PATTERN CLASSIFICATION ANALYSIS OF NON-POINT SOURCE POLLUTION USING NEASURED RUNOFF AND WATER QUALITY DATA

Young-Hoon Jin[1], Akira Kawamura[2], Jonas Olsson[3], and Sung-Chun Park[4]

[1] Research Professor, Institute of Industrial Research and Technology, Dongshin University, Korea
[2] Professor, Department of Civil and Environmental Engineering, Tokyo Metropolitan University, Japan
[3] Senior Researcher, Swedish Meteorological and Hydrological Institute, Sweden
[4] Professor, Department of Civil Engineering, Dongshin University, Korea

ABSTRACT

In the present study, Self-Organizing Map (SOM) method which is one of the pattern classification methods was applied for the analysis of water quality concentration and stormwater runoff data measured from a commercial district in Gwangju Metropolitan City, Korea. In particular, the SOM in the study was combined with a hierarchical cluster analysis using Ward's linkage method and Euclidean distances to obtain more accurate classification result. Biochemical Oxygen Demand (BOD), Total Organic Carbon (TOC), Suspended Solids (SS), Total Nitrogen (TN), Total Phosphorus (TP) concentration and stormwater runoff data were irregularly measured from the study area between May and September in 2009, according to five rainfall events. The results of SOM application showed that the six variables were classified into six groups with respect to the variation of each variable in the study area. Interpretation of the mean standardized values of reference vectors for the respective variables revealed that five clusters except for cluster-4 were associated with low stormwater runoff with various patterns of water quality concentration data, and the cluster-4 included the reference vector corresponding to the highest runoff with the lowest water quality concentration data. Cluster-1 was clearly classified by considerably low concentration values of water quality and, conversely, the extremely high concentration pattern of water quality data was presented in cluster-6. Low water quality concentration pattern with slightly higher TOC value than those of other items was shown in cluster-2. Relatively low SS, TN and TP concentration patterns were commonly found in cluster-3 and cluster-5, however, the cluster-3 included the mean standardized values of reference vectors corresponding to much higher BOD and TOC concentration than those of the cluster-5. Consequently, the applied method of SOM combined with a hierarchical cluster analysis revealed it could be used to analyze the data by showing the exclusively different patterns and give better knowledge of the non-point source pollution based on the variation of water quality patterns classified in the present study.

1. INTRODUCTION

During the last several decades, cities in Korea have been highly industrialized, overpopulated and urbanized, and accordingly most policies for water quality improvement have been focused on how to control and reduce point source pollution. However, non-point source pollution has significantly increased due to the intensive land use in the cities, so it has been realized that to reach relevant target of water quality in watersheds is impossible by treating only point source pollution.

Non-point source pollution has been recently considered as one of the most important factors to manage and improve streams, rivers and natural environment in watersheds because it has significantly influenced on the water quality in various ways. As a result of such consideration and understanding of non-point source pollution, the regulation of Total Maximum Daily Loads (TMDLs) in watersheds has been executed step by step in Korea focusing on the management and control for non-point source pollution as well as point source pollution.

It is obvious that the better understanding of non-point source pollution can give more appropriate and supportive information for the TMDLs. Understanding the characteristics of non-point source pollution is critical to deal with water quality problems and keep rivers from pollution. Therefore, it is considerably important to understand its characteristics to prevent the rivers from being polluted, and it is also critical to develop and apply appropriate methods for the purpose.

Recently, Self-Organizing Map (SOM) has been used as a powerful tool for pattern classification and the results showed it can classify multi-dimensional data into a number of clusters, delineating two-dimensional visualization. The application of SOM has reported in diverse fields including hydrology (Hsu *et al*., 2002; Jain and Srinivasulu, 2006; Srinivasulu and Jain, 2006), wastewater treatment (Gracia and González, 2004), meteorology (Nishiyama *et al*., 2007), geomorphology (Hentati *et al*., 2010), ecology (Faggiano *et al*., 2010), water quality (Su *et al*., 2010), and others (Jeong *et al*., 2010).

Surveying the details of recent studies on the application of SOM, Hetati *et al*. (2010) developed two geomorphologic indices and applied SOM combined with a hierarchical agglomerative clustering for evaluation of vulnerability to erosion using six variables including the two indices. Four groups according to the different vulnerability to erosion were classified with a global trend based on the geomorphologic characteristics. Faggiano *et al*. (2010) also utilized SOM and a hierarchical cluster analysis for ecological risk assessment, and classified sampling sites into four clusters representing similar toxic assemblages respectively.

Su *et al*. (2010) applied SOM based on the non-hierarchical K-means algorithm to determine spatial patterns and significant variables in water quality. The SOM grouped the data used into three different groups according to similarity of water quality. Joeng *et al*. (2010) surveyed numerous sites to collect data and develop Stream Modification Index (SMI), and constructed SOM model to investigate the relationship between stream modification and socio-geographical data. The high applicability of SOM can be seen in the above studies using various kinds of data.

The primary objective of the study was to understand the characteristics of measured water quality and stormwater runoff data and provide basic and supportive information for TMDLs in Korea. For the purpose, in the present study, SOM combined with a hierarchical cluster analysis using Ward's linkage method and Euclidean distances was applied to the data measured from a commercial district in Korea, and the respective clusters classified by SOM was analyzed in detail using a series of graphical visualization and its interpretation.

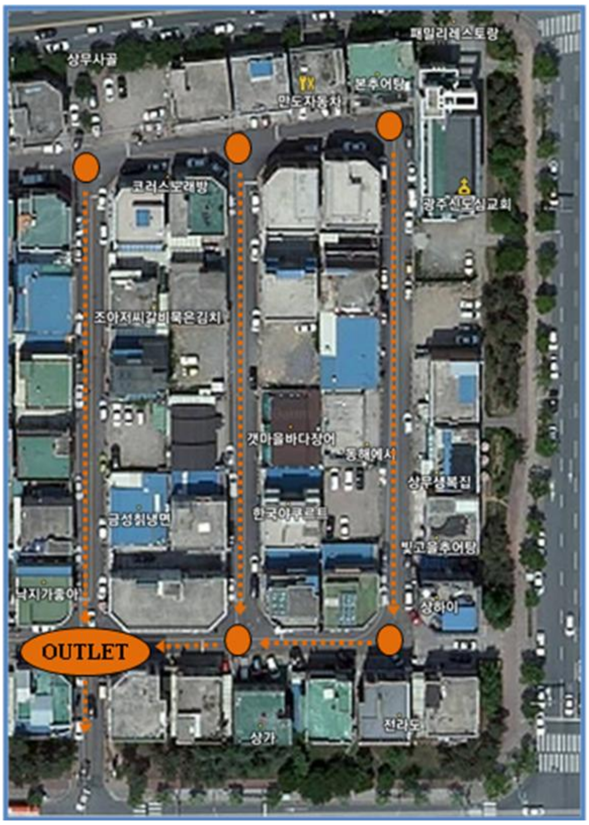


Figure 7. Study area and its stormwater drainage system

2. STUDY AREA AND DATA USED

2.1 Study area

A commercial district mainly composed of restaurants, parking lots, and offices was chosen to measure various water quality items and stormwater runoff data for the present study. The study area is situated in the downtown of Gwangju Metropolitan City which is located in southwestern region of Korea. Figure 1 shows the study area and its stormwater drainage system. The study area of 0.0126 km^2 is covered by impervious layer of 85% and pervious layer of 15%. Its stormwater drainage system is composed of the pipes with 450mm, 550mm and 600mm diameters, and the diameter of outlet pipe is 700mm.

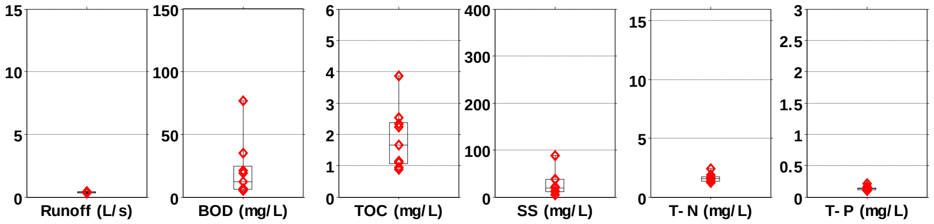
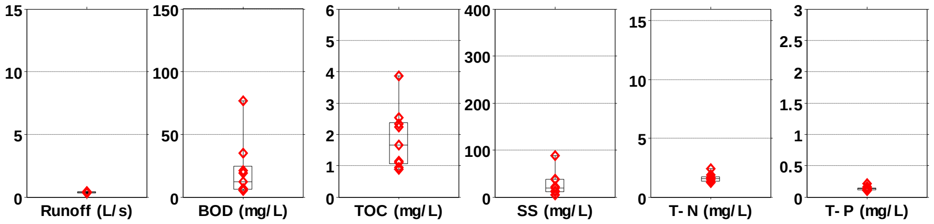
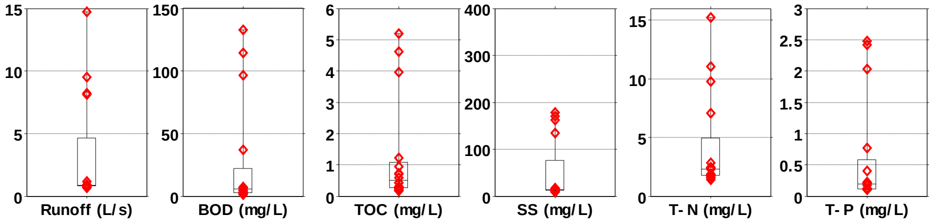
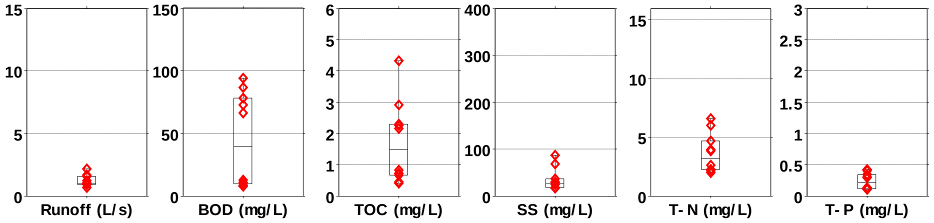
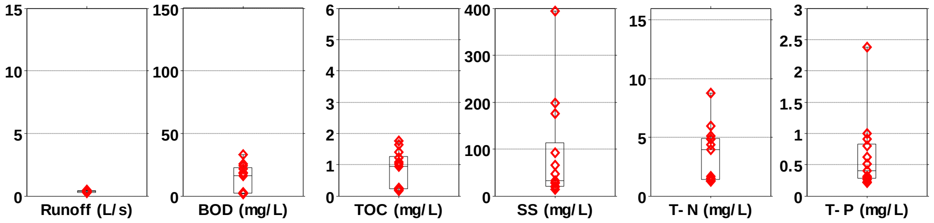
2.2 Data used

Stormwater runoff and Biochemical Oxygen Demand (BOD), Total Organic Carbon (TOC), Suspended Solids (SS), Total Nitrogen (TN), Total Phosphorus (TP) concentration data were irregularly measured from the study area between May and September in 2009, according to rainfall occurrence. The input dataset for SOM in the present study included 63 training data consisting of the above six variables. The five water quality concentration data were obtained by laboratory test using the water sampled from a stormwater pipe in the outlet of the study area and the velocity of flow to calculate stormwater runoff was measured by a flowmeter installed inside the pipe. For the six different data, the measurement was carried out with 15-minute interval for the initial 8 samples and 30-minute interval for the next 5 samples. Since then, one hour interval was applied for the measurement.

Table 1 shows measurement periods for five rainfall events and box-whisker plots representing minimum, the 1st to 3rd quartiles, and maximum values with black lines for the six variables. The respective data were also plotted on the box-whisker plots with the marks of red diamond. The highest stormwater runoff and the worst water quality data except for SS were found in the third event between 28th and 29th of June in 2009, resulting in the widest data ranges for the five variables during the measurement period. Meanwhile, the fifth event included the highest SS concentration value.

Low stormwater runoff data were generally measured except for the high value of runoff data in the third event and hence the distribution of runoff data represented highly positive skewness. On one hand, the BOD and TOC concentration from the 1st to the 4th events showed similar data ranges and the fifth event revealed distinctively narrow data ranges for low BOD and TOC concentration. On the other hand, the data range of the fifth event for SS concentration was wider than the ranges of different events, showing similar ranges for the data. Wider ranges of data than other periods for TN concentration were seen in the third event and the third and the fifth events for TP concentration.

Table 8. Data measurement period and box whisker plots with minimum, the 1st to 3rd quartiles, maximum (black lines), and measured (marks with red diamond) for the respective events

<table>
<tr><th>Period</th><th>Box whisker plots for the measured data of the six variables</th></tr>
<tr><td>1st Measurement
2009-05-16 02:10
~
2009-05-16 10:25
(15 data)</td><td> </td></tr>
<tr><td>2nd Measurement
2009-06-09 16:40
~
2009-06-09 18:55
(9 data)</td><td> </td></tr>
<tr><td>3rd Measurement
2009-06-28 21:10
~
2009-06-29 04:25
(16 data)</td><td> </td></tr>
<tr><td>4th Measurement
2009-07-09 19:00
~
2009-07-09 21:45
(10 data)</td><td> </td></tr>
<tr><td>5th Measurement
2009-09-27 04:50
~
2009-09-27 09:35
(13 data)</td><td> </td></tr>
</table>

3. SELF-ORGANIZING MAP (SOM)

SOM is a kind of the Artificial Neural Networks (ANNs) and is categorized into an unsupervised algorithm. One of its advantages is it can project high-dimensional data onto two-dimensional regularly-arranged units (Nishiyama *et al.*, 2007). Therefore, it can provide a useful tool to understand complex data by visualization and classification. Two-dimensional visualization of trained SOM is carried out by U-matrix method which can identify cluster boundaries among clusters. In the next step, K-means method is utilized for fine-tuning the clusters identified from the U-matrix.

Kohonen network adopts the concept that "Winner takes all" and only its neighbors are allowed to adjust their connection weights. For the process, connection weight vectors of nodes should be properly initialized with random values. In general, linear initialization for weight vectors of SOM is used for classification of patterns with the small number of data and the initialization is known to accelerate training phase (Jeong *et al.*, 2010).

Each node adjusts connection weight while the following three important steps are in progress: competitive process, cooperative process, and adaption process. In the first process, each node is competitive with others in order to award privilege to be trained and the nearest node to input vectors wins. The winner is the only node which can send output signal. The winner and neighbors are allowed for adjustment of weight with presented input vectors in the second process. All of nodes in the network adjust connection weights by training of a repetitive process through the last process.

It is important to determine the number of appropriate output nodes and the ratio between the side lengths of a map to detect the deviation of the data (Hentati *et al.*, 2010). The heuristic rule (m=5√n, m for the map size and n for the number of samples) suggested by Vesanto *et al.* (2000) is widely used to determine the number of output nodes, which can be multiplied by the ratio between the two maximum eigenvalues of training data in order to get the side lengths of a map.

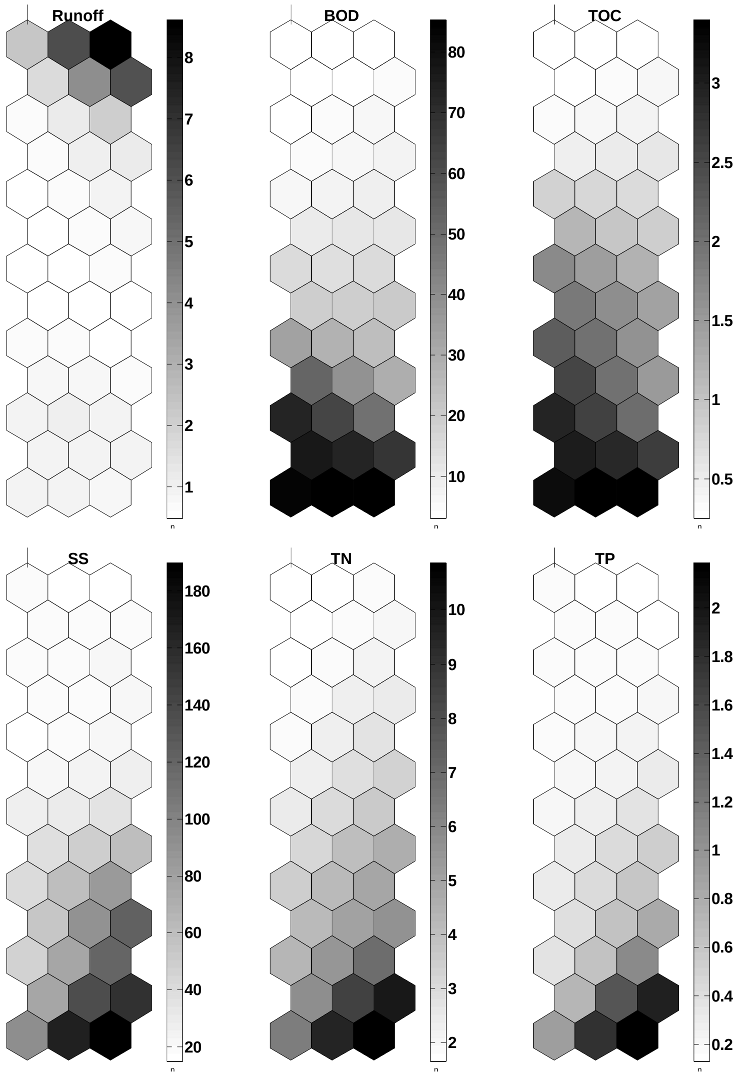


Figure 8. Component planes for the six variables after denormalization

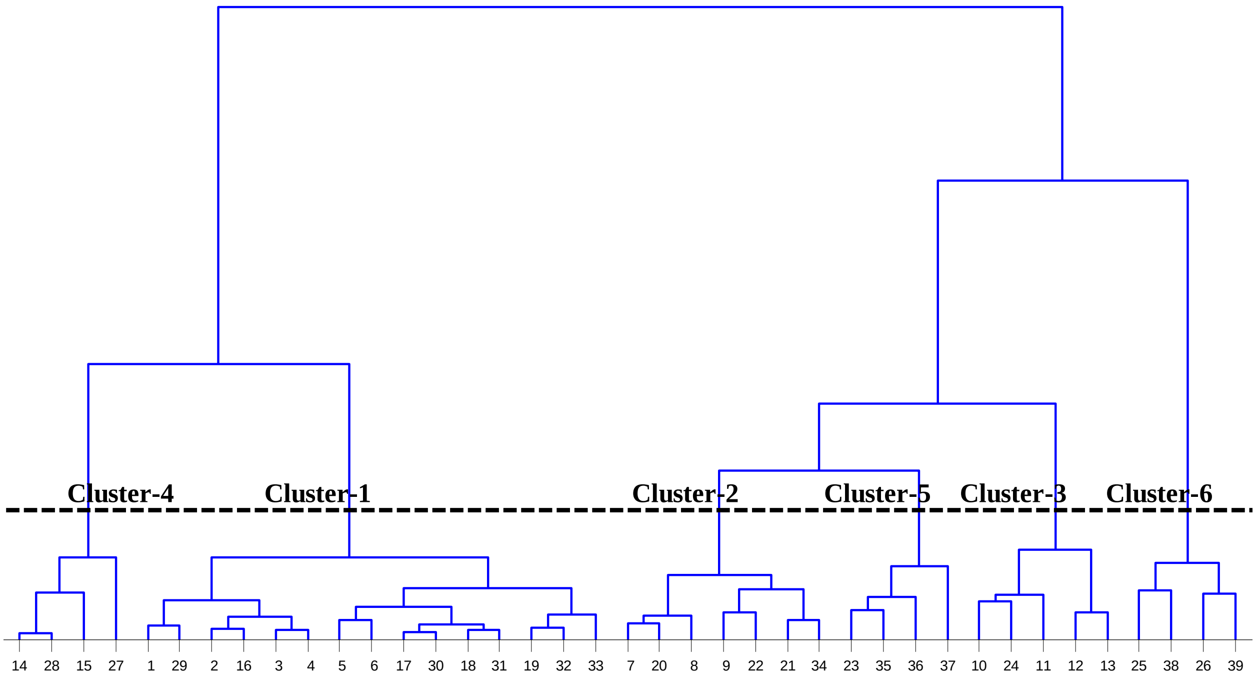


Figure 3. Hierarchical cluster tree for the reference vectors of the SOM

In the present study, SOM was combined with a hierarchical cluster analysis to classify reference vectors more accurately after its training. The hierarchical cluster analysis in the present study used Ward's linkage method and Euclidean distances, as Hentati *et al*. (2010) and Faggiano *et al*. (2010) applied for their works in order to obtain classification in which the reference vectors of a cluster get closer to each other in the cluster and get further from the reference vectors of other clusters. In other words, the two closest clusters were aggregated according to Ward's linkage method and Euclidean distances, which can create a hierarchical cluster tree as a graphical visualization. The hierarchical cluster tree is also known as dendrogram.

4. RESULTS

The SOM in the present study was constructed so as to have the map size of 13 × 3 with hexagonal arrays, and it was trained in a batch mode with the six variables of 63 measurement data. Figure 2 shows the component planes of the six variables in gray scale with the denormalized data after training for SOM. Black, gray and white indicate high, moderate and low values, respectively.

According to the component planes, three groups could be clearly classified. One group included only one variable, stormwater runoff data, showing totally different shading pattern from others. Another group displayed that shading directions of BOD and TOC were considerably similar, and the other group included SS, TN, and TP with very similar shading directions. Correlations of the variables in the same group can be considered to be strong, qualitatively.

SOM training phase produced 39 reference vectors corresponding to the respective nodes. In order to have more accurate classification of the six variables, Ward's linkage method and Euclidean distances were applied to the 39 reference vectors. As shown in Figure 3, Ward's linkage method created a hierarchical cluster tree using Euclidean distances between reference vectors.

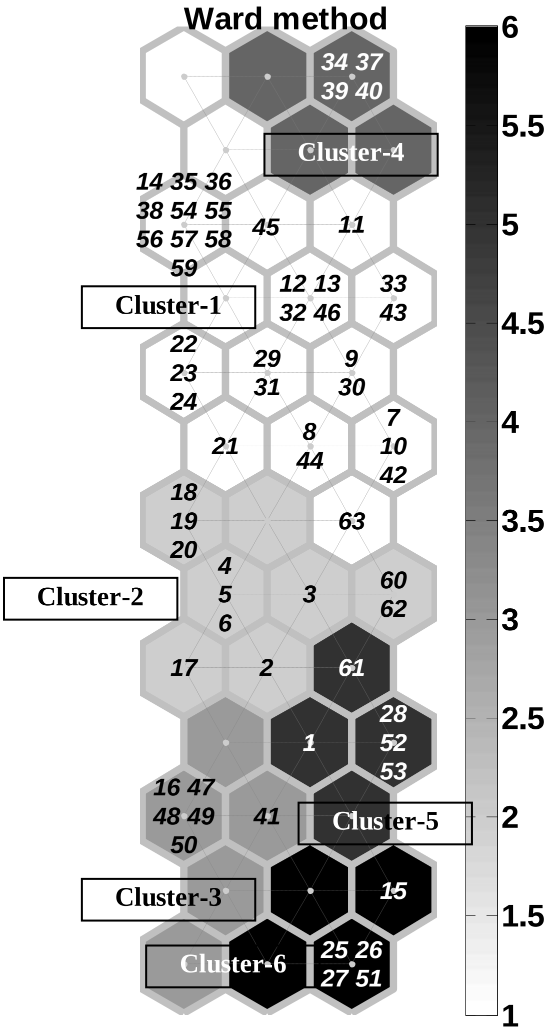


Figure 4. SOM combined with a hierarchical cluster analysis

The classification result from the application of SOM combined with the hierarchical cluster analysis showed that the six variables were classified into six different groups, as shown in Figure 4, with respect to the variation of each variable in the study area. According to Figure 3 and Figure 4, 15 nodes were classified into cluster-1 with 32 samples which is corresponding to 50.8% of total number of data used for the study. Cluster-2 included 7 nodes with 11 data (17.5%) and 5 nodes were classified into cluster-3 with 6 data (9.5%). Cluter-4 (4 samples), cluster-5 (5 samples) and cluster-6 (5 samples) have commonly 4 nodes.

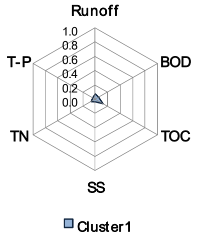
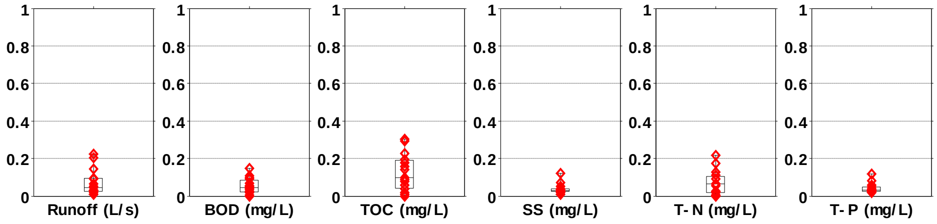
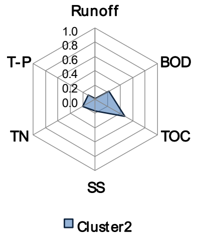
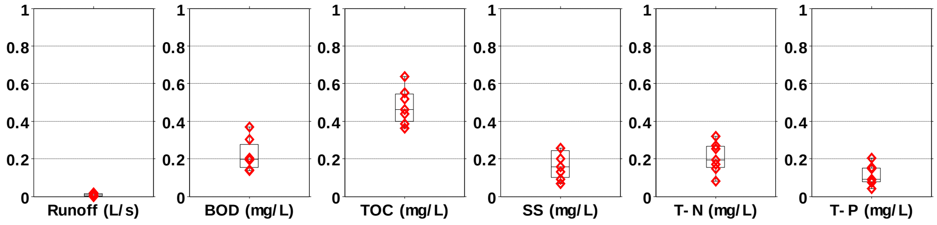
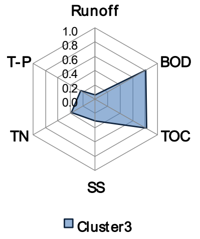
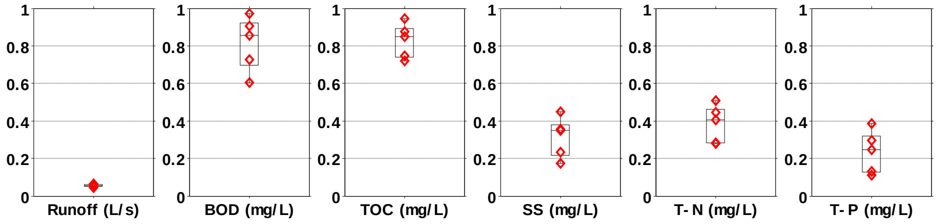
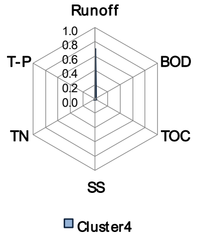
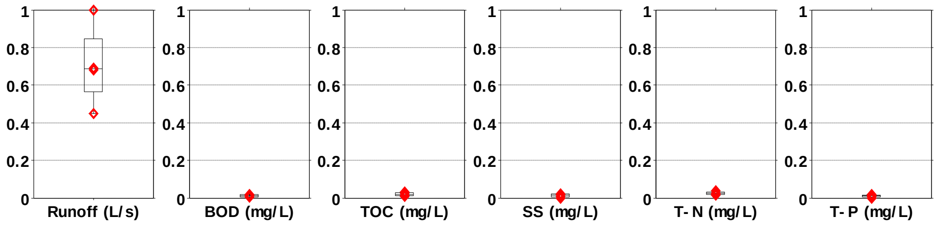
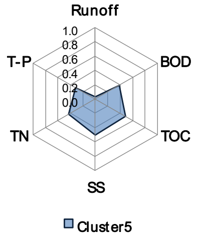
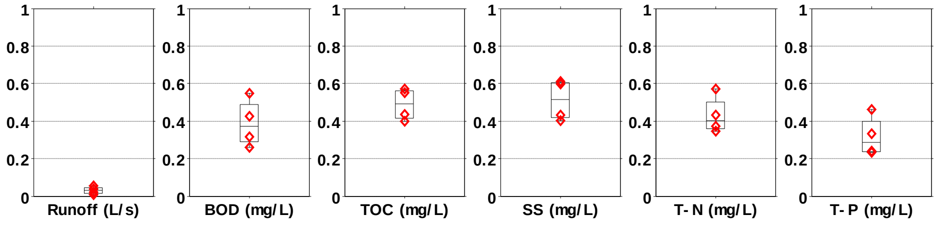
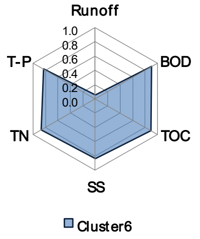
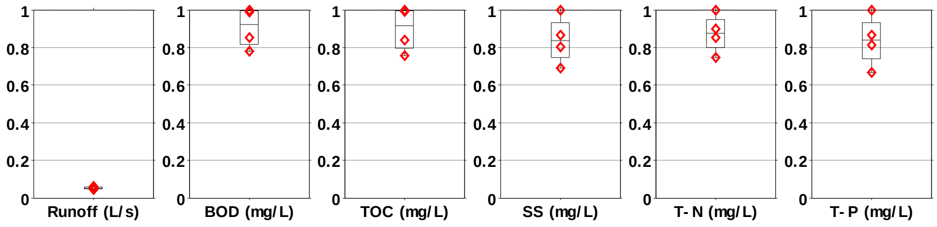
The reference vectors for the six variables were standardized into the range of [0 1] and the standardized values were used to calculate the mean values of each variable according to the respective clusters. The mean standardized values of the six variables were plotted on the hexagrams as shown in the left side of Table 9. Box-whisker plots representing the standardized reference vectors, and their minimum, the 1st to 3rd quartiles, and maximum values were shown in the right side of Table 9.

General interpretation of the hexagrams for the respective clusters, as shown in Table 9, revealed that three clusters (cluster-1, cluster-4 and cluster-6) were clearly separated by their distinguishable characteristics such as extremely low/high water quality or stormwater runoff pattern. Another three clusters including cluster-2, cluster-3 and cluster-5 were respectively classified based on different water quality patterns.

Cluster-1 was obviously classified by considerably low values of water quality concentration and stormwater runoff data and, conversely, cluster-6 was characterized by the extremely high concentration pattern of water quality with low stormwater runoff. Meanwhile, cluster-4 was associated with the highest stormwater runoff data and the lowest the water quality concentration data. Except for the cluster-4, another five clusters were related to low stormwater runoff data.

As mentioned above, cluster-2, cluster-3 and cluster-5 were grouped by different patterns with moderate values of water quality concentration. Low water quality concentration pattern with slightly higher TOC value than those of other items in the cluster was shown in cluster-2. Relatively low SS, TN and TP concentration patterns were commonly classified into cluster-3 and cluster-5. However, the cluster-3 was marked by much higher BOD and TOC concentration pattern than that of cluster-5.

Table 9. Hexagrams for the respective clusters and box whisker plots with minimum, the 1st to 3rd quartiles, maximum (black lines), and respective vectors (red diamond) for the respective variables

Hexagram	Box whisker plots for standardized reference vectors of six variables:
	
	
	
	
	
	

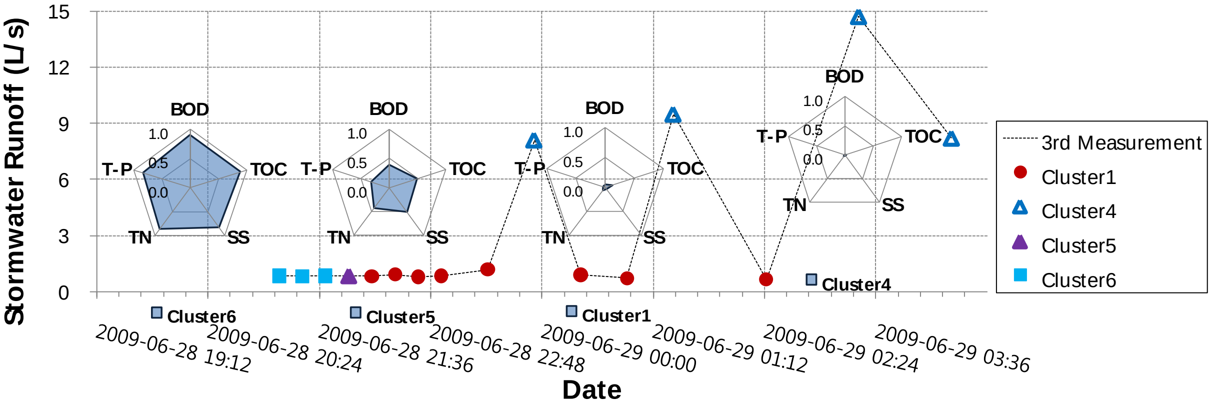


Figure 9. Variation of water quality patterns with respect to the stormwater runoff data during the third measurement period

5. CONCLUSIONS

Five water quality concentration and stormwater runoff data measured from a commercial district in the Gwangju Metropolitan City which is located in the southwestern region of Korea were used for the application of SOM. The primary objective of the utilization of the data and SOM was to obtain better understanding of the characteristics inherent in the measured water quality and stormwater runoff data, and to provide basic and supportive information for TMDLs in Korea.

As methodology to achieve the purpose, SOM combined with a hierarchical cluster analysis, which is one of the pattern classification methods and recently shows frequent application in a variety of research fields, was used in the present study, since it has remarkable ability for classification and visualization and accordingly its classification results can be interpreted and understood better.

SOM with hexagonal arrays of 13 × 3 map size classified six clusters according to the various water quality patterns and two types of stormwater runoff data. Initial interpretation of the six hexagrams showing the mean standardized values of reference vectors for the six variables divided the six clusters into two conceptually different groups based on the magnitude of each variable.

One group was characterized by extreme patterns such as extremely low values of all variables for cluster-1, the highest stormwater runoff with the lowest water quality concentration for cluster-4, and the worst water quality condition for cluster-6. The other group was marked by moderate patterns of water quality concentration expressing low water quality concentration with slightly high TOC for cluster-2, relatively high BOD and TOC concentration for cluster-3, and relatively low concentration of all water quality items for cluster-5.

In order to investigate the variation of water quality with respect to the stormwater runoff, the water quality patterns classified in the present study were presented in Figure 5 with the data measured during the third measurement period which includes the widest range of the runoff. The mean standardized values of the reference vectors for water quality items were drawn in the figure by pentagrams to show the variation of water quality only. It can be clearly seen in the figure that the dynamics of water quality patterns during the third event confirms the typical variation of pollutants showing the first-flush effect with a gradual decreasing trend of the five water quality concentration data.

As aforementioned, the better understanding of the characteristics inherent in the non-point source pollution data including water quality and stormwater runoff is crucial to manage and keep an environmental policy such as the TMDLs in Korea. In the present study, SOM classified the data into the six clusters according to exclusively different characteristics involved in the classified data. In other words, the method proposed in the present study provided informative knowledge on the water quality and stormwater runoff measured in a commercial area through pattern classification. Consequently, the applied method revealed it can be used to analyze the data used in the present study and obtain better knowledge of the non-point source pollution in the commercial district with highly impervious area.

Further study can be carried out with supplementary data between high and low stormwater runoff used in the present study. Additional measurement for water quality and stormwater runoff from various sites with different types of land use and cover is expected to give another representation of pattern classification in order to understand the relationship between non-point source pollution and land use/cover types. It also might be worth investigating the pattern classification including rainfall data.

6. ACKNOWLEDGEMENT

The present study was carried out as a part of the research project, “Solutions for the water related problems in Asian Metropolitan areas”, supported by the Tokyo Metropolitan Government, Japan. Field data were provided by the project, “Research on long-term monitoring and BMPs for the non-point source discharge”, financed by the Yeongsan River Environment Research Center, Ministry of Environment in Korea.

7. REFERENCES

Faggiano, L., Zwart, D., García-Berthou, E., Lek, S., and Gevrey, M., 2010. Patterning ecological risk of pesticide contamination at the river basin scale. *Science of the Total Environment* 408, 2319-2326.

Garcia, H.L., and González, I.M., 2004. Self-organizing map and clustering for wastewater treatment monitoring. *Engineering Applications of Artificial Intelligence* 17, 215-225.

Hentati, A., Kawamura, A., Amaguchci, H., and Iseri, Y., 2010. Evaluation of sedimentation vulnerability at small hillside reservoirs in the semi-arid region of Tunisia using the Self-Organizing Map. *Geomorphology* 122, 56-64.

Hsu, K., Gupta, H.V., Sorooshian, S., and Imam, B., 2002. Self-organizing linear output map (SOLO): An artificial neural network suitable for hydrologic modeling and analysis. *Water Resources Research* 38(12), 1302 (doi:10.1029/2001WR000795).

Jain, A., and Srinivasulu, S., 2006. Integrated approach to model decomposed flow hydrograph using artificial neural network and conceptual techniques. *Journal of Hydrology* 317, 291-306.

Jeong, K.-S., Hong, D.-G., Byeon, M.-S., Jeong, J.-C., Kim, H.-G., Kim, D.-K., and Joo, G.-J., 2010. Stream modification patterns in a river basin: Field survey and self-organizing map (SOM) application. *Ecological Informatics*, doi:10.1016/j.ecoinf.2010.04.005.

Nishiyama, K., Endo, S., Jinno, K., Uvo, C.B., Olsson, J., and Berndtsson, R., 2007. Indentification of typical synoptic patterns causing heavy rainfall in the rainy season in Japan by a Self-Organizing Map. *Atmospheric Research* 83, 185-200.

Srinivasulu, S., and Jain, A., 2006. A comparative analysis of training methods for artificial neural network rainfall-runoff models. *Applied Soft Computing* 6, 295-306.

Su, S., Zhi, J., Lou, L., Huang, F., Chen, X., and Wu, J., 2010. Spatio-temporal patterns and source apportionment of pollution in Qiantang River (China) using neural-based modeling and multivariate statistical techniques. *Physics and Chemistry of the Earth*, doi:10.1016/j.pce.2010.03. 021.

Versanto, J., Himberg, J., Alhoniemi, E., and Parahankangas, J., 2000. *SOM Toolbox for Matlab 5*. Report A57, Helsinki University of Technology, Finland.