서울特別市 保健環境硏究院報 第54號 (2018)
Report of S.I.H.E., 54:229~244 (2018)

통계모형을 이용한 수질자동측정망 자료 활용방안 연구 (감시기준 설정을 중심으로)

물환경생태팀 · 서울시립대학교 환경공학부*

이진효 · 윤호균 · 하현주 · 조수석 · 양인혜 · 이승구 · 천정완 · 이태호
양진영 · 김지희 · 김진아 · 길혜경 · 이목영 · 정 권 · 구자용*

Utilization Plan of Automatic Water Quality Monitoring Networks Data with Statistical Models (Focusing on Setting the Monitoring Standard)

Aquatic Ecology Team ·
*Department of Environmental Engineering, University of Seoul**

Jin-hyo Lee, Ho-kyun Yoon, Hyun-ju Ha, Soo-seock Cho, In-heai Yang, Seung-koo Lee, Chung-wan Chun, Tae-ho Lee, Jin-young Yang, Ji-hui Kim, Jin-a Kim, Hye-kyung Kil, Mok-young Lee, Kweon Jung and Ja-yong Koo*

Abstract

Automatic water quality monitoring networks(AWQMNs) are restricted in their use because measurable items are limited and because their reliability is low owing to the nature of the data and mechanical defects. Therefore, when the original data is used as it is, it may contain many errors. In order to improve the quality of the automatic measurement data, it is necessary to statistically select outliers and to judge the effectiveness of these selection methods. Accordingly, in this study, we used local regression models(LOESS), box-plots, and Z-scores to select outliers in the hourly data (DO, TN, TP) of Seonyu AWQMNs for five years from 2013 to 2017. From the results of the three abovementioned methods, the outliers were as follows: 1) LOESS: DO 640(1.5%), TN 2,863(6.8%), TP 922(2.2%); 2) Box-plot: DO 1(0.0%), TN 1,458 (3.5%), TP 269(0.6%); and 3) Z-score: DO 1(0.0%), TN 1,052(2.5%), TP 163(0.4%). Regarding the concentrations of each item, DO was 9.9 mg/L, TN was 3.991 mg/L, and TP was 0.139 mg/L before outlier removal. When outliers were removed, the following

concentrations were obtained: 1) LOESS: DO = 10.0 mg/L, TN = 3.940 mg/L, TP = 0.138 mg/L; 2) Box-plot: DO = 9.9 mg/L, TN = 3.752 mg/L, TP = 0.138 mg/L; and 3) Z-score : DO = 9.9 mg/L, TN = 3.795 mg/L, TP = 0.138 mg/L. As a result of comparing the mean values after outlier removal by the three methods, there were no differences among DO results before outlier removal, after box-plot outlier removal, and after Z-score outlier removal, but there was a statistically significant difference in the results after outlier removal by local regression analysis. In the case of TP, there were no statistically significant differences between the results before and after outlier removal according to the three methods. Conversely, in the case of TN, both the outcomes before and after outlier removal according to the three methods were statistically significant. These results show that the effect of outlier selection cannot be ignored.

Key words : AWQMNs, ARIMA, time series model, R 3.4.3

서 론

수질자동측정망은 자동화된 시스템을 통하여 연속적이고, 실시간으로 측정이 가능하며, 막대한 현장 자료를 축적함으로써, 하천이나 상수원의 수질 경향을 정확하게 파악하고, 수질오염사고 예방 및 상수원 관리에 중요한 역할을 한다(1~2). 서울시는 공공수역의 수질오염을 예방하기 위하여 1975년 노량진에 수질측정소를 처음 설치한 이래 한강과 주요 지천에 수질자동측정망을 설치·운영하여 수질 실태를 파악하고 있다. 특히 우리 연구원에서는 잠실수중보 하류지역을 관리·운영함으로써 생태계 보호 및 시민에게 안전하고 깨끗한 물환경을 제공하고 있으며, 이를 위해 한강본류에 노량진, 선유 2개 측정소와 주요 지천인 안양천, 중랑천, 탄천에 3개 측정소를 운영하고 있다. 각 측정소에서는 용존산소(DO), 총질소(TN), 총인(TP), 전기전도도(EC), 총유기탄소(TOC) 등을 실시간으로 분석하여, 급격한 수질변동, 유해물질 유입 등에 대해 통합적으로 감시하며 수질오염사고에 대처하고 있으며, 이러한 수질 측정자료는 2015년 3월부터 『열린 데이터 광장(http://data.seoul.go.kr)』에서 공개되고 있다.

하지만 수질자동측정망은 현장여건, 복잡한 전처리 과정 등으로 측정 가능 항목이 제한되고, 또한 기계적 결함으로 인한 데이터 불량 등의 문제점도 갖고 있다. 만약 이러한 원(raw) 자료를 그대로 사용할 경우 많은 오류가 내포되어 실제 측정값과 편차가 큰 통계량이 발생할 수 있기 때문에 자동측정 자료의 품질개선을 위하여 통계적으로 이상치(outlier)를 선별하는 작업, 유효성을 판단하는 작업이 선행되어야 한다(3). 여기서 이상치란 측정값 분포에서 극단적으로 크거나 작은 값 즉, 극단치(extreme value)를 말하는데 데이터에 극단치가 포함되어 있으면 분석결과가 왜곡될 수 있다.

이에 본 연구에서는 수질자동측정망 자료의 이상치 선별을 위해서, 최근 환경자료 분석에 많이 소개되고 있는 국소회귀모형(local regression, LOESS)을 이용하여 이상치를 선별하고, 이상치 제거 전·후 자료의 통계량을 살펴보았다(3). 동시에 일변량 데이터에서 흔히 사용되고 있는 Box-plot (box and whisker plot, 상자그림) 기법 및 Z-score(표준점수) 방법에 의한 이상치 선별결과를 비교하여 향후 수질점검이 필요한 농도를 설정하는 등 수질자동측정망 상시 상황관제 운영 및 감시기준 설정의 참고자료로 활용하고자 하였다.

재료 및 방법

1. 자료수집

상대적으로 결측치(missing value)가 가장 적고, 특히 사전에 점검·동작불량 등에 의한 기계적 결함이 내포된 자료를 사전에 배제한 한강수계 선유 수질자동측정망으로부터(그림 1) 2013년 1월부터 2017년 12월까지 5년간 시간단위로 측정된 자료를 사용하였다. 수질항목은 대표적 수질인자로서 하천수에 녹아있는 산소의 양을 나타내는 DO와 조류발생과 밀접한 관련이 있는 TN, TP 등 3개 항목을 선정하였다.

2. 자료 처리방법

본 연구에서는 풍부한 수리 방법론과 작업도구뿐 아니라 최신식 시각화 방법들, 특화된 자료 분석도구 등을 포함한 수백 개의 패키지를 기본적으로 제공하는 오픈소스 언어인 R 3.4.3을 이용하여 국소회귀분석, Box-plot, Z-score 방법을 통한 이상치를 선별하였다(그림 2~3).

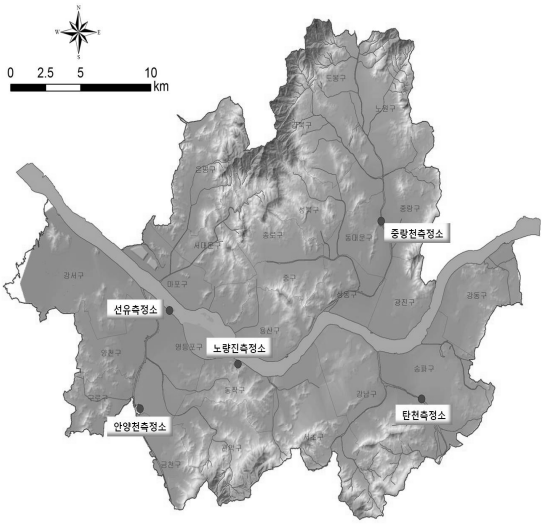


Fig. 1. The location of 5 automatic water quality monitoring networks.

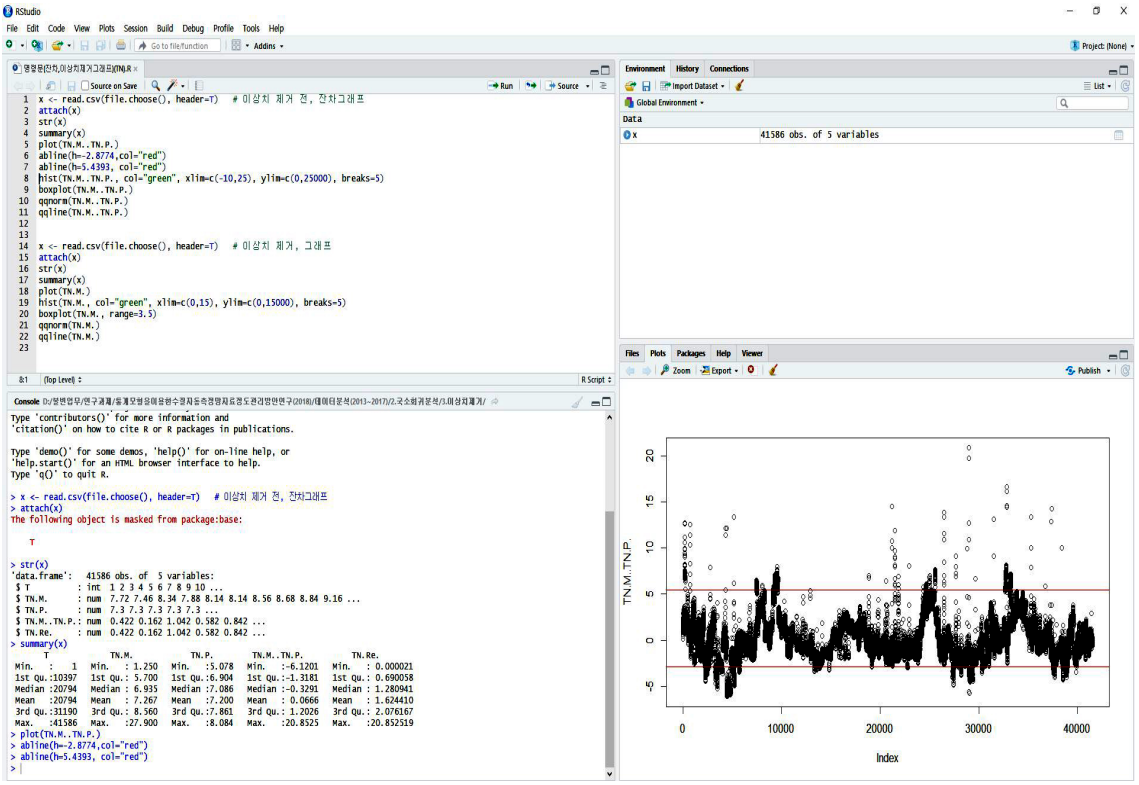


Fig. 2. The screen of RStudio.

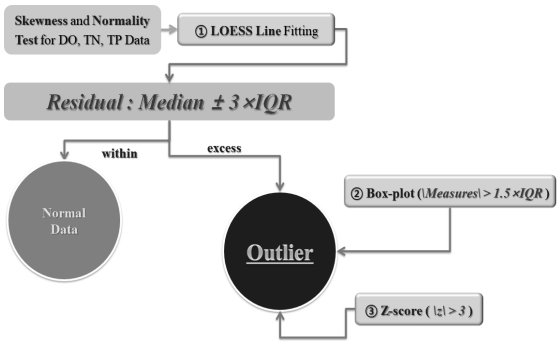


Fig. 3. Scheme of method.

1) 국소회귀분석

국소회귀분석은 전체적으로 모든 자료를 적합시키지 않고, 최근 자료에 가중치를 높여서 중요도를 높이고, 반면에 목표점에서 먼 자료들은 가중치를 낮추는 등 비선형함수를 유연하게 적합시키는 매우 유연한 모형을 통해 예측이 가능한 통계적 분석방법이다(4~5). 국소회귀분석을 하는 이유는 대부분 자료에 존재하는 시간적 추세(temporal trend), 계절성(seasonality) 등을 제거하고, 계산상의 부담을 줄일 수 있기 때문이다(6~8). 따라서 이상치에 둔감하여 환경자료와 같이 이상치가 많은 경우 모수적 시계열 분석에 비하여 신뢰도가 높은 결과를 얻을 수 있다(3). 본 연구에서는 수질자동측정망 자료의 특성을 좀 더 정확하게 파악하기 위해서 우선 정규성 검정(비정규분포 확인), 왜도(skewness, 자료의 좌우분포 확인)를 통해서 원 자료에 국소회귀선(loess line)을 적합시켰다. 이 국소회귀선 예측값과 원 자료의 차이인 잔차의 분포로부터 중위수(median), IQR (interquatile range, 사분위범위)을 구하여, 잔차가 중위수로부터 3 × IQR 이상으로 벗어난 경우를 이상치로 결정하였다.

2) Box-plot

Box-plot은 최대값, 최소값, 중위수, 사분위수를 사용하여 자료의 측정값들이 어떤 모양으로 분포되어 있으며, 극단치 존재 여부, 개수 등을 쉽게 알 수 있도록 하는 그림을 말하며, 측정값들의 중심위치와 산포도의 척도로 사용할 수 있다. 일반적으로 Box-plot 기법을 이용한 이상치는 제1사분위와 제3사분위인 Q1, Q3의 차이 즉, 사분위범위의 1.5 또는 3배가 넘는 측정값을 이상치로 결정하는데(9~10), 본 연구에서는 측정치 절대값이 1.5 × IQR 보다 클 때, 이상치로 판정하였다.

3) Z-score

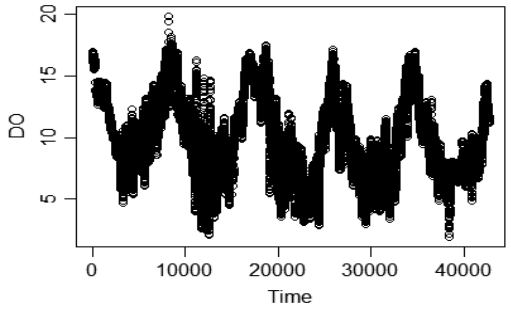
Z-score는 측정값이 평균으로부터 얼마만큼 떨어져 있는지를 확인하는 것으로서, 측정값과 평균의 차이를 표준편차로 나누어서 구한다. Z-score는 측정의 단위로 표준편차를 사용하여 해당분포의 평균과 관련된 점수의 위치를 나타내기 때문에 서로 다른 분포로부터 나온 값들을 비교 가능토록 해주는 역할을 한다. 일반적으로 Z-score 절대값이 2.5~4 이상이면 이상치로 판정하며, 본 연구에서는 Z-score 절대값이 3 이상일 때, 이상치로 판정하였다(11).

결과 및 고찰

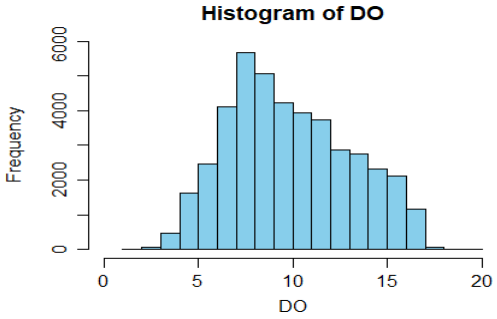
1. 자료분포 특성

2013~2017년 5년간 선유 수질자동측정망에서 수집된 자료 개수는 DO 총 42,705개, TN 총 42,132개, TP 총 41,920개이며, 이 수치에는 기기점검, 채수불량 등에 따른 무효자료는 제외되었다. 측정 자료의 분포와 특성은 그림 4~6과 같다.

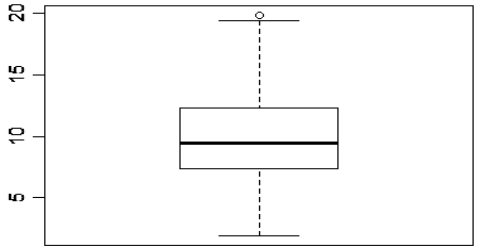
2013~2017년 5년간의 선유 수질자동측정망 DO, TN, TP 평균값은 각각 9.9 mg/L(1.9 mg/L~19.9 mg/L), 3.991 mg/L(1.000 mg/L~15.840 mg/L), 0.139 mg/L(0.017 mg/L~0.500 mg/L)로 나타났다. 시계열 그림(time series plot)을 살펴보면, DO는 전반적으로 측정치 추세에서 벗어나지 않았지만 반면에 TN, TP는 측정치 추세에서 벗어난 자료가 일부 나타났다. 이는 히스토그램(histogram)과 상자그림(box plot)을 통해 확인할 수 있는데, 특히 TN은 여름(2016년 7월, 2017년 7월)과 겨울(2016년 1~2월)에 비정상적으로 농도값이 튀는 현상을 보였는데, 이는 여름철 장마 및 대조기(동일시기에 인천조위관측소 관심 및 주의단계 수시 발령) 등에 의한 일시적인 영향과 겨울철 낮은 온도에서 안정적인 질소



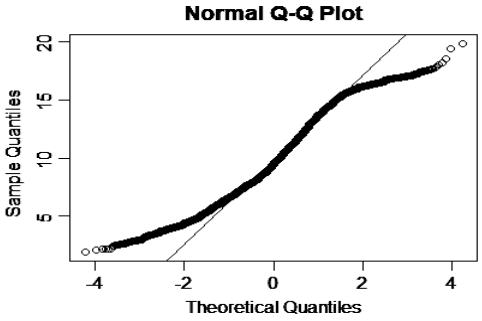
Time series plot of DO



Histogram of DO

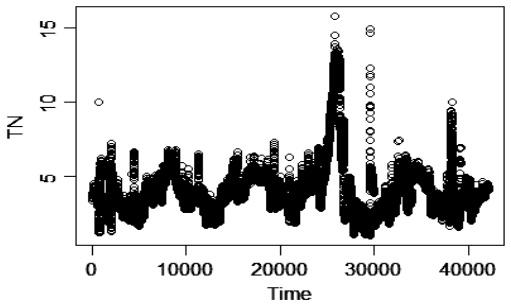


Box plot of DO

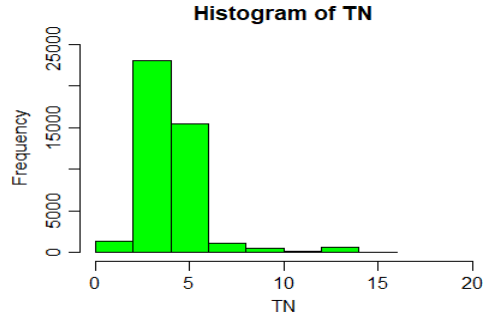


Normality check of DO

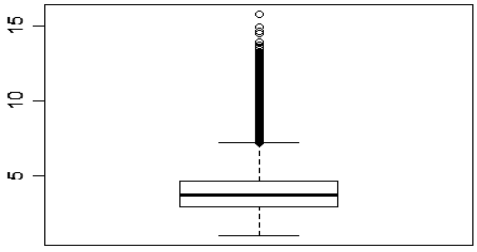
Fig. 4. Statistical characteristics of DO.



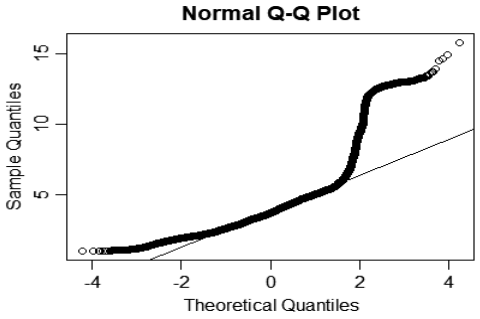
Time series plot of TN



Histogram of TN

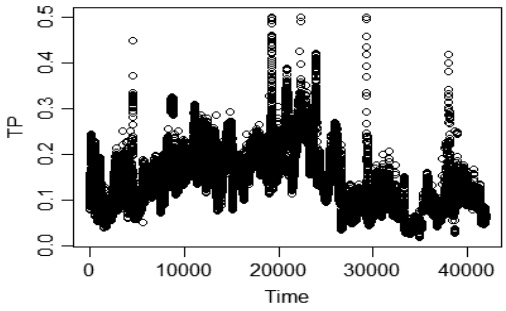


Box plot of TN

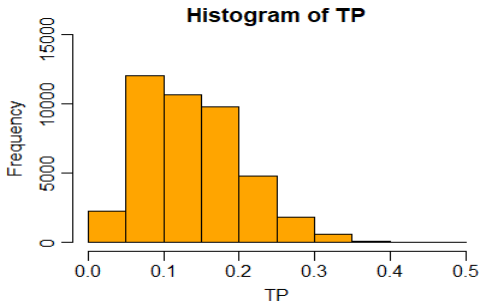


Normality check of TN

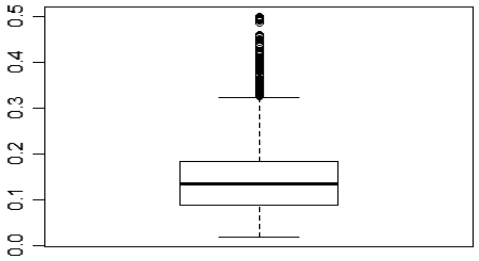
Fig. 5. Statistical characteristics of TN.



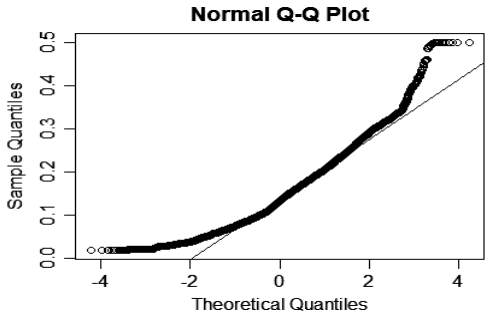
Time series plot of TP



Histogram of TP



Box plot of TP



Normality check of TP

Fig. 6. Statistical characteristics of TP.

처리가 되지 않은 방류수 등(12) 고농도의 질소가 유입되었기 때문이라고 판단된다. 또한 정규확률그림(Q-Q plot)을 통해 자료들이 정규분포를 따르는지 확인할 수 있는데, 그림에서 보듯이 각 항목별 분포가 직선상에서 크게 벗어난 경우가 존재하였고, 동시에 Anderson-Darling normality test, Pearson chi-square normality test 결과, 모두 유의확률 p값이 유의수준 5% 보다 작게 나타났기 때문에 DO, TN, TP 모두 정규분포하지 않는다는 것을 확인하였다. 한편 수질항목별 왜도는 DO 0.26502, TN 2.55525, TP 0.69391로 TN을 제외한 DO, TP는 대체로 대칭분포에 가까웠다. 일반적으로 왜도는 0에 가까울수록 대칭분포에 따르고, 절대값 2를 초과하면 대칭분포에 벗어난 것으로 판단을 하는데, 이에 따라 원 자료에 대해서 적절한 조정(예를 들면 자연로그 변환 등) 실시여부를 판단하는 기준이 된다. 따라서 본 연구에서는 국소회귀분석의 경우, DO, TP는 원 자료 그대로를 사용하고, TN은 자연로그 변환 후 조정된 자료를 사용하여(변환 후 왜도 0.43407) 국소회귀분석을 실시하였으며, 한편 Box-plot, Z-score 방법에서는 DO, TN, TP 모두 원 자료 그대로 사용하여 이상치를 선별하였다.

2. 이상치 선별 및 처리

선유 수질자동측정망에서 측정된 DO, TN, TP 자료를 이용하여 이상치를 선별하고 이를 통해 이상치 선별 전·후 통계량을 비교하였다.

1) 국소회귀분석

이상치는 자료 처리방법에서 언급했듯이 잔차가 중위수로부터 $3 \times IQR(median \pm 3 \times IQR)$ 이상으로 벗어났을 때, 이상치로 결정하였다. 항목별 이상치 기준은 표 1과 같고, 이상치 선별결과, DO는 42,705 중 640개(1.5%), TN은 42,132개 중 2,863개(6.8%), TP는 41,920개 중 922개(2.2%)

Table 1. Outlier criteria by items(LOESS)

Outlier criteria	DO residual		log TN residual		TP residual	
	-5.5127 <	> 10.0564	-0.5638 <	> 1.0012	-0.0789 <	> 0.1372

Table 2. Comparison with statistics about outlier removal(DO, TN, TP).

	DO(mg/L)		TN(mg/L)		TP(mg/L)	
	Before	After	Before	After	Before	After
Min	1.9	3.4	1.000	1.800	0.017	0.017
1st Qu.	7.4	7.5	2.950	3.060	0.088	0.088
Median	9.5	9.6	3.710	3.780	0.133	0.133
3rd Qu.	12.3	12.3	4.670	4.670	0.182	0.181
Max	19.9	19.9	15.840	10.940	0.500	0.338
Mean	9.9	10.0	3.991	3.940	0.139	0.138
SD	3.2	3.2	1.691	1.204	0.066	0.062
IQR	4.9	4.8	1.720	1.610	0.094	0.093

로 나타났으며, 이상치 제거 후, 자료의 분포특성은 그림 8, 10, 12와 같다. 이상치를 제거하였을 때 전반적으로 DO는 5~15 mg/L 사이에, TN은 2~6 mg/L 사이에, TP는 0.05~0.25 mg/L 사이에서 변동하는 등 이상치 제거 전보다 변폭이 좀 더 작아짐을 볼 수 있었다.

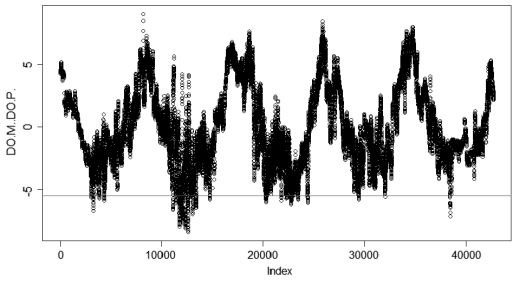
DO, TN, TP 이상치 제거 전·후의 기술통계량을 표 2로 나타냈으며, 표 3~5에서와 같이 Wilcoxon Matt Whitney test 검정결과, DO, TN, TP의 유의확률 p값이 각각 0.0002412, 1.151e-14, 0.306으로 나타났다. DO, TN의 경우, 유의수준 5% 이하에서 이상치 제거 전·후의 평균값의 변화가 발생하는 등 기술통계량에서 유의한 차이가 나타났으나 TP의 경우는 뚜렷한 차이를 볼 수 없었다. 이는 DO, TN의 경우, 일반적인 측정값과 이상치와의 편차가 너무 크거나 또는 이상치 비율이 상대적으로 높아서 이상치 제거 전·후 유의한 차이가 발생한 것으로 판단된다. TP는 이상치가 평균치 위아래에서 비교적 고르게 분포되어 있어 그 영향이 적게 나타났고, 측정된 자료가 40,000개 이상으로 방대하여 이상치 제거에 대한 효과는 적었을 것으로 판단된다. 이상치를 결정하기 위하여 국소회귀모형의 값과 측정값의 차이인 잔차를 구하며, 각 항목별 잔차분포 특성을 그림 7, 9, 11로 나타냈다. 시계열 그래프(time series plot)에 나타난 실선이 이상치 판단을 위한 기준선으로 이 범위를 벗어난 경우 이상치로 판단하게 된다.

2) Box-plot, Z-score

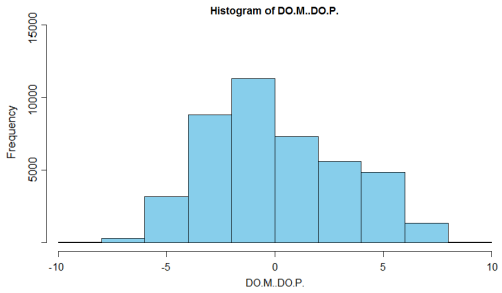
일변량 데이터에서 사용되고 있는 Box-plot 기법, Z-score 방법에 의한 이상치를 선별하여 국소회귀분석에 의한 이상치 선별결과와 비교·분석하였다. 항목별 이상치 기준은 표 6과 같고, Box-plot에 의한 이상치 선별결과, DO는 42,705중 1개(0.0%), TN은 42,132개 중 1,458개(3.5%), TP는 41,920개 중 269개(0.6%)가 이상치로 나타났으며, Z-score 경우, DO는 1개(0.0%), TN은 1,052개(2.5%), TP는 163개(0.4%)가 이상치로 나타났다(그림 13~15).

결론적으로 3가지 분석방법에 따른 이상치 제거 전·후의 기술통계량은 표 7~9와 같다.

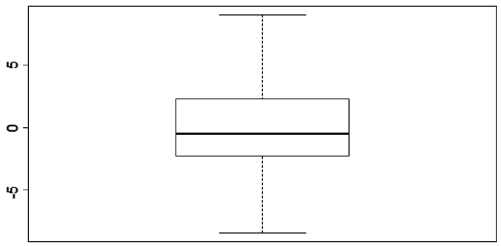
DO, TN, TP 3항목에 대한 이상치 제거 전과 3가지 방법에 따른 평균값이 서로 차이가 있는지



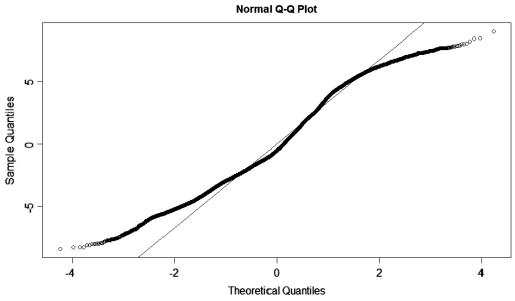
Time series plot of DO residual



Histogram of DO residual

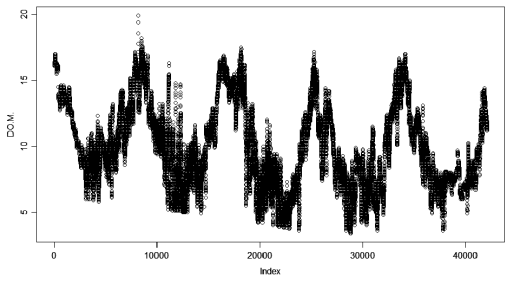


Box plot of DO residual

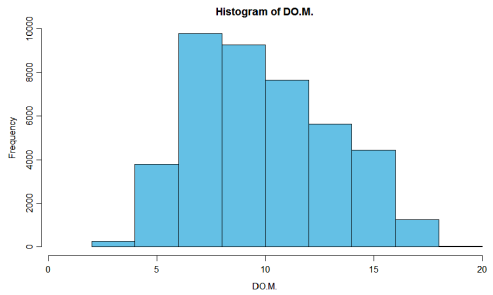


Normality check of DO residual

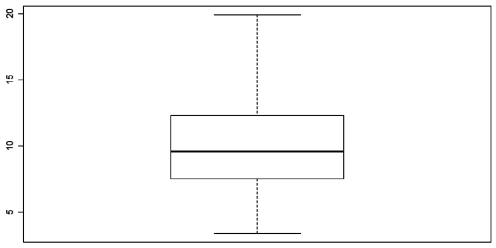
Fig. 7. Plots of DO residual.



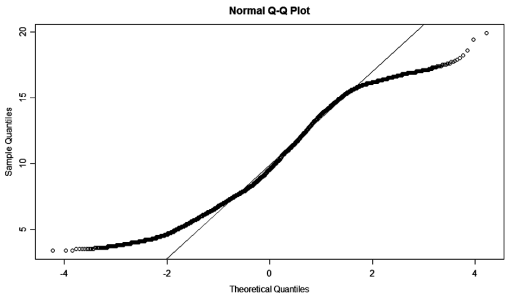
Time series plot of DO



Histogram of DO

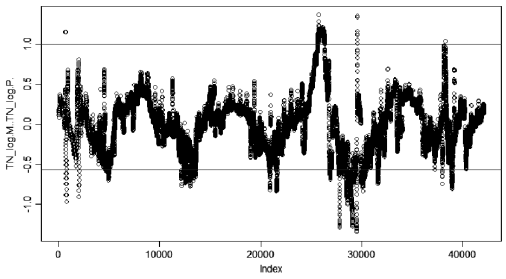


Box plot of DO

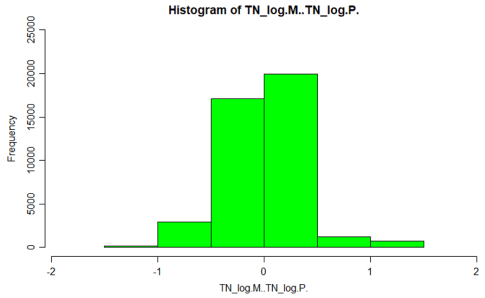


Normality check of DO

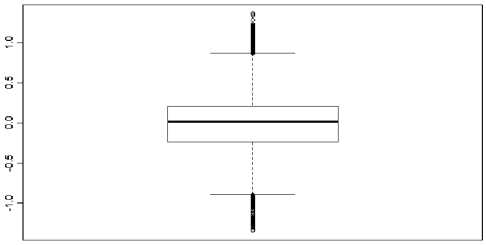
Fig. 8. Statistical characteristics of DO after outlier detection and omission.



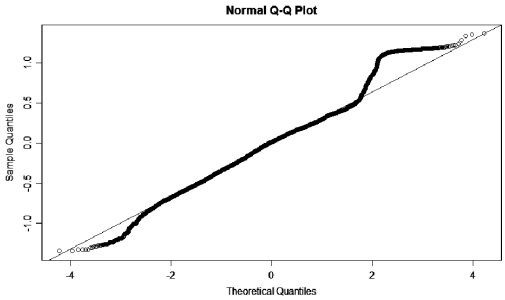
Time series plot of TN_ln residual



Histogram of TN_ln residual

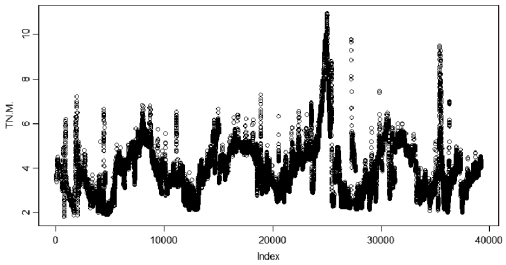


Box plot of TN_ln residual

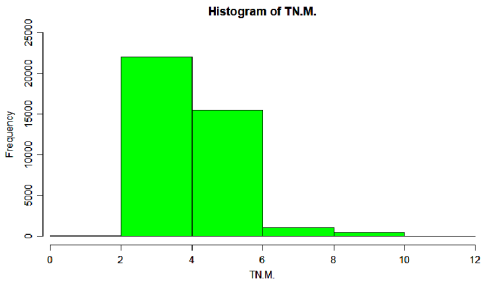


Normality check of TN_ln residual

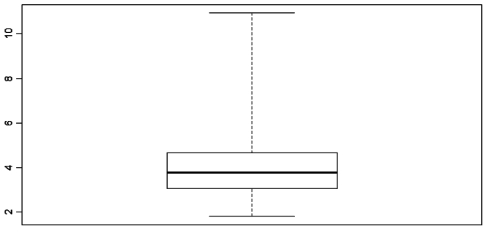
Fig. 9. Plots of TN_ln residual.



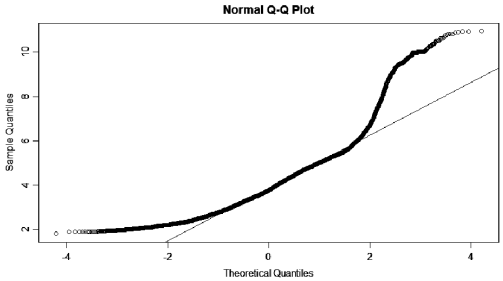
Time series plot of TN



Histogram of TN

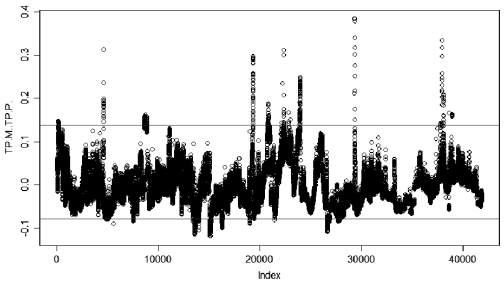


Box plot of TN

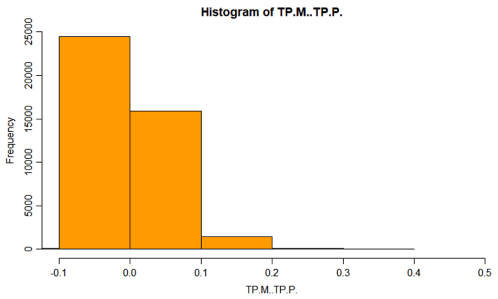


Normality check of TN

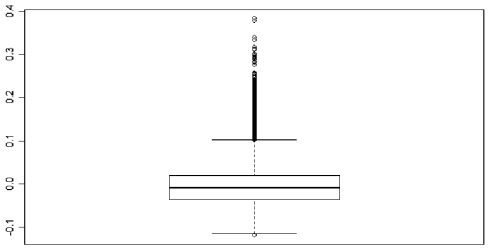
Fig. 10. Statistical characteristics of TN after outlier detection and omission.



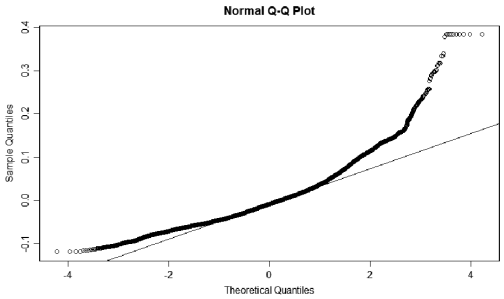
Time series plot of TP residual



Histogram of TP residual

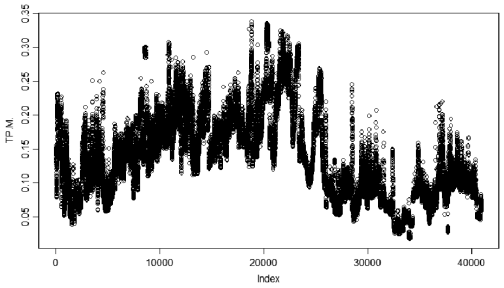


Box plot of TP residual

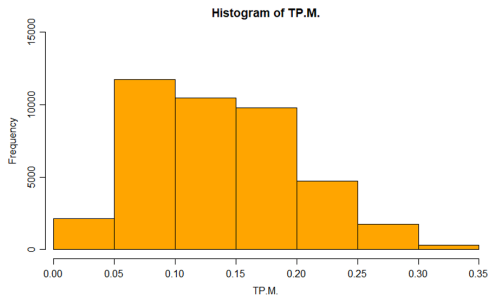


Normality check of TP residual

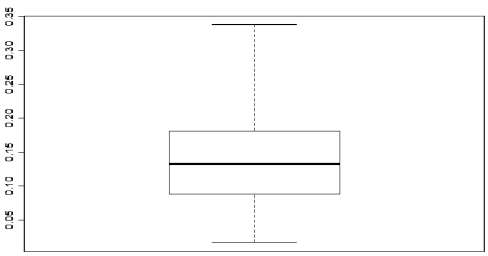
Fig. 11. Plots of TP residual.



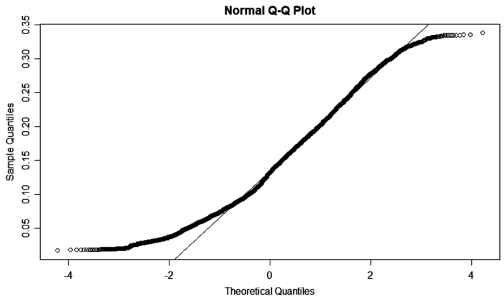
Time series plot of TP



Histogram of TP



Box plot of TP



Normality check of TP

Fig. 12. Statistical characteristics of TP after outlier detection and omission.

Table 3. DO result of independent two-sample Wilcoxon Matt Whitney test

```
> var.test(DO.M..B., DO.M..A.)   # 두 집단 등분산성 확인

        F test to compare two variances

data:  DO.M..B. and DO.M..A.
F = 1.0373, num df = 42704, denom df = 42064, p-value = 0.0001615
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.017769 1.057276
sample estimates:
ratio of variances
          1.037335

> wilcox.test(DO.M..B., DO.M..A., var.equal=FALSE)   # var.equal = FALSE(분산이 다름), TRUE(분산이 같음)

        Wilcoxon rank sum test with continuity correction

data:  DO.M..B. and DO.M..A.
W = 885110000, p-value = 0.0002412
alternative hypothesis: true location shift is not equal to 0
```

Table 4. TN result of independent two-sample Wilcoxon Matt Whitney test

```
> var.test(TN.M..B., TN.M..A.)   # 두 집단 등분산성 확인

        F test to compare two variances

data:  TN.M..B. and TN.M..A.
F = 1.9708, num df = 42131, denom df = 39268, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.932841 2.009481
sample estimates:
ratio of variances
          1.970795

> wilcox.test(TN.M..B., TN.M..A., var.equal=FALSE)   # var.equal = FALSE(분산이 다름), TRUE(분산이 같음)

        Wilcoxon rank sum test with continuity correction

data:  TN.M..B. and TN.M..A.
W = 801370000, p-value = 1.151e-14
alternative hypothesis: true location shift is not equal to 0
```

Table 5. TP result of independent two-sample Wilcoxon Matt Whitney test

```
> var.test(TP.M..B., TP.M..A.)   # 두 집단 등분산성 확인

        F test to compare two variances

data:  TP.M..B. and TP.M..A.
F = 1.1034, num df = 41919, denom df = 40997, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.082347 1.124838
sample estimates:
ratio of variances
          1.103389

> wilcox.test(TP.M..B., TP.M..A., var.equal=FALSE)   # var.equal = FALSE(분산이 다름), TRUE(분산이 같음)

        Wilcoxon rank sum test with continuity correction

data:  TP.M..B. and TP.M..A.
W = 862850000, p-value = 0.306
alternative hypothesis: true location shift is not equal to 0
```

알아보기 위해서 비모수 검정의 일원배치 분산분석의 한 종류인 Kruskal-Wallis 분석을 실시하였다. 또한 각 항목에 대하여 이상치 제거 전과 방법별 평균값이 다른 경우, 어디에서 차이가 있는지 확인하기 위해 Bonferroni adjustment를 이용한 사후분석(post analysis)을 실시하였다(표 10~12). 여기서 GROUP 1은 이상치 제거 전 집단, GROUP 2는 Box-plot을 이용한 이상치 제거 후 집단, GROUP 3은 Z-score를 이용한 이상치 제거 후 집단, GROUP 4는 LOESS를 이용한 이상치 제거 후 집단을 말한다.

Kruskal-Wallis 분석결과, 유의확률 p값이 0.0001553으로 나와 유의수준 5% 이하에서 적어도 한 쌍의 집단에서 유의한 차이가 발생하였으며, 사후분석 결과, DO는 GROUP 1, 2, 3은 서로 차이가 없고, 동시에 GROUP 1, 2, 3과 GROUP 4는 DO 평균값에 유의한 차이가 있는 것으로 나타났다.

마찬가지로 TN 역시 유의확률 p값이 2.2e-16보다 작게 나와 집단별 차이가 존재하였고, 유의수준 5% 이하에서 GROUP 2, 3과의 관계를 제외한 GROUP 1, 2, 3, 4 모두 유의한 차이를 보였다. 즉, TN은 계절별 하천으로 유입되는 질소농도 및 처리효율 변화 등에 의한 영향이 커서 상대적으로 다른 항목보다 이상치 개수가 많이 나타나 이상치 선별에 따른 영향을 무시할 수 없음을 보여주고 있다.

반면 TP의 경우, 유의확률 p값이 0.446으로 나타나, 유의수준 5% 이하에서 집단별로 유의한 차이를 볼 수 없었다.

3. 고찰

수질자동측정망에서 이상치를 선별해야 하는 중요한 이유는 생성된 자료의 신뢰성을 높여 정보공개에 대한 시민의 기대수준에 부응하고, 수질점검이 필요한 농도 설정 등을 통해 상시 수질상태를 모니터링하며, 동시에 이상치 발생여부 자체를 신속히 파악하여 시스템 점검 등 수질자동측정망 시스템이 최적의 상태에서 운영될 수 있도록 하기 위함이다. 효율적인 이상치 선별을 위해서는 연구대상, 분석항목 및 이상치 선별방법에 따라 이상치 제거 전·후 결과가 통계적으로 유의한 차이를 보일 수 있기 때문에 자료분포 특성에 맞는 이상치 선별방법을 적용해야 한다, 특히 본 연구에서

Table 6. Outlier criteria by items(Box-plot, Z-score)

Outlier criteria	DO measures		TN measures		TP measures		Z-score
	0.0500 <	> 19.6500	-0.3700 <	> 7.2500	-0.0530 <	> 0.3230	3

Table 7. Comparison with statistics about outlier removal according to 3 methods(DO)

	DO(mg/L)			
	Before	LOESS	Box-plot	Z-score
Min	1.9	3.4	1.9	1.9
1st Qu.	7.4	7.5	7.4	7.4
Median	9.5	9.6	9.5	9.5
3rd Qu.	12.3	12.3	12.3	12.3
Max	19.9	19.9	19.4	19.4
Mean	9.9	10.0	9.9	9.9
SD	3.2	3.2	3.2	3.2
IQR	4.9	4.8	4.9	4.9

Table 8. Comparison with statistics about outlier removal according to 3 methods(TN)

	TN(mg/L)			
	Before	LOESS	Box-plot	Z-score
Min	1.000	1.800	1.000	1.000
1st Qu.	2.950	3.060	2.910	2.920
Median	3.710	3.780	3.670	3.690
3rd Qu.	4.670	4.670	4.560	4.590
Max	15.840	10.940	7.250	9.060
Mean	3.991	3.940	3.752	3.795
SD	1.691	1.204	1.084	1.161
IQR	1.720	1.610	1.650	1.670

Table 9. Comparison with statistics about outlier removal according to 3 methods(TP)

	TP(mg/L)			
	Before	LOESS	Box-plot	Z-score
Min	0.017	0.017	0.017	0.017
1st Qu.	0.088	0.088	0.088	0.088
Median	0.133	0.133	0.132	0.132
3rd Qu.	0.182	0.181	0.181	0.181
Max	0.500	0.338	0.323	0.336
Mean	0.139	0.138	0.138	0.138
SD	0.066	0.062	0.063	0.064
IQR	0.094	0.093	0.093	0.093

Table 10. Kruskal-Wallis results according to before and after removal of outlier(DO)

```
> kruskal.test(DO ~ GROUP, data = x)     # 비모수 분산분석

        Kruskal-Wallis rank sum test

data:  DO by GROUP
Kruskal-Wallis chi-squared = 20.186, df = 3, p-value = 0.0001553

> pairwise.wilcox.test(DO, GROUP, p.adj='bonferroni', exact=F)     # 각 그룹별 평균비교 (차이유무 확인)

        Pairwise comparisons using Wilcoxon rank sum test

data:  DO and GROUP

  1      2      3
2 1.0000 -      -
3 1.0000 1.0000 -
4 0.0014 0.0014 0.0014

P value adjustment method: bonferroni
```

Table 11. Kruskal-Wallis results according to before and after removal of outlier(TN)

```
> kruskal.test(TN ~ GROUP, data = x)    # 비모수 분산분석

        Kruskal-Wallis rank sum test

data:  TN by GROUP
Kruskal-Wallis chi-squared = 324.31, df = 3, p-value < 2.2e-16

> pairwise.wilcox.test(TN, GROUP, p.adj='bonferroni', exact=F)    # 각 그룹별 평균비교 (차이유무 확인)

        Pairwise comparisons using Wilcoxon rank sum test

data:  TN and GROUP

  1       2       3
2 < 2e-16 -       -
3 2.7e-09 0.086   -
4 6.9e-14 < 2e-16 < 2e-16

P value adjustment method: bonferroni
```

Table 12. Kruskal-Wallis results according to before and after removal of outlier(TP)

```
> kruskal.test(TP ~ GROUP, data = x)    # 비모수 분산분석

        Kruskal-Wallis rank sum test

data:  TP by GROUP
Kruskal-Wallis chi-squared = 2.6663, df = 3, p-value = 0.446

> pairwise.wilcox.test(TP, GROUP, p.adj='bonferroni', exact=F)    # 각 그룹별 평균비교 (차이유무 확인)

        Pairwise comparisons using Wilcoxon rank sum test

data:  TP and GROUP

  1    2    3
2 0.65 -    -
3 1.00 1.00 -
4 1.00 1.00 1.00

P value adjustment method: bonferroni
```

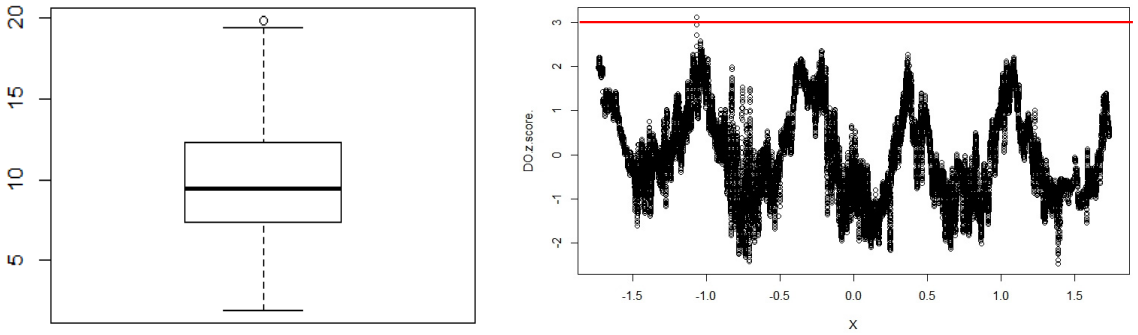


Fig. 13. Box plot(left) and Z-score plot(right) of DO.

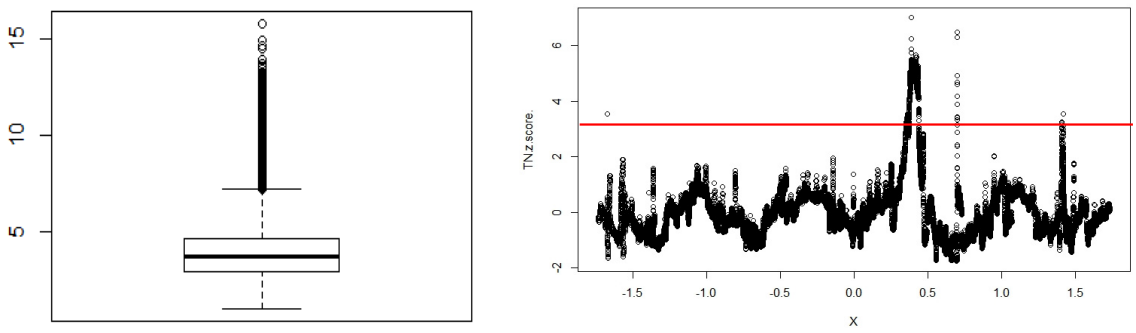


Fig. 14. Box plot(left) and Z-score plot(right) of TN.

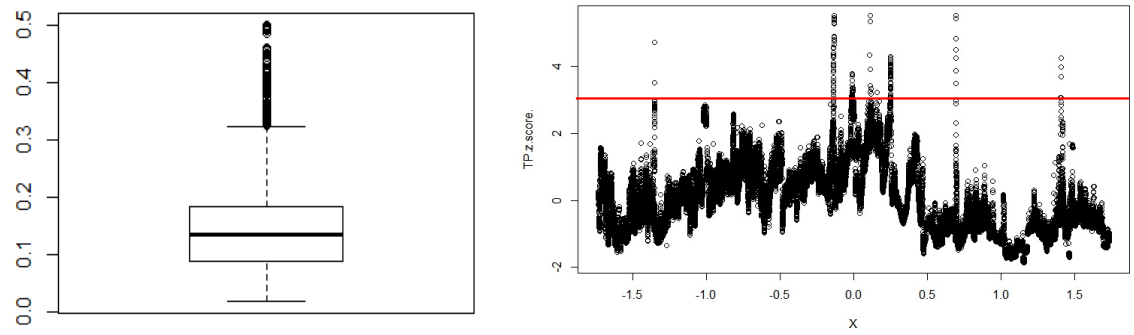


Fig. 15. Box plot(left) and Z-score plot(right) of TP.

보듯이 TN과 같이 이상치 제거 전·후 결과가 3가지 방법 모두에서 차이가 나타나는 항목이 있다면, 지속적으로 이상치 원인분석 및 해결책을 도출하여 데이터 신뢰성 제고 등 수질자동측정망 자료의 품질을 개선해야 할 것이다.

이처럼 계절 및 월별 주기 등 시간적 패턴을 가지고 변하는 수질자동측정망 자료, 환경자료의 경우 일반적으로 사용되고 있는 Box-plot 기법, Z-score 방법과 함께 본 연구에서 소개된 국소회귀분석을 적용함으로써 적절한 이상치를 선별할 수 있을 것으로 기대된다.

결 론

본 연구에서는 수질자동측정망 자료의 이상 유무를 신속하게 파악하기 위해 통계적 분석방법으로 이상치 선별 및 이상치 제거 전·후 자료의 기술통계량을 살펴보았으며, 다음과 같은 결론을 도출하였다.

1. 선유 수질자동측정망에서 시간단위로 측정된 DO, TN, TP 유효자료는 각각 42,705개, 42,132개, 41,920개이며, 이상치 제거 전, 항목별 결과는 DO 9.9 mg/L(1.9 mg/L~19.9 mg/L), TN 3.991 mg/L(1.000 mg/L~15.840 mg/L), TP 0.139 mg/L(0.017 mg/L~0.500 mg/L)으로 나타났다.
2. 국소회귀분석, Box-plot, Z-score 방법에 따른 항목별 이상치 선별결과, DO는 각각 640개(1.5%), 1개(0.0%), 1개(0.0%), TN은 2,863개(6.8%), 1,458개(3.5%), 1,052개(2.5%), TP는 922개(2.2%), 269개(0.6%), 163개(0.4%)가 이상치로 나타났으며, 상대적으로 장마, 대조기, 수온 등 계절 및 월별 주기 등의 영향을 많이 받는 항목인, 이에 원 자료에서 왜도가 높

았던 TN에서 이상치가 많이 발생하였다.

3. 이상치 제거 전과 제거 후 평균값을 비교한 결과, DO는 Box-plot 및 Z-score에 의한 이상치 제거 후 결과와는 서로 차이가 없었지만, 국소회귀분석에 따른 이상치 제거 후 결과와는 통계적으로 유의수준 5% 이하에서 유의한 차이를 보였다.
4. TP는 이상치 제거 전과 3가지 방법에 따른 이상치 제거 후 결과 모두 유의수준 5% 이하에서 통계적으로 유의한 차이가 없었지만, TN은 모두 유의한 차이를 보여, 이상치 선별에 따른 영향을 무시할 수 없음을 확인할 수 있었다.
5. 향후 수질자동측정망을 활용한 수질감시 시, 본 연구에서 제시된 이상치 선별방법을 통해서 이상치를 선별하고, 이상치를 기준으로 수질자동측정망 수질감시기준(주의보, 경보 등)을 설정하는데 활용될 수 있을 것으로 사료된다.

참고문헌

1. Byungjin, L, Eunyoung, H and Insung, Y : Comparative analysis on the outlier data of each parameter in automatic water quality monitoring networks, Journal of *Korean Society on Water Quality*, 26(4):700~706, 2010.
2. 환경부: 물환경측정망 운영계획, 2017
3. Ho, K, Okjin, K and Dohyeon, P : A study on outlier detection method of automatic water quality monitoring Data, *Journal of Korean Society of Urban Environment*, 12(3):197~203, 2012.
4. 박노진: 국소 회귀 모형을 활용한 변화 감지 기법, *Journal of the Korean Data Analysis Societ*, 17(4):1889~1896, 2015.
5. Paul, G and Radia, MJ : Mastering Scientific Computing with R. 성안당, 2016.
6. Joel, S, Francine, L and Antonella, Z : The concentration-response relation between PM2.5 and daily death, *Environ Health Perspect*, 110(10):1025~1029, 2002.
7. Reena, BK, Singh, SH, Neil de W, Rishi, R, Mark, H and Phil, W : The influence of climate variation and change on diarrheal of climate variation and change on diarrheal disease in the pacific islands. Environ Health Perspect, 109(2): 155~159, 2001.
8. Richard, TB, Marc, SD, Dave, S, Mark, E, Raizenne, Jeffrey, RB and Robert, ED : Am J Epidemiol. 153:444~452, 2001.
9. 김영우 : 쉽게 배우는 R 데이터 분석. 이지스퍼블리싱, 2017.
10. 서울특별시상수도사업본부: 수질자동 감시시스템 운영과정, 2010
11. Taemi, Y, Songhee, B and Jaesung, L : Outlier detection methods for monthly rate of housing price, *Journal of the Korea Real Estate Analysts Association*, 19(3):153~164, 2013.
12. Eunyoung, J, Seungmin, P, Inseol, Y, Joengsik, M, Juyoung, P, Jongcheol, K, Yangseob, K and Changyu, P : Study on the removal efficiency of nitrogen and phosphorus in wastewater treatment system using magnetite powder, *Journal of Korean Society for Fluid Machinery*, 18(2):43~47, 2015.