J. Korean Soc. Environ. Eng., 37(3), 137~144, 2015
Original Paper
http://dx.doi.org/10.4491/KSEE.2015.37.3.137
ISSN 1225-5025, e-ISSN 2383-7810

다변량 통계기법을 활용한 실시간 수질이상 유무 판단 시스템 개발

Development of Real-Time Water Quality Abnormality Warning System for Using Multivariate Statistical Method

허태영 · 전항배 · 박상민* · 이영주**,†
Tae-Young Heo · Hang-Bae Jeon · Sang-Min Park* · Young-Joo Lee**,†

충북대학교 정보통계학과 · *새만금지방환경청 측정분석과 · **K-water연구원
Department of Information & Statistics, Chungbuk National University
**Monitoring and Analysis Division, Saemangeum Regional Environmental Office*
***Water Research Center, K-water Institute, K-water*

(Received June 3, 2014; Revised August 7, 2014; Accepted February 3, 2015)

Abstract : The purpose of this study is to develop an warning system to detect real-time water quality abnormality using a multivariate statistical approach. In this study, we applied principal component analysis among multivariate data analyses which was used for the correlation between water quality parameters considering the real-time algorithm to determine abnormality in water quality. We applied our approach to real field data and showed the utilization of algorithm for the real-time monitoring to find water quality abnormality. In addition, our approach with Korea Meterological Adminstration database identified heavy rain data due to climate change is one of the most important factors to explain water quality abnormality.
Key Words : *Multivariate statistical method, Warning system, Principal component Analysis, Real-Time monitoring*

요약 : 본 연구는 다변량 통계기법 중 하나인 주성분분석을 활용하여 실시간으로 수질이상 유무를 판단할 수 있는 경보시스템 개발을 목적으로 하였다. 본 연구에서는 다변량 분석 방법 중 수질항목 간의 상관성을 고려한 주성분 분석 방법을 실시간으로 수질이상 유무를 판단하는 알고리즘에 적용시켰다. K-water에서 제공하는 실제 자료를 이용하여 수질 이상에 대한 실시간 감시 알고리즘의 활용성을 검증하였으며, 집중호우 등과 같은 기후변화에 따른 수질이상에 대해서는 기상청 자료와의 비교를 통해 검증하였다.
주제어 : 다변량 통계기법, 경보시스템, 주성분 분석, 실시간 수질 모니터링

1. 서 론

현재 대부분의 정수장에서는 운영관리 자동화 및 먹는물의 안전성을 확보하기 위해 실시간으로 수질을 감시할 수 있는 수질계측기를 설치하여 운영 중에 있다. 실시간 수질 모니터링 시스템은 센서로부터의 데이터 전송과 처리에 있어 다양한 정보통신 기술들을 활용하고 있다. 이러한 실시간 수질 측정값은 효율적인 운전과 수질 변화에 대응 가능하고 현장 운영자의 의사결정의 정확성과 신속성을 가능하게 하는 장점을 가지고 있다.[1)]

정수장에서 실시간으로 측정 가능한 수질항목은 수온, pH, 탁도, 알칼리도, 전기전도도, 잔류염소 등이 있으며, 정수처리 각 공정(취수장-착수정-혼화지-응집지-침전지-여과지-정수지)에서의 수질변화를 1분 단위로 측정하여 공정관리의 적정성 및 수질 이상 유무 확인에 활용하고 있다(http://www. kwater.or.kr). K-water의 경우 정수장 한 개소 당 수질계측기 약 34개를 설치 · 운영하고 있으며, 1분 단위 데이터를 기준으로 하루에 약 5,000개의 수질자료가 축적되고 있다. 그러나 이러한 많은 자료의 축적에도 불구하고 수질 급변 시에는 신속한 대처가 어려워, 기후변화에 대응한 실시간 위기대응 능력 향상이 필요한 실정이다.

정수처리공정에서 수질 오염물질의 실시간 모니터링에 대한 필요성이 급격히 증가하고 있으나, 측정된 자료는 수질기준 만족 여부 또는 수질 변화를 단순 감시하는데 주로 활용되고 있기 때문에 수질오염에 대한 조기경보 및 대응방법에 대한 연구 또한 미흡한 실정이다.

최근 센서 및 컴퓨터의 발달로 현실에서 실시간으로 수많은 자료를 얻고 관리할 수 있게 됨에 따라 이러한 실시간 자료를 이용한 다변량 통계 공정관리 기법(multivariate statistical process control, MSPC)이 각광을 받고 있다.[2,3)] 통계적 공정관리 기법은 기존의 연구처럼 모형이라는 매개를 거치지 않고 실시간으로 얻어지는 자료만으로 손쉽게 자료의 이상여부를 알아낼 수 있다는 장점이 있다. 이러한 통계 공정 관리 방법에서는 실시간으로 얻어지는 수많은 자료 중에서 유용한 정보를 찾아내 강건성(robustness)과 민감성(sensitivity)을 모두 갖춘 시스템을 구현하는 것이 중요하다. 따라서 모니터링과 이상치 판단을 위한 이러한 다변량 통계기법의 사용에 대한 연구가 많이 이루어지고 있다.

다변량 통계 공정관리기법은 1990년 초에 시작하여 화학적 연속 중합 공정 및 공장의 품질관리 등 여러 산업 분야

† Corresponding author E-mail: yjlee1947@kwater.or.kr Tel: 042-870-7532 Fax: 042-870-7549

에서 공정 모니터링과 고장 검출을 위해 적용되어 이에 대한 많은 연구가 이루어지고 있다.[4~7)]

다변량 통계 방법 이전에는 단변량 통계 방법이 많이 연구되어 왔으나 단변량 방법을 변수가 여러 개인 다변량 자료에 적용했을 때 변수간의 상관성을 전혀 고려하지 않기 때문에 비정상적인 상태를 잘 감지하지 못하는 단점을 가지고 있다.[8,9)] 수질분야에서 다변량 통계 방법을 이용한 연구로는 Yang 등[10)]이 수질 센서로부터 얻어진 자료를 이용하여 수질오염에 대한 실시간 검출 분야 알고리즘에 대한 연구를 수행하였으며, Lennox[11)]와 Baggiani[12)]는 생물학적 폐수처리 공정에서 실시간 오류 검출에 대한 연구를 수행하였다.

본 연구에서는 다변량 통계 방법을 적용하여 실시간 수질에 대한 이상징후 판단 알고리즘을 개발하고, 실제 운영 데이터를 통하여 성능 평가를 수행하였으며, 강우 정보와의 연동을 통해 강우가 수질에 미치는 영향을 연구하였다.

본 논문은 총 4장으로 구성되어 있다. 제2장에서는 연구방법으로 주성분 분석기법의 정의 및 방법을 소개하고 수질이상 판단을 검증하기 위해 산정되는 가중 스코어 제곱합인 T^2-통계량과 제곱예측오차인 Q-통계량(Squared Prediction, Error, SPE)의 산정방법에 대해 설명하였다. 제3장에서는 정수장에서 취득한 수질 센서자료를 이용하여 다변량 통계 분석알고리즘을 제시하였으며, 분석 결과 및 강우정보와의 관계를 통해 활용성에 대해 논의하였다. 제4장에서는 본 연구에서 새롭게 제안한 정수장 수질 센서자료에 대한 다변량 통계분석 적용에 대한 결론을 종합적으로 기술하였다.

2. 연구방법

실시간 이상 발생 검출에 대한 성공적인 다변량 통계방법들의 응용은 다음과 같은 세 가지 단계를 포함한다. 첫 번째 단계는 정상적인 과정(stationary process)에 대한 모형설정이며, 두 번째 단계는 모형을 통해 정상적인 과정을 벗어나는 이상치(outlier)를 식별하는 것이다. 그리고 세 번째 단계는 이상치에 대한 이유 또는 원인을 확인하는 단계이다.

여기에서 첫 번째 단계는 정상적 운영 조건에서 얻어진 자료를 이용하여 모형을 구축하는 과정으로 통계적 방법 중 주성분 분석에 의해 수행되어 진다. 두 번째 단계인 이상치 식별은 다변량 방법을 통한 호텔링의 T^2-통계량과 Q-통계량(SPE)으로 수행된다. 그리고 세 번째 단계인 이상 사건에 대한 이유 또는 원인을 확인하는 단계는 기여도 그림(contribution plot)으로 수행될 수 있다.

2.1. 주성분 분석

다변량 통계방법의 첫 번째 단계는 주성분 분석을 이용하여 정상적인 조건을 기반으로 모형을 구축하는 것이다. 실시간으로 계측되는 센서 네트워크의 특성상 새로운 자료가 추가됨에 따라 모형을 갱신(update)하여 사용할 수도 있어 본 연구에서는 실시간으로 적응(adaptive)이 가능한 주성분 모형을 활용하였다. 즉, 센서로부터 실시간으로 측정되는 수질 변수들은 시간에 따라 변화하는 특성을 가지고 있어, 시간에 따라 변하지 않는 주성분 분석의 결과는 잘못된 결과를 초래할 수 있기 때문에, 이동창(moving window)을 반영한 실시간 적응(adaptive) 주성분 분석을 제안하였다.

새로운 값이 센서로부터 얻어져 이용 가능하면 가장 과거의 자료는 삭제되고 새로운 값이 반영된 새로운 창이 만들어져 새로운 공분산 함수(covariance function)을 통해 주성분 분석을 실시하며 간략한 설명은 다음과 같다[13)].

차수가 $n \times j$인 임의의 자료행렬 X는 차수가 $n \times k$인 스코어 행렬(score matrix) T와 차수가 $j \times k$인 로딩 행렬(loading matrix) P의 선형결합으로 표현될 수 있다. 여기서, n은 전체 측정값의 개수, k는 주성분의 개수, j는 변수의 개수를 의미한다. 첫 번째 주성분(first component)은 자료를 가장 많이 설명해 주는 성분을 의미하며, 두 번째 주성분은 두 번째로 자료를 많이 설명하는 성분을 의미한다. 이와 같이 세 번째, 네 번째 등과 같은 주성분들을 해석할 수 있다. 주성분의 개수인 k는 n 또는 j 둘 중 하나보다 작거나 같은 개수를 가진다.

일반적으로 자료행렬 X는 식 (1)과 같이 몇 개의 주성분으로 구성될 수 있으며, E는 차수가 $n \times j$인 잔차행렬(residual matrix)을 의미한다.

$$X = TP^T + E \tag{1}$$

일단 주성분 분석을 통해 주성분의 개수가 결정되고 최종 모형이 구축되고 나면 새로운 측정값들에 대한 스코어 값은 식 (2)를 이용하여 계산할 수 있다.

$$t_{new}^T = x_{new}^T P(P^T P)^{-1} \tag{2}$$

여기서 소문자는 벡터, 대문자는 행렬로 정의되며, x_{new}는 새로 관측된 벡터화된 자료, P는 기존의 과거 이력 자료로부터 계산된 로딩행렬, $\boldsymbol{t}_{new}^T$는 새로 관측된 자료를 통한 스코어(score)의 벡터를 나타낸다.

본 연구에서는 주성분 분석 모형 구축 시, 실시간으로 계측되는 상시 계측시스템의 센서 네트워크 자료의 특성상 새로운 자료가 관측되면 새로운 관측치를 포함하여 주성분 분석을 실시하였다. 이와 같이 실시간으로 관측되는 새로운 관측치를 포함하여 갱신된 주성분 분석 모형을 구축하여 고유값(eigenvalue)과 고유벡터(eigenvector) 또는 로딩값(loading vector)을 통해 새롭게 갱신된 주성분 분석 모형을 얻을 수 있다. 따라서 실시간 자료가 갱신될 때마다 고유값, 고유벡터 역시 갱신되는 적응형(adaptive) 모형을 구축하였으며 누적기여율을 80%로 설정하여 주성분의 개수를 결정하도록 하였다.

2.2. 호텔링의 T^2-통계량과 Q-통계량

다변량 통계기법은 여러 개의 변수들이 서로 높은 상관관계를 가지고 있을 경우, 효과적으로 다룰 수 있는 분석 방법이다. 따라서 다변량 통계기법을 이용할 경우 여러 개의 변수들을 함께 이용하여 공정을 보다 정확하게 모니터링 할 수 있게 된다.

다변량 통계기법의 역할은 센서로부터 얻어지는 실시간 자료가 일정 기준에서 벗어남에 따라 어떤 사건 또는 고장과 같은 이상 징후가 발생했음을 감지하고 판단하는 것이다. D-통계량으로 불리는 호텔링의 T^2-통계량은 새로운 측정값이 정상상태에서 얻어진 과거 이력자료와 얼마나 비슷한지를 판단하는 통계량으로 활용된다.[14~19] 호텔링의 T^2-통계량은 과거 이력자료와 새로운 측정값과의 마하라노비스 거리(Mahalanobis distance)를 의미하며 식 (3)과 같이 계산된다.

$$D_{new} = t_{new}^T S^{-1} t_{new} \tag{3}$$

여기서 S는 과거이력자료에 대한 주성분 분석의 스코어 값에 대한 공분산행렬을 나타내며, t_{new}는 새로운 측정값을 위한 평균으로 중심화된 예측된 스코어 벡터값을 의미한다.

새로운 측정값이 정상적인 모형과 매우 다르다고 판단할 수 있는 것은 센서자료에 대한 이상여부를 판단하는데 있어 매우 중요한 측면으로 고려될 수 있다. 따라서 호텔링의 T^2-통계량에 근거한 임계값(critical limit)을 찾는 것이 중요하고, 호텔링의 T^2-통계량은 F 분포를 따른다고 알려져 있으며, 식 (4)에서와 같이 임계값을 계산할 수 있다.

$$D_{UCL} = \frac{k(n-1)}{(n-k)} F_{(k,n-k,\alpha)} \tag{4}$$

여기서, $F_{(k,n-k,\alpha)}$는 α 분위수에서 자유도가 k와 $n-k$인 F 분포의 임계값(critical value)을 나타내며, n은 전체 관측값의 개수, k는 주성분 분석 모형에서 사용된 주성분의 개수를 의미한다.

각각의 새로운 추정값을 과거 이력자료와 비교하면서 호텔링의 T^2-통계량을 실시간으로 모니터링 하면, 호텔링의 T^2-통계량은 사건 또는 고장과 같은 이상여부의 판단을 알려주는 사건 지시자(event indicator)로 활용될 수 있다.

호텔링의 T^2-통계량과 비슷하게 이상여부 판단을 위해 Q-통계량(SPE)을 활용할 수도 있다. 호텔링의 T^2-통계량과는 달리 Q-통계량은 식 (5)와 같이 실제 측정값과 추정값(x_{new})과의 차이를 모니터링 하면서 이상여부를 판단할 수 있다. 즉, Q-통계량은 식 (6)과 같이 실제 측정값과 추정값의 차이의 제곱을 통해 계산할 수 있으며 Q-통계량은 추정값이 과거 이력자료로부터 얻어진 정상적인 모형에 얼마나 잘 맞는지 확인하는 척도로 사용될 수 있다.

$$e_{new} = x_{new} - t_{new}^T P^T \tag{5}$$

$$Q_{new} = e_{new} e_{new}^T \tag{6}$$

정상 조건 하에서 Q-통계량은 다변량 정규분포를 따르며 임계값은 가중된 카이제곱 분포를 통해 추정될 수 있다. Jackson and Mudholkar에 의해 제안된 근사식은 Q-통계량의 상한값(upper control limit, UCL)에 대한 임계값을 계산하는데 사용되었으며 식 (7)과 같이 표현된다.[20]

$$Q_\alpha = \theta_1 \left[1 + \frac{\theta_2 h_0 (1-h_0)}{\theta_1^2} + \frac{z_\alpha (2\theta_2 h_0^2)^{1/2}}{\theta_1} \right]^{1/h_0} \tag{7}$$

여기서, z_α는 정규분포의 $(1-\alpha)$분위수에 해당하는 값을 의미하며, h_0, θ_1, θ_2, θ_3는 식 (8)과 같은 값을 의미한다.

$$h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2},\ \theta_1 = \sum_{i=k+1}^{j} \lambda_i,\ \theta_2 = \sum_{i=k+1}^{j} \lambda_i^2,\ \theta_3 = \sum_{i=k+1}^{j} \lambda_i^3 \tag{8}$$

여기서, λ_i는 주성분 분석에서 얻어진 고유값(eigenvalue)을 의미하며, k는 변수의 개수를 의미하며, j는 주성분 분석에서 사용된 주성분의 개수를 의미한다.

2.3. 기여도

본 연구에서는 수질 이상에 대한 판단 여부에 대해 초점을 맞추었으나 어떤 수질 오염물질에 의해 영향을 받는지에 대한 기여도 평가가 중요하다. 기여도 차트는 어떤 비정상적인 사건발생에 대한 원인 또는 근원에 대해 매우 가치있는 식견을 제공해주는 역할을 할 수 있다. 예를 들어, 호텔링의 T^2-통계량이 임계값을 넘어 비정상적인 사건으로 판단하자고 할 때, 호텔링의 T^2-통계량은 단지 어떤 사건이 발생했음을 알려줄 뿐 그 사건이 어느 센서로부터의 원인 또는 근원으로부터 발생했는지의 여부에 대해서는 정보를 제공해 주지 못한다. 그러나 기여도 차트는 어떤 센서 또는 변수에 의해 이상 징후가 발생했는지를 확인할 수 있는 도구로 활용될 수 있다. 호텔링의 T^2-통계량에 대해 개별 변수에 대한 기여도를 계산하기 위해 식 (9)를 이용할 수 있다.

$$c_{new,j}^D = t_{new}^T S^{-1} \left[x_{new,j} p_j \left(P^T P \right)^{-1} \right]^T \tag{9}$$

여기서, t_{new}는 새로운 관측값에 대한 예측된 스코어값, S는 주성분모형으로부터의 스코어행렬에 대한 공분산행렬, p_j는 j번째 변수에 대한 $1 \times N$의 차수의 로딩벡터, P는 주성분모형으로부터의 로딩행렬을 의미한다. $c_{new,j}^D$는 양 또는 음의 값을 가질 수 있지만 호텔링의 T^2 통계량의 기여도의 합은 항상 양의 값을 가진다.

3. 결과 및 고찰

3.1. 실시간 원수 수질자료 분석 결과

본 연구에서는 2010년 7월 1일부터 8월 31일까지 G정수장에서 15분 간격으로 측정된 실시간 자료를 이용하여 실시간 수질 이상 판단을 위해 다변량 알고리즘을 적용하였다. 정수장에 유입되는 원수의 기본 수질 항목인 pH, 수온, 알카리도, 전기전도도, 탁도에 대한 시계열적 패턴분석 결과는 Fig. 1과 같다.

또한 Table 1은 Fig. 1의 탁도, pH, 수온, 알카리도, 전기전도도에 대한 피어슨 상관분석 결과로서 탁도는 pH, 수온, 알카리도, 전기전도도와 음의 상관관계를 가지며, pH는 수온, 알카리도, 전기전도도와 양의 상관관계를 가지고 있고, 괄호 안은 P-값을 의미한다. 수온은 알카리도, 전기전도도와 양의 관계를 가지고 있으며, 알카리도는 전기전도도와 양의 상관관계를 가지고 있는 것으로 나타났다. 따라서 5가지 수질 항목의 경우 서로 독립적이지 않고 상관성을 가지고 있는 것으로 판단되어 개별 항목에 대한 분석 보다는 주성분 분석을 통해 새로운 주성분을 통한 분석이 더 유용함을 알 수 있다.

Table 1. Correlation of real-time monitoring parameters

Parameter	Turbidity	pH	Temperature	Alkalinity	Conductivity
Turbidity	1.0000	-0.4582 (<0.001)	-0.5146 (<0.001)	-0.6094 (<0.001)	-0.7061 (<0.001)
pH	-0.4582	1.0000	0.6360 (<0.001)	0.4513 (<0.001)	0.3754 (<0.001)
Temperature	-0.5146	0.6360	1.0000	0.4585 (<0.001)	0.5207 (<0.001)
Alkalinity	-0.6094	0.4513	0.4585	1.0000	0.6612 (<0.001)
Conductivity	-0.7061	0.3754	0.5207	0.6612	1.0000

3.2. 다변량 통계방법 적용 결과

다변량 통계방법은 하나의 센서에서 측정되는 여러 변수들의 측정값을 동시에 고려하여 이상여부를 판단한다. 상시 계측시스템의 경우 시스템 특성상 실시간으로 여러 변수에 대한 데이터가 한 번에 측정되어 다변량 통계방법을 이용하면 새로운 측정값이 실시간으로 측정될 때마다 그 측정값을 반영하여 갱신된 새로운 분석모형을 통해 데이터의 이상여부 판단이 가능하다. 즉, 데이터가 새롭게 관측됨에 따라 분석모형 역시 매번 갱신하여 최신 모형을 통해 분석할 수 있다.

따라서 본 연구에서는 데이터기반의 다변량 통계분석을 통해 계측데이터를 활용한 상태분석 및 이상징후 판정기법을 개발하고자 한다. 본 연구에서는 이와 같은 과정을 R 프

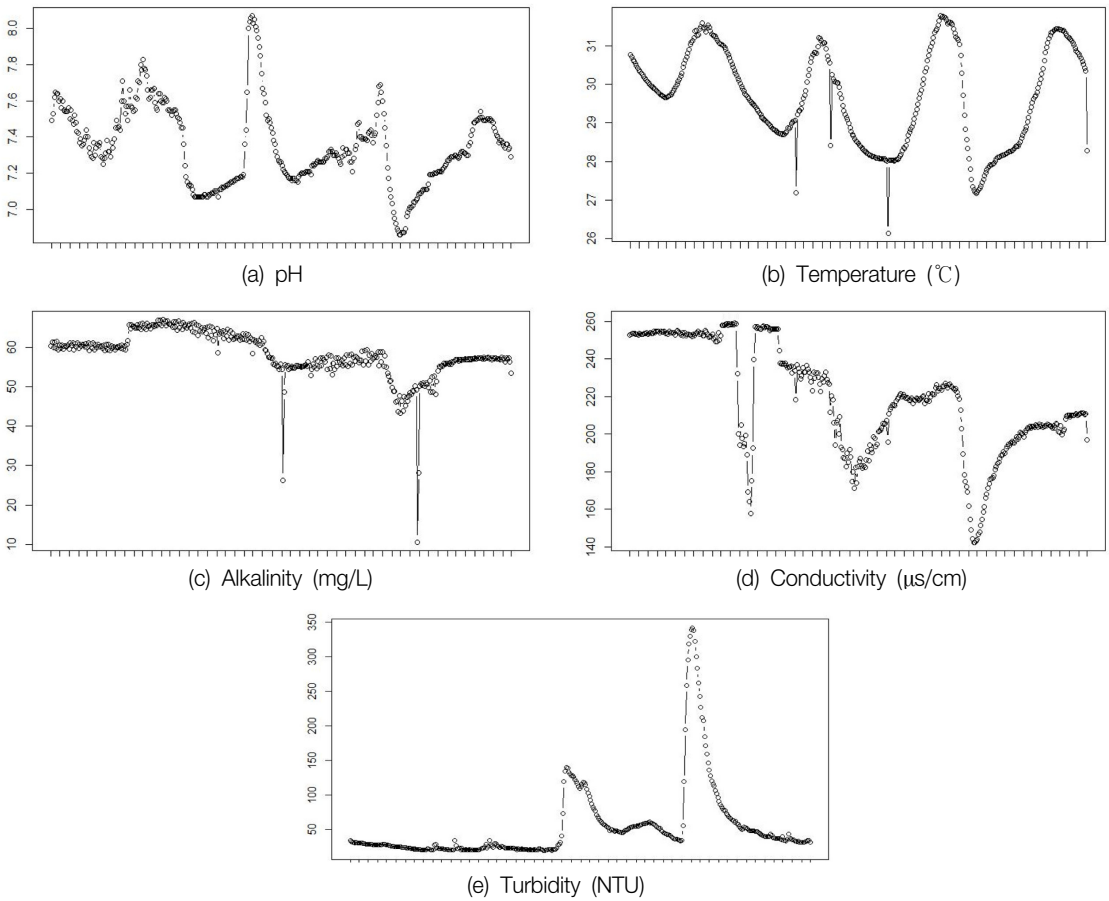

Fig. 1. Time series plot of pH, temperature, alkalinity, conductivity and turbidity.

로그램을 이용한 분석알고리즘을 개발하여 자동화되어 수행되었다. R 프로그램은 통계분석용 프로그램으로서 프로그래밍 언어이기 때문에 사용자가 원하는 새로운 방법으로 프로그램 하기가 편리하다. 따라서 추후 분석모형의 데이터 범위 지정 및 관리한계값 등을 사용자의 판단에 의해 조정하여 분석하기 용이할 것으로 판단된다.

본 연구에서 제안한 다변량 통계방법을 통한 수질이상에 대한 이상 여부를 판단하기 위해 95%, 99%, 99.99%와 같은 세 등급의 신뢰도를 통해 등급별 수질이상 여부를 평가하였다.

Fig. 2에서와 같이 호텔링의 T^2-통계량과 Q-통계량의 관리한계선은 모두 95%, 99% 그리고 99.99%의 3단계로 구분하여 제시하였으며, 3개의 관리한계선 중 녹색선은 95%, 파란선은 99%, 빨간선은 99.99%의 신뢰도를 의미한다. 여기서 95.0%의 관리한계선은 보수적인 측면의 한계선으로 1단계 관리한계선이며, 99.99% 관리한계선은 엄격한(restrict) 측면의 한계선으로 3단계 관리한계선을 의미한다.

관리한계선은 이상 데이터 판단을 위한 것으로 한계선 내에 값이 분포하면 정상상태로 판단하며 한계값을 초과하게 되면 이상발생으로 판단한다. 이러한 신뢰도 기반의 경계수준 구분은 수질이상 여부에 대한 경보 수준을 제시할 수 있다는 장점을 가지고 있으며, 이러한 경보 수준의 설정은 현재 정수장에서 수질이상 상황을 근무자에게 신속히 전파하여 단계적인 대응조치를 통해 수질이상으로 인한 피해를 최소화할 수 있도록 도와줄 수 있다.

3.3. 이상치 여부 판단 결과

본 연구에서는 K-water에서 운용중인 정수장 자료를 이용하여 다변량 통계방법을 적용하기 위해 정상상태의 과거 자료를 통해 호텔링의 T^2-통계량과 Q-통계량(SPE)을 계산한 후 실시간으로 자료가 측정될 때마다 가장 과거의 값을 제외하고 새로운 값을 반영한 호텔링의 T^2-통계량과 Q-통계량(SPE)을 갱신하여 실시간으로 새로운 호텔링의 T^2-통계량과 Q-통계량(SPE)을 계산할 수 있도록 적응형 알고리즘을 구축하였다.

Fig. 3과 4의 검은색 점(●)은 정상과정에서의 호텔링의 T^2-통계량과 Q-통계량(SPE)을 나타내고 회색 점(●)은 이상발생에 대한 호텔링의 T^2-통계량과 Q-통계량(SPE)을 나타낸다. 또한 초록색, 파란색, 빨간색의 선은 95%, 99%, 99.99%의 관리한계선을 의미한다.

본 연구에서는 실시간 이상 여부 판단 알고리즘을 통해 게릴라성 폭우 등으로 인한 원수의 급격한 수질변동이 정수장 수질관리에 어떤 영향을 미치는지를 평가하였다. 이를 위해, 기상청의 지역별 상세 강우 관측 자료와 G정수장에서 운용중인 수질계측기 자료을 이용하여 다변량 통계알고리즘을 적용하였고, 그 결과를 통해 얻어진 이상 자료 판별을 통해 강우가 수질이상 여부에 미치는 영향을 확인하였다.

강우에 대한 수질 오염물질의 영향력 및 영향 도달 시간을 확인하기 위해 G정수장과 가장 거리가 가까운 무인자동기상관측장비(Automatic Weather Station, AWS)에서 관측된 기상자료를 살펴보면 2010년 8월 6일 오후 16:43분에 강우가 산발적으로 시작되어 16:47분부터 본격적으로 강우가 시작되었으며, 2010년 8월 6일 밤 21:18분에 강우가 멈춘 것으로 나타났으며 약 0.5~8 mm까지의 양으로 강우가 지속된 것으로 측정되었다.

강우 시작 시점과 멈춤 시점에 대해 다변량 통계방법을 통해 수질이상을 판단한 결과, Fig. 3에서와 같이 2010년 8

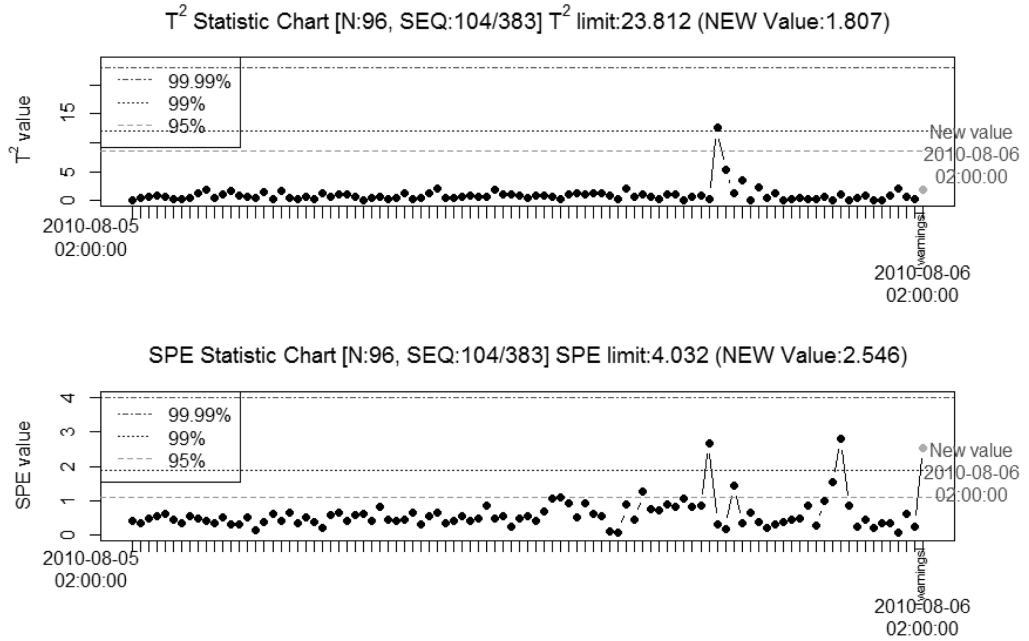

Fig. 2. Example of water quality evaluation by real time T^2 and SPE value.

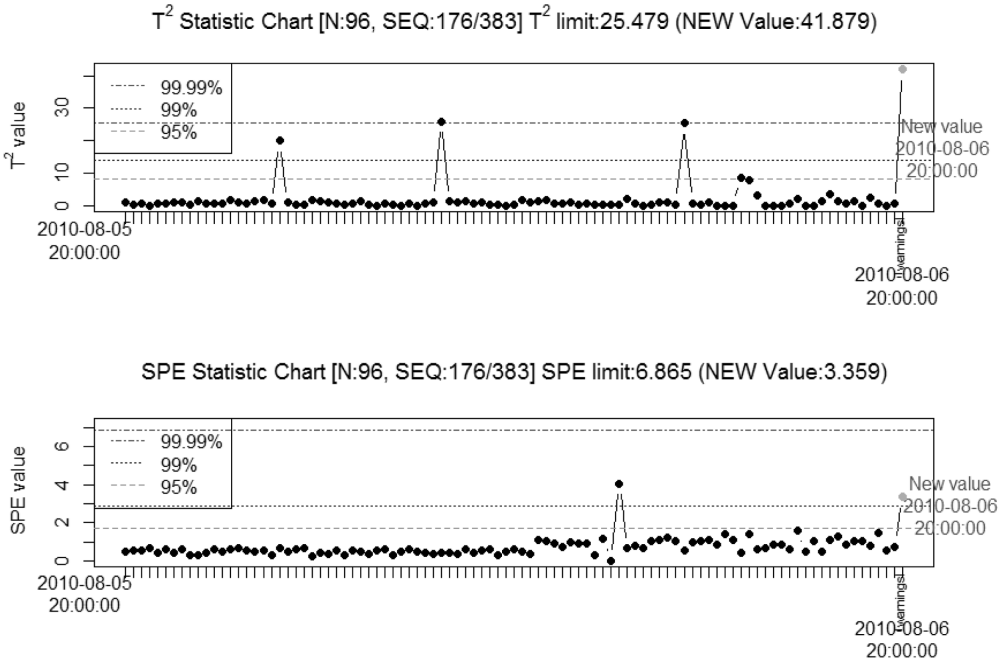

Fig. 3. The results of real time T^2 and SPE from starting anomaly data at 20:00 pm on Aug. 6, 2010.

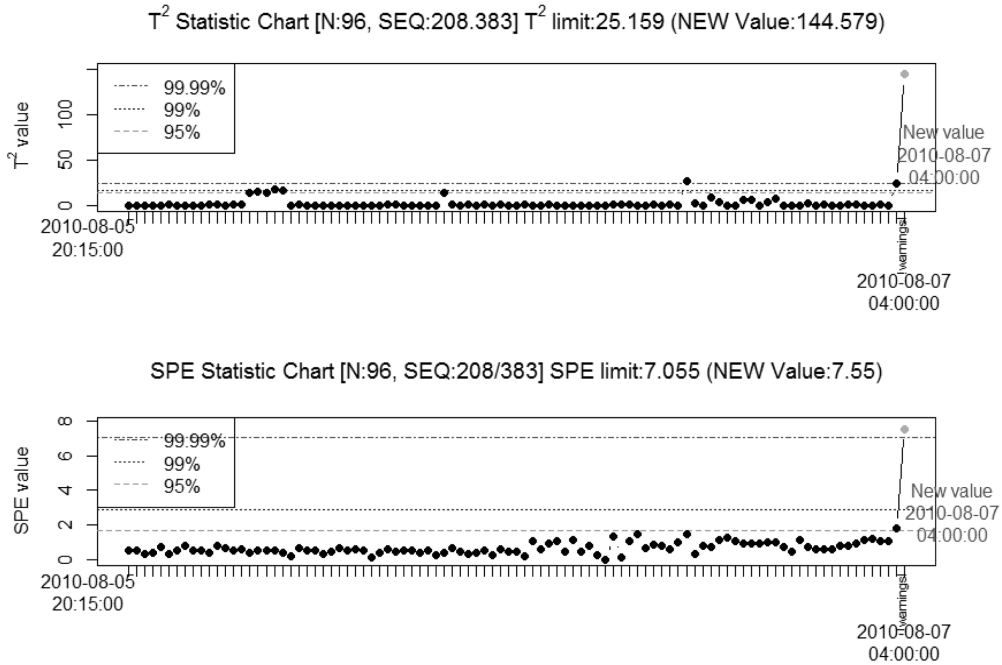

Fig. 4. The results of real time T^2 and SPE from ending anomaly data at 04:00 am on Aug. 7, 2010.

월 6일 오후 16:43분에 시작된 강우의 영향으로 2010년 8월 6일 저녁 8:00경에 호텔링의 T^2-통계량의 경우 41.879로 3단계 관리한계선인 99.99%의 임계값인 25.129보다 커서 모든 관리한계선 밖에 위치하게 되어 3등급의 이상 상태로 판단하였으며, Q-통계량(SPE)의 경우 3.359로 가장 높은 관리한계선인 99.99%의 임계값인 6.865보다는 작아 2등급의 이상상태로 판단하였다. 이러한 결과는 강우가 일어날 경우 해당 정수장의 수질감시장치에 이상치가 나타나기까지 약 2시간이 소요되어 향후 강우 시 정수장에서 조기 대응을 위한 유용한 정보로 활용될 수 있다고 판단된다.

Fig. 4의 결과를 보면 2010년 8월 7일 새벽 04:00경에 마지막으로 이상치로 판단하고 그 이후는 정상으로 다시 회복된 결과를 보여주고 있어 강우가 멈춘 시점인 8월 6일 밤 21:18분 이후에 2010년 8월 7일 04:00까지 산발적으로 이상치가 포착된 것은 강우가 끝난 시점 이후부터 약 7시간 정도 정수장 수질에 영향을 미치는 것으로 판단할 수 있다.

마지막으로 Fig. 5는 2010년 8월 6일 20시에 이상치로 판

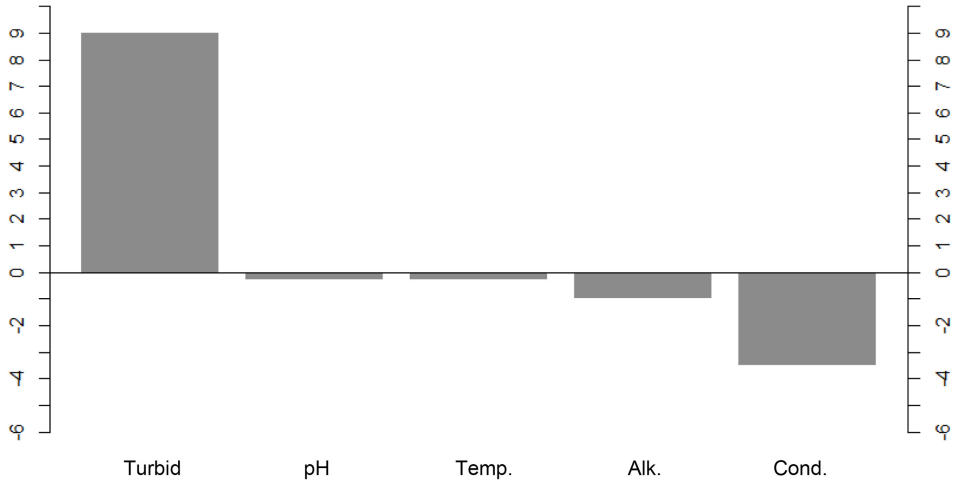

Fig. 5. Contribution plot for 20:00 pm on Aug. 6, 2010.

단된 자료에 대한 기여도 그림(contribution plot)을 나타낸 것으로 탁도와 전기전도도에 의해 이상치로 판단되었음을 알 수 있다. 따라서 기여도 분석을 통해 그 기여도가 가장 큰 변수들을 선정하여 이상치에 대한 원인을 찾을 수 있다는 장점을 가지고 있다.

4. 결 론

본 연구는 다변량 통계기법 중 하나인 주성분분석을 활용하여 실시간으로 수질이상 유무를 판단할 수 있는 경보 시스템 개발을 목적으로 하였다.

실시간 이상 발생 검출에 대한 다변량 통계방법 응용은 세 단계로 이루어졌다. 첫째는 정상적인 과정(stationary process)에 대한 모형설정으로 통계적 방법 중 주성분 분석법을 이용하였으며, 둘째는 모형을 통해 정상적인 과정을 벗어나는 이상치(outlier)를 식별하는 과정으로 호텔링의 T^2-통계량과 Q-통계량(SPE)을 활용하였다. 마지막 세 번째는 이상치에 대한 이유 또는 원인을 확인하는 단계로 기여도 그림(contribution plot)으로 수행하였다.

실시간 감시 알고리즘의 활용성을 검증하기 위해, 2010년 7월 1일부터 8월 31일까지 G정수장에서 15분 간격으로 측정된 실시간 원수 자료를 이용하였으며, 게릴라성 폭우 등으로 인한 원수의 급격한 수질변동이 어느 정도 영향을 미치는지를 확인하기 위해 기상청에서 운용중인 무인자동기상관측장비(Automatic Weather Station, AWS)에서 관측된 기상 자료도 함께 활용하였다.

그 결과, 첫째 정수장에 유입되는 원수의 기본 수질 항목인 pH, 수온, 알카리도, 전기전도도, 탁도에 대한 피어슨 상관분석 결과, 5가지 수질 항목의 경우 서로 독립적이지 않고 상관성을 가지고 있는 것으로 판단되어 개별 항목에 대한 분석 보다는 주성분 분석을 통한 분석이 더 유용함을 알 수 있었다.

둘째, 수질이상 여부를 판단하기 위해 호텔링의 T^2-통계량과 Q-통계량을 이용한 결과, 초기 강우가 정수장의 수질 감시장치에 이상치로 나타나기까지 약 2시간이 소요되었으며, 강우가 끝난 시점부터 약 7시간 정도까지 정수장 수질에 영향을 미치는 것으로 나타났다.

마지막으로 기여도 분석 결과, 이상치에 대한 기여도가 가장 큰 변수를 실시간으로 찾아내어 이상치에 대한 원인을 규명할 수 있었다.

본 연구자가 개발한 수질이상에 대처할 수 있는 수질감시 알고리즘을 기상청 AWS 자료와 연계한다면, 집중강우와 같은 급격한 수질변동에도 정수장에서 사전 대응이 가능할 것으로 판단된다. 또한, 본 연구에서 제안한 기법은 국내 주요 하천에 설치・운영 중인 수질자동측정망이나 지하수 오염 등과 같은 다양한 수질오염에 대한 조기경보 알고리즘으로 확장할 수 있다는 장점을 가지고 있으며, 지자체의 간이상수도 또는 소규모 급수시설에 대해서도 최적의 효과를 기대할 수 있다고 판단된다.

Acknowledgement

이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2012R1A1A1040358).

KSEE

References

1. Oh, H. J., Hwang, T. M., Jung, J. H. and Bahn, S. H., "Development of Web-Based Auto Monitoring System for the Control of Conventional Water Treatment System," Proceeding of KSEE, Chonbuk University, pp. 1086~1090(2004).
2. Mason, R. L., Tracy, N. D. and Young, J. C., "Monitoring a multivariate step process," *J. Qual. Technol.*, **28**, 39~50(1996).

3. Mason, R. L. and Young J. C., "Multivariate Statistical Process Control with Industrial Applications," ASA-SIAM(2002).
4. Nomikos, P. and MacGregor, J. F., "Multivariate SPC Charts for Monitoring Batch Processes," *Technometrics*, **37**(1), 41~58(1995).
5. Wise, B. M. and Gallagher, N. B., "The Process Chemometrics Approach to Process Monitoring and Fault Detection," *J. Proc. Control*, **6**(6), 329~348(1996).
6. Chen, Q., Wynne, R. J., Goulding, P. and Sandoz, D., "The Application of Principal Component Analysis and Kernel Density Estimation to Enhance Process Monitoring," *Control Eng. Prac.*, **8**(5), 531~543(2000).
7. Kourti, T., "Multivariate dynamic data modeling for analysis and statistical process control of batch processes, start-ups and grade transitions," *J. Chemometrics*, **17**(1), 93~109(2003).
8. Montgomery, D. C., Introduction to Statistical Quality Control, 3rd Edition, *John Wiley & Sons Inc.*, USA(1996).
9. Kourti, T. and MacGregor, J. F., "Process Analysis, Monitoring and Diagnosis Using Multivariate Projection Methods," *Chemomet. Intelligent Laboratory Syst.*, **28**(1), 3~21(1995).
10. Yang, Y. J., Haught, R. C. and Goodrich, J. A., "Real-time contaminant detection and classification in a drinking water pipe using conventional water quality sensors: techniques and experimental results," *J. Environ. Manage.*, **90**(8), 2494~2506(2009).
11. Lennox, J. and Rosen, C., "Adaptive multiscale principal components analysis for online monitoring of wastewater treatment," *Water Sci. Technol.*, **45**(4-5), 227~235(2002).
12. Baggiani, F. and Marsili-Libelli, S., "Real-time fault detection and isolation in biological wastewater treatment plants," *Water Sci. Technol.*, **60**(11), 2949~2961(2009).
13. Choi, S. W., Martin, E. B., Morris, A. J. and Lee, I., "Adaptive Multivariate Statistical Process Control for Monitoring Time-varying Processes," *Ind. Eng. Chem. Res.*, **45**(9), 3108~3118(2006).
14. Hotelling, H., "The Generalization of Student's Ratio," *Annal. Mathematical Statistics*, **2**, 360~378(1931).
15. Hotelling, H., Multivariate Quality Control, McGraw-Hill, New York(1947).
16. Mason, R. L., Tracy, N. D. and Young, J. C., "Decomposition of T^2 for multivariate control chart interpretation," *J. Qual. Technol.*, **27**(2), 99~108(1995).
17. Mason, R. L., Tracy, N. D. and Young, J. C., "A practical approach for interpreting multivariate T^2 control chart signals," *J. Qual. Technol.*, **29**(4), 396~406(1997).
18. Mason, R. L. and Young, J. C., "Improving the sensitivity of the T^2 statistic in multivariate process Control," *J. Qual. Technol.*, **31**(2), 155~165(1999).
19. George, J. P., Chen, Z. and Shaw, P., "Fault detection of drinking water treatment process using PCA and Hotelling's T^2 chart," *World Acad. Sci. Eng. Technol.*, **3**, 733~738(2009).
20. Jackson, J. E. and Mudholkar, G. S. "Control procedures for residuals associated with principal component analysis," *Technometrics*, **21**(3), 341~349(1979).