Desalination and Water Treatment
www.deswater.com

doi:10.5004/dwt.2018.22883

124 (2018) 202–210
August

Comparison of classification and supervised learning algorithms in assessing the hydraulic conditions of sewer collection systems: A case study of local sewer networks in Jinju City, Korea

Seo Jin Ki, Chun-Sik Lee, Won-Hee Jung, Hyun-Geoun Park*

Department of Environmental Engineering, Gyeongnam National University of Science and Technology, 33 Dongjin-ro, Jinju-si, Gyeongsangnam-do 52725, Korea, email: hgpark@gntech.ac.kr (H.-G. Park)

Received 16 April 2018; Accepted 24 July 2018

ABSTRACT

Accurate screening of sewer conditions from monitoring data contributes to maintaining their operations (in terms of water quality and quantity) safe as well as reducing their associated costs (for operation and maintenance). This study was designed to assess the performance deterioration in sewer systems using a series of data classification tools, namely classical classification and novel supervised learning algorithms. The hydraulic data available for four sewer systems at Jinju City in Korea in a daily format during the monitoring period of 2013–2017 were provided as example data sets to those algorithms, which were evaluated independently with 70% training and 30% test data sets randomly divided. A self-organizing map (SOM) with a specialty in extracting hidden patterns in data was used to classify the data sets into three warning levels in the absence of any definite warning criteria for individual parameters. Our findings showed that three supervised learning algorithms achieved comparable performance in predicting warning levels defined from SOM to exiting classification algorithm in terms of accuracy and error rate. The network architecture optimized for supervised learning algorithms, in fact, varied significantly depending on the data sets, including that with additional variables on top of the original data set. In contrast, exiting classification algorithm unexpectedly produced high error rates in case that the hydraulic parameters had low coefficient of variation values reaching as high as 16%. Overall, these results demonstrated that novel supervised learning algorithms were more universally applicable for the assessment of hydraulic and/or water quality conditions in sewer systems than classical classification algorithm, regardless of the amount of variability in the data sets.

Keywords: Supervised learning algorithms; Classification algorithm; Self-organizing map; Hydraulic parameters; Water quality; Sewer systems

1. Introduction

Public health and the environment depend on the sustainability of sewer infrastructure [1–3]. Sewer deficiencies resulted from pipe deterioration and failures often led to overflows and flooding, exposing the public to diverse pollutants (e.g., bacteria, viruses, and inorganic ions) that caused water-borne disease outbreaks as well as contaminating nearby environmental media (e.g., surface waters, soils,and groundwaters) [1,3]. Proper sanitation services also enhance the national economy and social welfare by reducing the costs for rehabilitation (i.e., repair and renewal) and property damages as well as the incidents of untreated sewer release to recreational water bodies [3,4]. Previous study in the United Kingdom has discovered that a gross replacement cost reached as much as 104 billion for 302,000 km of sewer pipes [4]. The cost spending on water and wastewater infrastructure in the United States for 2007 has more than doubled since the 1956 estimate of 12.5 billion for about 1,300,000 km-long sewer line as of 2009 [3].Therefore,

*Corresponding author.

operation and maintenance (O&M) programs are of significant importance to sustainable sewer infrastructure.

Numerical simulation models are capable of describing the deterioration or capacity deficiencies of sewer pipes, thus providing guidance on implementing O&M and safety planning [5]. Factors to affect sewer replacement and/or rehabilitation included pipe characteristics (e.g., shape, materials, depth, and slope), soil and groundwater conditions (e.g., infiltration and groundwater table), sewer location (e.g., street and private property), service area characteristics (e.g., the number of connections, tributary area and meteorological conditions), and so on (https://www.resourcerecoverydata.org/weffactsheets.htm). Simulation accuracy was strongly affected by the availability of those data, otherwise all important parameter settings in the models needed to be optimized through inverse modeling techniques [5]. On the other hand, statistical models, in addition to machine learning models, relate monitoring data to environmental information available (or even themselves) to provide valuable insight into the deterioration processes of sewer lines in the absence of full site-specific parameters [5–10]. More specifically, statistical models (e.g., Cohort survival and Markov chain models as well as logistic regression and discriminant analyses) played an important role in explaining the uncertainty associated with asset deterioration and failure [5]. Conversely, machine learning models (e.g., random forest and artificial neural networks) were best adopted to identify the sewer line condition in discrete classes, which were useful for the short and mid-term planning processes [5,6]. Note that both models can be applicable to either network or pipe levels as well as either hydraulic or structural aspects of assessment [5–10].

As compared to earlier studies, this study was motivated to assess the hydraulic conditions of sewer pipes using classical classification and novel supervised learning algorithms released recently as independent toolboxes which were implementable in MATLAB software [11–17]. From the (five-year) monitoring data sets observed for four sewer systems at Jinju City in Korea, the specific objectives of this research were 1) to classify the dynamic hydraulic behavior of sewer systems into discrete classes (i.e., warning levels) using unsupervised learning algorithm (either provided as a parallel tool or available within supervised learning algorithms), to assess the predictive ability of defined classes by 2) classification and 3) supervised learning algorithms, and 4) to compare their performance under conditions of different dimensions of variables as well as different levels of variable variability. We hope that this preliminary study help address sewer pipes in critical condition for O&M and safety planning from the hydraulic and water quality data measured in real-time.

2. Materials and Methods

2.1. Data sets at local sewer systems

Fig. 1 shows sewer service areas (i.e., Insa-dong, Chiram-dong, and Sangpyeong-dong) operated by a particular private company under the Build-Transfer-Lease (namely, BTL) scheme in Jinju City (see Fig. 1a), including one example sewer network installed at Insa-dong (see Fig. 1b). The company monitored hydraulic and water quality parameters at the final outlet of four sewer systems (IS-1, CA-1, CA-2, and SP-1) to ensure proper functioning and operational control of the pipe networks as well as to implement a successful maintenance strategy using the simulation model Stormwater Management Model (namely, SWMM) with monitored data [18]. During the monitoring, the hydraulic parameters such as flow rate, velocity, and water level were recorded at 10-min intervals, whereas biochemical oxygen demand measured every 2 h was selected to rapidly detect and prevent water quality degradation in the sewer networks. Table 1 presents the total number of data collected at individual sewer networks during the five-year monitoring period (between 2013 and 2017). Out of the parameters monitored, only hydraulic parameters were provided as inputs to both unsupervised [19–22] and supervised learning algorithms [14–17], in addition to

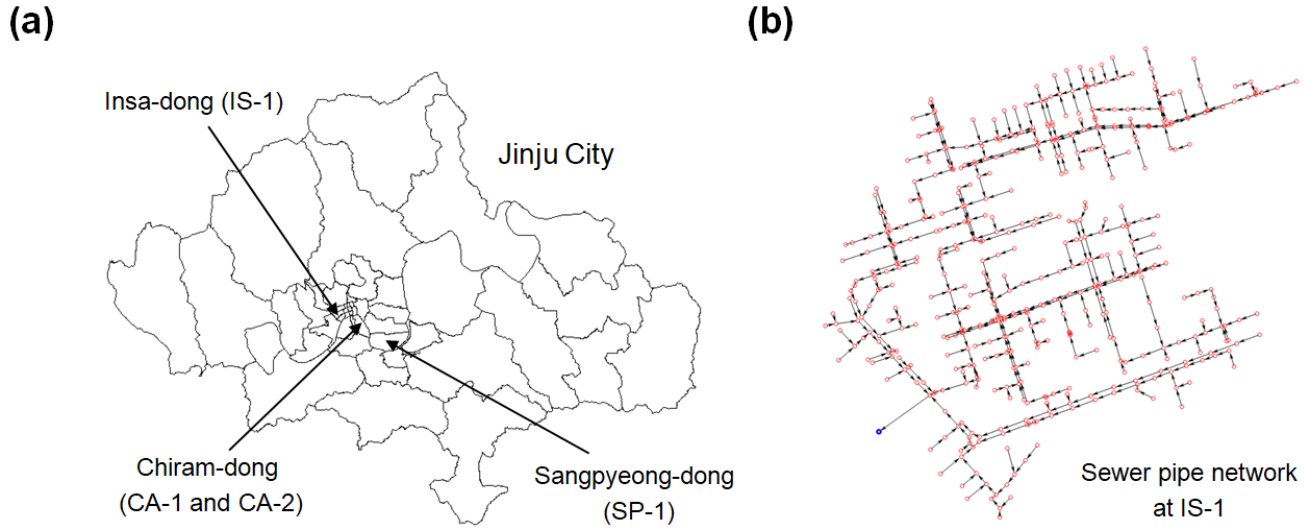

Fig. 1. (a) Four sewer systems monitored in Jinju City and (b) example sewer pipe network at the sewer system MS-1. Note that an open circle with blue color in (b) indicates the final outlet of the sewer pipe network, which monitors hydraulic parameters (i.e., flow rate, velocity, and water level) in discrete time intervals (i.e., 10 min), at IS-1.

classification tool [11–13] which was evaluated separately to compare its performance against supervised learning. While the data set collected from IS-1 recorded the highest flow rate and water level among sewer systems, the largest velocity was observed in CA-1 (see Table 1). Interestingly, IS-1 also showed the largest amount of variation in water level, as compared to other monitoring parameters in different sewer systems (compare the coefficients of variation). Detailed information on time series data patterns, including monitoring devices, is available in the operational reports of sewer systems released quarterly [18].

2.2. Data partitioning from unsupervised learning

The data sets compiled from four sewer systems were analyzed in two different steps. The first procedure was to partition an entire dataset (of each sewer system) into smaller subgroups, i.e., different warning levels, using unsupervised learning algorithm. In this step, each data set was classified into three warning levels (i.e., the strongest, moderate, and weakest warnings) based on similarities of hydraulic properties rather than classes assigned arbitrarily (by end-users). Note that as there are no strict criteria for determining the degree of warning from individual hydraulic parameters, warning levels defined from this process play a role in assessing the prediction performance of classification and supervised learning algorithms described in Section 2.3. However, in case clear standards exist, the original data set should be split according to the criteria for multiple parameters that are intended to diagnose the hydraulic performance (or capacity) of local sewer systems. SOM toolbox (version 2.0), downloadable at http://www.cis.hut.fi/somtoolbox/, was specifically used for grouping hydraulic data into three warning levels [19]. The self-organizing map (SOM), in fact, has gained popularity in robust data classification without relying on any prior knowledge on data structure as well as in the presence of large noise and outliers [19–22]. Default parameter settings, except for the range normalization of the raw data, linear initialization, and batch training, were adopted in the training of SOM with MATLAB software (version 2016b). More information on theory, algorithms and applications of the SOM, including data handling procedures, is introduced well in earlier studies [19–22].

2.3.Class prediction from supervised learning

In the second step, two additional tools, classification toolbox (version 5.0) as well as Kohonen and CP-ANN toolbox (version 3.8), were used to compare the predictive ability of warning levels defined from SOM. Those toolboxes, which were also implementable in MATLAB, contained a series of modules for classifying data patterns in a supervised manner. The current classification toolbox included 8 different multivariate models [11–13], whereas 3 supervised learning algorithms were embedded in the Kohonen and CP-ANN toolbox, along with unsupervised learning algorithm Kohonen Map (namely, SOM) [14–17]. Specifically, partial least squares-discriminant analysis (PLS-DA) in classification toolbox, a variant of PLS regression which projected (categorical) dependent and (continuous) independent variables into new orthogonal axes, was selected for comparison to supervised learning algorithms [11]. Counter propagation artificial neural network (CP-ANN), supervised Kohonen network (SKN), and XY-fused network (XY-F), all of which effectively predicted class membership with non-linear boundaries, were employed to assign warning levels according to hydraulic data monitored [14]. Three supervised

Table 1
The quality of self-organizing map (SOM) for data sets collected from different sewer systems from January 1, 2013 to December 31, 2017

Sewer systems	Total number of data	Mean ± standard deviation	Coefficient of variation	Quantization error	Topographic error	Map size
CA-1	1,824	2,776.49 ± 865.82 (F, m^3/d)[b]	0.31 (F)	0.019	0.019	18 × 12
		2.06 ± 0.24 (V, m/s)	0.11 (V)			
		7.15 ± 6.65 (L, cm)	0.93 (L)			
CA-2	1,826	2,569.38 ± 409.76 (F, m^3/d)	0.16 (F)	0.010	0.035	21 × 10
		0.58 ± 0.04 (V, m/s)	0.06 (V)			
		13.01 ± 1.46 (L, cm)	0.11 (L)			
IS-1	1,816	6,526.46 ± 1,489.48 (F, m^3/d)	0.23 (F)	0.012	0.046	19 × 11
		1.14 ± 0.18 (V, m/s)	0.16 (V)			
		21.67 ± 47.64 (L, cm)	2.20 (L)			
SP-1	1,819	1,472.22 ± 241.55 (F, m^3/d)	0.16 (F)	0.007	0.229	23 × 9
		0.55 ± 0.07 (V, m/s)	0.13 (V)			
		12.30 ± 4.94 (L, cm)	0.40 (L)			
IS-1_9[a]	1,816	–	–	0.057	0.049	21 × 10

[a]The data set included 6 additional variables (i.e., 1- and 2-day antecedent variables for each parameter) derived from 3 original variables (i.e., flow rate, velocity, and water level) during the study period at the sewer system IS-1. Note that descriptive statistics of this data set are not shown due to their similarity to the original data set.
[b]The capital letters F, V, and L indicate the parameters flow rate, velocity, and water level, respectively.

learning algorithms slightly differed in that how Kohonen and output layers were working together. For example, the output layer was attached to the Kohonen layer during the training phase for SKN, different similarity distances calculated separately in the Kohonen and output layers were merged to search winner neurons for XY-F, and winner neurons in the Kohonen layer were designed to select their corresponding position in the output layer for CP-ANN [14]. The data sets were randomly divided into two partitions, 70% for training data and 30% for test data. As described in unsupervised learning algorithm, all supervised learning algorithms were run using default parameter settings, except for the auto-scaling transformation of the raw data and cross-validation based on venetian blinds with 3 to 5 cancellation groups. Note that all supervised learning algorithms adopted for this study, including their theories or network architecture, are described extensively in literature [11–17]. The toolboxes for classification and Kohonen and CP-ANN are available at http://michem.disat.unimib.it/chm/download/classificationinfo.htm and http://michem.disat.unimib.it/chm/download/kohoneninfo.htm, respectively.

3. Results and discussion

3.1. Determining warning levels from unsupervised learning

Fig. 2 displays the results of data partitioning to determine warning levels in different sewer systems, CA-1 for (a), CA-2 for (b), IS-1 for (c), and SP-1 for (d).The left panels of individual figures show Davies–Bouldin (D-B) index-which is used to determine the optimal number of clusters in each data set, whereas subgroups (i.e., groups 1 to 3) separated by D-B index are exhibited in the right panels. Note that the SOM algorithm assigns random colors to individual subgroups that are ranked from weakest (for group 1) through moderate (for group 2) to strongest warning levels (for group 3) in terms of hydraulic parameters in the right panels of the figures. From the left panels of the figures, it was shown that while D-B index varied slightly according to the number of clusters from 2 to 7, three subgroups were sufficient enough to explain the variation in hydraulic parameters observed in different sewer systems during the monitoring period. This implied that individual data sets were successfully categorized into three warning levels

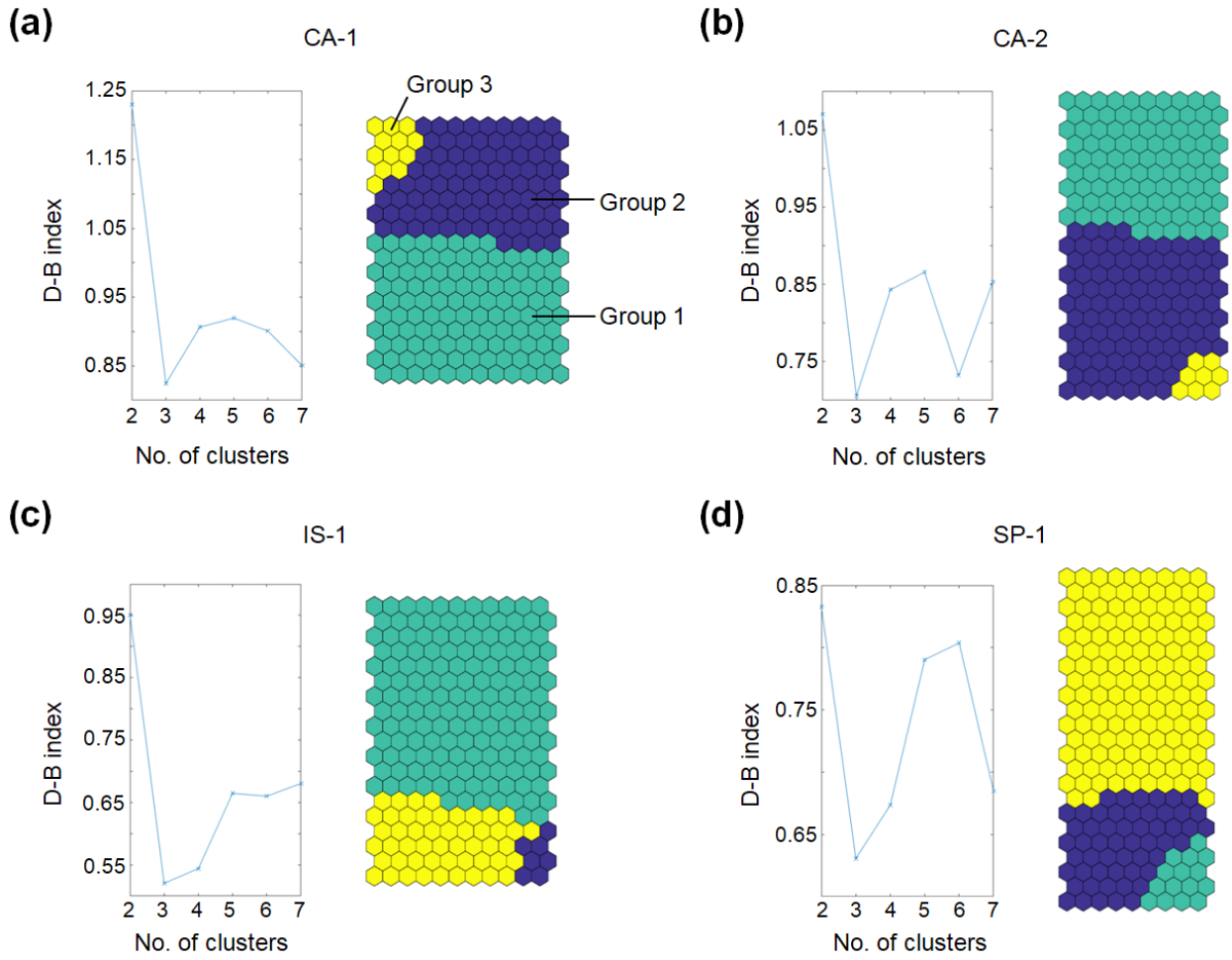

Fig. 2. The partitioning results of data sets for four sewer systems, (a) CA-1, (b) CA-2, (c) IS-1, and (d) SP-1, using the unsupervised learning algorithm SOM. Note that D-B index in the left panels of each figure is the abbreviation for the Davies–Bouldin index which assesses the quality of resulting clusters (i.e., the lower the better). Different colors in the right panels of each figure were randomly assigned to individual groups rather than relying on groups ranked from weakest to strongest based on three hydraulic parameters.

by SOM. In addition, from the right panels of the figures, total neurons (or cells) allocated to group 3 (i.e., the strongest warning level) were found to be always less than those of other groups in the component plane of SOM, although the relative amount of cells occupied differed from group to group. It should be noted that SOM yields the lowest quantification and topographic errors when the map size is determined automatically in the given data sets under three subgroups partitioned (see Table 1). There is no considerable change in the map size of SOM, except for quantification and topographic errors, when additional antecedent (hydraulic) variables were added to the original data set (compare IS-1 vs IS-1_9 in Table 1).

3.2. Prediction of warning levels by classification toolbox

Fig. 3 shows the performance of PLS-DA, embedded in classification toolbox, in two sewer systems, CA-2 and SP-1. The scaled data averaged for each class are exhibited in Fig. 3a, where warning levels are inversely assigned to class labels only in classification toolbox (that is, class 1 is equivalent to group 3). As displayed in the figure, the hydraulic parameters at CA-2 were effectively split into separate classes by auto-scaling process, among which class 1 (i.e., the strongest warning level) was characterized by high values of flow rate, velocity, and water level. This was also confirmed from receiver operating characteristic (ROC) curves that all classes (i.e., warning levels) were separated clearly when considering specificity and sensitivity (Fig. 3b). Fig. 3c illustrates score plot of the first principal component (PC) versus Q residuals at CA-2. Each neuron arranged in SOM was represented with colored circles according to classes in score plot, where asterisk symbols represented the test data sets. Except for very few outliers and overlap cases in the test data sets, all training and test data were allocated within class boundaries represented by solid line in differ-

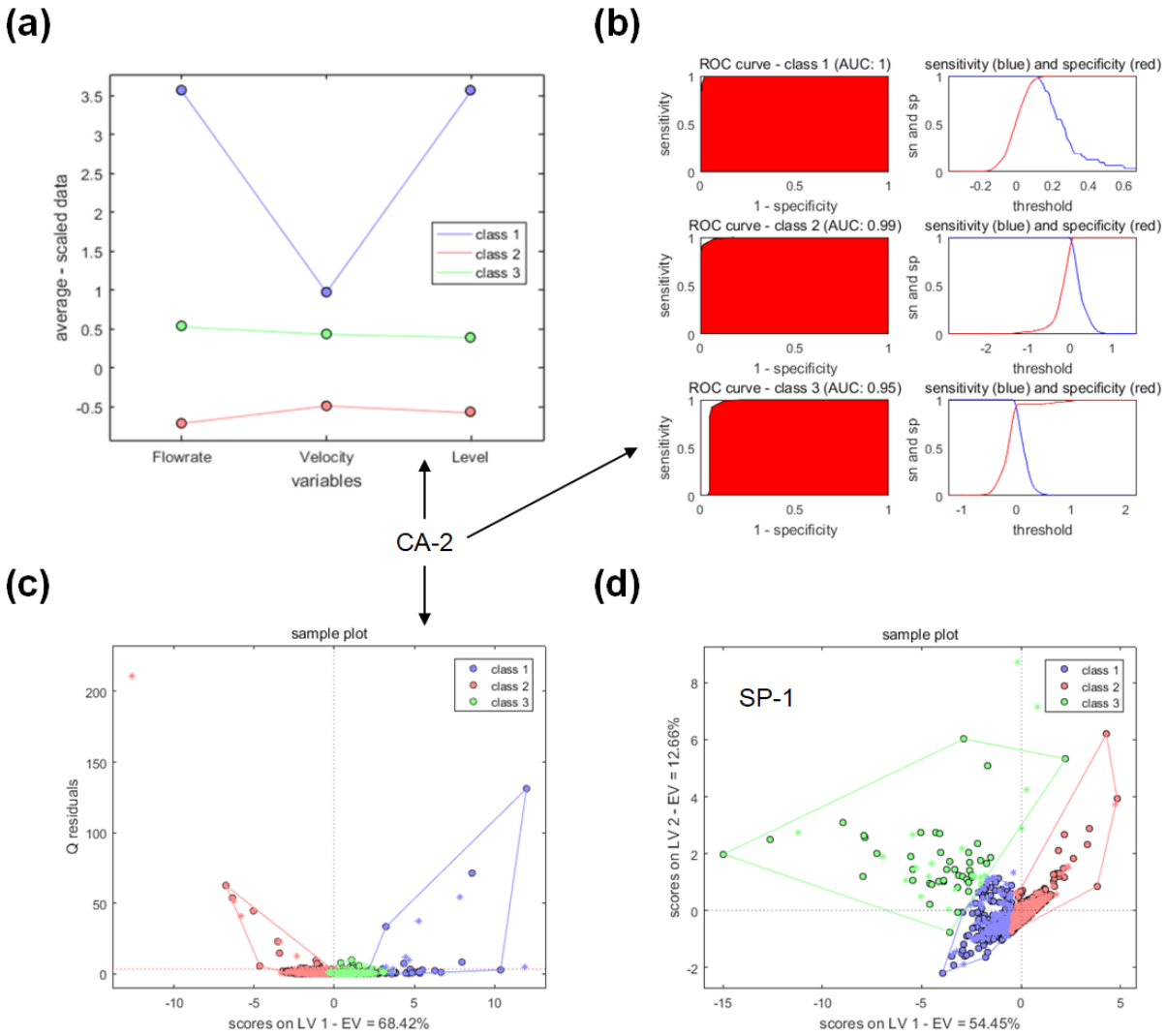

Fig. 3. The performance of the classification algorithm PLS-DA for two different data sets. Shown in (a) is the mean of scaled data for three monitored variables in CA-2 (used as inputs to the algorithm), whereas (b) and (c) display their corresponding outputs, the ROC curves of the classes and score plot, respectively. (d) is the score plot resulted from the data set SP-1. The solid circles and asterisk symbols in (c) and (d) indicate the training and test data sets, respectively.

ent colors, indicating that PLS-DA predicted warning levels correctly for the data set of CA-2. Shown in Fig. 3d is also score plot for the first two PCs at SP-1, which implies that warning levels are addressed accurately by PLS-DA, excluding extremely a few cases. Note that while only the first PC accounting for 68% of the total variance is selected for CA-2, 67.11% of the variance is explained by two PCs for SP-1.

3.3. Prediction of warning levels by Kohonen and CP-ANN toolbox

Fig. 4 illustrates the prediction results of warning levels by two supervised learning algorithms, CP-ANN (at CA-1) and XY-F (at CA-2), which are obtained based on optimization of network architecture (e.g., the number of neurons and epochs) from genetic algorithms. Note that unlike PLS-DA, class labels in supervised learning are correctly provided to individual warning levels (in other words, class 1 corresponds to group 1). As can be seen in Fig. 4a, the ROC curves for all classes are very closely located in the top-left corner, indicating ideal performance that three warning levels at CA-1 are accurately predicted by CP-ANN with little or no false alarms. However, the ROC curves in XY-F are slightly apart from the top-left corner for some classes (e.g., classes 2 and 3), implying that the prediction performance of XY-F at CA-2 is slightly lower than that of CP-ANN at CA-1. It turned out from the previous studies [14,17] that the network size and the number of training epochs required (for supervised learning algorithms) increased with the size and color intensity of bubbles (i.e., circles), respectively. In addition, the best architecture ensuring relatively good prediction performance as well as high frequency selection (by genetic algorithms) should be found on the top-right side in the plot of optimization results (see bubbles in red color at Figs. 4c and d). Based on these findings, the optimal network size and epoch value determined were 14 × 14 and 100 for CP-ANN at CA-1 as well as 16 × 16 and 50 for XY-F at CA-2, respectively.

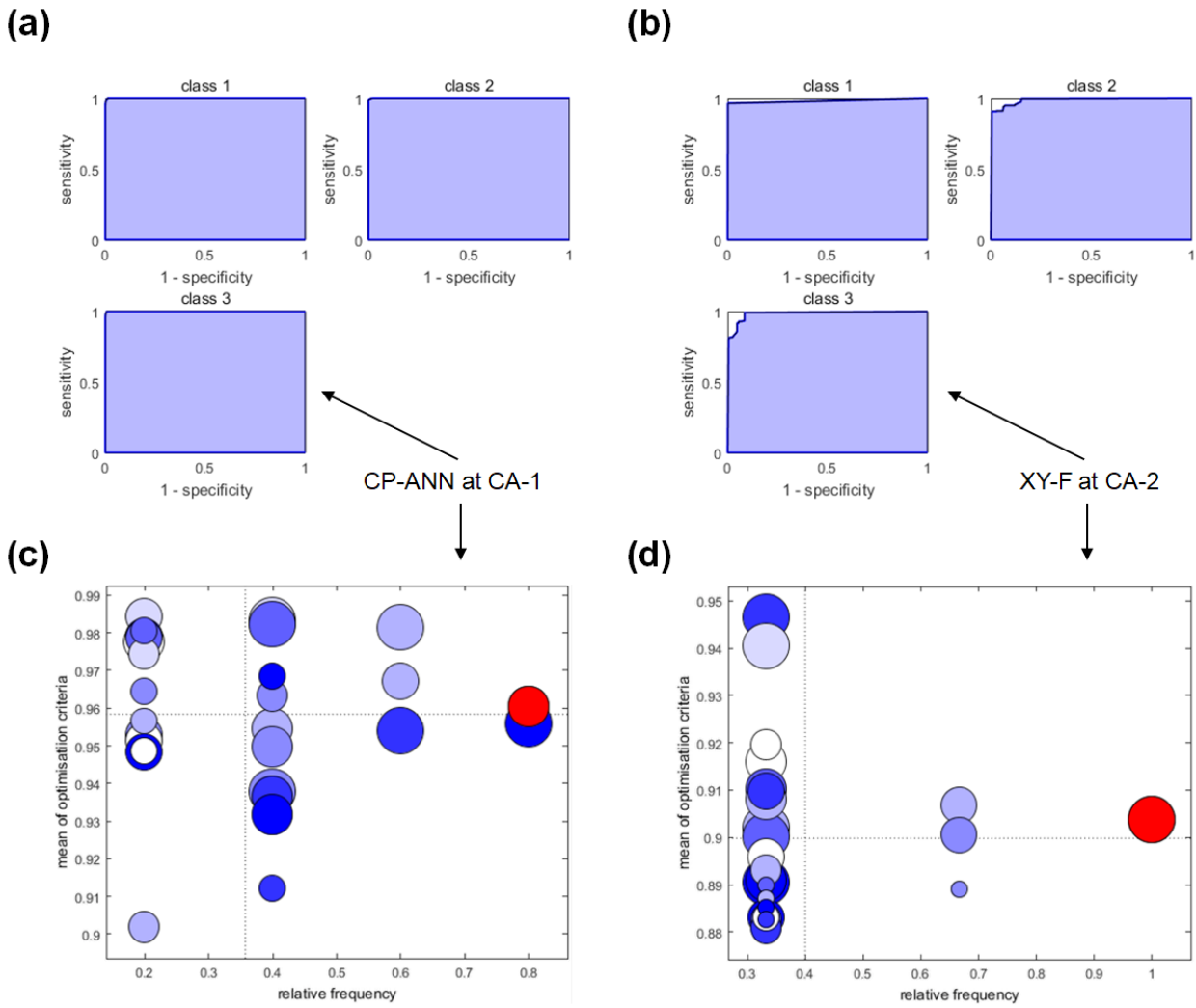

Fig. 4. The performance of two supervised learning algorithms in different data sets. Shown in (a) and (b) are the ROC curves of the classes and optimization results of CP-ANN for the data set CA-1, respectively. (b) and (d) represent those of XY-F for the data set CA-2.

3.4. Performance comparison between different tool boxes

Table 2 compares the performance of PLS-DA and three supervised learning algorithms (i.e., CP-ANN, SKN, and XY-F) at all monitored sewer systems. In the table, the percentage of not-assigned samples is only given for PLS-DA because of the difference in output results (directly provided) between tool boxes. The table also includes the performance assessment results of all prediction algorithms for the data set (namely, IS-1_9) which incorporates 6 antecedent variables on top of the original data set (IS-1, see also Table 1). From the table, it was found that all tested algorithms showed excellent prediction performance (of three warning levels), reaching generally over 95% accuracy, during the fitting and either validation (for PLS-DA) or cross-validation processes (for CP-ANN, SKN, and XY-F), except for a few cases. In addition, their performance appeared to be maintained successfully during the prediction step (i.e., (external) test data) although they suffered from a slight decrease in performance (see Section 2.3). The highest error rate only occurred for PLS-DA at CA-2, which reached a maximum of 0.360 (36%) during the prediction phase. In contrast, the remaining supervised learning algorithms showed slightly lower error rates, which reached as high as 0.170 (17%) for XY-F at SP-1, than PLS-DA. The high error rates of PLS-DA at CA-2 appeared to be associated with low coefficients of variation in the hydraulic parameters, as compared to those noticed in other data sets (see Table 1). All these results demonstrated that while the performance of all prediction algorithms slightly varied depending on the data sets, 1) we observed an abrupt increase in error rate for PLS-DA and 2) CP-ANN generally outperformed its peers (i.e., SKN and XY-F) in terms of accuracy and error rate.

Fig. 5 also provides more detailed information on two example outputs for PLS-DA and SKN at the modified data set IS-1_9. As shown in Fig. 5a, the error rate and percentage of not-assigned samples were not significantly reduced when PLS-DA included more additional variables. Only a few outliers were observed for PLS-DA (see Fig. 5c), but those minor errors seemed to be truncated in the error rate observed during the prediction phase (see Table 2). Like PLS-DA, SKN also showed outstanding prediction performance of three warning levels (see Fig. 5b) when its architecture was optimized with respect to the network size of 12 × 12 and epoch value of 100 (see Fig. 5d). The optimal architectures of PLS-DA as well as CP-ANN, SKN, and XY-F for all data sets are summarized in Table 3. Inclusion of additional variables did not help improve the prediction performance in terms of accuracy and error rate (for all tested algorithms, see Table 2), but appeared to reduce the number of neurons and epochs (required for three supervised learning algorithms) remarkably, as displayed in the table.

Table 2
Performance assessment of classification versus supervised learning algorithms for data sets collected from different sewer systems plus that with additional variables

Sewer systems	Algorithms	Fitting			Validation/Cross-validation			Prediction		
		Error rate	Accuracy	Not-assigned	Error rate	Accuracy	Not-assigned	Error rate	Accuracy	Not-assigned
CA-1	PLS-DA	0	0.990	0.070	0.010	0.990	0.070	0.010	0.990	0.080
	CP-ANN	0.021	0.992	–	0.054	0.978	–	0.077	0.923	–
	SKN	0.015	0.990	–	0.075	0.971	–	0.098	0.902	–
	XY-F	0.080	0.973	–	0.090	0.942	–	0.115	0.885	–
CA-2	PLS-DA	0.350	0.970	0.070	0.350	0.980	0.080	0.360	0.960	0.070
	CP-ANN	0.028	0.974	–	0.079	0.974	–	0.020	0.980	–
	SKN	0.065	0.950	–	0.070	0.943	–	0.103	0.897	–
	XY-F	0.059	0.929	–	0.124	0.937	–	0.129	0.871	–
IS-1	PLS-DA	0	1	0	0	1	0	0	1	0.010
	CP-ANN	0	0.999	–	0.001	0.999	–	0.042	0.958	–
	SKN	0.001	0.999	–	0.007	0.995	–	0.003	0.997	–
	XY-F	0.003	0.998	–	0.005	0.996	–	0.042	0.958	–
SP-1	PLS-DA	0.040	0.970	0.080	0.040	0.970	0.080	0.050	0.960	0.080
	CP-ANN	0.011	0.995	–	0.050	0.987	–	0.063	0.937	–
	SKN	0.056	0.972	–	0.071	0.974	–	0.071	0.929	–
	XY-F	0.102	0.978	–	0.072	0.985	–	0.170	0.830	–
IS-1_9[a]	PLS-DA	0	1	–	0	1	0	0	1	0
	CP-ANN	0	1	–	0	0.999	–	0.048	0.952	–
	SKN	0	1	–	0.002	0.998	–	0.101	0.899	–
	XY-F	0	1	–	0	0.999	–	0.101	0.899	–

[a]The data set included 6 additional variables on top of 3 original variables at IS-1 (see Table 1).

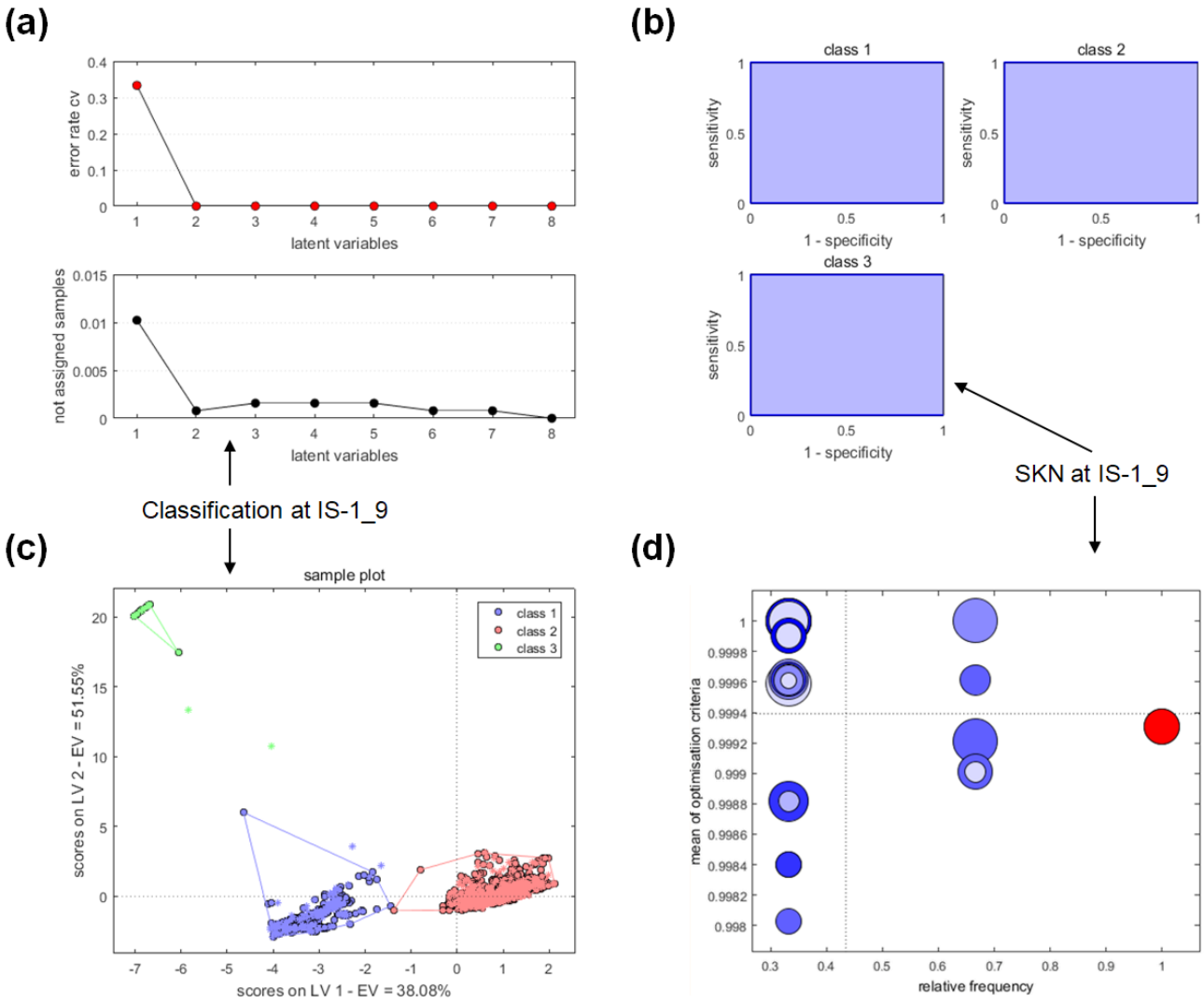

Fig. 5. The performance of classification (PLS-DA for (a) and (c)) and supervised learning algorithms (SKN for (b) and (d)) for the data set IS-1_9 (see Table 1). The error rate and ratio of non-assigned samples changed in response to the number of variables are exhibited in top and bottom panels at (a), respectively. Also, refer to captions for Figs. 3 and 4 for interpretation of (b), (c), and (d).

4. Conclusion

This study aimed to identify the best classification algorithm(s) in screening the hydraulic conditions at local sewer systems. The data sets, recorded at 10-min intervals for four sewer systems (CA-1, CA-2, IS-1, SP-1) at Jinju City in Korea during the five years between 2013 and 2017, served as inputs for classification (PLS-DA) and supervised learning algorithms (CP-ANN, SKN, and XY-F) after aggregated into the daily average. The aggregated data sets were also provided to the unsupervised learning algorithm SOM to classify each data set into three warning levels (from weakest through moderate to strongest), in the absence of any guidelines for classification criteria for individual hydraulic parameters (i.e., flow rate, velocity, and water level). A summary of our main findings are follows.

- The unsupervised learning algorithm SOM allowed the data sets to be separated into three homogenous sub-groups (namely, warning levels) based on similarities of hydraulic parameters. The map size in SOM was still maintained even if the data set included more antecedent variables.
- PLS-DA achieved good classification results in predicting three warning levels even though some outliers and overlapping samples between different classes (i.e., warning levels) were observed for 30% of the test data. A couple of PCs were selected to elucidate a large amount of variation in the data sets.
- The prediction performance of CP-ANN (at CA-1) was superior to that of XY-F (at CA-2), when considering ROC curves for three classes (i.e., warning levels). The network size and epoch value determined by the optimization process were sensitive to supervised learning algorithms as well as data sets.
- Outstanding prediction performance was typically observed in all tested algorithms (PLS-DA, CP-ANN,

Table 3
Summary of the optimization results for three supervised learning algorithms evaluated at data sets collected from different sewer systems plus that with additional variables

Sewer systems	Algorithms	Neurons	Epochs	Frequency	Optimal criterion
CA-1	CP-ANN	14 × 14	100	0.800	0.960
	SKN	16 × 16	350	1.000	0.911
	XY-F	16 × 16	300	1.000	0.898
CA-2	CP-ANN	12 × 12	200	1.000	0.918
	SKN	14 × 14	50	1.000	0.908
	XY-F	16 × 16	50	1.000	0.904
IS-1	CP-ANN	16 × 16	100	1.000	0.998
	SKN	16 × 16	350	1.000	0.998
	XY-F	16 × 16	100	1.000	0.997
SP-1	CP-ANN	10 × 10	350	0.667	0.978
	SKN	14 × 14	250	0.667	0.930
	XY-F	12 × 12	50	1.000	0.887
IS-1_9 [a]	CP-ANN	12 × 12	150	0.667	1.000
	SKN	12 × 12	100	1.000	0.999
	XY-F	12 × 12	50	1.000	0.999

[a]The data set included 6 additional variables on top of 3 original variables at IS-1 (see Table 1)

SKN, and XY-F), except for one particular case that PLS-DA yielded high error rates of about 36% at CA-2. Antecedent variables added to the original data set did not affect their prediction performance. Rather, those played a role in decreasing their network sizes and epoch values.

Acknowledgements

This work was supported by Gyeongnam National University of Science and Technology Grant in 2017–2018.

References

[1] US Environmental Protection Agency (EPA), EPA Enforcement: Preventing Backup of Municipal Sewage into Basements, Report No. 325-N-06-001, Office of Enforcement and Compliance Assurace, Washington, DC, 2006.
[2] US Environmental Protection Agency (EPA), Collection Systems O&M Fact Sheet: Sewer Cleaning and Inspection, Report No. 832-F-99-031, Office of Water, Washington, DC, 1999.
[3] K. Miller, K. Costa, D. Cooper, How to Upgrade and Maintain Our Nation's Wastewater and Drinking-Water Infrastructure, Center for American Progress, Washington, DC, 2012. Available from: <https://cdn.americanprogress.org/wp-content/uploads/2012/10/MillerWaterInfrastructureReport.pdf>.
[4] J.P. Davies, B.A. Clarke, J.T. Whiter, R.J. Cunningham, Factors influencing the structural deterioration and collapse of rigid sewer pipes, Urban Water, 3 (2001) 73–89.
[5] G. Kley, N. Caradot, Review of sewer deterioration models, Project acronym: SEMA, Report D 1.2, Kompetenzzentrum Wasser Berlin gGmbH, Berlin, Germany, 2013. Available from: <http://www.kompetenz-wasser.de/wp-content/uploads/2017/05/d12_sema_review_of_sewer_deterioration_models.pdf>.
[6] N. Caradot, M. Riechel, M. Fesneau, N. Hernandez, A. Torres, H. Sonnenberg, E. Eckert, N. Lengemann, J. Waschnewski, P. Rouault, Practical benchmarking of statistical and machine learning models for predicting the condition of sewer pipes in Berlin, Germany, J. Hydroinformatics (2018) doi:10.2166/hydro.2018.217.
[7] R. Baur, R. Herz, Selective inspection planning with ageing forecast for sewer types, Water Sci. Technol., 46 (2002) 389–396.
[8] F. Chughtai, T. Zayed, Infrastructure condition prediction models for sustainable sewer pipelines, J. Perform. Constr. Facil., 22 (2008) 333–341.
[9] E.V. Ana, W. Bauwens, Modeling the structural deterioration of urban drainage pipes: The state-of-the-art in statistical methods, Urban Water J., 7 (2010) 47–59.
[10] D.H. Tran, A.W.M. Ng, B.J.C. Perera, S. Burn, P. Davis, Application of probabilistic neural networks in modelling structural deterioration of stormwater pipes, Urban Water J., 3 (2010) 175–184.
[11] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: Linear models. PLS-DA, Anal. Meth., 5 (2013) 3790–3798.
[12] M. Alvarez-Guerra, D. Ballabio, J.M. Amigo, J.R. Viguri, R. Bro, A chemometric approach to the environmental problem of predicting toxicity in contaminated sediments, J. Chemometrics, 24 (2010) 379–386.
[13] M. Alvarez-Guerra, D. Ballabio, J.M. Amigo, R. Bro, J.R. Viguri, Development of models for predicting toxicity from sediment chemistry by partial least squares-discriminant analysis and counter-propagation artificial neural networks, Environ. Pollut., 158 (2010) 607–614.
[14] D. Ballabio, M. Vasighi, A MATLAB toolbox for self organizing maps and supervised neural network learning strategies, Chemomet. Intell. Lab. Syst., 118 (2012) 24–32.
[15] J. Zupan, M. Novic, I. Ruisánchez, Kohonen and counterpropagation artificial neural networks in analytical chemistry, Chemomet. Intell. Lab. Syst., 38 (1997) 1–23.
[16] W. Melssen, R. Wehrens, L. Buydens, Supervised Kohonen networks for classification problems, Chemomet. Intell. Lab. Syst., 83 (2006) 99–113.
[17] D. Ballabio, M. Vasighi, V. Consonni, M. Kompany-Zareh, Genetic algorithms for architecture optimisation of counter-propagation artificial neural networks, Chemomet. Intell. Lab. Syst., 105 (2011) 56–64.
[18] Modeling analysis report on XP-SWMM operated under the Chemomet. Intell. Lab. Syst. (BTL) scheme for sewer systems in Jinju City, Quarterly report for the financial year 2015–2017, Jinju City, Korea, 2018.
[19] 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.
[20] T. Kohonen, Self-organizing Maps, 3rd ed., Springer Series in Information Sciences, Vol. 30, Springer-Verlag, Berlin, Heidelberg, New York, USA, 2001, p. 502.
[21] 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.
[22] 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.