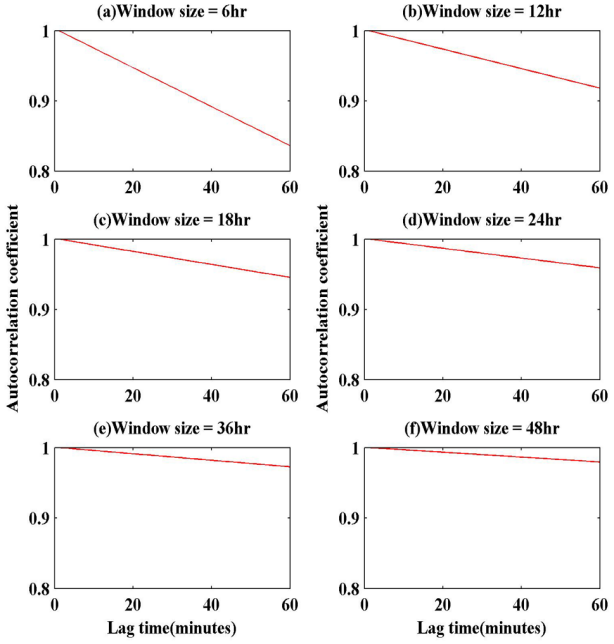

Fig. 5. Variation of autocorrelation coefficient for Turbidity data set.

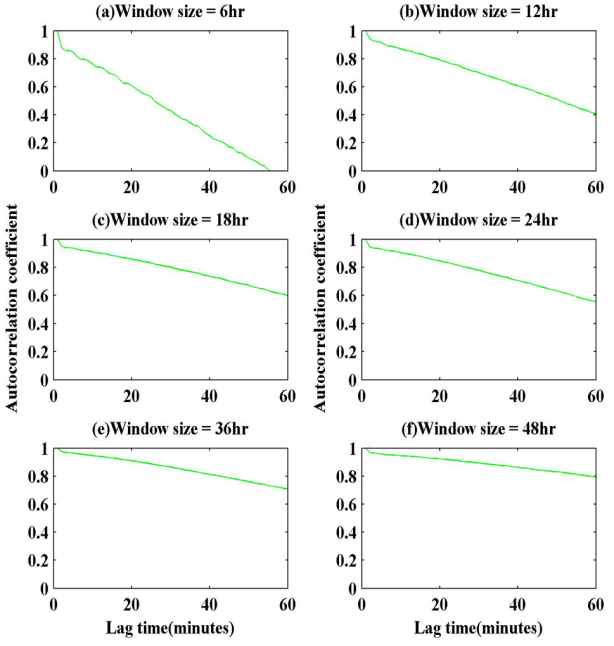

Fig. 6. Variation of autocorrelation coefficient for Free residual chlorine data set.

Fig. 5는 탁도 데이터의 자기상관계수 변화를 도시한 것으로 탁도 데이터의 자기상관계수는 전 시계열분석 윈도우에서 0.8이상의 높은 값을 나타내었다. 그리고 각 시계열분석 윈도우에서 자기상관계수들은 시차의 증가에 따라 선형적으로 감소하는 구조를 나타내었으며, 이러한 자기상관계수의 선형적인 감소폭은 시계열분석 윈도우 크기가 증가함에 따라 서서히 감소하여 시계열분석 윈도우 24시간에서 이후에는 일정한 자기상관계수 구조를 나타내었다.

Fig. 6는 잔류염소 데이터의 자기상관계수 변화를 도시한 것이다. 잔류염소 데이터의 자기상관계수는 탁도 데이터의 자기상관계수 구조와 유사하게 선형적으로 감소하는 구조를 나타내었으며, 시계열분석 윈도우 6시간에서 자기상관계수는 시차의 증가에 따라 급격히 감소하여 시차 55분 이후에는 '0' 값을 나타내었다. 하지만 시계열분석 윈도우의 크기가 증가함에 따라 자기상관계수의 감소폭은 서서히 감소하여 시계열분석 윈도우 48시간에서는 최대 시차에서도 자기상관계수는 0.8이상의 높은 값을 나타내었다.

본 연구에서는 국내 C_정수장의 수질상태 추정을 위해 시계열 모형인 $AR(1)$, $AR(2)$ 모형을 근간으로 한 선형필터 모델을 적용하였다. 시계열 모형에 따른 예측성능 및 시계열분석 윈도우 크기에 따른 민감도를 비교·평가하기 위해 시계열분석 윈도우 크기를 6시간~72시간의 범위로 변화하여 국내 C_정수장의 2014년 11월01일~11월30일 기간의 pH, 탁도 및 잔류염소의 수질상태 변화를 모의하였다.

Fig. 7은 $AR(1)$ 및 $AR(2)$ 모형을 근간으로 한 선형필터 모델을 이용해 pH, 탁도 및 잔류염소 데이터들의

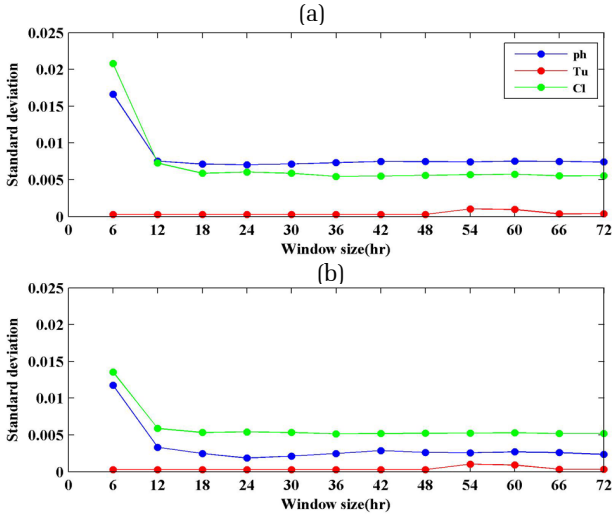

Fig. 7. Standard deviation of the prediction error for linear filter model. (a)linear filter model based on $AR(1)$, (b)linear filter model based on $AR(2)$.

상태변화를 모의한 후 예측값과 관측값의 잔차(Residual)를 나타내는 예측오차들의 표준편차를 나타낸 것이다. 각 선형필터모델들로 부터 모의된 pH 및 잔류염소 데이터들에 대한 예측오차들의 표준편차는 시계열분석 윈도우 6시간에서 가장 높은 값을 나타내었으며, 시계열분석 윈도우 12시간에서 크게 감소하여 이후 72시간 사이에는 비교적 일정한 값을 나타내었다. 그리고 pH와 잔류염소 데이터들은 $AR(1)$ 모형에 비해 $AR(2)$ 모형을 근간으로 한 선형필터 모델에서 예측오차들의 표준편차가 상대적으로 낮은 범위에서 분포하는 것으로 나타났다. 반면, 탁도 데이터들의 예측오차들에 대한 표준편차는 전 시계열분석 윈도우에서 '0'에 근접한 일정한 값을 나타내었으며, $AR(P)$ 모형의 차이에 따른 예측오차의 변동은 미소한 것으로 나타났다.

5. 결 론

본 연구에서는 국내 상수도 공급 시스템에서 잠재적으로 발생될 수 있는 자연적 또는 인위적인 수질사고를 Online으로 감지하고 조기에 대응할 수 있는 CWS의 국내 구축 타당성을 조사하기 위해 국내 C_정수장을 대상으로 하여 현재 실시간으로 측정되고 있는 수질 데이터들에 대한 품질을 분석하고 각 수질 데이터들에 대한 예측 가능성을 검토하고자 하였다. 그리고 그 결과를 요약하면 다음과 같다.

1) 현재 국내 C_정수장에서 실시간 측정되고 있는 pH, 탁도 및 잔류염소 데이터들의 연평균 측정값은 95% 신뢰구간 하에서 pH 7.24~7.25, 탁도 0.04~0.05NTU, 잔류염소 0.78~0.79mg/L로서 각 수질 데이터들의 연평균 관측값들은 국내 C_정수장의 수질 목표수준 및 국내 상수도 수질기준을 만족하는 것으로 나타났다. 그리고 pH, 탁도 및 잔류염소의 월평균 관측값들은 상대적으로 낮은 범위이긴 하나 매월 지속적으로 변화하였으며 이러한 월별 평균값의 변동에 대해 독립 t-검정을 수행한 결과 통계적으로 유의미한 것으로 나타났다.

2) 고주파통과필터를 이용해 수질 데이터들의 측정잡음들을 분리한 후 자료 내에 분포하는 이상치들의 특성을 고찰한 결과 pH, 탁도 및 잔류염소 데이터들 내에는 정상적인 측정과정에서 발생되는 백색잡음(White-noise)들과 더불어 다수의 상·하한 이상치들의 분포가 발견되었다. 이러한 이상치들의 특성을 살펴보면 pH, 탁도 및 잔류염소 데이터 모두 결측으로 인한 하한 이상치가 대부분이었으며, 탁도의 경우 상수도 수질 목표수준인 0.5NTU를 초과하는 상한 이상치들의 분포도 일부 발견되었다. 그리고 pH 및 탁도의 측정 잡음에 비해 잔류염소의 측정 잡음들이 상대적으로 강한 밀도와 빈도를 나타내었는데 이는 잔류염소에 대한 측정이 pH 및 탁도의 측정에 비해 해당 정수장의 운영조건 및 측정 프로세서에 상대적으로 민감하게 반응하기 때문인 것으로 판단된다.

3) pH, 탁도 및 잔류염소 데이터들의 결측값 발생 특성을 살펴보면 년간 발생 빈도는 pH 782회(13hr), 탁도 2785회(46.5hr) 및 잔류염소 1365회(22.8hr)로서 자료결측은 매월 지속적으로 발생되었으며, 특히 3가지 항목 모두 동시에 결측되는 빈도가 매우 높은 것으로 나타났다. 현재 단일통신망을 이용해 다항목의 수질 데이터를 전송 및 저장하는 국내 수도 시스템을 고려 할 때 통신장비의 오류 등으로 의해 정상적인 데이터의 송신과 저장이 이루어지지 않는 경우 측정되는 수질 데이터 전항목에 대한 결측은 필수적일 수밖에 없다. 따라서 단일통신망에 의존하는 국내 수질측정 프로세스에 내제된 근본적인 문제점과 이로 인해 발생되는 수질 데이터의 결측 문제는 추후 지속적인 논의와 연구가 필요할 것으로 판단된다.

4) 수질 데이터들의 상태추정을 위해 본 연구에는 $AR(P)$ 모형을 근간으로 한 선형필터 모델을 적용하였다. 선형필터 모델의 구축에 앞서 시계열분석 윈도우 크기 변화에 따른 각 수질 데이터들의 자기공분산 구조의 변동성을 분석하기 위해 1시간~48시간의 범위를 갖는 시계열분석 윈도우를 대상으로 자기상관계수의 변화를 분석하였다. 그 결과 전반적으로 시계열분석 윈도우의 크기가 24시간을 초과하는 경우 자기상관계수의 구조는 안정적으로 유지되는 것을 확인 할 수 있었다. 그리고 $AR(1)$ 및 $AR(2)$ 모형을 근간으로 한 수질예측 모의에 있어서도 예측오차들은 시계열분석 윈도의 크기 12시간이후에서부터 크게 감소하는 것을 알 수 있었다. 상기의 결과를 종합해볼 때 정수장을 대상으로 한 pH, 탁도 및 잔류염소 데이터의 수질상태 추정을 위해 선형필터 모형을 적용하는 경우 최소 12시간 이상의 시계열분석 윈도우가 필요할 것으로 판단된다.

사사의 글

본 연구는 환경부 “차세대 에코이노베이션사업(글로벌탑 환경기술개발사업)”의 지원에 의해 수행되었으며 이에 감사드립니다.(GT-SWS-11-02-007-8)

References

Alfredo Hua-Sing. Ang and Wilson H. Tang (2006). *Probability concepts in Engineering 2nd Ed*., John Wiley & Sons. New York, pp.406.

Hart, W.E., Murray, R. (2010). Review of Sensor Placement Strategies for Contamination Warning Systems in Drinking Water Distribution Systems. Journal of Water Resources Planning and Management, ASCE, 136(6), pp. 611-619.

Janke, R., Murray, R. Uber, J., Taxon, T. (2006). Comparison of Physical Sampling and Real-Time Monitoring Strategies for Designing a Contamination Warning System in a Drinking Water Distribution System. Journal of Water Resources Planning and Management, 132(4), pp. 310-313.

Kim, Y.H, (1990), *Time series analysis*, FREEACADEMY, pp 735

Kim, S.J. (2010), *Essential Kalman Filter*, A-jin, pp. 270.

Kim, S., Cho, N.W., Kang, S.H. (2010), Density-based Outlier Detection for Very Large Data , Journal of the Korean Operations Research and Management Science Society, 35(2), pp. 71-88.

Kim, D.J. (2011). Solutions for Missing Values in Categorical Data, The Korean Association For Comparative Government, 15(2), pp.319-342.

Lee, T.H. (2014), A Comparison of Full Information Maximum Likelihood, Multiple Imputation, and Bayesian Approach in Overall Goodness of Fit Assessment of Structural Equation Modeling with Missing Data, Korean Journal of Psychology: General, 33(2), pp.507-536

Lim, S.H., Kim, S.W., Park, H.J (2006). Optimal Construction of Multiple Indexes for Time-Series Subsequence Matching, Journal of KISS: Databases, 33(2), pp. 201-2013.

Tangena, B. H., Janssen, P. J. C. M., Tiesjema, G., van den Brandhof, E.J., Klein Koerkamp, M., Verhoef, J. W., Filippi, A., (2011), A Novel approach early warning of drinking water contamination events, *the Water Contamination Emergencies Conference 4 in Mullheim*, October 11-13, 2010.

U.S. EPA (2005), *Water Sentinel System Architecture Draft*, *Version 1.0*, U.S EPA Water Security Division.

U.S. EPA (2010), *Water Quality Event Detection Systema for Drinking Water Contamination Warning System, Development, Testing, and Application of CANARY.*

U.S. EPAa (2014), *Water Security Initiative: System Evaluation of the Cincinnati Contamination Warning System Pilot,* U.S EPA Water Security Division.

U.S. EPAb (2014), *Water Security Initiative: Evaluation of the Water Quality Monitoring Component of the Cincinnati Contamination Warning System Pilot*, U.S EPA Water Security Division.

Watkinson, John (1998), *The Art of Sound Reproduction*. Focal Press. Massachusetts, pp. 576.