Desalination and Water Treatment
www.deswater.com
doi: 10.5004/dwt.2019.23899

149 (2019) 157–163
May

Introducing new outlier detection method using robust statistical distance in water quality data

Sukmin Yoon[a], Seong-Su Kim[b], Seon-Ha Chae[b], No-Suk Park[a,*]

[a]*Department of Civil Engineering and Engineering Research Institute, Gyeongsang National University, 501, Jinju-Daero, Jinju 52828, Korea, Tel. +82 55 772 1798; Fax: +82 772 1799; emails: nspark@gnu.ac.kr (N.-S. Park), gnuysm@gamil.com (S. Yoon)*
[b]*Korea Water Resources Corporation, K-water Institute, Daejeon 33045, Korea, emails: kssman@kwater.or.kr (S.-S. Kim), shchae@kwater.or.kr (S.-H. Chae)*

Received 24 August 2018; Accepted 1 February 2019

ABSTRACT

Various water qualities are currently being measured in real time in order to monitor source water as well as drinking and waste water processed by treatment plants. However, there are likely to be various potential outliers in the water quality dataset due to replacement of consumables and equipment calibration; and missing data from mechanical malfunctions, etc. Outlier detection method based on multivariate analysis, which has been generally used, is an approach to detecting outliers using chi-squared distribution and Mahalanobis distance derived from multivariate Gaussian distribution. However, Mahalanobis distance is sensitive to the effects of potential outliers and extreme values distributed outside the cluster mean. Accordingly, we adopted robust distance based on minimum covariance determinant estimators to minimize the effects of potential outliers and extreme values. In addition, the modified cutoff point of chi-squared distribution and the cutoff point calculation methodology were applied to reduce the effects of data size in detecting outliers using robust distance and chi-squared distribution.

Keywords: Water quality; Outliers; Multivariate analysis; Mahalanobis distance; Chi-squared distribution; Robust distance

1. Introduction

Currently, various water quality (WQ) parameters (e.g., pH, BOD, TP, SS, NH_3–N) are being measured in real time in order to monitor source water and drinking and waste water processed by treatment plants; the appropriateness of such plants for water and waste water is constantly evaluated. There are close relationships among the WQ data measured in water sources and treatment plants, that is, the WQ data are not mutually independent.

Over the past 10 years, the U.S. Environmental Protection Agency (EPA) has made numerous efforts to develop a contamination warning system (CWS) for water supply system security [1,2]. A CWS can be defined as a system that collects and analyzes data from multiple components or sources of information to detect an intentional or unintentional contamination event early enough to reduce or minimize potentially devastating consequences. In developing an online WQ monitoring system, which is a kind of brain comprising CWS components, the key task is to develop software that can analyze the data produced by the monitoring equipment to see if there are abnormal WQ conditions present that may indicate contamination. The software should include a predictive model to analyze the large volume of data from online monitors, to differentiate normal WQ patterns from anomalous conditions, and to alert the operator to these situations [3]. However, there are likely to be various potential outliers in the WQ dataset due to replacement of consumables and equipment calibration; and missing data from communication

* Corresponding author.

problems, mechanical malfunctions, blackouts, etc. These potential outliers distributed in the WQ dataset may have a major impact on the performance of the prediction model for WQ data and may lower the accuracy and meaning of the dataset [4].

In statistics, outliers are objects that are located far away from the main data cluster (or other data clusters). Hawkins also defines an outlier as an observation that deviates so much from the other observations as to arouse suspicion that it was generated by a different mechanism [5]. In this context, outliers are measurements that have high variability and could be caused by experimental error. Outliers are usually removed from the dataset because they cause a negative effect on data analysis and can seriously bias or influence estimates [6]. Therefore, outlier detection methods have been applied in various fields and have been categorized into four classes depending on the method of detection: the deviation-based method, the distance-based method, the density-based method, and the statistical-based method [7,8].

The deviation-based outlier detection method uses a smoothing factor that indicates how much a dissimilarity can be reduced by removing a subset of data from the original dataset [9]. However, it is not very effective and fails to identify many exceptions when applied to real, complicated data. The distance-based outlier detection method determines outliers based on the distances between the objects in the dataset. The easiest approach to calculating the distances between the objects is to use the *k*-th nearest neighbor (*k*-NN) algorithm. The density-based outlier detection method classifies objects that represent a low density as outliers in the dataset. It is closely related to the distance-based outlier detection method as the density of the objects is calculated on the basis of the distance between them. Accordingly, the density-based outlier detection method also uses the *k*-NN algorithm. However, the *k*-NN algorithm requires excessive computation time and the *k* value is a cutoff point that is the same for all datasets. Therefore, it is hard to identify local outliers with outlier detection methods based on the *k*-NN algorithm. In recent years, outlier detection approaches using the local outlier factor, fuzzy clustering and fuzzy *k*-NN algorithms have been proposed to overcome these disadvantages [10–13].

Statistics-based outlier detection methods either assume a known underlying distribution of the observation or, at least, are based on statistical estimates of unknown distribution parameters [5]. They can be classified into univariate methods, traditionally applied to identify outliers of water quality data, and multivariate methods, which form most of the current body of research. Most of the univariate methods for outlier detection are some type of discordancy test, such as the extreme value test, the discordance test, Rosner's test or Walsh's test [14,15]. The choice of discordancy test depends on whether the data are normally distributed, the sample size, and the existence of multiple outliers. All of the discordancy tests except Walsh's test require the data to be normally distributed. In a univariate method, the variables are considered to be independent of each other and not correlated, but the reality is that the data are often correlated, for example, raw wastewater ammonia has higher values when the carbonaceous biochemical oxygen demand is higher, or stream pH is higher when the alkalinity is higher. A multivariate approach that accounts for the correlation between variables is more statistically correct. As will be shown later, a univariate approach is likely to miss some outliers when the data are correlated [4].

The main purpose of this study is to find an effective way of detecting potential outliers present in the WQ dataset measured from source water and treatment plants, and to introduce a new methodology for removing the outliers. As mentioned above, it has been confirmed that the statistics-based outlier detection methods using the univariate method have fundamental problems: the assumption of regularity, the limited number of data, and ignoring interrelationships in the WQ dataset. In this study, we introduced multivariate analysis for effectively detecting outliers of WQ dataset. Outlier detection method based on multivariate analysis, which has been used generally, is an approach in detecting outliers using the χ^2-distribution (chi-squared distribution) and the Mahalanobis distance (MD) derived from the multivariate Gaussian distribution and the sample covariance of the dataset [16]. However, MD is sensitive to the effects of potential outliers and extreme data distributed away from the main cluster of dataset. It is a disadvantage that outlier detection may not be efficient in cases where MD is not robust due to potential outliers and extreme data. In addition, the χ^2-distribution is known to be dependent on data set size, and as such there is a lack of objectivity in determining whether an extreme value is normal or whether it is an outlier for the data over the cutoff point. Accordingly, the authors introduced robust distance (RD) based on minimum covariance determinant (MCD) estimators to minimize the effects of the potential outliers and extreme data. In addition, the adjusted cutoff point of the χ^2-distribution and the calculation methodology, suggested by Filzmoser et al. [17], were applied to reduce the effects of dataset size in detecting outliers using RD.

2. Theoretical background

The outlier detection approach, which is based on the multivariate Gaussian distribution, searches for the outlier using MD derived from the estimated multivariate location and the estimated covariance matrix. The probability density function of the multivariate Gaussian distribution for a vector $X = (x_1, \ldots, x_n)$ of p dimensional multivariate dataset is represented as follows:

$$f(X) = \frac{1}{\left(\sqrt{2\pi}\right)^p |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(X-\mu)\Sigma^{-1}(X-\mu)\right) \quad (1)$$

where μ is the mean vector for X; and Σ is the covariance matrix for X.

The MD for X can be defined as Eq. (2), and the squared MD (MD^2) is known to follow the χ^2-distribution with p degree of freedom (χ^2_p) [18,19]:

$$MD_i = \left((x_i-\mu)' \Sigma^{-1}(x_i-\mu)\right)^{1/2} \quad \text{for} \quad i = 1,\ldots,n \quad (2)$$

Usually, μ is the multivariate arithmetic mean, the centroid, and Σ is the sample covariance matrix.

The outliers have relatively higher MD^2 than the surrounding neighboring dataset, which can be detected using a cutoff point defined as Eq. (3).

$$c = \chi^2_{P;1-\alpha} \tag{3}$$

where c is cutoff point for outlier detection, $\chi^2_{p;1-\alpha}$ is the quantile χ^2-distribution with p degree of freedom and α confidence level.

However, this approach is not efficient if the multivariate arithmetic mean and the sample covariance matrix are not robust, because of candidate outliers and the extreme data. In order to overcome the above shortcomings, Rousseeuw and Van Zomeren suggested a modified outlier detection methodology using RD as follows [20]:

$$\mathrm{RD}_i = \left(\left(x_i - \mu_{\mathrm{MCD}} \right)' \Sigma^{-1}_{\mathrm{MCD}} \left(x_i - \mu_{\mathrm{MCD}} \right) \right)^{1/2} \quad \text{for} \quad i = 1,\ldots,\mathrm{n} \tag{4}$$

where μ_{MCD} and $\Sigma^{-1}_{\mathrm{MCD}}$ are the robust location and the robust covariance matrix, respectively, derived from the MCD estimator.

The squared RD (RD^2) is also known to follow the χ^2-distribution with p degree of freedom ($\chi^2_{p;1-\alpha}$). Therefore the cutoff point of RD^2 for outlier detection is the same as the MD^2 shown in Eq. (3). The MCD estimators are derived from the subset of X of size h, which minimizes the determinant of the sample covariance matrix [19]. The robust location estimator is the average of the subsets with size h. The selection of h determines the robustness of the estimator. Filzmoser et al. [17] employ a value of $h = 0.75n$ (n is size of X) to consider efficiency and robustness. The same value of subset size h is adopted in this study.

Although RD derived from MCD estimators has minimized the effect of candidate outliers and extreme data, there are still significant problems. In other words, it does not consider multivariate dataset size, and could not distinguish between outliers and extreme data. For this reason, Filzmoser et al. [17] proposed a methodology to minimize the influence of the multivariate dataset and objectively separate the outliers and the extreme values as follows.

Let $G_n(u)$ denote the empirical distribution function of RD^2 and $G(u)$ is the theoretical distribution function of χ^2_p. If the multivariate dataset is normally distributed, $G_n(u)$ converges to $G(u)$. Therefore, the tails of $G_n(u)$ and $G(u)$ are compared with identify outliers and the tail is defined by $\delta = \chi^2_{p;1-\alpha}$ for α confidence level. The distance between $G_n(u)$ and $G(u)$ is measured only in the tails and $p_n(\delta)$ is calculated as in Eq. (5) in order to objectively distinguish outliers and extreme data:

$$p_n(\delta) = \sup_{u \geq \delta} \left(G(u) - G_n(u) \right)^+ \tag{5}$$

where the superscript, "+" means positive difference.

The $p_n(\delta)$ can be considered as a measure of outlier detection, but will not be directly adopted for outlier detection. The cutoff point should be infinity in the case of a multivariate normally distributed datasets. Therefore Filzmoser et al. [17] introduced the new critical values $p_{\mathrm{crit}}(\delta,n,p)$ for separating the outliers from the extreme data (Eqs. (6) and (7)). They can be derived by simulation (in the simulation $\delta = \chi^2_{p;0.98}$ is applied) [17].

$$p_{\mathrm{crit}}(\delta,n,p) = \frac{0.24 - 0.003p}{\sqrt{n}} \quad \text{for} \quad \delta = \chi^2_{p;0.98} \quad \text{and} \quad p \leq 10 \tag{6}$$

$$p_{\mathrm{crit}}(\delta,n,p) = \frac{0.252 - 0.0018p}{\sqrt{n}} \quad \text{for} \quad \delta = \chi^2_{p;0.98} \quad \text{and} \quad p > 10 \tag{7}$$

If $p_n(\delta)$ exceeds $p_{\mathrm{crit}}(\delta,n,p)$, the extreme data are declared as outliers and the cutoff point should be modified using Eqs. (8) and (9).

$$\alpha_n(\delta) = \begin{cases} 0 & \text{if} \quad p_n(\delta) \leq p_{\mathrm{crit}}(\delta,n,p) \\ p_n(\delta) & \text{if} \quad p_n(\delta) > p_{\mathrm{crit}}(\delta,n,p) \end{cases} \tag{8}$$

$$c_n(\delta) = G_n^{-1}\left(1 - \alpha_n(\delta)\right) \tag{9}$$

where $c_n(\delta)$ is the new cutoff point for outlier detection and it is called the 'adjusted quantile'.

3. Methods for multivariate outlier detection

This study was conducted to find a way of effectively detecting potential outliers distributed in WQ dataset and to introduce a new methodology for removing the outliers. To these ends, the C_water treatment plant (WTP) located in South Korea was selected and investigated. The C_WTP supplies 263,800 $m^3\ d^{-1}$ of drinking water. The target WQ of C_WTP is free residual chlorine (F–Cl) 0.1–4.0 mg L^{-1}, pH 5.8–8.5 and turbidity less than 0.5 NTU. These WQ datasets are monitored online continuously every minute. In this study, the measured F–Cl and pH dataset from 1st to 30th November 2014 were used (n = 43,183).

Fig. 1 summarizes the procedure of this study. The first step was to search for existing potential outlier candidates using a univariate outlier detection method (i.e., z-score). The second step was to calculate MD and RD for multivariate outlier detection. The RD was estimated using MCD estimators based on a subset size of $h = 0.75n$ and the estimated RD was compared with MD to examine outlier candidates in more detail. In the third step, the empirical distribution function of RD^2 and the theoretical distribution function of χ^2_p was derived. Next, the $p_n(\delta)$ was calculated only in the tail and was compared to $p_{\mathrm{crit}}(\delta,n,p)$ in order to distinguish outliers and extreme data. The dataset in the case of $p_n(\delta)$ being higher than $p_{\mathrm{crit}}(\delta,n,p)$ were determined to be outliers. Conversely, data in the opposite case were determined to be extreme data. In the fourth and final step, initial cutoff point for outlier detection based on RD was adjusted to detect the ultimate outliers.

4. Results and discussion

The potential outlier candidates of the collected F–Cl and pH data were preferentially searched by using univariate outlier detection methods (i.e., z-score). The cutoff point for

outlier detection was *z*-score >3 or < –3, which corresponded to a *p*-value <0.014 [21].

Fig. 2(a) is a scatter plot of F–Cl and pH dataset which are clustered over the range of target water quality. Fig. 2(b) is a scatter plot of *z*-scores of F–Cl and pH datasets. As shown in Fig. 2(b), the *z*-scores of pH are uniformly distributed over the total observation time. There are no upper potential outlier candidates whose *z*-score is higher than 3 but there are some sparsely distributed lower potential outlier candidates whose *z*-score is lower than –3. However, *z*-scores of F–Cl are more irregularly distributed relative to the pH data. Both upper and lower potential outlier candidates are partially found over the total observation time.

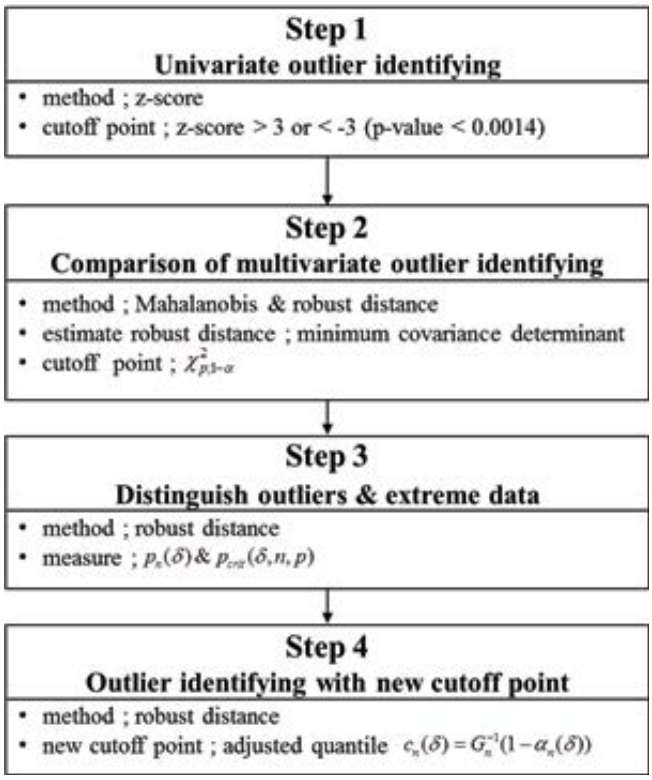

Fig. 1. Procedure for multivariate outlier detection in the current study.

The mean and standard deviation of MD were calculated as 0.846 and 1.156, respectively. As a result of analysis with semi-log scale graph, MD^2 derived from multivariate dataset follows the χ^2-distribution, and the multivariate outlier detection using MD^2 is known to be more effective than the univariate detection method (refer to Fig. 3(a)) [18,19]. Accordingly, the outlier candidates of F–Cl and pH dataset were searched for using the derived MD^2 with cutoff point $\chi^2_{2;0.975} = 7.378$.

Figs. 3(b) and (c) show multivariate outlier detection results using MD^2 for F–Cl and pH data. For detecting outlier candidates, the cutoff point was set as $\chi^2_{2;0.975} = 7.378$; the confidence level α and the degree of freedom p were 0.025 and 2, respectively. The results of detecting outlier candidates were as follows: From MD^2 with $\chi^2_{2;0.975} = 7.378$ the ratio of outlier was 1.63%, which corresponds to 703 outliers (Table 1).

Even though the test using MD^2 looks more efficient than the *z*-score test for detecting outlier candidates, there is still a serious problem: MD is too sensitive to the potential outliers distributed in the raw dataset. Further, extreme data departing from the main data cluster could obviously affect the MD. In order to overcome these weaknesses of MD, it is necessary that the multivariate location μ and covariance matrix Σ should be robust. Accordingly, the MCD estimator was derived to enhance robustness and efficiency using computationally fast algorithms suggested by Rousseeuw

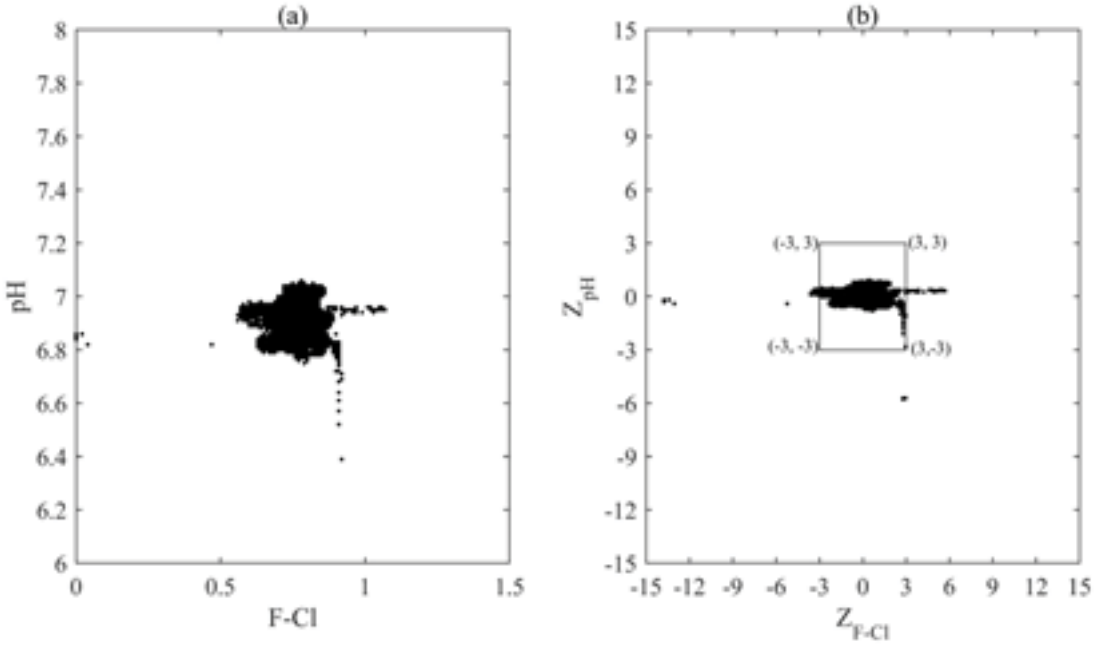

Fig. 2. Scatter plot of F–Cl vs. pH dataset ((a) scatter plot of raw dataset and (b) scatter plot of *z*-score).

and Van Zomeren [20]. In this study, a subset size of $h = 0.75n$ (n = 43,183) was applied to derive the MCD estimator.

Table 2 summarizes the results of MCD estimators. The difference in the means between the raw WQ dataset and the subset data is small, while the sample covariance of F–Cl and pH dataset was reduced by 37.5%.

Fig. 4 illustrates the relationship between MD and RD in analyzing the effect of enhancing robustness by means of the MCD estimator. As shown in Fig. 4, the difference of MD and RD was insignificant when the range of MD was less than 5. However, in a range of over 13, the difference increased significantly, up to twice the distance. The rapid

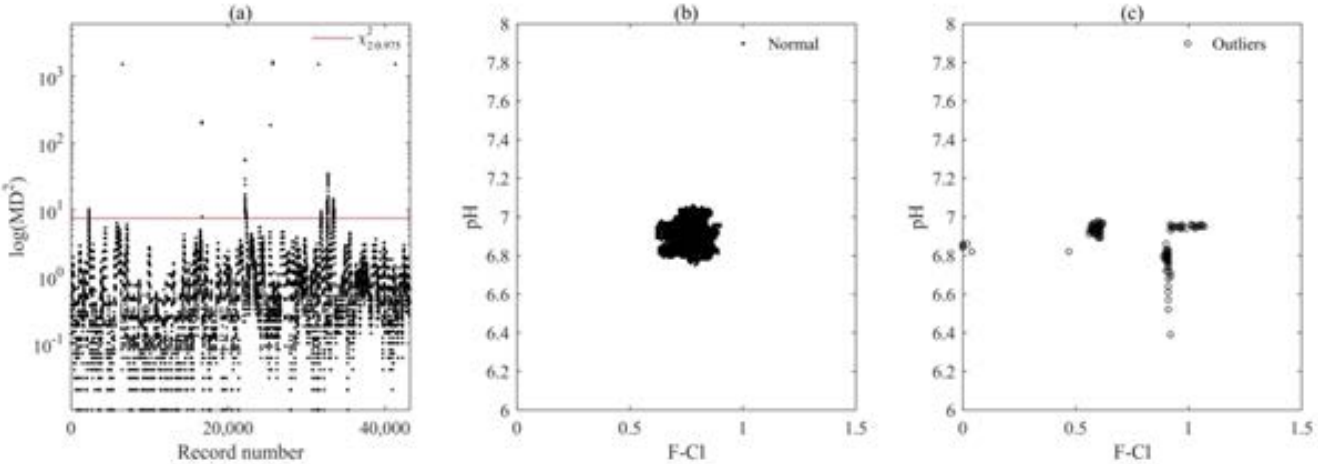

Fig. 3. Plot of derived MD² and detected outlier candidates for F–Cl and pH dataset ((a) derived MD² (point) and cutoff point (solid line), (b) classified normal dataset, and (c) detected outlier candidates).

Table 1
Results of the outlier candidates detection using z-score and MD

Method	Outlier candidates results for water qualities		
	Test	F–Cl	pH
z-score	>3	29	203
	< –3	None	27
MD	$MD^2 \geq \chi^2_{2;0.975} = 7.378$	703	

Table 2
Mean vector and covariance matrix of the water quality dataset

WQ	Mean vector		Covariance matrix			
	Dataset	Subset	Dataset		Subset	
			F–Cl	pH	F–Cl	pH
F–Cl	6.89	6.90	0.0015	–0.0008	0.0020	–0.0005
pH	0.75	0.76	–0.0008	0.0051	–0.0005	0.0040

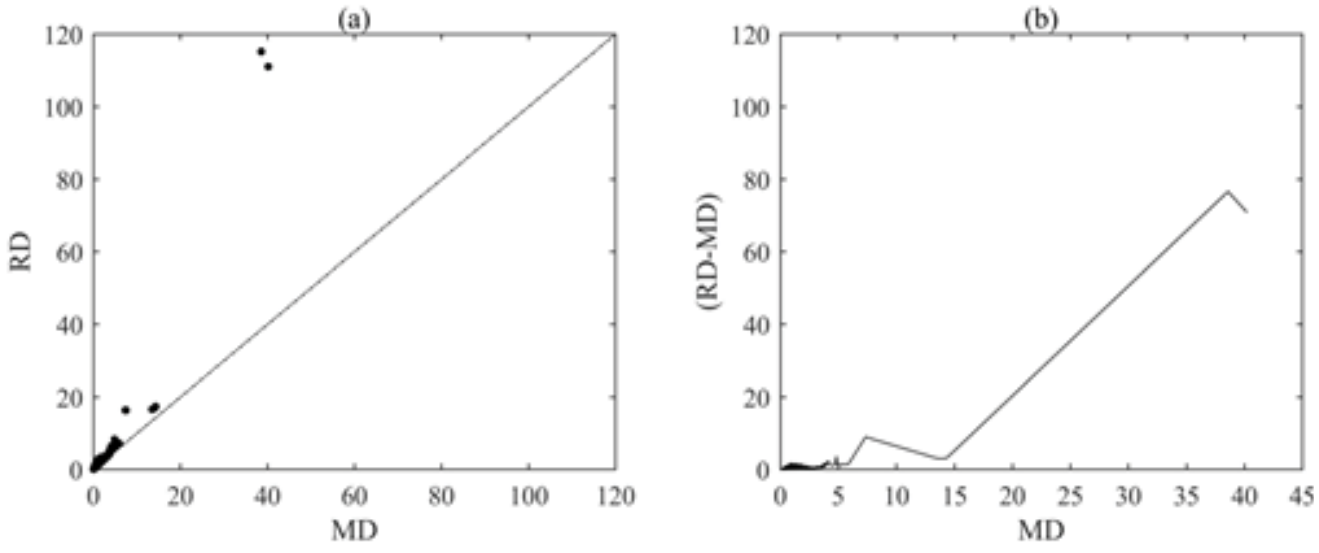

Fig. 4. Comparison of derived MD and RD ((a) scatter plot of MD vs. RD and (b) difference between MD and RD).

increase in difference in the region of relatively higher MD was primarily due to the reduced sample covariance derived from the MCD estimators minimizing the impact of the potential outliers.

The derived RD^2 from multivariate dataset followed the χ^2-distribution. The cutoff point of RD^2 for detecting outliers was identical with that of MD^2. Therefore, the cutoff point was set as $\chi^2_{2;0.975} = 7.378$ to detect outlier candidates using RD^2.

The mean and standard deviation of RD were calculated as 1.349 and 2.904, respectively. Fig. 5(a) shows the distribution of the derived RD^2 using semi-log scale graph. Figs. 5(b) and (c) show multivariate outlier detection results using RD^2 for F–Cl and pH data. As a result, a total of 1,926 (4.46%) outlier candidates were found using RD^2 which was approximately 174% higher than the outlier detection using MD^2. However, there are serious problems related to outlier detection. In other words, the distinction between the outlier and the extreme data is a subjective problem. Therefore, the value $p_n(\delta)$ and the critical value $p_{crit}(\delta,n,p)$ were calculated to objectively determine whether the dataset over the cutoff point was part of the extreme data or were true outliers.

According to Eq. (5), the task is to find the supremum of the difference between $G_n(u)$ and $G(u)$ in the tail. With $\delta = \chi^2_{2;0.975} = 7.378$ a supremum of $p_n(\delta) = 0.0208$ was calculated. Eq. (7) gives a critical value $p_{crit}(\delta,n,p) = 0.0011$, which was clearly lower than the $p_n(\delta)$. For this reason, it can be assumed that the dataset with a large RD^2 comes from at least one different distribution, in which case it can be classified as outliers. The new cutoff point $c_n(\delta)$ was calculated using Eqs. (8) and (9) to declare the outliers for the F–Cl and pH datasets and $c_n(\delta)$ is called the 'adjusted quantile'.

Fig. 6(a) shows derived RD^2 (point in Fig. 6(a)) with $c_n(\delta)$ (solid line in Fig. 6(a)) and $c_n(\delta) = 7.746$ was a little bit higher than the initial cutoff point $\chi^2_{2;0.975} = 7.378$. Figs. 6(b) and (c) shows multivariate outlier detection results using RD^2 with the adjusted quantile $c_n(\delta)$. From RD^2 with $c_n(\delta)$, the ratio of outlier was 3.79%, corresponding to 1,637 outliers (Table 3). These results imply that multivariate outlier detection based

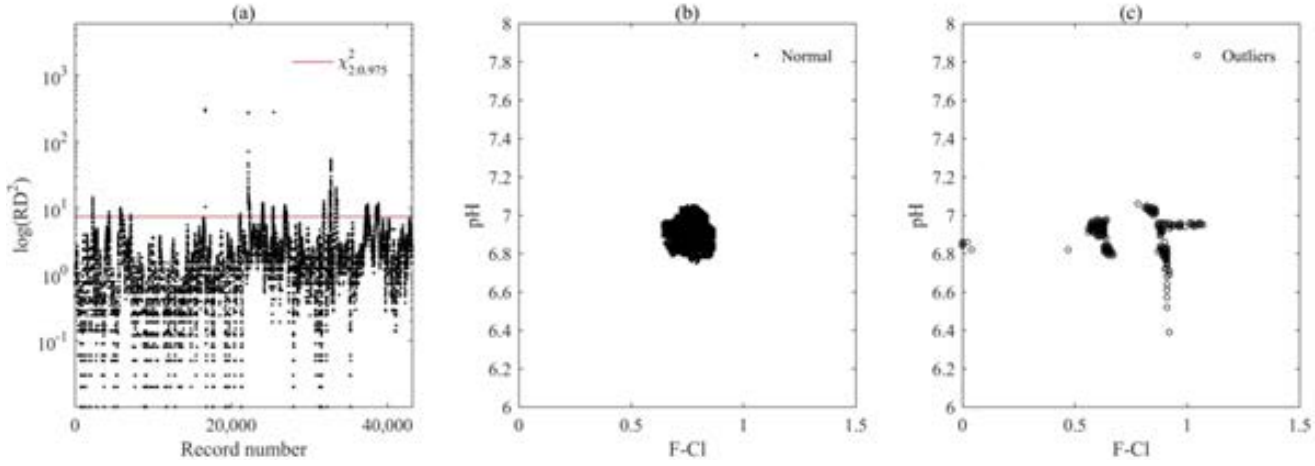

Fig. 5. Plot of derived RD^2 and detected outlier candidates for F–Cl and pH dataset ((a) derived RD^2 (point) and cutoff point (solid line), (b) classified normal dataset, and (c) detected outlier candidates).

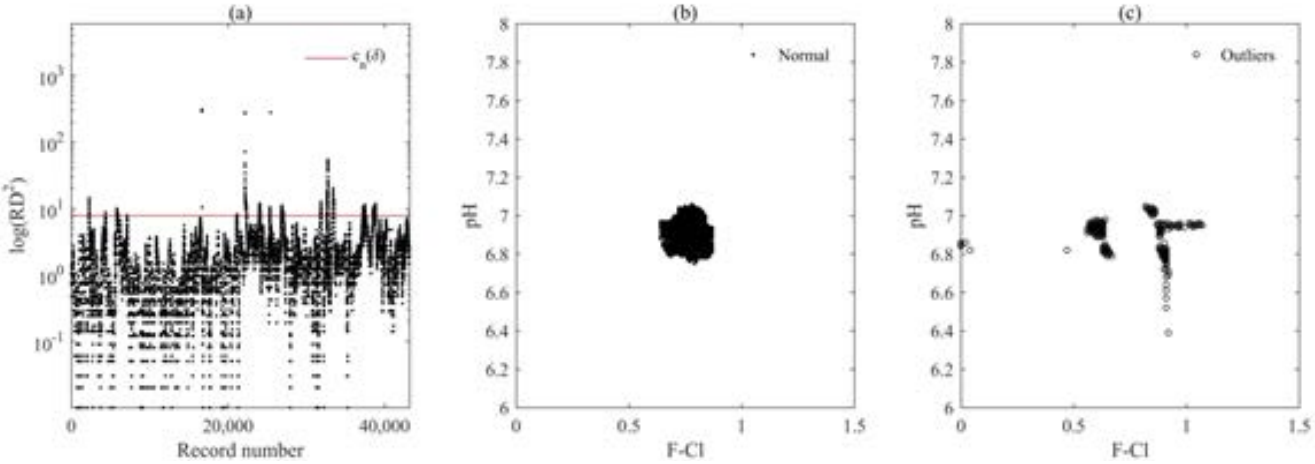

Fig. 6. Plot of derived RD^2 and detected outlier for F–Cl and pH data ((a) derived RD^2 (point) and new cutoff point (solid line), (b) classified normal water qualities and (c) detected outliers).

Table 3
Summary of the multivariate outlier detection results

Method	Outlier results for water qualities dataset	
	Test	Outliers
MD	$MD^2 \geq \chi^2_{2;0.975} = 7.378$	1.63% (703)
RD	$RD^2 \geq \chi^2_{2;0.975} = 7.378$	4.46% (1,926)
	$RD^2 \geq c_n(\delta) = 7.746$	3.79% (1,673)

on RD^2 with the new cutoff point was stricter in approach than using the MD^2 for identifying outliers distributed in WQs datasets.

5. Conclusions

This study was conducted to find a way of effectively detecting multivariate outliers present in WQs dataset measured from source water and treatment plants, and to introduce a new methodology for removing such multivariate outliers. The concept of RD derived MCD estimators was introduced to minimize the effect of the potential outliers and extreme data. Moreover, the modified cutoff point of χ^2-distribution, called the 'adjusted quantile', was applied to reduce the effects of sample size and to distinguish between outliers and extreme data. The findings of the study can be summarized as follows:

- 259 potential outlier candidates (both lower and upper potential outliers) were detected using the univariate method for outlier detection (i.e., *z*-score). Over three times as many outlier candidates were detected using the multivariate method of outlier detection based on MD^2 and RD^2. This implies that the multivariate method for outlier detection is more effective for identifying the true outliers distributed in the WQs dataset.
- Comparing RD and MD, the difference between MD and RD is small when the range of MD is less than 5. However, for in higher ranges, the maximum difference increases by more than twice the distance. These phenomena occurred because of the reduced sample covariance derived from the MCD estimators.
- From the results of outlier detection using RD^2 with the adjusted quantile $c_n(\delta)$ which was suggested by Filzmoser et al. [17] twice as many outliers were detected than for outlier detection using the MD^2 test. It can be concluded that the methodology of outlier detection using RD^2 derived from MCD estimators is stricter than the MD^2 test for detecting the multivariate outliers distributed in the WQs dataset.

Acknowledgment

This research was supported by a grant (16AWMP-B113766-01) from the Research & Development (R&D) Program on Water Management funded by the Ministry of Environment of the Korean Government.

References

[1] U.S. EPA, Water Security Initiative: System Evaluation of the Cincinnati Contamination Warning system Pilot, U.S. EPA Water Security Division, Washington, D.C., 2014.
[2] U.S. EPA, Water Security Initiative: Evaluation of the Water Quality Monitoring Component of the Cincinnati Contamination Warning System Pilot, U.S. EPA Water Security Division, Washington, D.C., 2014.
[3] U.S. EPA, Water Quality Event Detection Systems for Drinking Water Contamination Warning System (EPA/600/R-010/036), Office of Research and Development, National Homeland Security Research Center, Washington, D.C., 2010.
[4] R.B. Robinson, C.D. Cox, K. Odom, Identifying outliers in correlated water quality data, J. Environ. Eng., 131 (2005) 651–657.
[5] D. Hawkins, Identification of Outlier, Chapman and Hall, London, 1980.
[6] J.W. Osborne, A. Overbay, The power of outliers (and why researchers should always check for them), Pract. Assess. Res. Eval., 9 (2004) 1–12.
[7] H.P. Kriegel, P. Kroger, A. Zimk, Outlier Detection Techniques, Tutorial at the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 10, 2009.
[8] O. Malmon, L. Rockach, Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Kluwer Academic Publishers, Boston, MA, 2005.
[9] A. Arning, R. Agrawal, P. Raghavan, A Linear Method for Deviation Detection in Large Databases, In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, AAAI Press, Portland, Oregon, 1996, pp. 164–169.
[10] M.M. Breunig, H.P. Kriegel, R.T. Ng, J. Sander, LOF: Identifying Density-Based Local Outliers, In ACM SIGMOD Record, 29 (2000) 93–104.
[11] N.A. Yousri, M.A. Lsmail, M.S. Kamel, Fuzzy Outlier Analysis: A Combined Clustering – Outlier Detection Approach, Proceedings of IEEE International Conference on Systems, Man and Cybernetics, 2007, pp. 412–418.
[12] M.B. Al-Zoubi, A.D. Ali, A.A. Yahya, Fuzzy Clustering-Based Approach for Outlier Detection, Proceedings of the 9th WSEAS International Conference on Applications of Computer Engineering, 2008, pp. 192–197.
[13] R. Ostermark, A fuzzy vector valued kNN-algorithm for automatic outlier detection, Appl. Soft Comput., 9 (2009) 1263–1272.
[14] V. Barnett, T. Lewis, Outliers in Statistical Data, 3rd Ed., John Wiley & Sons, New Jersey, 1994.
[15] U.S. EPA, Data Quality Assessment: Statistical Methods for Practitioners, EPA QA/G-9S, 2006.
[16] K.P. Murphy, A Probabilistic Perspective, The MIT Press, London, 2012.
[17] P. Filzmoser, R.G. Garrett, C. Reimann, Multivariate outlier detection in exploration geochemistry, Comput. Geosci., 31 (2005) 579–587.
[18] J.D. Jobson, Applied Multivariate Data Analysis, Springer-Verlag, New York, 1992.
[19] A.C. Rencher, Multivariate Statistical Inference, Wiley, New York, 1998.
[20] P.J. Rousseeuw, B.C. van Zomeren, Unmasking multivariate outliers and leverage points, J. Am. Stat. Assoc., 85 (1990) 633–639.
[21] D.R. Anderson, D.J. Sweeney, T.W. Williams, Statistics for Business and Economics, West, Minneapolis/St. Paul, 1993.