Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2017.20681

77 (2017) 156–162
May

Addressing water pollution hotspots in the tributary monitoring network using a non-linear data analysis tool

Seo Jin Ki[a], Sanghwan Song[b], Tae-Woo Kang[b], Sangdon Kim[b], Taegu Kang[b], Seung Gwon Baek[b], Jong Hun Baek[b], Joon Ha Kim[a,*]

[a]*School of Environmental Science and Engineering, Gwangju Institute of Science and Technology (GIST), 123 Cheomdangwagi-ro, Buk-gu, Gwangju 61005, Korea, email: joonkim@gist.ac.kr*
[b]*Yeongsan River Environment Research Center, National Institute of Environmental Research, 5 Cheomdangwagi-ro 208beon-gil, Buk-gu, Gwangju 61011, Korea*

Received 14 November 2016; Accepted 1 January 2017

ABSTRACT

Successful data analysis is an essential component of any environmental monitoring programs. This study introduces an effective data analysis method to identify water pollution hotspots as well as to drop redundant monitoring parameters and samples using a self-organizing map (SOM), which has a strong specialty in pattern extraction from complex monitoring data. A full data set consisted of nine parameters that were obtained on a monthly basis from 83 sites in various tributary streams along the Yeongsan River, Korea, from May 2011 to December 2015. The given data set was further partitioned into a number of subsets to examine their effect on variable importance and temporal pattern analysis. We found that water pollution hotspots were more clearly addressed in load-based SOM analysis than in concentration-based SOM analysis due to strong correlation between variables resulted from variability reduction by combining two variables into a single one for load analysis. In addition, the variables chemical oxygen demand and electrical conductivity and the parameters discharge and total nitrogen were found to participate most and least actively in describing spatial and temporal variation of the observed variables, respectively. About 35% of the sampling locations showed high similarity among the monthly data extending from November in the previous year to February in the following year. We believe that the proposed methodology can be useful in revising the upcoming water monitoring study by clarifying several issues related to monitoring parameters and frequency in the existing program.

Keywords: Self-organizing map; Tributary monitoring; Water pollution hotspots; Sampling frequency; Temporal variability; Variable selection

1. Introduction

Surface water quality and quantity contribute to enriching public health and ecosystem integrity [1]. The current total maximum daily loads (TMDL) program, which has been adopted to address and restore impaired water bodies worldwide, required the authorizing agency and its partners to monitor water quality and quantity conditions on predetermined spatial and temporal scales [1,2]. Watershed management plan should be revisited and amended if the existing source loads estimated directly from the monitoring data (or indirectly from modeling with alternative future scenarios) were expected to exceed water quality standards allowed for individual pollutants [1]. Using those data, statistical analysis assisted in redesigning the present water monitoring program by assessing its efficiency in terms of monitoring parameters and frequency as well as by detecting water pollution hotspots, although its role was not often specified explicitly (like predictive models) under TMDL or watershed management plan [3–5].

* Corresponding author.

A self-organizing map (SOM), one of the non-linear data analysis tools, is superior to other conventional statistical methods in terms of tolerance to outliers and noisy data as well as data abstraction from large data sets [5–12]. For example, the previous study of Park et al. [13] showed that a large number of species abundance data collected from 836 sites were effectively reduced to medium to small amounts of data (for instance, from 941 through 353 to potentially 200) without losing much information of the original data using a particular index computed from unique properties of the SOM output map. Another study of Ki et al. [5] also revealed that SOM successfully captured heterogeneous water quality (with respect to trace metals) and quantity signals, which varied considerably during storm events, in addition to determining the appropriate sample size required for pollutant load estimation in each event. Other study of Tudesque et al. [14] demonstrated that SOM was sufficiently robust to address monitoring locations that experienced significant changes in water physicochemistry (e.g., cations and anions) during a long-term monitoring period reaching 3 decades. All these representative examples confirmed that SOM analysis could be applied to various types of environmental monitoring data; even in cases the relationship between them was highly non-linear. A readily understandable visualization for correlation between variables is also a plus. Various application examples of SOM, including its fundamental theory, can be found elsewhere [6–8,15].

By applying the versatile tool SOM to spatially and temporally sparse data sets, the main objective of this study was to provide an in-depth diagnosis of water quality and quantity conditions in a tributary monitoring network. More specifically, we used SOM in this study: (1) to compare water pollution hotspots between concentration- and load-based analyses, (2) to identify informative and redundant parameters in describing spatial and temporal behaviors of water quality and quantity, and (3) to assess temporal data repeatability for individual monitoring locations. It is our hope that the proposed methodology can be used to refine the existing water monitoring programs in terms of monitoring parameters and frequency, along with various simulation models, which apply for regulatory purposes.

2. Materials and methods

2.1. Tributary monitoring network and observed parameters

Fig. 1 shows 83 sampling locations that are selected to assess water quality and quantity conditions at the tributary monitoring network in the Yeongsan River, Korea. In fact, the parent river receives water from a total of 170 small streams classified into four types, from primary through secondary to later order (i.e., tertiary and quaternary) tributaries. Among them, only 74 candidate streams were included in our tributary monitoring network based on the initial screening process. Specifically, the screening criteria excluded tributaries that not only had minimal influence on the main channel (with respect to water quality) but also provided intermittent (or discontinuous) or slower-moving (or stagnant) water discharges. Along those candidate streams, we finalized the design of the tributary monitoring network for 83 sites. Note that Jiseokcheon (53.00 km in terms of main channel length),

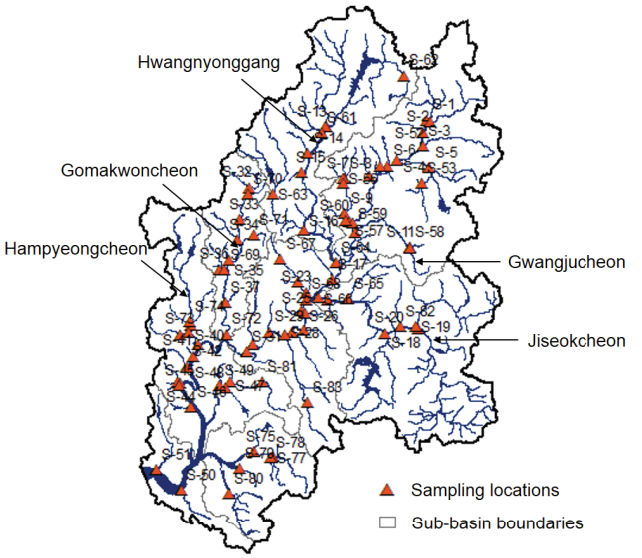

Fig. 1. 83 sampling locations at various tributaries in the Yeongsan River, Korea.

Hwangnyonggang (58.60 km), Gomakwoncheon (34.30 km), Hampyeongcheon (28.80 km), and Gwangjucheon (22.80 km) are the top five large tributaries that flow directly into the mainstream of the Yeongsan River (129.50 km). Diffuse pollution was found to be largely responsible for water quality degradation in the Yeongsan River, because the basin was mainly dominated by a combination of forestland (46.4%), agricultural land (24.6%), and rangeland (14.8%).

In those locations, monthly field studies to measure water quality and quantity were conducted from May 2011 to December 2015. The water quality measurements were made by two different methods: using a real-time instrument YSI-650 MDS (Xylem Inc., Rye Brook, New York, USA) for water temperature, pH, dissolved oxygen (DO), and electrical conductivity (EC) as well as using standard test methods for water pollution (released by the Ministry of Environment in Korea) for biochemical oxygen demand (BOD), chemical oxygen demand (COD), total organic carbon (TOC), total nitrogen (TN), total phosphorus (TP), chlorophyll-a, and suspended solids (SS), once sent to the laboratory under 4°C. In general, discharge was estimated from cross-sectional area (i.e., a product of channel depth and width) and velocity (recorded manually at individual subsections). However, when (continued) access to field sites was technically impossible, an indirect estimation method such as water balance approach was adopted for discharge estimation. Table 1 displays summary statistics for nine parameters collected from the tributary monitoring network. Note that we only provide nine variables (out of them) as inputs to SOM analysis for illustrative purposes. Another reason is that the variables excluded either involve many missing data (for chlorophyll-a) or do not contain significant information to characterize water quality and quantify variation along the tributaries (for water temperature and pH). In the table, discharge, ranked first in terms of missing data, was found to exhibit the highest variability across the tributaries among the input variables (compare coefficient of variation

values). More detailed information on the (drainage) basin characteristics such as soils, topography, and climate as well as assessment of water quality and quantity is documented well in our recent works [2,5,10,16,17].

2.2. Non-linear pattern analysis

SOM is a data analysis tool that efficiently retrieves concise (spatial and temporal) profiles from the complex data set (of high dimensionality) in a non-linear manner. SOM has two distinct features, vector quantification property that makes the codebook vectors (i.e., representative samples) approach to the input probability density as well as topology preserving mapping that still retains the relative distances between the raw data points in a new output space (of low dimensionality). Unlike other conventional statistical analyses, the tool is found to be free from outliers and noisy data and also capable of restoring, in part, the lost data. Two main algorithms involved in SOM are initialization and training, which assign initial values to the codebook vectors and adjust the codebook vectors with their neighbors toward the input vectors, respectively. After this step, the codebook vectors updated during iterative training are eventually arranged and visualized in two-dimensional neurons (in our case), where similar data are located closely and dissimilar data are far apart from each other.

When the full data set containing nine variables was provided as inputs to SOM, we did not specify the size (i.e., the number of neurons) of the output map (namely concentration-based analysis). However, we adjusted the map size for the reduced data set consisting of only six variables (by multiplying each pollutant concentration, except for two unnecessary variables, by discharge for load-based analysis) to that of the full data set to provide a consistent view between two output maps. Nine separate data sets, which excluded a particular variable sequentially one by one from the full data set, were additionally prepared to assess the importance of individual variables in the tributary monitoring study. Finally, SOM received a total of eighty-three data sets to review temporal data patterns per sampling location. These data sets were made of nine variables, but only included the monitoring data for each sampling location. While the map size of nine additional data sets was equal to the original size of the full data set, that of eighty-three data sets was set to twelve based on the assumption that the data were significantly modulated by month rather than by year in the absence of any anthropogenic activities. Note that in each SOM run, we use the default options of initialization (through linear initialization mode), and training (through batch training mode) to avoid any influence of (learning) algorithms on the codebook vectors produced. SOM toolbox, which can be embedded in MATLAB 5 or higher, is available for download at http://www.cis.hut.fi/somtoolbox/.

Table 1
Summary statistics for monthly water quality and quantity data obtained from 83 sampling locations at various tributaries in the Yeongsan River, Korea, from May 2011 to December 2015 (n = 4,648)

Parameters	Mean	CV[a]	Missing data (%)
Dissolved oxygen (DO), mg/L	10.10	0.29	0.95
Electrical conductivity (EC)	287.39	0.78	0.95
Biochemical oxygen demand (BOD), mg/L	3.07	0.99	0.71
Chemical oxygen demand (COD), mg/L	6.57	0.72	0.75
Total organic carbon (TOC), mg/L	4.46	0.68	0.77
Total nitrogen (TN), mg/L	3.41	0.92	0.77
Total phosphorus (TP), mg/L	0.15	1.33	0.75
Suspended solids (SS), mg/L	19.58	1.79	0.77
Discharge, cm^3/s	0.81	2.50	4.58

[a]CV = the coefficient of variation (a ratio of the standard deviation to the mean).

3. Results

3.1. Concentration- and load-based water pollution hotspots

Fig. 2 presents the difference of water pollution hotspots between concentration- and load-based SOM analyses using

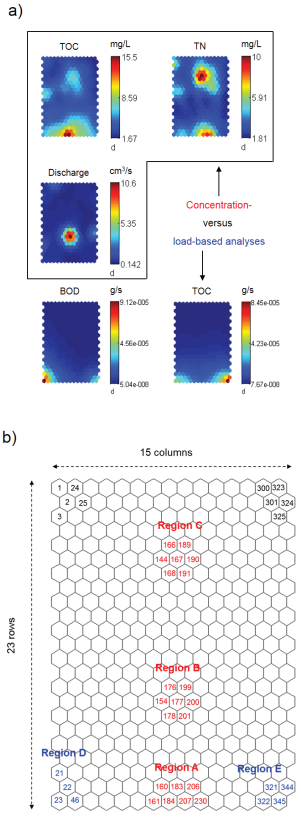

Fig. 2. (a) Concentration- and load-based SOM analyses and (b) major areas of concern (i.e., a collection of neurons) in the SOM output map. Note that in Fig. 2(b), individual numbers indicate the sequence of neurons arranged in the SOM output map.

the monthly data sets obtained from 83 sampling locations at various tributaries in the Yeongsan River of Korea for almost 5 years, along with the correlation among measured variables. In Fig. 2(a), color bars indicate the range of individual variables, and component planes enclosed by the solid lines represent the results of concentration-based analysis only. Note that both concentration- and load-based analyses do not include (spatial and temporal) variation of all variables, provided as inputs to SOM, for brevity.

It was found from the figure that while concentration-based analysis showed three distinct data distributions, two clear patterns were mainly observed in load-based analysis. For example, in concentration-based analysis, TOC concentrations were only high at a series of neurons around Region A (see the bottom-middle of the SOM output map in Fig. 2(b)). In contrast, high discharge was observed mostly in six neurons around Region B. TN component plane showed moderate to high levels of the target variable in two different areas, Regions A and C, respectively. In fact, with respect to concentration-based analysis, TOC showed a strong positive correlation with EC, COD, TP, and SS, whereas no relationship was observed between either DO or BOD and the remaining parameters such as TN and discharge (data not shown). For load-based analysis, BOD and TOC loads were high in Regions D and E, respectively. Also, data distribution (i.e., contaminant loads) in TOC component plane was almost or exactly identical to those of other component planes, except for BOD (data not shown). These results revealed that (1) little or no relationship existed among either concentration- or load-based variables presented and (2) key areas for concern were different among variables of interest as well as between concentration- and load-based analyses. These implied that concentration- and load-based analyses might address different water pollution hotspots depending on whether discharge was included as a separate variable or not in the data sets. Note that two variables DO and EC are removed from load-based analysis due to the absence of contaminant mass load units.

Table 2 shows a detailed list of sampling locations assigned to individual areas of concern (in the SOM output map), which were obtained from concentration- and load-based analyses (see Fig. 2(b)). From the table, it was shown that the total number of data assigned from concentration-based analysis was considerably larger than those of load-based analysis. Also, water pollution hotspots (i.e., sampling locations) identified were significantly different between concentration- and load-based analyses, as discussed above. However, load-based analysis still addressed and shared six sampling locations S-1, S-39, S-59, S-74, S-76, and S-83 in Regions D and E as potential hotspot locations, which were also included and spread in concentration-based analysis. Therefore, load-based SOM analysis appeared to be superior to concentration-based SOM analysis in addressing apparent water pollution hotspots from complex spatial and temporal data sets of water quality and quantity.

3.2. Important variables for tributary water monitoring program

The effects of individual variables on SOM analysis was assessed by eliminating each variable at a time from the full data set, which included nine variables for the entire sampling period (see Table 3). In the table, the percentage change of the remaining variables due to the absence of one particular variable was estimated by dividing the difference of the median values between modified and original codebook vectors by the median value of the original ones and then multiplying this quantity by 100. Note that in this calculation, we use the median rather than the mean, which is more influenced by skewed data. The codebook vectors assigned to individual neurons in SOM denote denormalized values transformed back into the original range of variables (for visualization of component planes, see Fig. 2(a)). The table should be read as the following example. When the variable DO was eliminated from the full data set, discharge and TOC, ranked first and second in terms of absolute values in descending order, recorded a decrease of 10.35% and an increase of 9.06% from the original codebook vectors, respectively (see the first row). This implied that if we did not measure DO parameter fully in the tributary water monitoring program, this led to significant bias in (produced) data for discharge and TOC (component planes in SOM). The same applies to the remaining variables, i.e., consecutive rows of the table. Note that in the table, the values on the diagonal line are left empty because the modified codebook vectors cannot be generated and compared with the original ones when a particular variable is excluded.

Table 2
A detailed list of sampling locations identified by major areas of concern using concentration-based SOM analysis (regions A, B, and C) and load-based SOM analysis (regions D and E; see Fig. 2(b))

Regions	Sampling locations[a]	Total number of data
A	S-8[b], S-14, S-15, S-27, S-36, S-37, S-38, S-45, S-47, and S-67	89
B	S-1, S-3, S-5, S-7, S-16, S-17, S-19, S-20, S-26, S-30, S-35, S-39, S-59, S-69, S-71, S-74, S-75, S-76, S-77, S-79, and S-83	79
C	S-18, S-20, S-23, S-24, S-25, S-26, S-27, S-28, S-34, S-36, S-38, S-39, S-40, S-41, S-45, S-49, S-50, S-58, S-62, S-63, S-64, S-76, and S-82	71
D	S-1, S-5, S-30, S-39, S-59, S-69, S-70, S-74, S-76, S-78, and S-83	16
E	S-1, S-7, S-16, S-20, S-39, S-59, S-74, S-76, and S-83	19

[a]Refer to Fig. 1 for individual sampling locations.
[b]Note that for simplicity, sampling locations are presented only after excluding temporal information.

Table 3
Percentage change of the median values for representative samples (i.e., codebook vectors) in the SOM output map when individual variables are removed rotationally one at a time (%)[a]

	DO	EC	BOD	COD	TOC	TN	TP	SS	Discharge
DO[b]		−1.68	4.32	3.59	9.06	−3.43	2.30	−1.27	−10.35
EC	−0.35		6.26	2.60	3.32	4.12	6.95	8.52	5.32
BOD	−0.17	1.55		−2.20	0.14	−2.70	0.47	−11.16	−19.44
COD	−0.45	−0.74	9.86		5.77	5.60	5.73	10.59	−6.59
TOC	−0.11	−1.06	2.30	1.57		2.96	3.28	4.39	−1.29
TN	−0.26	−1.28	6.16	−1.33	−0.10		−0.83	−0.04	3.24
TP	−0.09	1.73	2.65	0.85	1.98	3.46		4.99	−1.64
SS	0.04	−1.41	3.86	−1.02	1.32	−0.44	2.74		−1.61
Discharge	0.13	0.19	7.69	−0.96	1.45	1.06	2.96	2.99	

[a]See Table 1 for full names of individual parameters.
[b]The percentage change estimated was arranged in a row direction.

Table 4
Rank order of significant variables for generating representative samples (i.e., codebook vectors) in the SOM output map

Rank	Parameters	Median values[a]
1	COD	5.75
2	EC	4.72
3	DO	3.51
4	TOC	1.94
5	BOD	1.88
6	TP	1.86
7	SS	1.37
8	Discharge	1.26
9	TN	1.06

[a]The median values were estimated by the absolute changes of individual variables arranged in a row (see Table 3).

Table 4 shows a list of variables arranged in order of importance by taking the median of their absolute values of the percentage change (for the remaining variables) in the absence of a particular variable (see each row in Table 3). The table confirmed that the variables COD and EC were most useful to elucidate spatial and temporal data patterns observed, whereas discharge and TN were identified as the least influential variables. These results indicated that (1) those important variables, at least, should be monitored ahead of the others in the tributary water monitoring campaign and (2) redundant (or unimportant) parameters could be dropped for the following program in this way, if needed.

3.3. Sampling frequency for different tributary monitoring stations

We also observed temporal data patterns of individual sampling locations by allowing the monthly data in each site to one of twelve neurons in the SOM output map (see Fig. 3). In the figure, color bars indicate the number of data assigned to individual neurons in the output map, and only half of

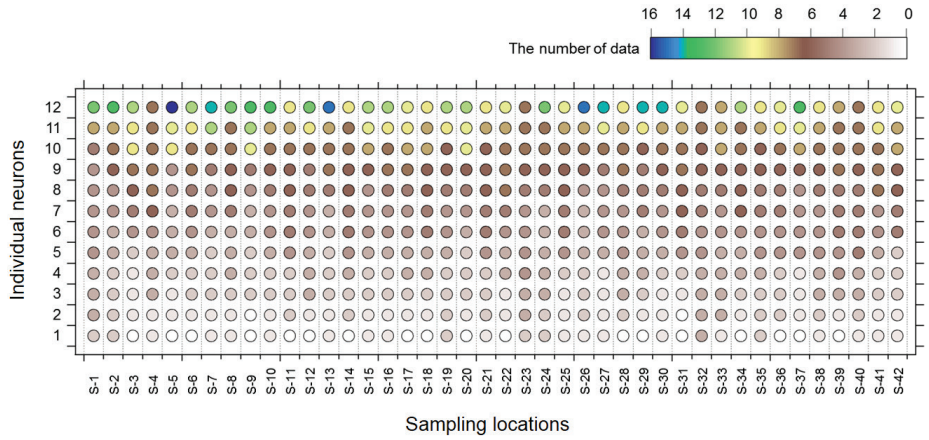

Fig. 3. Data distribution (pattern) assigned to individual SOM neurons for sampling locations from S-1 to 42. Note that concentration-based SOM analysis is conducted separately for each sampling location to evaluate temporal data repeatability per site.

the entire sampling locations from S-1 to S-42 are shown as an example of temporal pattern analysis. Note that twelve neurons are specifically selected for this analysis because the monthly data are assumed to vary meaningfully from month to month as the general (or intended) purpose of the tributary water monitoring program, regardless of monitoring years. Thus, if the monitoring data are modulated fairly well on monthly base (rather than on yearly base) without additional pollutant loads, they are evenly distributed in the twelve neurons. Otherwise, some neurons receive more data than expected. For example, a set of four to five monthly data should be assigned to individual neurons (in the normal case) when considering the entire monitoring period (from May 2011 to December 2015). As displayed in the figure, data distribution patterns across the neurons were highly variable depending on sampling locations. For example, extremely skewed distributions (i.e., a large number of data was assigned to a particular neuron) were observed for the sampling locations S-5 and S-13, whereas the monthly data were evenly spread in the twelve neurons for S-4, S-23, S-32, S-33, S-39, and S-40. Also, the skewness was moderate for the remaining sampling locations.

Table 5 provides a complete list of sampling locations that contain more than twelve data in a particular neuron in this way. A total of 29 sampling stations were found to have at least twelve monthly data. Among them, S-5 and S-13 were ranked first and second in terms of a number of data, respectively. In addition, two neurons in S-48 were in excess of the twelve monthly data. These results revealed that data repeatability in three sampling locations was high (in other words, temporal variability of the monthly data was low), indicating that sampling frequency for these locations could be reduced preferentially to minimize the cost of monitoring ahead of other sites. Specifically, the monthly data between November in the previous year and February in the following year were very similar to each other for most of sampling locations, except for some sites.

Table 5
Sampling locations of concern that show high temporal data repeatability in (concentration-based) SOM analysis (see Fig. 3)

Sampling locations[a]	Monthly data in specific neuron(s)[b]	No. of data
S-1	1, 2, 11, and 12	12
S-2	1, 2, 3, 11, and 12	13
S-5	1, 2, 3, 4, 10, 11, and 12	16
S-7	1, 2, 3, 4, 11, and 12	14
S-8	1, 3, 4, 5, 6, 8, 9, and 10	12
S-9	1, 2, 3, 5, 11, and 12	13
S-10	1, 2, 3, 10, 11, and 12	13
S-12	1, 2, 3, 4, 11, and 12	12
S-13	1, 2, 3, 11, and 12	15
S-24	1, 2, 3, 11, and 12	12
S-26	1, 2, 4, 6, 10, 11, and 12	14
S-27	1, 2, 3, and 12	14
S-29	1, 2, 3, 10, and 12	14
S-30	1, 2, 3, 11, and 12	14
S-37	1, 2, 3, 11, and 12	13
S-48	1, 2, 3, 4, 11, and 12; and 3, 5, 7, 8, 9, and 10	13 and 12
S-51	4, 5, 6, 7, 8, and 9	14
S-52	1, 2, 3, 4, and 12	12
S-58	1, 2, 3, 4, 11, and 12	14
S-60	1, 2, 4, 11, and 12	12
S-61	1, 2, 4, 5, 11, and 12	13
S-63	1, 2, 3, 4, 10, 11, and 12	14
S-65	1, 2, 11, and 12	12
S-68	1, 2, 5, 6, 7, 8, 9, 10, and 11	14
S-69	1, 2, 3, 11, and 12	13
S-70	1, 2, 3, 4, and 12	12
S-71	1, 2, 3, and 12	14
S-79	1, 2, 4, 11, and 12	12
S-81	1, 2, 3, 4, and 12	12

[a]Refer to Fig. 1 for individual sampling locations.
[b]Note that monthly data are presented regardless of monitoring year.

4. Conclusion

The present study describes the methodology to elucidate complex spatial and temporal variation of the monthly data set obtained from the tributary water monitoring study using the non-linear data analysis tool, SOM. The full data set measured for almost 5 years included nine parameters, from which separate data sets were provided to SOM to effectively screen water pollution hotspots, important variables, and temporal patterns. From this study, we obtained the following results:

- Water pollution hotspots addressed by concentration-based analysis was much larger than those from load-based analysis. In contrast, the correlation between measured variables was stronger in load-based analysis than concentration-based analysis. Out of the two, load-based analysis showed superior performance in apparently detecting potential water pollution hotspots.
- Removing a particular variable (rotationally) from the full data set had a significant influence on the codebook vectors of the remaining variables, which represented their spatial and temporal variation over the tributaries. The SOM analysis was most sensitive to elimination of COD and EC; and least sensitive to variables such as discharge and TN.
- A total of 29 sampling locations among all 83 sites investigated exhibited repeatable data patterns on the monthly time scale. We also observed similar patterns of the monthly data from November in the previous year to February in the following year for most of these sampling locations. Specifically, three sites S-5, S-13, and S-48 were found to have low temporal variability during the entire sampling period.

Acknowledgments

This research was supported by the 2016 Basic Environmental Survey Projects funded by the Watershed Management Committee at the Yeongsan and Seomjin Rivers in Korea. We greatly acknowledge the financial support of the committee.

References

[1] U.S. Environmental Protection Agency, Handbook for Developing Watershed Plans to Restore and Protect Our Waters, Report No. 841-B-08-002, Office of Water, Washington, D.C., USA, 2008.

[2] S.J. Ki, Y.G. Lee, S.W. Kim, Y.J. Lee, J.H. Kim, Spatial and temporal pollutant budget analyses toward the total maximum daily loads management for the Yeongsan watershed in Korea, Water Sci. Technol., 55 (2007) 367–374.

[3] H. Lee, X. Swamikannu, D. Radulescu, S.-j. Kim, M.K. Stenstrom, Design of stormwater monitoring programs, Water Res., 41 (2007) 4186–4196.

[4] P.L. Brezonik, T.H. Stadelmann, Analysis and predictive models of stormwater runoff volumes, loads, and pollutant concentrations from watersheds in the Twin Cities metropolitan area, Minnesota, USA, Water Res., 36 (2002) 1743–1757.

[5] S.J. Ki, J.-H. Kang, S.W. Lee, Y.S. Lee, K.H. Cho, K.-G. An, J.H. Kim, Advancing assessment and design of stormwater monitoring programs using a self-organizing map: characterization of trace metal concentration profiles in stormwater runoff, Water Res., 45 (2011) 4183–4197.

[6] T. Kohonen, Self-organizing Maps, 3rd ed., Springer Series in Information Sciences, Vol. 30, Springer-Verlag, Berlin, Heidelberg, New York, USA, 2001, p. 502.

[7] J. Vesanto, Data Exploration Process Based on the Self-organizing Map, Acta Polytechnica Scandinavica, Mathematics and Computing Series No. 115, Finnish Academies of Technology, Espoo, Finland, 2002.

[8] J. Vesanto, J. Himberg, E. Alhoniemi, J. Parhankangas, SOM Toolbox for Matlab 5, Report A57, SOM Toolbox Team, Helsinki University of Technology, Espoo, Finland, 2000.

[9] K.H. Cho, J.-H. Kang, S.J. Ki, Y. Park, S.M. Cha, J.H. Kim, Determination of the optimal parameters in regression models for the prediction of chlorophyll-a: a case study of the Yeongsan Reservoir, Korea, Sci. Total Environ., 407 (2009) 2536–2545.

[10] S.J. Ki, S.W. Lee, J.H. Kim, Developing alternative regression models for describing water quality using a self-organizing map, Desal. Wat. Treat., 57 (2016) 20146–20158.

[11] G. Loganathan, S. Krishnaraj, J. Muthumanickam, K. Ravichandran, Chemometric and trend analysis of water quality of the South Chennai lakes: an integrated environmental study, J. Chemom., 29 (2015) 59–68.

[12] A. Astel, S. Tsakovski, P. Barbieri, V. Simeonov, Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets, Water Res., 41 (2007) 4566–4578.

[13] Y.-S. Park, J. Tison, S. Lek, J.-L. Giraudel, M. Coste, F. Delmas, Application of a self-organizing map to select representative species in multivariate analysis: a case study determining diatom distribution patterns across France, Ecol. Inf., 1 (2006) 247–257.

[14] L. Tudesque, M. Gevrey, G. Grenouillet, S. Lek, Long-term changes in water physicochemistry in the Adour–Garonne hydrographic network during the last three decades, Water Res., 42 (2008) 732–742.

[15] M. Bieroza, A. Baker, J. Bridgeman, Exploratory analysis of excitation–emission matrix fluorescence spectra with self-organizing maps as a basis for determination of organic matter removal efficiency at water treatment works, J. Geophys. Res., 114 (2009) G00F07.

[16] J.-H. Kang, Y.S. Lee, S.J. Ki, Y.G. Lee, S.M. Cha, K.H. Cho, J.H. Kim, Characteristics of wet and dry weather heavy metal discharges in the Yeongsan Watershed, Korea, Sci. Total Environ., 407 (2009) 3482–3493.

[17] Yeongsan River Environment Research Center (YRERC), The Second Final Report on Water Quality Monitoring on Tributaries in the Yeongsan River Basin, Korea, YRERC, National Institute of Environmental Research, Gwangju, Republic of Korea, 2014.