Desalination and Water Treatment
www.deswater.com

doi: 10.1080/19443994.2015.1112981

57 (2016) 20146–20158
September

Taylor & Francis
Taylor & Francis Group

Developing alternative regression models for describing water quality using a self-organizing map

Seo Jin Ki[a], Seung Won Lee[b], Joon Ha Kim[a,c,*]

[a]*School of Environmental Science and Engineering, Gwangju Institute of Science and Technology (GIST), Gwangju 500-712, Korea, Tel. +82 62 7153277; Fax: +82 62 7152434; email: joonkim@gist.ac.kr (J.H. Kim)*
[b]*Environmental and Plant Engineering Research Institute, Korea Institute of Civil Engineering and Building Technology, Gyeonggi-do 10223, Korea*
[c]*Sustainable Water Resource Technology Center, GIST, Gwangju 500-712, Korea*

Received 15 September 2015; Accepted 8 October 2015

ABSTRACT

Statistical models play an important role in elucidating the dynamic behaviors of surface water quality, given limited data on a large scale. In this study, we examine alternative approaches to develop regression models that predict fecal coliform (FC) concentrations in a river using different methods for selecting important variables provided by a self-organizing map (SOM). The raw data used as input to the SOM included 11 water quality, 6 meteorological, and 7 land use parameters that were monitored along the Yeongsan River in Korea on various time scales (from daily to half a decade) during 1996–2008. In both test and validation data sets, (multiple) regressions using backward elimination were compared against regression models via forced entry, which included a set of ranked variables simultaneously based on four indices in the SOM (i.e. structuring index, relative importance, cluster description, and Spearman's rank correlation). Results showed that the SOM effectively illustrated the complex relationship between FC and the remaining variables in the entire data set. This relationship was seen more clearly in homogeneous clusters, indicating that the regression models became more robust in each subdivided group. While the original backward elimination model ($R^2 = 0.66$) had much better performance than the models with four indices ($R^2 = 0.40$–0.45) in the test data set, its performance ($R^2 = 0.42$) was quite comparable to the relative importance model ($R^2 = 0.38$) in the validation data set. Based on this preliminary study, we recommend further investigation of these indices for a reliable regression analysis, as the t values currently used for the variable selection in regressions provide only a locally optimal solution for the final model. The proposed methodology, if verified successfully, would be useful in developing early warning models that control mortality or disease rates of fishes in high-density aquafarms via water quality.

Keywords: Regression models; Self-organizing map; Variable selection; Relative importance; Fecal coliform; Water quality data sets

*Corresponding author.

1. Introduction

The protection of surface water quality is essential for promoting the benefits of a healthy watershed, i.e. both human and ecosystem health and associated services required for economic growth [1]. Among the multiple pollutants (e.g. nutrients, toxic metals, pesticides, and pathogens) detected in surface waters, fecal contamination has increasingly become a global concern due to its threat in causing disease and restrictions on water use for recreation and irrigation [1–6]. Potential sources of fecal contamination are storm water run-off, untreated sewage, failing septic systems, and landfill leachate, which enter surface waters through overland flow and groundwater discharge [1,3,4]. Multiple sources of fecal contamination have then amplified its cumulative impact at several hotspots in the watershed, including nearby coastal areas [3–6]. Thus, correctly assessing such locations in terms of the level of fecal pollution is important for developing effective microbial pollution reduction scenarios, specifically for increasing the beneficial uses of surface waters.

Statistical tools are effectively used to address the spatial and temporal characteristics of surface water quality, when the data for calibrating a set of parameters in complex simulation models are not completely available [1,5–8]. Previous studies have shown that various types of environmental data consisting of physical, chemical, and biological parameters were successfully analyzed using both linear and nonlinear statistical approaches [5–10]. In particular, multiple linear regression (MLR) has been widely used as a fundamental tool for exploring the relationship between variables in a linear fashion, because of its simple formula and straightforward computation [5,6]. In contrast, a self-organizing map (SOM) enabled a more robust analysis of complex, nonlinear data patterns in an unsupervised manner without requiring the complete understanding of a given data structure, such as the number of clusters [7–13]. More precisely, SOM have been preferred over linear data reduction methods (e.g. principal component and discriminant analyses) due to its strong capabilities for classification and discrimination, even in the presence of noise and outliers [11–14]. Therefore, both tools can be applied to describe water quality degradation and its interactions with environmental variables, if used properly [6–10]. However, the (prediction) performance of SOM tends to be more reliable than the MLR as the data sets violate the underlying assumptions of parametric methods, such as normality, linearity, and independence (of residuals).

As compared to previous studies, this study aims to develop regression models that predict fecal coliform (FC) concentrations in a river by combining the advantages of SOM (to characterize data patterns and variable importance) and MLR (to generate a simple prediction equation). Note that a series of variables are sequentially selected or removed in the current MLR based on the *t*-test values that represent the contribution of each variable to the model without sound theoretical reasons [15]. A raw data set consisting of the water quality, climate, and land use parameters from the Yeongsan River (Basin) in Korea was used for illustrative purposes. From the data set, this study specifically: (1) describes the relationship between FC and all other variables (using SOM), (2) identifies significant variables in organizing or classifying data in a map structure (using SOM), and (3) assesses the prediction performance of a reference regression and the models constructed from recommended variables (using both MLR and SOM). It is our hope that the proposed methodology offers not only a rationale for determining physically meaningful variables when developing statistical models, but also can provide new insights into fingerprinting microbial pollution (hotspots) in surface waters.

2. Materials and methods

2.1. Field site description

The study area is located on the Yeongsan River (Basin) in the South Jeolla Province of Korea (Fig. 1). The river is relatively short (at 135 km) and covers a small drainage area (3,500 km^2), when compared to other major rivers in Korea. In total, 13 tributaries join the mainstream of the river, which passes though the urban residential areas of Gwangju City in its midstream and then drains into the West Sea. Historically, the river has played an important role as a waterway transportation route due to its large range of tidal currents that reached around 73 km inland from its mouth during a flood tide. After the construction of an estuary dyke in the early 1980s, the river has been used as an irrigation water source as well as for drainage and flood control. However, from that time its water quality has regularly been a primary concern in Korea as the water degraded from both intense agricultural activities and moderate urban growth [2,16].

Fig. 1 illustrates the processes used to prepare the input data set for subsequent statistical analyses, i.e. selecting the subbasin outlets (i.e. monitoring locations), overlaying the land use map onto the target region, and subdividing the entire basin into multiple subbasins. The basin was subsequently delineated into

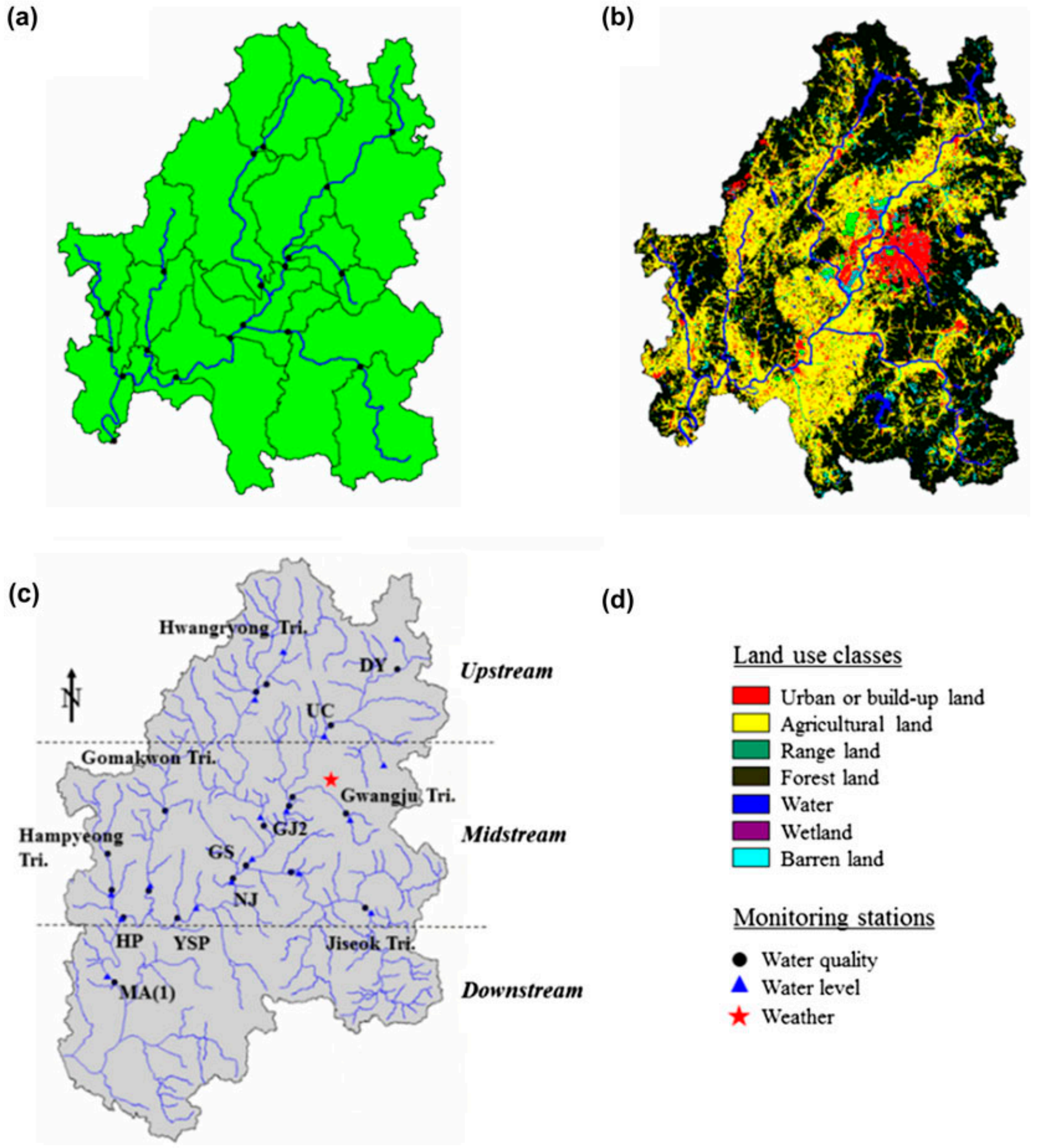

Fig. 1. Study area of the Yeongsan River (Basin) in Korea: (a) 19 discretized subbasins, (b) land use distribution for 2005, and (c) network of water quality, (water) level, and weather monitoring stations. In (c), the basin is divided into three main zones from the headwater to the river mouth: up, mid, and downstream sectors. The following abbreviations denote representative water quality monitoring points along the river: DY = Damyang, UC = Uchi, GJ(2) = Gwangju 2, GS = Gwangsan, NJ = Naju, YSP = Yeongsanpo, HP = Hampyeong, and MA(1) = Muan 1.

19 drainage areas based on the routine water quality monitoring stations along the river (i.e. 8 sites from the mainstream and 11 sites from the tributaries) such that each subbasin encompassed different combinations of land use (Fig. 1(a) and (b)). Similarly, water level gauging stations, at the closest points to the nearby water quality monitoring stations, were selected to obtain data for the subbasin outlets (Fig. 1(c)). Note that some areas in the downstream are not presented in the maps during the discretization process (Fig. 1(a) and (b)).

Table 1 presents the historical land use patterns of the selected drainage areas in the basin from 1975 to 2005, defined according to the classification system used by the United States Geological Survey (USGS) [17]. In the table, each land use class is subdivided into three spatial sections, i.e. up, mid, and downstream (Fig. 1(c)). From the table, the basin

Table 1
Historical changes of land use (and land cover) at three sections of the Yeongsan River (Basin) in Korea from 1975 to 2005 (unit: km^2)

Sublevel classes[a]	Sections	1975 (60 m)[b]	1980 (60 m)	1985 (30 m)	1990 (30 m)	1995 (30 m)	2000 (30 m)	2005 (30 m)	Total change[c]
Urban or built-up land	Upstream	0.8	2.6	12.1	14.2	18.3	16.3	27.6	26.8 (0.9%)
	Midstream	15.9	41.5	56.0	94.9	106.6	139.0	161.1	145.1 (5.1%)
	Downstream	1.9	7.5	8.3	16.1	19.2	27.5	27.0	25.0 (0.9%)
Agricultural land	Upstream	186.9	182.5	184.8	189.2	198.5	183.7	185.8	−1.1 (0%)
	Midstream	622.5	615.9	638.9	603.3	648.9	589.1	590.2	−32.3 (−1.1%)
	Downstream	221.2	235.5	234.4	231.6	230.9	237.4	236.0	14.8 (0.5%)
Rangeland	Upstream	3.5	14.0	2.0	1.9	19.7	15.0	14.2	10.7 (0.4%)
	Midstream	61.9	36.3	25.5	31.7	34.3	31.1	45.6	−16.3 (−0.6%)
	Downstream	13.6	6.9	10.1	15.9	19.3	2.9	7.9	−5.7 (−0.2%)
Forest land	Upstream	502.9	489.9	485.1	478.5	443.5	472.7	448.0	−54.9 (−1.9%)
	Midstream	874.3	866.6	873.1	860.5	805.3	811.9	776.3	−98.0 (−3.4%)
	Downstream	247.0	244.4	240.9	231.2	221.7	214.9	216.3	−30.7 (−1.1%)
Water	Upstream	5.4	10.7	15.3	15.7	12.7	14.3	20.6	15.2 (0.5%)
	Midstream	23.8	32.5	27.6	26.7	21.2	29.7	37.7	13.9 (0.5%)
	Downstream	17.8	9.2	14.0	11.5	12.9	17.6	18.9	1.2 (0%)
Wetland	Upstream	0.6	1.2	0.1	0.1	0.1	0.1	2.1	1.5 (0.1%)
	Midstream	3.9	2.0	0.3	0.3	0.2	0.2	7.4	3.5 (0.1%)
	Downstream	3.2	0.8	0.1	0.1	0.1	0.2	1.6	−1.6 (−0.1%)
Barren land	Upstream	4.1	3.5	4.9	4.6	11.6	2.2	6.0	1.9 (0.1%)
	Midstream	38.5	46.1	19.4	23.6	24.4	39.7	22.7	−15.9 (−0.6%)
	Downstream	6.5	6.8	3.4	5.0	7.1	10.8	3.5	−3.0 (−0.1%)
Total area		2,856.4	2,856.4	2,856.4	2,856.4	2,856.4	2,856.4	2,856.4	

[a] Sublevel classes are defined by the land use and land cover classification system (at level 1) from the United States Geological Survey (USGS) [17].

[b] Values in parentheses (m) denote the spatial resolution of each digital map for different years.

[c] Total change indicates the difference in area between 1975 and 2005. Percentage change in area (%) at both positive and negative directions during this period is shown in parentheses.

experienced a steady, moderate urbanization process over the last three decades, including the period 1996–2008 that was specifically analyzed in this study. The dominant land use in the basin was forest land, followed by agricultural and urban or built-up land. Forested areas mainly decreased with increases in other land use activities, though the most significant change occurred in the midstream sector, regardless of land use classes (see total amount of land use change in the last column).

2.2. Input variables and data pre-processing

Table 2 lists the environmental variables (i.e. 11 water quality, 6 weather, and 7 land use parameters) used in the SOM and MLR analyses. Water quality parameters, except for water level data that were recorded daily by the Ministry of Land, Infrastructure, and Transport (MLIT), were measured by the Ministry of Environment (ME) in Korea on a monthly basis. For convenience, the water quality and water level variables were grouped into a single category. Daily weather data were obtained at one representative location from the Korea Meteorological Administration, whereas both the MLIT and ME provided time-series land use data sets (i.e. raster maps during 1975–2000 and vector map for 2005, respectively) at 5 year intervals.

As these variables were compiled over different time scales, appropriate data aggregation or disaggregation methods were required so that all variables were analyzed simultaneously. In this study, three data aggregation schemes (i.e. geometric and arithmetic means as well as the sum total) were applied to different water quality and weather variables to perform data analyses on an annual basis (Table 2). For instance, monthly FC data were averaged for each year using the geometric mean, whereas the arithmetic mean was used for other variables as they did not vary considerably between months or between days. Instead of using the mean value, the 1 year accumulated daily rainfall depth and sunshine duration were used; land

Table 2
List of input variables used for SOM and MLR analyses

Types	Variable name[a]	Units	Data aggregation[b]
Water quality	Fecal coliform (FC)*	CFU/100 mL	G-mean
	Water temperature (WT)	℃	A-mean
	pH	–	A-mean
	Dissolved oxygen (DO)	mg/L	A-mean
	Biochemical oxygen demand (BOD)	mg/L	A-mean
	Chemical oxygen demand (COD)	mg/L	A-mean
	Suspended solids (SS)	mg/L	A-mean
	Total nitrogen (TN)	mg/L	A-mean
	Total phosphorus (TP)	mg/L	A-mean
	Conductivity (COND)	μS/cm	A-mean
	Water level (WL)	m	A-mean
Weather	Air temperature (AT)	℃	A-mean
	Rainfall depth (RD)	mm	Sum total
	Wind speed (WS)	m/s	A-mean
	Relative humidity (RH)	%	A-mean
	Sunshine duration (SD)	hr	Sum total
	Cloud amount (CA)	%	A-mean
Land use[c]	Water*	km^2	–
	Rangeland*	km^2	–
	Wetland*	km^2	–
	Urban or built-up land (Urban)*	km^2	–
	Barren land (Barren)*	km^2	–
	Forest land (Forest)*	km^2	–
	Agricultural land (Agriculture)*	km^2	–

[a] Log transformed variables (i.e. log[x + 1]) are marked with an asterisk (*).

[b] G-mean and A-mean indicate the geometric and arithmetic means, respectively.

[c] No data aggregation is required for the land use variables as they are regularly updated every 5 years.

use variables, once updated, were fixed at 5-year periods due to their regular publication schedule.

After constructing a complete matrix of input values, variables marked with an asterisk (*) were log transformed (i.e. log[x + 1]) again to reduce the difference in their magnitudes between years or between locations. Note, however, that this type of data preprocessing is not required for the SOM (though necessary for regression analyses) because all variables are converted on a linear scale from 0 to 1 during the analysis phase (Section 2.3). Again, the main purpose of this study was to evaluate the effectiveness of alternative approaches for developing different statistical models rather than an accurate analysis and interpretation of the whole data set. Thus, the aggregation and transformation processes of data should be undertaken carefully in future studies. As a further consideration, we simply used the water level parameter rather than the river discharge because the stage-discharge relationship at each site did not often capture the dynamic flow behavior in extreme conditions, i.e. low and high discharges. In total, 13 years of annual data from 1996 to 2008 were used in this study.

2.3. Data analysis tools

ArcSWAT (version 2012) and ArcGIS (version 10.0) software was used for the spatial discretization of subbasin units (i.e. attribute tables consisting of 7 land use variables and 19 subbasins; Section 2.1), and the visualization of the summarized output in the digital map, respectively [18,19]. To delineate the basin at the subbasin level, a digital elevation map (at 30 m resolution) was introduced to the ArcSWAT, along with the river network, followed by the superposition of the land use (and land cover) maps (at 30 or 60 m resolutions) within the basin boundary. Initially, the land use maps published over different years had various land use classes and subclasses, but they were reorganized based on the classification system suggested by the USGS (Table 1) [17]. The land use attributes derived from this discretization process were finally combined with water quality and weather records for subsequent data analyses. After the main SOM analysis, the distribution of FC values in the reduced clusters was displayed along the river network in the basin boundary through the ArcGIS software.

We applied the SOM (version 2.0) to provide different types of important variables for use in statistical regression models as well as to explore spatial and temporal patterns in the data set [13]. Since the initial release of its algorithm in 1981, the SOM has been widely applied to environmental research due to its many benefits (e.g. noise tolerance, data abstraction, and visualization) over conventional methods [7–14]. Topology preservation and vector quantization are two leading features in this tool, which enable the data input in multidimensional domains to be visualized in a reduced data space [11–13]. After completing the desired data aggregation, the raw data set was then normalized to between zero and one, followed by a map training phase using linear initialization and batch training algorithms [13]. Next, the trained data set was visualized in component planes that exhibited the distribution of the values for each variable in two-dimensional map units, along with a unified distance matrix (U-matrix) that presented the Euclidean distance in the map unit itself and between adjacent map units [13]. The theoretical background and environmental applications are well documented in literature [7–14]; the indices associated with variable selection provided by the SOM are described in detail in Section 2.4.

The MLR was performed using popular statistical software (SPSS version 15.0) to construct statistical models for predicting FC concentrations in a river [15]. The regression model illustrates the relationship between the dependent and independent variables, providing a measure to assess the prediction performance estimated from two or more predictor variables. For this study, a regression using the backward method (hereinafter referred to as MLR-Backward) was selected as the reference model, assuming that all variables were equally important in the model. However, only variables having a *p*-value less than 0.05 were retained as predictors in the model. Different regression models were also constructed based on the four indices recommended by the SOM (Section 2.4), which were compared with MLR-Backward. When developing various regression models, the data set was divided into two subdata sets, the test (70% of entire data set) and validation data (30%) sets, using a random splitting method provided by the program [15].

2.4. Various indices for variable selection

Previous studies suggested several measures to identify significant variables that had notable effects on the SOM structure: structuring index (SI [9]), relative importance (RI [7,13]), and cluster description (CD [12]). Park et al. [9] introduced SI to detect important variables as the number of samples in the data set decreased. If a certain variable shows a low SI value in the reduced data set, it is assumed to have a small influence on the map organization. In other words, its removal does not cause any significant information loss from the original data set. The SI value of variable *i* can be computed as follows:

$$\mathrm{SI}_i = \sum_{j=1}^{S} \sum_{k=1}^{j-1} \frac{\left| w_{ij} - w_{ik} \right|}{\left\| r_j - r_k \right\|} \tag{1}$$

where the denominator and numerator indicate the weight and topological differences between map units *j* and *k* in the total number of map units *S*, respectively.

In the tool itself, another index RI was suggested that showed the relative importance of variables in arranging map units [7,13]. To estimate RI values for different variables, the distance matrix was initially computed with respect to the map structure, which then adjusted the size of the pie charts located in each map unit. The borders between different clusters were mainly determined by the size of the pie chart (if designed to become bigger), where the composition of important variables was higher than for the others. In this study, we selected the five largest pie charts, normalized the contributions of individual variables in each pie chart (to obtain a total of 100%) regardless of its size, and finally averaged their relative contributions in the selected pie charts [7]. Accordingly, the average composition of a significant variable *i* should be high.

Vesanto [12] provided an index for CD that described the internal properties of clusters using a statistical value of the data set. Unlike other indices, CD assesses the variation of a variable in each cluster, such that:

$$CC_i = \sum_{l=1}^{C} S_{li}^{D} = \sum_{l=1}^{C} \frac{(C-1)S_{li}^{C}}{\sum_{m=1,\, m\neq l}^{C} S_{mi}^{C}} \tag{2}$$

$$\text{where } S_{li}^{C} = \frac{\sigma_{li}}{\sigma_i} \tag{3}$$

where σ_{li} and σ_i represent the standard deviations of variable i in cluster l and the entire data set, respectively. C indicates the total number of clusters partitioned in the map. The variable is considered important if the estimated index value CD is high.

Finally, a simple and intuitive approach is to use a correlation analysis that investigates the relationship between the dependent and independent variables [15]. Note that although this method is not recommended in the SOM tool, it provides a rough idea of how variables interact with each other. We used the Spearman's rank correlation (SRC) to rank variables in descending order based on the absolute value of the correlation coefficient, as most variables did not clearly follow a normal distribution even after the data transformation process. Note also that the correlation coefficient between variables in the trained data set obtained from the SOM analysis is expected to be slightly higher than from the original data set, due to the removal of data noise and outliers.

3. Results and discussion

3.1. Relationship between FC and explanatory variables

Identifying the relationship among variables is an essential step in understanding the structure of the final (regression) models. Fig. 2 shows the U-matrix and 24 component planes that illustrate the cluster structure of the trained data set and the relationship between FC and explanatory variables, respectively. In the figure, the color bar displays the range of values of the 24 variables (for component planes) as well as the topological distance between adjacent map units (for the U-matrix). Note that the average quantization and topographic errors attain at a minimum (0.68 and 0.07, respectively), indicating that the original data set is reduced successfully to the trained data set after removing noise and/or outliers from the SOM. In the U-matrix, large topological distances were observed in several locations (i.e. map units), around which the trained data set could be divided into a small number of clusters (see red, orange, and yellow dots). In fact, the trained data set was initially partitioned into five different groups (i.e. clusters 1a1, 1a2, 1b1, 1b2, and 2) based on the minimum Davies–Bouldin (DB) index (i.e. 0.81 at 5 clusters, figure not shown) which confirmed the validity of these groupings. Among them, clusters 2 and 1b1 represented the highest and the lowest contamination levels with respect to FC, respectively. However, no clear patterns were observed as the variables showed a wide range of distributions in individual clusters. From the component planes, it was found that most water quality variables, except for a few variables such as DO and WL, showed a good correlation with FC in both positive and negative directions. However, a weak correlation was generally observed between FC and either climate or land use variables, implying that these variables are less helpful in elucidating the variation of FC concentration in the river. Only AT and RD (among climate variables) and Urban and Barren (out of land use variables) displayed a low correlation with FC. From these results, it can be concluded that although the SOM well provides the nonlinear relationship between variables, there are some inherent drawbacks to the use of this analysis: (1) in describing the distinct characteristics of the partitioned clusters, and (2) in directly selecting important variables for the FC prediction.

3.2. Cluster characterization

To better illustrate the characteristics of variables in the clusters, we regrouped a series of clusters from five to three groups (namely, clusters 1a, 1b, and 2). The intent of this regrouping of small clusters was to provide a convincing visualization in the map while increasing the discriminating properties of the clusters. Fig. 3 illustrates the spatial and temporal distributions of FC concentrations for 1997, 2002, and 2007, showing the benefit of the reduced clusters in the visualization along with data reduction (derived from the SOM). In the figure, the color of circles indicates different concentration levels of FC in the river, whereas their size represents the reduced clusters, i.e. cluster 1b (low-pollution group), cluster 1a (intermediate pollution group), and cluster 2 (high-pollution group). Note that temporal changes in FC for three representative years are only shown as examples, as their patterns for 1997, 2002, and 2007 are quite similar to those in 1996–1999, 2000–2004, and 2005–2008, respectively.

From the figure, the intermediate pollution group was shown to be dominant in 1997 (or between 1996 and 1999). The water quality then deteriorated greatly at some monitoring locations for the subsequent five years (2000–2004), while others did not. The water quality has improved again in recent years (2005–2008), displaying relatively good water quality conditions that were less than 250 CFU/100 mL along the river. However, FC was routinely found in Uchi (UC, upstream of Gwangju City) and the Gwangju Tributary monitoring stations regardless of the temporal changes, which was consistent with our previous studies [2,16]. The results also showed that the monitoring stations in the Hwangryong and Hampyeong Tributaries were considered as potential hotspots since they

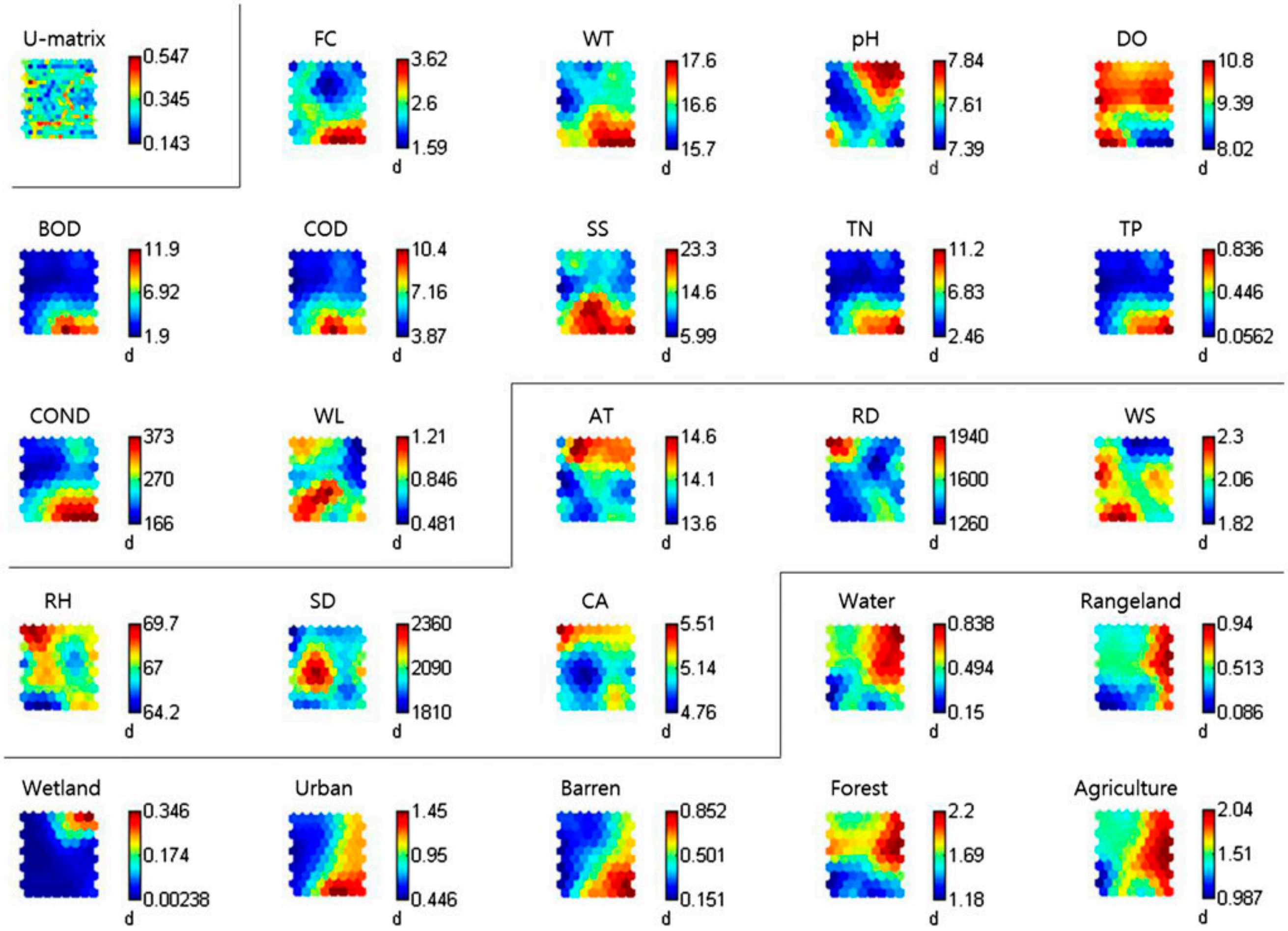

Fig. 2. Net output of a SOM analysis: a unified distance matrix (U-matrix) and component planes for 24 variables. The variables under different categories are divided by solid (black) lines. Changes in individual variables are indicated by the (jet) color bar.

were partially responsible for water quality degradation during 2000–2004.

To describe the overall variation of variables in individual clusters in further detail, we employed a new indicator, referred to as the concentration index [10]. The concentration index is analogous to the CD index, except that it uses the average values of the group and whole data rather than the standard deviation, and does not estimate the sum (of the discriminating property) of individual groups (Eqs. (2) and (3)). Similar to the CD, the concentration index of a variable can be estimated using the ratio of the mean of the cluster to that of the whole data, which is then converted to a percent after multiplying it by 100. The distribution of the concentration index for 24 variables is illustrated in Fig. 4, where individual clusters (i.e. clusters 2, 1a, and 1b) represent (a) high-, (b) medium-, and (c) low-pollution groups, respectively. In the high-pollution group for FC (i.e. cluster 2), some water qualities (e.g. BOD, COD, SS, TN, TP, and COND) and land use variables (e.g. Urban and Barren) achieved high concentration index values (%). This finding was very similar to the relationship shown using the component planes (Section 3.1), indicating that these variables were mainly responsible for the increased levels of FC. However, no distinct relationship between FC and other variables was observed in the intermediate pollution group (i.e. cluster 1a). In the low-pollution group (i.e. cluster 1b), a few land use variables such as wetland, rangeland, and water contributed, to some extent, to decreasing FC concentrations. These results confirm the complexity of constructing a single regression model from the entire data set, illustrating why the models developed from more homogeneous groups best account for the amount of variation of FC in the river.

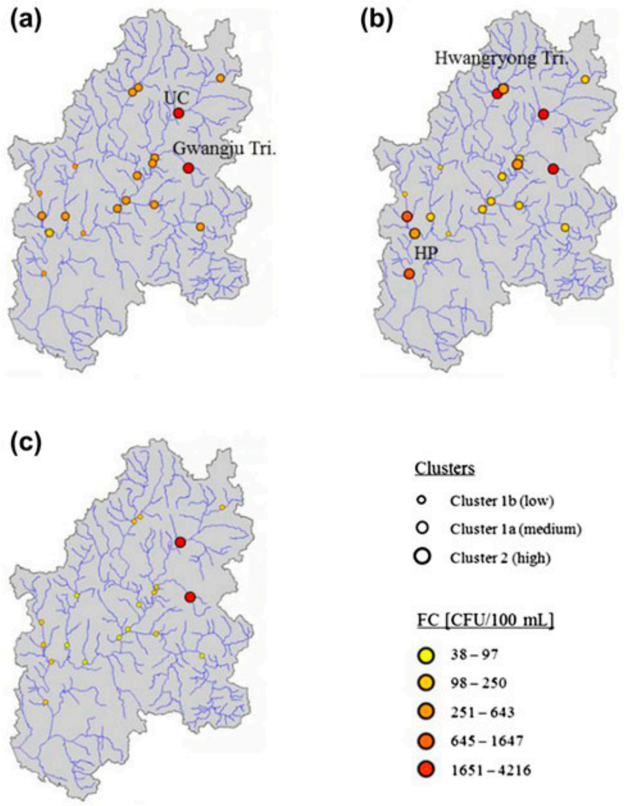

Fig. 3. Spatial and temporal distribution of FC concentrations in the Yeongsan River in Korea for (a) 1997, (b) 2002, and (c) 2007. The sizes and colors of the circles reveal three main clusters and five different levels of FC concentrations, respectively.

3.3. Multiple regression models

Previous studies have proposed several indices (i.e. SI, RI, and CD) that could be employed to identify significant variables from the data set in the SOM, in addition to the widely used SRC index. Table 3 presents a list of significant variables and their corresponding values, which are estimated based on these indices (Section 2.4). When we considered the top 10 variables only, each index provided various types of important variables due to the differences in the implemented algorithms. Of these, land use variables were considered as the most important parameters in SI, whereas RI selected weather variables—which had a significant effect on the map organization. Also, the dominant parameters in CD consisted of a mixture of water quality and climate variables; SRC was mainly characterized by water quality variables. From these results, there seems to be an apparent discrepancy between these indices in determining significant variables, as their performance substantially varied depending on data sets.

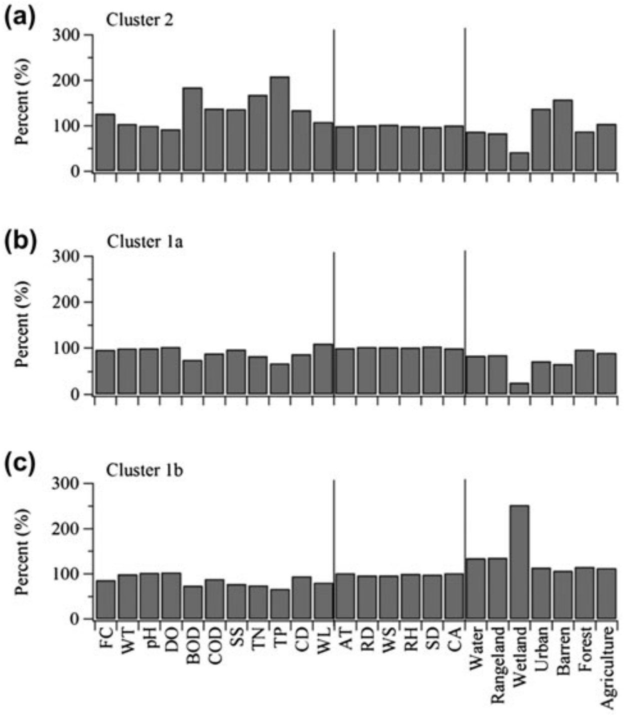

Fig. 4. Overall variation of 24 variables in three different clusters obtained using the concentration index: (a) cluster 2 (high-pollution group), (b) cluster 1a (intermediate pollution group), and (c) cluster 1b (low-pollution group).

Five different regression models were constructed to determine the best model for predicting FC concentrations in a river. Four different models were developed using the forced entry of the top 10 variables from the indices recommended above. MLR-Backward was used as the reference model to allow for a comparison of the prediction performance between them. Table 4 presents the overall performance summary of the regression models evaluated in both the test and validation data sets (divided into a ratio for 7:3). In the table, R indicates the correlation coefficient between the observed and predicted values of FC, whereas R^2 denotes the total amount of variation of FC that is explained by the different models. In general, the R value approaches 1 as the predicted data matches well with the observed data, and the amount of variance accounted for (i.e. R^2) increases as more variables are included in the models. From the table, it was found that MLR-Backward (66.1%) best explained the variation of FC in the test data set, followed by MLR-SRC (44.8%), MLR-RI (41.7%), MLR-SI, (41.4%), and MLR-CD (39.7%). However, the best prediction performance in MLR-Backward could partially be attributed to the small sample size, as 16.2% of the water level data that was included in the model was,

Table 3
List of significant variables that are ranked based on four indices (SI—structuring index, RI—relative importance, CD—cluster description, and SRC—Spearman's rank correlation) in a SOM

Rank	SI[a]	Values	RI	Values	CD	Values	SRC[b]	Values
1	Barren	337.28	RH	10.35	Wetland	8.29	TN	**0.68**[c]
2	Urban	333.43	SD	9.72	BOD	3.83	BOD	**0.65**
3	WS	328.12	WS	7.82	SS	3.76	TP	**0.63**
4	SD	308.44	Barren	7.66	SD	3.75	SS	**0.62**
5	CA	290.62	AT	6.28	DO	3.54	COD	**0.61**
6	Rangeland	287.03	Agriculture	5.58	COD	3.37	pH	**−0.59**
7	Forest	284.82	CA	5.45	TP	3.34	AT	**−0.56**
8	RH	280.64	Water	5.44	CA	3.30	COND	**0.56**
9	Agriculture	274.19	RD	5.35	RD	3.23	Water	**−0.54**
10	Water	273.73	Urban	5.25	AT	3.22	DO	**−0.54**
11	RD	263.41	Forest	3.56	TN	3.19	Forest	**−0.51**
12	AT	254.68	Rangeland	3.18	WT	3.18	WT	**0.49**
13	Wetland	240.21	FC	3.17	Rangeland	3.12	WS	**0.47**
14	TN	224.90	TN	3.10	FC	3.12	Wetland	**−0.47**
15	FC	196.40	SS	3.05	WS	3.10	Barren	**0.45**
16	COND	182.00	TP	2.21	Urban	3.09	Urban	**0.31**
17	TP	164.80	WT	2.19	COND	3.08	WL	**0.30**
18	SS	162.98	DO	2.14	RH	3.08	CA	−0.23
19	WT	157.16	COND	1.86	Barren	3.07	Rangeland	−0.20
20	pH	131.73	pH	1.82	pH	3.05	RD	0.13
21	WL	119.25	COD	1.70	Forest	3.05	Agriculture	−0.05
22	BOD	105.01	WL	1.49	WL	3.04	SD	0.04
23	COD	96.73	BOD	1.41	Water	3.03	RH	−0.02
24	DO	89.38	Wetland	0.22	Agriculture	3.01	–	–[c]

[a] For explanation of variable abbreviations, refer to Table 2.

[b] SRC omits one variable in the list since the correlation between FC and itself is always 1.

[c] Bold letter indicates *p*-value is less than 0.05.

Table 4
Evaluation of the predictive ability of different regression models for the test and validation data sets

Model[a]	R		R^2	Adjusted R^2	Durbin–Watson	
	Test	Validation			Test	Validation
MLR-Backward	0.813	0.645	0.661	0.633	1.933	1.740
MLR-SI	0.643	0.591	0.414	0.379	1.471	1.810
MLR-RI	0.645	0.613	0.417	0.382	1.337	1.832
MLR-CD	0.630	0.473	0.397	0.361	1.669	2.037
MLR-SRC	0.669	0.491	0.448	0.414	1.861	2.093

[a] Abbreviations: MLR = multiple linear regression, SI = structuring index, RI = relative importance, CD = cluster description, and SRC = Spearman's rank correlation.

in fact, lost. In the validation data set, MLR-RI showed a remarkable improvement of the prediction ability ($R = 0.613$ and $R^2 = 0.376$); as compared to the other models, its performance was found to be quite comparable with MLR-Backward ($R = 0.645$ and $R^2 = 0.416$). The Durbin–Watson values fell between 1 and 3 in the test and validation data sets, indicating that no serial correlation existed in the constructed models. Given all of the above, RI appears to provide the best variables for predicting the FC concentration in a river among the tested indices, although its performance is slightly lower than MLR-Backward.

Table 5 shows important variables that should be retained in the regression models for the test data set.

Table 5
Important variables from five regression models for the test data set

Model[a]	Variable[b]	Unstandardized coefficient		Standardized coefficient
		B	SE[c]	β
MLR-Backward	(Constant)	−6.65	3.12	
	CA	−3.88	0.96	−1.14*[d]
	SD	−0.01	0.00	−1.08*
	RH	0.33	0.06	0.76*
	TN	0.16	0.03	0.51*
	WS	2.18	0.33	0.45*
	COND	0.00	0.00	0.38*
	AT	0.82	0.19	0.37*
	COD	−0.16	0.05	−0.31*
	WL	−0.41	0.13	−0.21*
	Urban	0.60	0.26	0.21*
	Agriculture	−0.56	0.20	−0.21*
MLR-SI	(Constant)	−1.52	3.55	
	Urban	1.33	0.27	0.48*
	CA	−1.59	0.78	−0.45*
	RH	0.18	0.06	0.39*
	WS	1.83	0.38	0.36*
	Forest	−0.40	0.19	−0.18*
MLR-RI	(Constant)	−3.77	3.60	
	CA	−2.77	1.03	−0.79*
	SD	0.00	0.00	−0.75*
	RH	0.27	0.09	0.59*
	Urban	1.28	0.27	0.47*
	WS	1.85	0.38	0.36*
	Agriculture	−0.55	0.26	−0.22*
	AT	0.37	0.17	0.18*
	Water	−0.57	0.28	−0.17*
MLR-CD	(Constant)	5.06	3.66	
	TP	1.61	0.41	0.48*
	CA	−0.78	0.50	−0.22*
	Wetland	−1.56	0.58	−0.19*
	SD	0.00	0.00	−0.19*
MLR-SRC	(Constant)	3.13	2.20	
	BOD	0.12	0.05	0.45*
	COD	−0.10	0.05	−0.29*
	COND	0.00	0.00	0.28*
	pH	−0.66	0.23	−0.20*
	Water	−0.62	0.21	−0.18*
	DO	0.14	0.06	0.18*

[a] Abbreviations: MLR = multiple linear regression, SI = structuring index, RI = relative importance, CD = cluster description, and SRC = Spearman’s rank correlation.

[b] For abbreviation and transformation of variables, refer to Table 2.

[c] SE denotes the standard error of the unstandardized coefficient B.

[d] Asterisk (*) indicates p-value is less than 0.05.

In the table, the unstandardized coefficient B explains the degree of influence of a variable on the predicted values when other variables remain constant. In contrast, the standardized coefficient β represents the relative importance of each variable in the model. From the table, though MLR-Backward included as many variables as possible in the model, only a small number of variables (among the 10 variables entered intentionally from the 4 indices) were retained in the remaining models. Interestingly, both MLR-Backward and MLR-RI selected the same climate variables (i.e. CA, SD, RH, WS, and AT) as important predictors, although the number and types of variables varied across the models. In addition, their relative importance was found to be similar between the two models when we compared the ranking of those variables in order of significance. In other words, not only were the variables CA and SD identified as the most significant variables in MLR-Backward and MLR-RI, but they were almost equally important in each model (i.e. see β coefficient values). However, the variance inflation factor that assessed the degree of multicollinearity increased significantly when these variables were simultaneously involved in any of the models under consideration (i.e. MLR-Backward, MLR-RI, and MLR-CD). Therefore, one of the two should be carefully removed in further regression models. The standardized coefficients of all important variables considerably decreased in the remaining models, indicating that these variables played a minor role in the FC prediction.

4. Conclusions

In this present study, we describe a new method for constructing reliable regression models using important variables recommended by four different indices (SI, RI, CD, and SRC) from the SOM. Overall, 13 years of water quality, climate, and land use data were compiled from various monitoring locations along the Yeongsan River in Korea, which were used as inputs for the SOM and MLR analyses. The performance of the reference model MLR was compared to the models developed from the four indices in the SOM. The main conclusions of this study are as follows.

(1) The nonlinear relationship between FC and other variables was fairly well addressed by the SOM. However, the tool itself did not directly describe the properties of individual clusters and influence of variables on the FC prediction, unless specifically modified to do so.

(2) Reducing the number of clusters obtained from the SOM by default increased the discriminating ability of the partitioned clusters, from which various water pollution hotspots could then be readily identified. The concentration index effectively described variable patterns in the clusters, implying the necessity of regression models at the cluster level (rather than the entire data set) to increase its performance.

(3) The contributions of variables in the four indices were assessed differently due to the various algorithms implemented in the SOM. On average, a model combining MLR and RI (i.e. MLR-RI) displayed a good prediction performance compared to the original model (i.e. MLR-Backward). Although both models included the same climate variables as predictors, however, there was a need to carefully remove highly correlated variables to avoid the multicollinearity issue.

It should be noted that the regression analysis results described above are provisional and may be subject to change in rivers having different environmental settings. Therefore, further research is required to verify the effectiveness of these indices for the development of a more robust model using updated or new data sets. As an excellent example of the proposed methodology, statistical models including real-time data provide timely information to control mortality or disease rates of fishes in high-density aquafarms via water quality.

Acknowledgments

This research was a part of the project entitled “Development of an Integrated Control System of Eel Farms based on Short-Range Wireless Communications (grant number 20130283),” funded by the Ministry of Oceans and Fisheries, Korea. The authors would also like to thank the Ministry of Land, Infrastructure, and Transport, the Ministry of Environment, and the Meteorological Administration in Korea for providing environmental monitoring and geospatial data.

References

[1] US Environmental Protection Agency (EPA), Handbook for Developing Watershed Plans to Restore and Protect Our Waters, US EPA, Office of Water, Washington, DC, Report No. 841-B-08-002, 2008.

[2] 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(11) (2009) 3482–3493.

[3] A.B. Boehm, S.B. Grant, J.H. Kim, S.L. Mowbray, C.D. McGee, C.D. Clark, D.M. Foley, D.E. Wellman, Decadal and shorter period variability of surf zone water quality at Huntington Beach, California, Environ. Sci. Technol. 36(18) (2002) 3885–3892.

[4] R.L. Reeves, S.B. Grant, R.D. Mrse, C.M.C. Oancea, B.F. Sanders, A.B. Boehm, Scaling and management of fecal indicator bacteria in runoff from a coastal urban watershed in Southern California, Environ. Sci. Technol. 38(9) (2004) 2637–2648.

[5] D.S. Francy, R.A. Darner, E.E. Bertke, Models for predicting recreational water quality at Lake Erie beaches, US Geological Survey, Virginia, Scientific Investigations, Report 2006–5192 (2006) 13.

[6] R.G. Zepp, M. Cyterski, R. Parmar, K. Wolfe, E.M. White, M. Molina, Predictive modeling at beaches, Volume II: Predictive tools for beach notification, US Environmental Protection Agency, National Exposure Research Laboratory, Report No. 600-R-10-176, 2010.

[7] 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(14) (2011) 4183–4197.

[8] 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(3) (2008) 732–742.

[9] 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(3) (2006) 247–257.

[10] S. Tsakovski, B. Kudlak, V. Simeonov, L. Wolska, J. Namiesnik, Ecotoxicity and chemical sediment data classification by the use of self-organising maps, Anal. Chim. Acta 631(2) (2009) 142–152.

[11] T. Kohonen, Self-organizing Maps, third ed., Springer-Verlag, Berlin Heidelberg, New York, Springer Series in Information Sciences, 30, (2001), 502.

[12] J. Vesanto, Data exploration process based on the self-organizing map, Finnish Academies of Technology, Espoo, Acta Polytechnica Scandinavica, Mathematics and Computing Series No. 115, (2002).

[13] J. Vesanto, J. Himberg, E. Alhoniemi, J. Parhankangas, SOM toolbox for Matlab 5, Helsinki University of Technology, SOM toolbox team, Espoo Report, A57, (2000).

[14] 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 (19) (2007) 4566–4578.

[15] A. Field, Discovering Statistics Using SPSS, second ed., Sage Publications Inc., California, 2005, p. 779.

[16] 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 (1–2) (2007) 367–374.

[17] R. Anderson, E.E. Hardy, J.T. Roach, R.E. Witmer, A land use and land cover classification system for use with remote sensor data, US Geological Survey (GS), Washington, DC, GS Professional Paper 964, 1976.

[18] S.J. Ki, T. Sugimura, A.S. Kim, OpenMP-accelerated SWAT simulation using Intel C and FORTRAN compilers: Development and benchmark, Comput. Geosci. 75 (2015) 66–72.

[19] S.J. Ki, C. Ray, M.M. Hantush, Applying a statewide geospatial leaching tool for assessing soil vulnerability ratings for agrochemicals across the contiguous United States, Water Res. 77 (2015) 107–118.