Desalination and Water Treatment
www.deswater.com

doi: 10.5004/dwt.2020.25481

182 (2020) 127–134
April

Characterizing water quality and quantity profiles with poor quality data in a machine learning algorithm

Zhonghyun Kim[a], Heewon Jeong[b], Sora Shin[b], Jinho Jung[a], Joon Ha Kim[b], Seo Jin Ki[c,*]

[a] *Division of Environmental Science and Ecological Engineering, Korea University, Seoul 02841, Republic of Korea*
[b] *School of Earth Sciences and Environmental Engineering (SESE), Gwangju Institute of Science and Technology (GIST), Gwangju 61005, Republic of Korea*
[c] *Department of Environmental Engineering, Gyeongnam National University of Science and Technology, Jinju 52725, Republic of Korea, email: seojinki@gntech.ac.kr*

Received 10 September 2019; Accepted 3 January 2020

ABSTRACT

Statistical analyses are often subject to misinterpretation due to poor data quality which is inaccurate, incomplete, or unavailable. This study describes how incomplete data diminishes the screening accuracy of water pollution hotspots using a self-organizing map (SOM), a popular algorithm in reducing the dimension of complex data in a nonlinear fashion. A full data set consisting of 12 water quality and quantity parameters monitored monthly over 3 years at the Yeongsan River in Korea was provided to SOM as a reference input. For purposes of comparison, SOM was further allowed to accept three incomplete data sets in terms of variable availability as well as data loss for single and multiple parameters and different pollution levels. We found that data loss of either single or multiple parameters exceeding 15% of the entire data set led to significant changes in spatial and temporal patterns of the original data. However, the variables intentionally unavailable in the given data set affected the screening performance of water pollution hotspots in SOM, to a less obvious extent, as long as the percentage of missing data fell below 10%. The same applied to data loss with three pollution levels, from high through moderate to low concentrations of one important variable. Therefore, we recommend the use of multiple approaches that couple dimensionality reduction algorithms with reasonable imputation methods for the data set with a high percentage (e.g. above 15%) of missing values.

Keywords: Non-linear data analysis; Dimensionality reduction; Water quality data; Pollution hotspots; Incomplete data; Self-organizing map

1. Introduction

Statistical analyses not only contribute to facilitating strategic planning but also improve informed decisions for sustainable watershed management. Statistical analyses were applied to a wide range of watershed management issues including, but not limited to, characterization of spatial and temporal data patterns, source apportionment, design of sampling locations, and impact assessment of complex relationships among variables [1–5]. For example, the previous study of Aguilera et al. [1] adopted a dynamic factor analysis to identify riverine nutrient patterns and associated environmental drivers. Principal component analysis (PCA) and its variant were employed in the study of Yang et al. [2] to allocate pollution sources across different zones such as urban, suburban, and rural watersheds. The modified Sanders approach was involved in the development of surface water quality monitoring network [3]. Structural equation modeling, a combination of confirmatory factor analysis and path analysis, was also used to address the direct and

* Corresponding author.

indirect effects of wastewater effluent on macroinvertebrate communities [4,5].

The performance of statistical analyses was, however, sensitive to data quality such as missing values, noise, outliers, censored and truncated data, and so on [6–9]. Alameddine et al. [6] introduced three robust outlier detection methods to detect and isolate outliers and found their effectiveness in analyzing multidimensional water quality data over conventional methods such as Mahalanobis distance. Imputations of missing values were examined in the study of Betrie et al. [7], which demonstrated that two imputation methods such as the iterative robust model-based and sequential imputation approaches were appropriate to recover missing water quality data at mine sites. Another study of He [8] studied the data set with censored observations and multiple detection limits and applied an expectation-maximization algorithm to both simulated and observed water quality data for statistical justification. Another study by Ruždjak et al. [9] found that the standard PCA algorithm was only tolerant of less than 4% of incomplete data such as missing values and outliers.

A self-organizing map (SOM) was capable of accomplishing the intended functionality, which discovered complex data characteristics, under limitations on data quality described above [10–18]. The SOM algorithm, in particular, provided outstanding capabilities in removing the number of data records with similar properties as well as in visualizing the reduced records in a low-dimensional map [10–12]. Due to all those distinct merits, the SOM studies had a wide span from surface and subsurface water quality [10–16] through hydrology [17] to ecology [11,18]. As opposed to those previous studies showing diverse application opportunities, the originality of this research lies in evaluating the efficiency of the SOM algorithm in response to intentional data loss in the data sets. Using the data sets with and without data loss, we specifically addressed the effects of (1) missing records occurred in single and multiple variables, (2) missing variables themselves, and (3) missing records made at different pollution levels in a fixed variable on the performance of the algorithm. We hope that the results derived from this study help promote correct use and interpretation of SOM as well as extend its applications in areas of poor data quality.

2. Materials and methods

2.1. Monitoring network

For this study, we used only a small part of the published data which were available in the tributary monitoring program for a target watershed [19] as well as compiled from our previous study [20]. Fig. 1 presents a basin-wide monitoring network designed to measure water quality and quantity in the Yeongsan (YS) River, one of the four major rivers in South Korea. The monitoring network consisted of a total of 83 sampling sites: 2 are on the mainstream as well as 81 are on the tributaries. These observation stations were, in particular, selected from all available monitoring programs maintained by relevant authorizing agencies based on a

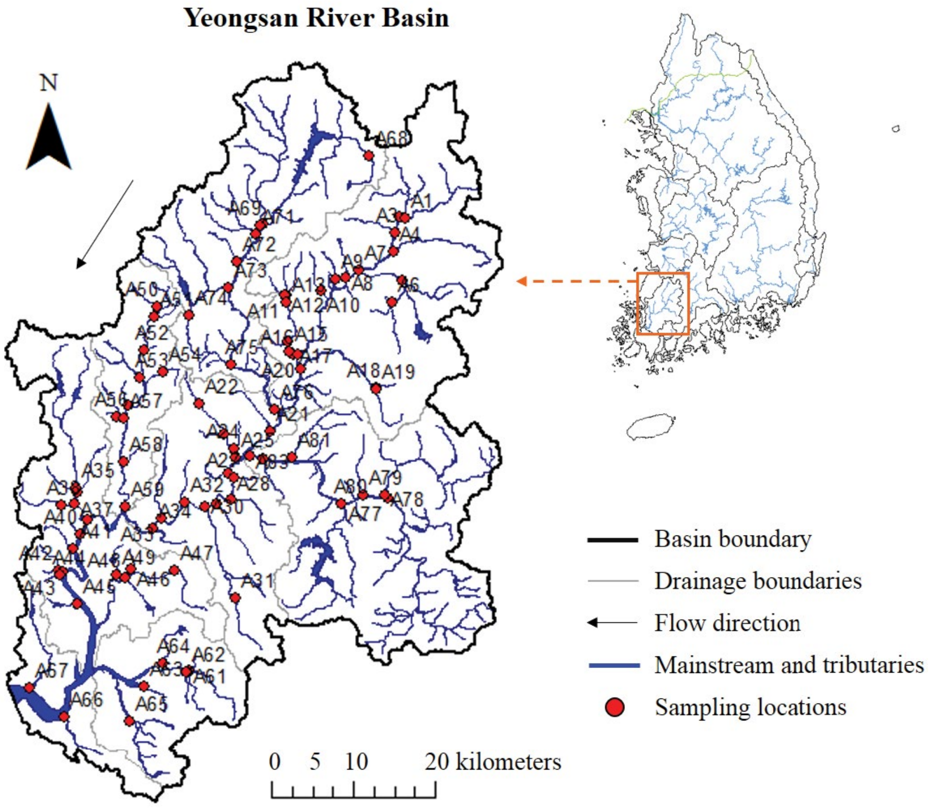

Fig. 1. Selected sampling locations for discharge and water quality at the mainstream and tributaries in the Yeongsan River Basin, Korea.

number of factors such as the size of the entire monitoring network, stream size, spatial representativeness, available budget, etc. A complete list of 83 stations, including their physical characteristics (i.e., drainage area and main channel length), is provided in Table 1. As shown in the table, individual observation stations varied greatly in drainage area and main channel length, regardless of their locations in the mainstream or tributaries. The established monitoring

Table 1
Monitoring stations for measurement of water quality and discharge in the mainstream and major tributaries of the Yeongsan River, Korea

Sections	Site ID[a]	Drainage area (km^2)	Channel length (km)	Sections	Site ID[a]	Drainage area (km^2)	Channel length (km)
	A1§	82.84	24.98		A43	5.05	5.08
	A2	47.29	14.13		A44	5.25	4.07
	A3	15.53	8.97		A45	9.01	5.6
	A4	59.2	26		**A46**	5.03	4.85
	A5	45.39	15.49		A47	137.51	29.4
	A6	32.21	16.7		**A48**	12.4	8.7
	A7	149.34	23.56		**A49**	9.08	7.85
	A8	15.39	12.91		**A50**	16.28	7.92
	A9	6.76	6.35		A51§	76.64	14.6
Upstream areas in the mainstream	A10	9.94	10.16		**A52**	7.08	3.68
	A11	11.97	8.23		**A53**	20.1	9.64
	A12	12.31	5.07	Tributary areas in the Gomakwon-cheon	**A54**	15.48	12.04
	A13§	68.93	16.11		**A55**	8.02	6.92
	A14	5.97	4.75		**A56**	15.01	9.9
	A15	N/A[b]	N/A		**A57**	5.75	4
	A16†	82.84	24.98		**A58**	11.45	6.73
	A17‡	68.93	16.11		A59‡	76.64	14.6
	A18	11.45	50.89		A60§	264.08	25.26
	A19§	25.05	11.27		**A61**	5.74	8.53
	A20‡	25.05	11.27	Tributary areas in the Yeongam-cheon	**A62**	9.45	8.45
	A21	33.76	13.31		**A63**	10.14	9
	A22§	14.37	8		**A64**	27.68	9.94
	A23	7.07	4.49		**A65**	91.47	19.48
	A24	8.84	6.3	Yeongsan Reservoir	A66	40.27	19
	A25‡	14.37	8		A67	37.79	23.1
	A26	6.92	7.83		**A68**	76.96	19.02
Midstream areas in the mainstream	A27	2.76	3.88		**A69**	116.03	26.15
	A28	41.36	19.15		A70§	22.23	8.93
	A29	35.34	12.4	Tributary areas in the Hwangnyong-gang	**A71**	37.9	9.96
	A30	102.53	17.51		**A72**	17.08	10.83
	A31	53.45	14.2		**A73**	8.11	4.48
	A32	10.25	7.57		**A74**†	116.48	30.56
	A33	40.78	24.32		**A75**‡	116.48	30.56
	A34	19.48	6.45		A76‡	22.23	8.93
	A35§	85.89	14.89		A77§	141.36	19
	A36	25.89	11.73		**A78**	126.24	28.64
	A37	18.47	8.74		**A79**	22.26	12.62
Downstream areas in the mainstream	**A38**	27.28	10.6	Tributary areas in the Jiseok-cheon	**A80**	122.11	31.22
	A39‡	85.89	14.89		**A81**	34.02	21.3
	A40	3.51	4.11		**A82**	22.58	9.4
	A41	8.03	6.42		A83‡	141.36	19
	A42	16.46	7.72				

[a] Underlined, bold, and italic texts indicate the monitoring stations in the mainstream, secondary tributaries, and drainage channel, respectively. Also, shown as symbols §, †, and ‡ in superscript are the upstream, midstream, and downstream sites along each section, respectively.
[b] N/A represents values that are not available from the source [19].

network appeared to comprehensively cover most of the geographic areas of the YS River, except for a few large dams located in the upstream and midstream areas (see Fig. 1).

2.2. Experimental analyses

Monthly water sampling was carried out at all monitoring sites from January 2013 to December 2015 [19,20]. The water quality parameters measured directly in the field using multiple probes included water temperature (Temp), dissolved oxygen (DO), pH, and electrical conductivity (EC). Laboratory analyses of collected water samples were done for biochemical oxygen demand (BOD), chemical oxygen demand (COD), total organic carbon (TOC), total nitrogen (TN), total phosphorus (TP), suspended solids (SS), and chlorophyll-a (Chl). Note that water samples, kept at 4°C on in a cooling box as soon as after taken, are delivered to the laboratory, in which water quality testing is conducted based on official test methods of water pollution enacted by the Korean Ministry of Environment. The river discharge (Discharge) in the mainstream and tributary channels was derived directly from the velocity-area method (which was estimated from the measured velocity in discrete segments and verticals across the channel). However, the water balance approach (which was computed from precipitation and evapotranspiration, in addition to other fluxes of water) was adopted in cases direct measurements of the stream velocity along the cross-sectional area were extremely difficult or impossible. Table 2 shows descriptive statistics of 12 monitoring parameters in the mainstream and tributaries observed during the 3 years. Among monitored parameters, the variables Discharge (CV = 2.38), SS (CV = 1.69), and Chl (CV = 1.56) was found to be most changeable across the YS river and the parameters pH (CV = 0.06), DO (CV = 0.30), and Temp (CV = 0.53) was the least changeable.

2.3. Input data sets

The input data set submitted to SOM was initially constructed from all observations on the mainstream and tributaries (n = 2,419, see Table 2), which was used as a reference input. Three additional data sets were also prepared by modifying the reference data set. For instance, a specific amount of data records (i.e., rows) at single and multiple variables (i.e., columns) corresponding to 5% to 30% of the entire data set was replaced with null values, hereinafter referred to as group 1. The second extra data set was developed by intentionally removing some variables from the reference data set, hereinafter referred to as group 2. Similar to group 1, 10% of data loss occurred in the data set composed of a single variable based on its pollution levels such as low, moderate, and high pollutant concentrations, hereinafter referred to as group 3. A comparison was then made between the returned results of the reference and three modified data sets using SOM.

2.4. Self-organizing map

We adopted a popular algorithm SOM, which provided new insights into complex data sets, to examine its ability to detect water pollution hotspots, including changes in water quality and quantity profiles, from both reference and modified data sets. SOM was specifically selected for this study among various classification and clustering algorithms because it was highly resistant to poor data quality (e.g., heterogeneity, outliers, noise, missing data, etc) [10–18]. This implies that SOM still maintains superior performance in characterizing spatial and temporal patterns in the data set of poor quality. SOM also had a notable advantage that it returned a low-dimensional view of informative vectors (i.e., weight or codebook vectors), which were extracted from heterogeneous data set with large input variables (i.e., high dimensions) [10–12]. Those reduced vectors were then visualized in a figure with multiple subplots (named as component planes) so that the end-users readily described the relationship among variables as well as addressed areas of interest such as water pollution hotspots in our case. Before executing the SOM algorithm, the four input data sets were rescaled between 0 and 1 in terms of individual variables

Table 2
Summary statistics for major monitoring parameters recorded during 36 months (i.e., January 2013 to December 2015) in the mainstream and tributaries of the Yeongsan River, Korea (n = 2,419)

	Parameters	Units	Mean	SD	CV
1	Water temperature (Temp)	°C	15.07	7.96	0.53
2	Dissolved oxygen (DO)	mg/L	10.23	3.03	0.30
3	pH	–	7.36	0.47	0.06
4	Electrical conductivity (EC)	µm hos/cm	282.68	212.37	0.75
5	Biochemical oxygen demand (BOD)	mg/L	2.97	2.94	0.99
6	Chemical oxygen demand (COD)	mg/L	6.31	4.86	0.77
7	Total organic carbon (TOC)	mg/L	4.44	3.17	0.71
8	Total nitrogen (TN)	mg/L	3.42	3.53	1.03
9	Total phosphorus (TP)	mg/L	0.15	0.21	1.36
10	Suspended solids (SS)	mg/L	18.37	31.00	1.69
11	Chlorophyll a (Chl)	mg/m^3	10.44	16.28	1.56
12	River discharge (Discharge)	m^3/s	0.83	1.97	2.38

SD = Standard deviation and CV = coefficient of variation.

using the embedded normalization method of range [20]. The algorithm was identically run with the transformed data sets in a combination of linear initialization and batch training modes. Note that the map size of SOM returned from three variant data sets is designed to always match that of the reference data set to provide a consistent view on the variation of spatial and temporal data profiles. More extended capabilities, as well as an in-depth review of SOM, are available elsewhere [10–18,20].

3. Results

3.1. Effect of data loss for single and multiple variables

Fig. 2 shows the influence of missing values of one and multiple parameters on the average values of the entire data set, which are assigned by SOM. Note that random data loss ranging from 5% to 30% was intentionally made in the modified data set of group 1, specifically for single variable (i.e., (a) pH, (b) TN, and (c) Discharge) as well as (d) multiple variables such as pH and TN (i.e. a plus b). In the figures, blue and red colors indicate the relative change of the mean values of individual variables as a percentage in the positive and negative directions, respectively. In contrast, the size of a circle signifies the absolute change of their mean values, regardless of their directions. In other words, a large red circle implies the mean value in the modified data set was much bigger than that of the original data set. It was generally expected that variables with lots of missing data, once provided to SOM, produced larger changes in the mean values of all observed variables in the entire data set than those with fewer missing data. This is particularly true for pH which elevates the mean values in other variables in both directions such as BOD, Chl, and Discharge (Fig. 2a) as well as for Discharge which only increases its average value itself (Fig. 2c). More importantly, the two variables (i.e., pH and Discharge) led to significant changes of the assigned values (of individual neurons in the given map size), including their average values, for other variables when a proportion of missing data went beyond 15% (Figs. 2a and c). However, neither were all variables sensitive to data loss occurred in the two variables, nor did a missing rate of 10% or less affect the average values of all monitored parameters considerably. Interestingly, data loss in the parameter TN altered the produced values of all variables in SOM to a lesser extent, irrespective of its missing rates (Fig. 2b). Similar results were also observed with the data sets which contained 5% of missing data for multiple variables (Fig. 2d). Note that only the data set with missing values in completely different samples rather than in identical samples for more than three variables causes noticeable change to the assigned values for a certain parameter such as Discharge. Collectively, data loss taken place in either single or multiple variables might modify

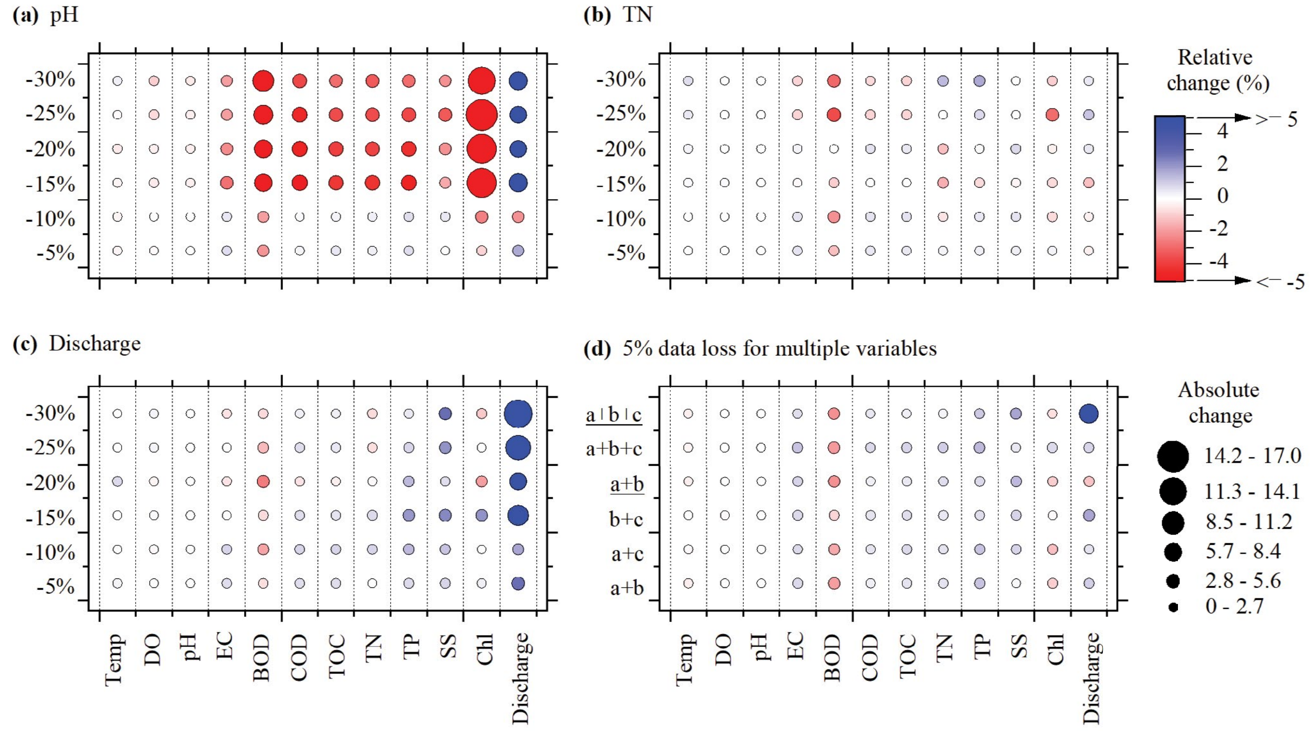

Fig. 2. Changes in the average value of individual variables when the original and modified data sets (namely, group 1) are provided to SOM. The group 1 data sets are prepared by increasing the number of missing values (from –5% to –30%) for a single variable such as (a) pH, (b) TN, and (c) discharge as well as (d) for multiple variables (i.e., 5% data loss for each variable) in the original data set. *Note*: in Fig. 2d, lowercase letters a, b, and c indicate the variables pH, TN, and discharge, respectively. Also, shown in plain and underlined texts are that missing values occur at the same and completely different samples (i.e., rows) in the modified data sets, respectively.

the allocated values of all tested variables, but produced a minimal change in the presence of a missing rate of 10% or less, regardless of variables.

3.2. Effect of data loss for variables

The screening accuracy of water pollution hotspots in SOM was also assessed in response to the number of available variables (Fig. 3). In Fig. 3a, the vertical axis on the left indicates the number of water pollution hotspots addressed by SOM based on the modified data sets of group 2 that either exclude a single variable rotationally from the original data set (see blue dotted line) or include only a few variables (see black dotted line). On the other hand, the vertical axis on the right represents quantification error observed with equivalent data sets (see red dots). The exact locations selected as water pollution hotspots from the corresponding data sets are also displayed in Fig. 3b. Note that those locations only show spatial information rather than sampling time when water pollution mainly occurs. As can be seen in Fig. 3a, the number of water pollution hotspots fluctuates between 28 and 48 in the absence of a particular variable from the original data set as well as in the presence of one or a couple of variables only in the modified data sets. However, some excluded variables such as EC, TOC, SS, and Discharge still maintained a similar number of pollution hotspots, as compared to the original data set (i.e., raw data). Out of the four variables, only the variables SS and Discharge appeared to correctly address pollution hotspots with respect to spatial location (Fig. 3b, compare the total number of samples as well as the locations addressed from each data set). Quantification error which measured the quality of the map in SOM in terms of the distortion (i.e., the Euclidian distance) between input and weight vectors decreased with the decrease in the number of variables, specifically becoming the lowest in the data sets with few variables (Fig. 3a). All these results implied that providing all monitored parameters simply as input to SOM did

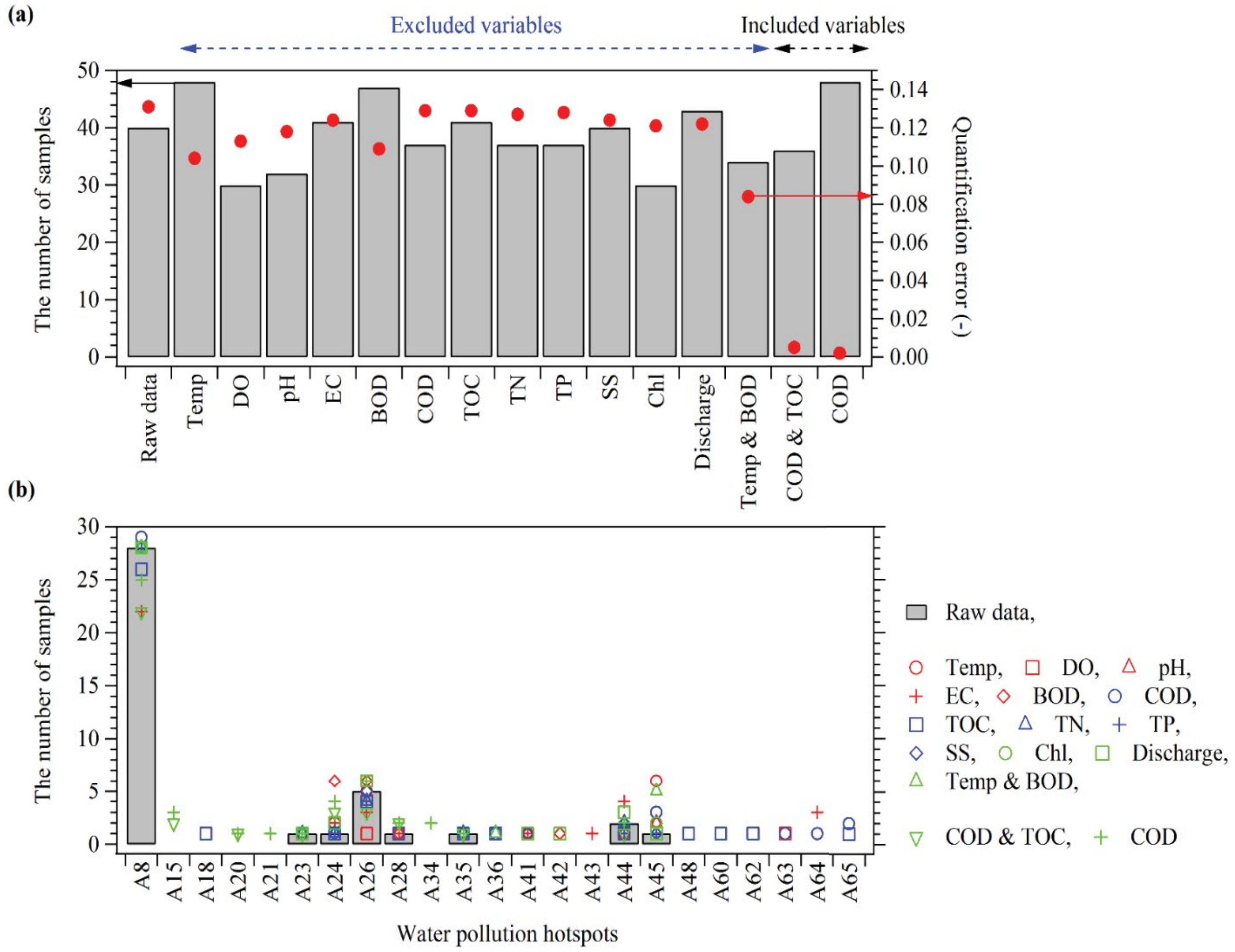

Fig. 3. Changes in water pollution hotspots in terms of COD when the original and modified data sets (namely, group 2) are provided to SOM: the effects of (a) excluded and included variables on the total number of pollution hotspots and quantification error and (b) on individual hotspot locations. The group 2 data sets are prepared not only by rotationally eliminating one variable at a time from the original data set (see excluded variables in Fig. 3a), but also by involving a couple of variables (see included variables in Fig. 3a).

not improve the screening performance of water pollution hotspots. Rather, they reduced the quality of the map as well as did not help classify water quality and quantity characteristics properly.

3.3. Effect of data loss in different pollution levels

Fig. 4 illustrates the screening performance of water pollution hotspots in SOM between the original and modified data sets (i.e., group 3). Note that only one single variable COD (labeled as raw data in Fig. 4a) is included in these modified data sets, from which intentional data loss reaching 10% occurs separately according to the level of water pollution (labeled as low, moderate, and high levels in Fig. 4a). Also, shown in Fig. 4a and b are component planes that visualize the variation of COD concentrations and exact locations of water pollution hotspots according to the data sets with and without data loss, respectively. It was shown in Fig. 4a that spatial and temporal patterns of COD in the original data set appeared to be quite similar to those of the modified data set with missing values only in high pollution level (compare the areas of concern exhibiting high COD concentrations at both component planes). However, those areas of concern in the original data set slightly shifted rightwards and completely moved to other locations in the modified data sets that contained 10% null values in moderate and low pollution levels, respectively. Missing values in data also decreased the number of water pollution hotspots, as compared to that of the original data set ($n = 40$). The reduction in the number of pollution hotspots screened was more apparent in the data set with missing records in high pollution levels ($n = 31$) than those of other levels ($n = 34$). Only the station where water pollution happened more frequently (i.e., A8) appeared to be correctly addressed by SOM, in the presence of 10% data loss at any pollution levels. This result was particularly consistent with the screening results with variable availability, as described in the previous section 3.2. Overall, 10% of data loss made in different pollution levels did not alter the screening accuracy of water pollution hotspots in SOM considerably. However, this is only true and valid for

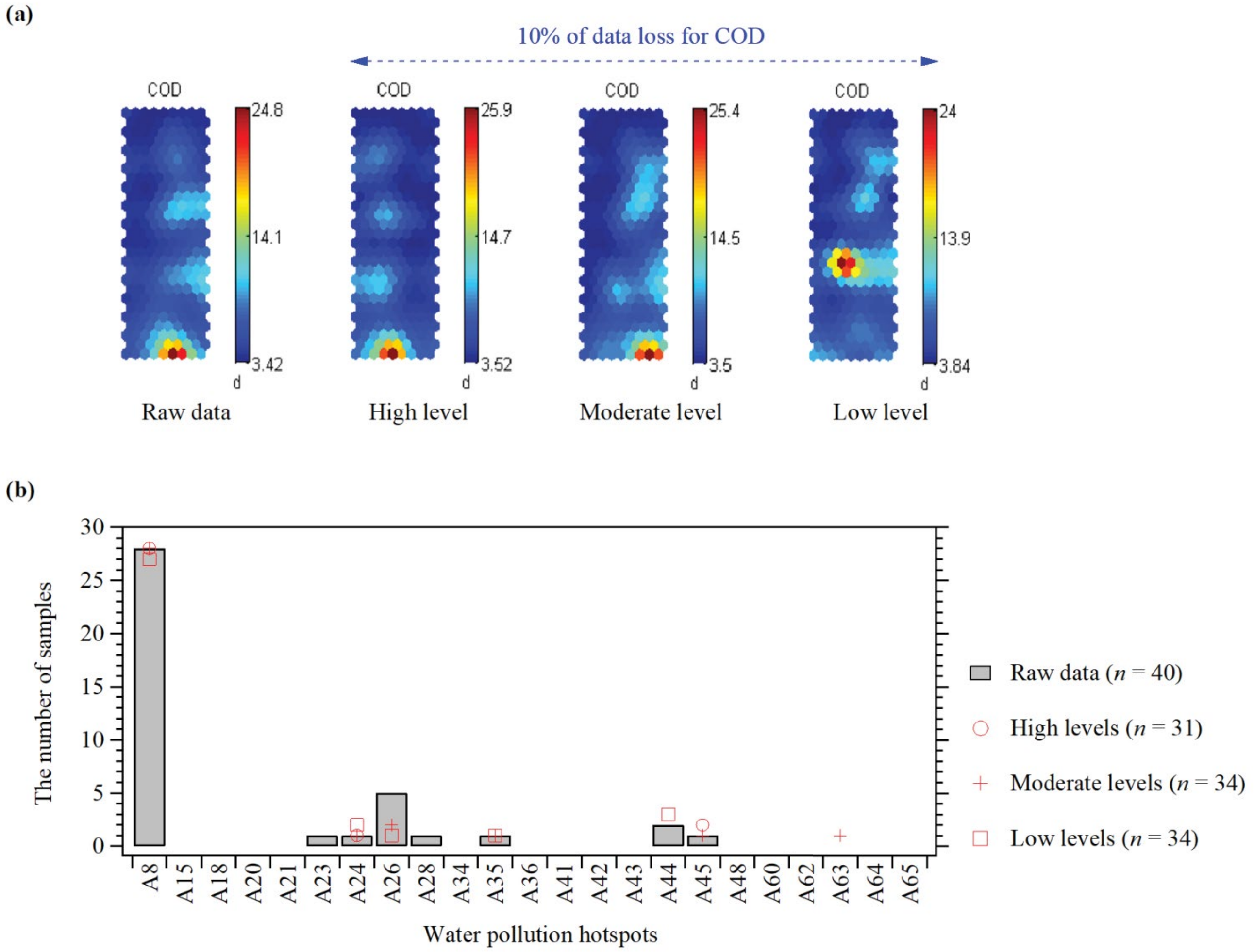

Fig. 4. Changes in water pollution hotspots in terms of COD when the original and modified data sets (namely, group 3) are provided to SOM: (a) component plane of COD and (b) individual hotspot locations. The group 3 data sets are prepared by removing a given amount of records randomly in different concentration levels of COD (i.e., 10% data loss of high, moderate, and low pollution samples) from the original data set.

stations with more frequent pollution records than less frequent pollution events.

4. Conclusion

This study aims to quantify the relative contribution of missing data to the screening performance of water pollution hotspots using a popular tool SOM in terms of dimension reduction and data visualization. The input data set was compiled from monthly water quality and quantity records for 3 years at the YS River in Korea. Three variant data sets were also derived from the reference input through the intentional deletion of variables as well as a certain amount of records in single and multiple variables, including those in a fixed variable according to different pollution levels. By running SOM multiple times with four different data sets, we arrived at the following conclusions.

- An increase in the number of missing data records in a single variable resulted in drastic changes in the weight vectors of some variables, specifically when data loss reached over 15%. Almost similar results appeared in the data sets with 15% data loss, made in different rows rather than in identical rows, for multiple variables.
- Removing one or a couple of variables, except for SS and Discharge, from the reference data set returned heterogeneous results in terms of the number of water pollution hotspots and their spatial location. SOM successfully captured only pollution hotspots with chronic contamination, regardless of the number of monitored variables retained in the data sets.
- The presence of null values imposed on different pollution levels (i.e., low, moderate, and high pollutant concentrations) in a fixed variable as high as 10% slightly modified the spatial and temporal patterns of water quality and quantity, as compared to the data set without data loss. The screening performance of water pollution hotspots was maintained solely for stations with more frequent pollution events, which coincided with the results derived from the absence of variables in the reference data set.

In summary, SOM adopting the non-linear projection seemed to more tolerant of variable availability than data loss in the data set. Therefore without proper integration with the imputation methods, the universal use of SOM as a prediction tool is not warranted, specifically when missing records exceed more than 15% of the entire data set.

Acknowledgment

This work was supported by Gyeongnam National University of Science and Technology Grant in 2018 - 2020.

References

[1] R. Aguilera, S. Sabater, R. Marcé, A methodological framework for characterizing the spatiotemporal variability of river water-quality patterns using dynamic factor analysis, J. Environ. Inf., 31 (2017) 97–110.

[2] L. Yang, K. Mei, X. Liu, L. Wu, M. Zhang, J. Xu, F. Wang, Spatial distribution and source apportionment of water pollution in different administrative zones of Wen-Rui-Tang (WRT) river watershed, China, Environ. Sci. Pollut. Res., 20 (2013) 5341–5352.

[3] V. Varekar, S. Karmakar, R. Jha, Seasonal rationalization of river water quality sampling locations: a comparative study of the modified Sanders and multivariate statistical approaches, Environ. Sci. Pollut. Res., 23 (2016) 2308–2328.

[4] R. Hatami, Development of a protocol for environmental impact studies using causal modelling, Water Res., 138 (2018) 206–223.

[5] Y. Fan, J. Chen, G. Shirkey, R. John, S.R. Wu, H. Park, C. Shao, Applications of structural equation modeling (SEM) in ecological studies: an updated review, Ecol. Process, 5 (2016) 19.

[6] I. Alameddine, M.A. Kenney, R.J. Gosnell, K.H. Reckhow, Robust multivariate outlier detection methods for environmental data, J. Environ. Eng., 136 (2010) 1299–1304.

[7] G.D. Betrie, R. Sadiq, S. Tesfamariam, K.A. Morin, On the issue of incomplete and missing water-quality data in mine site databases: comparing three imputation methods, Mine Water Environ., 35 (2016) 3–9.

[8] J. He, Mixture model based multivariate statistical analysis of multiply censored environmental data, Adv. Water Resour., 59 (2013) 15–24.

[9] A. Marinović Ruždjak, D. Ruždjak, Evaluation of river water quality variations using multivariate statistical techniques, Environ. Monit. Assess., 187 (2015) 215.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] T.T. Nguyen, A. Kawamura, T.N. Tong, N. Nakagawa, H. Amaguchi, R. Gilbuena, Clustering spatio–seasonal hydrogeochemical data using self-organizing maps for groundwater quality assessment in the Red River Delta, Vietnam, J. Hydrol., 522 (2015) 661–673.

[15] A. Astel, S. Tsakouski, 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.

[16] R. Chea, G. Grenouillet, S. Lek, Evidence of water quality degradation in lower mekong basin revealed by self-organizing map, PLoS ONE, 11 (2016) e0145527.

[17] C. Lennard, G. Hegerl, Relating changes in synoptic circulation to the surface rainfall response using self-organising maps, Clim. Dyn., 44 (2015) 861–879.

[18] W.-P. Tsai, S.-P. Huang, S.-T. Cheng, K.-T. Shao, F.-J. Chang, A data-mining framework for exploring the multi-relation between fish species and water quality through self-organizing map, Sci. Total Environ., 579 (2017) 474–483.

[19] 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.

[20] S.J. Ki, S. Song, T.W. Kang, S. Kim, T. Kang, S.G. Baek, J.H. Baek, J.H. Kim, Addressing water pollution hotspots in the tributary monitoring network using a non-linear data analysis tool, Desal. Wat. Treat., 77 (2017) 156–162.