Assessment through Statistical Methods of Water Quality Parameters (WQPs) in the Han River in Korea

Jae Hyoun Kim†
Department of Health Science, School of Natural Science, Dongduk Women’s University, Seoul 136-714 Korea

ABSTRACT

Objective: This study was conducted to develop a chemical oxygen demand (COD) regression model using water quality monitoring data (January, 2014) obtained from the Han River auto-monitoring stations.

Methods: Surface water quality data at 198 sampling stations along the six major areas were assembled and analyzed to determine the spatial distribution and clustering of monitoring stations based on 18 WQPs and regression modeling using selected parameters. Statistical techniques, including combined genetic algorithm-multiple linear regression (GA-MLR), cluster analysis (CA) and principal component analysis (PCA) were used to build a COD model using water quality data.

Results: A best GA-MLR model facilitated computing the WQPs for a 5-descriptor COD model with satisfactory statistical results ($r^2=92.64, Q^2_{LOO}=91.45, Q^2_{Ext}=88.17$). This approach includes variable selection of the WQPs in order to find the most important factors affecting water quality. Additionally, ordination techniques like PCA and CA were used to classify monitoring stations. The biplot based on the first two principal components (PCs) of the PCA model identified three distinct groups of stations, but also differs with respect to the correlation with WQPs, which enables better interpretation of the water quality characteristics at particular stations as of January 2014.

Conclusion: This data analysis procedure appears to provide an efficient means of modelling water quality by interpreting and defining its most essential variables, such as TOC and BOD. The water parameters selected in a COD model as most important in contributing to environmental health and water pollution can be utilized for the application of water quality management strategies. At present, the river is under threat of anthropogenic disturbances during festival periods, especially at upstream areas.

Keywords: CA, GA-MLR, Han River, PCA, Water quality parameter (WQP)

I. Introduction

Water quality can be thought of as a measure of the suitability of water for a particular use in water and potential water quality hazards or risk scenarios that can affect the WQP poses a risk to the health and safety of the users. Moreover, a Master Plan for Water Environment Management (2006-2015) by the Ministry of Environment - Korea Environment Institute was established in September 2006, as the follow-up of “Comprehensive Water Quality Improvement Measures on the Four Major Rivers” for the implementation of comprehensive water quality management.

The BOD and COD are the total measurement of all chemicals and is used widely to determine the amounts of organic pollutants in wastewater. The COD test is an accurate quantitative test for water quality prediction model for persistent organic compounds. Hence COD test is a better estimate of organic matter than the biochemical oxygen demand (BOD) test. The upward trend of COD was observed that there has been an increase in nonbiodegradable organic matters in Han River.

In Korea, the seasonal variability of rainfall is relatively high and there is low precipitation and
river flow during every dry season. The WQPs have lower concentrations in the dry season than the wet season. And the January samples represent the dry season samples without significant temperature variation. Excess nutrient and pollutant loadings are linked to serious problems in the downstream river aquatic system.4-7)

Although concentrations of BOD and COD showed a consistent decreasing trend for past years due to increasing water treatment facilities and decreasing input of non-degradable organic pollutants.8,9) the water quality of COD in the Han River recently showed an upward trend at more than 78% of monitoring stations. The upward trend of COD contrary to the BOD trend indicates that there has been an increase in nonbiodegradable organic matter and phosphorus concentrations in the upstream of the Han River that is not detectable by means of BOD.3,10) Hence, major tributaries of the Han River and characteristics of water quality for each target streams needed to be compared based on inter-annual water quality data.11)

In previous studies, the relationship between water quality parameters has been carried out. Significant linear relationships between TOC and COD and TOC were reliably estimated for the generic replacement of both COD in river12) and lake.13) The correlation of BOD and COD to TOC was r=0.772 and r=0.810 on the Nakdong river stream.14) Organic components influence correlation between COD and TOC.15) Very few papers attempted to describe the pollution indicator model for the Han River.

The Artificial Neural Network (ANN) technique is known to be better than the MLR model in estimating the COD.16) The ANN model was used for real-time water quantity and quality management, because it is easy to construct and adjusts well to changes through learning.17) Nevertheless, a predictive water quality model for chlorophyll-a was demonstrated to produce the best performance to develop the MLR model using GA.18) Furthermore, combined PCA/GA-MLR is likely to be a comparable method to identify potential pollution sources to the Han River. No attempt was made to develop the COD GA-MLR model of the Han River yet.

The objectives of this study are: 1) to develop a predictive COD model used to represent the overall level of organic contamination of the Han River water system as one of a primary water quality concern, using GA-MLR approach, and 2) to interpret the monitoring data for water quality assessment and quality management decisions.

![Fig. 1. Map of study area with the location of water quality monitoring stations.](http://www.kseh.org/)
II. Materials and Methods

1. Study area and data
The Han River basin is the home of 24 million people, including the densely populated Seoul metropolitan area and a major source of drinking, irrigation, industrial and recreational water. Some tributary stations show degrading water quality as a result of a large population growth and inadequate wastewater treatment facilities.\(^{19,20}\) The average temperature range throughout the river varies from -0.8 to 11.8°C in January in the dry season.

The final data set (i.e., the dataset with no missing values) for the samples in 2014 (January) includes 198 water quality monitoring sites (Fig. 1), consisting of 18 WQPs monitored (Table 1), were selected in the present analysis.

2. Statistical methods
GA-MLR analysis for variable selection was performed by the MobyDigs 1.1 package (TALETE srl-Milano, Italy). Principal component analysis (PCA) and cluster analysis (CA) of water quality data sets were made through XLStat-pro.\(^{21}\)

Regression analysis (GA-MLR). To perform multi-linear regression, GA was used to select, from among all the water quality data, the most relevant in obtaining models that yielded the highest predictive power for the response. The fitness statistics provided by the bootstrapping method are indicative of the actual model's predictivity. Model performance was described by means of descriptors related to model predictive capability (\(Q_{2\text{LOO}}\) and \(Q_{2\text{boot}}\)) and fitting power (\(r^2\)). To avoid models with multi-collinearity without prediction power, the search for the best models was performed using the QUIK rule.\(^{22}\) Moreover, to avoid models with chance correlation, the prediction capability of the best models was checked using the Y-scrambling technique. Only models with positive differences \(K_{XY} – K_{XX}\) was chosen as a best MLR. The Hat value is the measure of leverage to verify the applicability domain.\(^{23}\) Influential compounds were those with a leverage greater than the critical value (warning leverage) \(h^* = 3p'/n\), where \(p'\) is the number of model variables plus one, and \(n\) the number of objects used to calculate the model. A hat value greater than the warning leverage (\(h^*\)) means unreliable. The application of GA provided 100 comparable models with similar predictive

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Unit</th>
<th>N</th>
<th>Min.</th>
<th>Max.</th>
<th>Median</th>
<th>75th Percentile</th>
<th>95th Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td></td>
<td>198</td>
<td>6.5</td>
<td>8.8</td>
<td>7.6</td>
<td>7.9</td>
<td>8.22</td>
</tr>
<tr>
<td>Temp Water Temperature</td>
<td>°C</td>
<td>198</td>
<td>-0.8</td>
<td>11.8</td>
<td>2.3</td>
<td>4.3</td>
<td>8.72</td>
</tr>
<tr>
<td>DO Dissolved Oxygen</td>
<td>mg/L</td>
<td>198</td>
<td>4.8</td>
<td>22</td>
<td>14</td>
<td>15.1</td>
<td>16.63</td>
</tr>
<tr>
<td>BOD Biochemical Oxygen Demand (5 d)</td>
<td>mg/L</td>
<td>198</td>
<td>0.2</td>
<td>18.3</td>
<td>1.1</td>
<td>3.4</td>
<td>9.22</td>
</tr>
<tr>
<td>COD Chemical Oxygen Demand</td>
<td>mg/L</td>
<td>198</td>
<td>0.6</td>
<td>24.4</td>
<td>2.8</td>
<td>6</td>
<td>12.02</td>
</tr>
<tr>
<td>TOC Total Organic Carbon</td>
<td>mg/L</td>
<td>198</td>
<td>0</td>
<td>22.6</td>
<td>1.8</td>
<td>4.58</td>
<td>7.83</td>
</tr>
<tr>
<td>SS Suspended Solids</td>
<td>mg/L</td>
<td>198</td>
<td>0.2</td>
<td>102.1</td>
<td>2.1</td>
<td>5.18</td>
<td>23.67</td>
</tr>
<tr>
<td>T-N Total Nitrogen</td>
<td>mg/L</td>
<td>198</td>
<td>0.9</td>
<td>28.65</td>
<td>3.98</td>
<td>7.71</td>
<td>14.55</td>
</tr>
<tr>
<td>T-P Total Phosphorus</td>
<td>mg/L</td>
<td>198</td>
<td>0</td>
<td>2.19</td>
<td>0.03</td>
<td>0.11</td>
<td>0.53</td>
</tr>
<tr>
<td>DTN Dissolved Total Nitrogen</td>
<td>mg/L</td>
<td>198</td>
<td>0.82</td>
<td>26.21</td>
<td>3.8</td>
<td>7.43</td>
<td>13.68</td>
</tr>
<tr>
<td>DTP Dissolved total Phosphorus</td>
<td>mg/L</td>
<td>198</td>
<td>0</td>
<td>2.07</td>
<td>0.02</td>
<td>0.07</td>
<td>0.4</td>
</tr>
<tr>
<td>NH3-N Ammonia-Nitrogen</td>
<td>mg/L</td>
<td>198</td>
<td>0</td>
<td>21.22</td>
<td>0.13</td>
<td>1.53</td>
<td>8.14</td>
</tr>
<tr>
<td>NO3-N Nitrate-Nitrogen</td>
<td>mg/L</td>
<td>198</td>
<td>0.18</td>
<td>11.46</td>
<td>2.79</td>
<td>3.97</td>
<td>6.86</td>
</tr>
<tr>
<td>PO4-P Orthophosphate</td>
<td>mg/L</td>
<td>198</td>
<td>0</td>
<td>2.01</td>
<td>0.01</td>
<td>0.05</td>
<td>0.35</td>
</tr>
<tr>
<td>EC Electrical Conductivity</td>
<td>µS/cm</td>
<td>198</td>
<td>59</td>
<td>29315</td>
<td>259</td>
<td>491.75</td>
<td>1401.7</td>
</tr>
<tr>
<td>F-Coli Fecal Coliform</td>
<td>CFU/100 ml</td>
<td>198</td>
<td>0</td>
<td>56000</td>
<td>37</td>
<td>1400</td>
<td>12150</td>
</tr>
<tr>
<td>T-Coli Total Coliform</td>
<td>CFU/100 ml</td>
<td>198</td>
<td>0</td>
<td>380000</td>
<td>570</td>
<td>15750</td>
<td>79250</td>
</tr>
<tr>
<td>Chl Chlorophyll a</td>
<td>µg/L</td>
<td>198</td>
<td>107.1</td>
<td>2.6</td>
<td>7.78</td>
<td>24.92</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. List of 18 WQPs and average monthly data (Max, Ave and Min values) of WQPs in January 2014.
Multivariate analysis. Hierarchical cluster analysis (CA) was performed on the standardized data using Ward’s method with squared Euclidean distances as a measure of similarity. Variations in water quality are determined from hierarchical CA using the linkage distance. The dendrogram indicates pollution status as well as the effect of contamination at the sampling sites. It provides a visual summary of the clustering processes, presenting a picture of the groups and their proximity. Principal component analysis (PCA) was carried out to extract the most important factors and physicochemical parameters affecting the water quality.

### III. Results

1. Descriptive statistics of data

   Descriptive statistics including maximum, minimum, 75th and 95th percentile, median and range values of proposed monitoring stations according to WQPs are given in Table 1.

   Results of analysis of the water samples collected from 198 monitoring stations are given in Table 1. The river presents in monthly average, slight alkaline pH as 7.6 in the range of 6.5 and 8.8. The results showed that the DO concentration ranged from 4.8-22 mg/L, the COD concentrations varied from 0.6 to 24.4 mg/L, the BOD values ranged from 0.2 to 18.3 mg/L, the ammonia nitrogen (NH₃-N) concentrations varied from 0-21.22 mg/L, the concentrations of T-P ranged from 0-2.19 mg/L, concentration of T-N varied from 0.904-28.65 mg/L, and TOC ranged from 0-22.6 mg/L. The average value of the conductivity was calculated as about 761 μS/cm. The overall median value of F-coli was 37 MPN/100 mL and the highest was 56,000 MPN/100 mL, whereas T-coli had 570 MPN/100 mL and 380,000, MPN/100 mL respectively. Both parameters showed large standard deviations. The values of chlorophyll obtained, ranging from 0 and 107.1 mg/L.

2. The best GA-MLR model for COD

   Validation of the Final 5-descriptor COD Model. COD model validation implies quantitative assessment of model robustness and their predictive power to decide acceptability of a COD model. The predictive power is defined as its ability to predict accurately the biological activity of new chemical substances. The validation criteria for a COD model have already been proposed and discussed previously.öh

   Statistical parameters used to assess a COD model are shown in Table 2. The validation of a proposed 5-descriptor COD model includes the following steps: 1) Diagnostic statistics, 2) Internal validation, 3) External validation and 4) Applicability.

   1) Diagnostic statistics

   (i) n/p ratio: n ≥ 4 p, where n is the number of data points and p is the number of descriptors used in the COD model. The COD model obeys the ‘rule of thumb’.

   (ii) Fraction of the variance ($r^2$): A COD model is acceptable when it has a $r^2 > 0.6$ units (60%) and will be considered for validation. The % value of $r^2$ for the COD model is 92.64.

   (iii) Cross-validation Test ($Q^2_{LOO}$): A COD model with a value of $Q^2_{LOO}$>0.5 units (50%) is normally considered to possess significant predictive ability. The value of $Q^2_{LOO}$ for this COD model is 91.45.

   (iv) Standard Error (SE): The smaller, as possible, SE value (SE < 0.4 units) is always necessary for the predictive COD model. The value of SE for COD model is 1.01.

   (v) $r^2Q^2_{LOO} < 0.3$ units (30%): For good predictability,
this difference between $r^2$ and $Q^2_{LOO}$ for a COD model should never be exceeded by 30%. A large difference between $r^2$ and $Q^2_{LOO}$ suggests the following: (a) over-fitted model, (b) presence of outliers, or (c) presence of irrelevant variables in the data set. As can be seen from the statistics, the small difference (1.19%) between the $r^2$ and $Q^2_{LOO}$ values, and the difference (0.18%) between the $r^2$ and $r^2_{adj}$ for this model avoid an over-fitting in the regression.

(vi) Quality Factor (Q): Over fitting and chance correlation, due to excess number of descriptors, can be detected by Q value. Q is quality factor, where Q = $r/SE$ (here $r$ is correlation coefficient and SE is standard deviation). A high value of Q (9.53) for the COD model suggests its high predictive power and the lack of over-fitting.

(vii) Fischer Statistics (F): A high F-value (422.77) indicates that the model is statistically significant at the 95% level.

2) Internal Validation

(i) Cross-validations Test: LOO Cross-Validations are the most commonly used techniques for internal validation to ensure that the final model isn't biased. A high value (> 0.5) of $Q^2_{LOO}$ is often considered as an ultimate proof for the high predictive power of COD model. The value is 91.45.

(ii) Bootstrapping: To avoid the danger of overfitting and the possibility of overestimating the model predictivity by using $Q^2_{LOO}$ and $Q^2_{Train}$, the internal predictive ability of the models was also verified using the bootstrap $Q^2_{Boot}$ procedure. By this technique, validation is performed by randomly generating training sets with sample repetitions and then evaluating the predicted responses of the samples not included in the training set. High $Q^2_{Train}$ (> 0.5 units) is considered as an indicator of the highly predictive power of a COD model. The value of $Q^2_{Train}$ (90.51) is fairly close to $Q^2_{LOO}$ (91.45), implying that the equation is proper representative of the group of analogs. In view of these observations, we conclude that the COD model of Equation 1 is fairly robust.

(iii) Collinearity among variables and redundancy of information: COD data sets often contain redundancy (exact linear dependencies between subsets of the variables), and multi-collinearity (high multiple correlations between subsets of the variables). Both of these features inhibit the development of a COD model with the ability to generalize successfully to new objects.

Multi-collinearity: Applying the QUIK rule, the difference in the correlation between the block of X variables plus the response Y ($K_{XX} = 55.54$) and that of X ($K_{XX} = 45.54$) is sufficiently high (DK = 10).

Chances of redundancy are checked in correlation matrix in Table 2. All the nondiagonal elements were less than 0.53, indicating that the co-linear situation among different descriptors and redundant information included in the set of descriptors are low. It is clear that the descriptors used in the equation are almost independent.

Y-randomization (Scrambling) Test: Only high correlation coefficient is not enough to select the equation as a model and hence various statistical approaches were used to confirm the robustness and practical applicability of the equations. Y-scrambling techniques are employed to exclude the possibility of chance correlation and to check for reliability and robustness by permutation testing. There are no chance correlations if they satisfy the criteria $a(r^2) < 0.3$ and $a(Q^2) < 0.05$. Internal validation resulted in one or more measures of robustness of the model descriptors such as Y-scrambling [$a(r^2) = – 0.015$, $a(Q^2) = – 0.084$] and internal predictability such as $Q^2_{Boot}$ (90.51).

3) External validation

External validation refers to a validation exercise in which the chemical structures selected for inclusion in the validation (or test) set are different from those included in the training set. To validate the predictive power of the mathematical model more explicitly one needs to conduct validation on the external set of data.

The external validation of a COD model was carried out in two steps: (i) Random selection (Training set), and (ii) Predictive power of the COD model.

(i) Random selection (Training set): The original (whole) data set is randomly split into training (70%) and test (30%) sets. The COD model for the resulting training set is then generated by using the same descriptors as those of the original equation and validated on the basis of acceptance criteria ($r^2 > 0.6$ and $Q^2 > 0.5$ units).

(ii) Predictive power of a COD model: The true predictive power of a COD model is determined by
comparing the predicted and experimental activities of the test set compounds that are not used in the COD model development of training set.

For 52 external test set compounds, the prediction capability of the model combined conclusively proves the good predictive potency of the quantitative relationship constructed on the GIR activity. The prediction capability of the model in terms of explained variance ($Q^2_{\text{Ext}}$) and External Standard Deviation Error of Prediction ($\text{SDEP}_{\text{Ext}}$), evaluated including only those test data with reliable predictions according to the leverage approach, is satisfactory ($N_{\text{Ext}} = 52, r^2 = 92.64,$ $Q^2_{\text{Ext}} = 88.17, \text{SDEP}_{\text{Ext}} = 1.16$). The value of $Q^2_{\text{Ext}}$ is smaller than $Q^2_{\text{LOO}}$ (91.45), and not too much different. Thus the value of 88.17 for $Q^2_{\text{Ext}}$ is acceptable.

4) Applicability domain (AD)

The Williams plot of residual vs. leverage is shown in Fig. 3. The plot is important for the predictive ability of COD. The Williams plot (scatter) depicts the standardized cross-validated residuals (RES) versus leverage values (h), and is used to obtain a graphical detection of both the response outliers (Y outliers) and the structurally influential chemicals (X outliers) of a model. Compound with cross-validated standardized residual greater than 2.5$\sigma$ is recognized as Y outlier.

The warning leverage ($^*h$), was found to be, 0.13 for the developed COD model. The plot displays that compounds 1022A50 and 1023A25 had the largest leverage values (0.560 and 0.7) $> h^* (0.130)$. However, the influence of this compound is not critical for the model, and it has been retained. However, 1019A40 and 1022A55 were shown to be critical outliers exceeding $\pm$ 3 std. err. (Ypred.-Yexp), which affect intercept and slope.

The results of the best COD model with the highest $Q^2_{\text{LOO}}$ values and the lowest values of PRESS and SE using 5 descriptors based on GA-MLR are given in Table 2.

3. Multivariate analysis

1) Correlation matrix

The correlation matrix for different water quality variables is depicted in Table 3. It is evident that distribution of BOD, COD, TOC, T-N, NH$_2$N, T-P, F-coli, T-Coli, DTN, DTP and PO$_4$-P were significantly correlated ($r > 0.5$) with COD in most of the study stations. Highly positive correlation is observed between T-N and DTN (0.996), between T-P and DTP (0.968), between PO$_4$-P and DTP (0.964), between PO$_4$-P and T-P (0.934) while highly negative correlation coefficient is seen among pH and DO and other variables.
The results of CA based on station scores of PCA for the three regions are shown in a dendrogram (Fig. 4), showing clustering of the sampling stations. The dissimilarity defined by Euclidean distance and the combination of cluster is based on Ward method.

### Table 3. Correlations matrix (r) between WQPs (p<0.05)

<table>
<thead>
<tr>
<th>Variables</th>
<th>pH</th>
<th>DO</th>
<th>BOD</th>
<th>COD</th>
<th>TOC</th>
<th>SS</th>
<th>T-N</th>
<th>NH₃-N</th>
<th>NO₃-N</th>
<th>T-P</th>
<th>Temp</th>
<th>EC</th>
<th>F-Coli</th>
<th>T-Coli</th>
<th>DTN</th>
<th>DTP</th>
<th>PO₄-P</th>
<th>Chl</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>1</td>
<td>-0.009-0.145-0.089-0.119 0.029 0.269 0.238 0.122 0.182 0.078 0.023 -0.109 0.206 0.272 0.188 0.198 0.067 0.967</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DO</td>
<td>-0.009 1</td>
<td>-0.389-0.357-0.351 0.033 0.312 0.321 -0.095-0.425 0.620 0.128 -0.335 0.435 0.316 0.417 0.406 0.031</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOD</td>
<td>-0.145-0.389 1</td>
<td>0.859 0.744 0.405 0.627 0.580 0.235 0.616 0.426 0.106 0.653 0.576 0.629 0.515 0.437 0.330</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COD</td>
<td>-0.089 -0.357 0.859 1</td>
<td>0.880 0.464 0.729 0.710 0.204 0.571 0.402 0.134 0.505 0.532 0.729 0.422 0.374 0.553</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOC</td>
<td>-0.119-0.351 0.744 0.880 1</td>
<td>0.350 0.681 0.656 0.252 0.467 0.307 0.078 0.366 0.448 0.687 0.295 0.269 0.461</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>0.029-0.033 0.405 0.464 0.350 1</td>
<td>0.199 0.207 0.023 0.254 0.094 0.228 0.262 0.159 0.197 0.155 0.136 0.225</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-N</td>
<td>-0.269-0.312 0.627 0.729 0.681 0.199 1</td>
<td>0.900 0.419 0.416 0.392 0.033 0.345 0.413 0.996 0.347 0.286 0.212</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NH₃-N</td>
<td>-0.238-0.321 0.580 0.710 0.656 0.207 0.900 1</td>
<td>0.032 0.299 0.427 0.046 0.230 0.335 0.902 0.222 0.207 0.235</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NO₃-N</td>
<td>-0.122-0.095 0.255 0.204 0.252 0.023 0.419 0.032 1</td>
<td>0.391 0.004-0.067 0.287 0.302 0.422 0.400 0.298 0.039</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-P</td>
<td>-0.182-0.425 0.616 0.571 0.467 0.254 0.416 0.299 0.391 1</td>
<td>0.394 0.069 0.587 0.849 0.414 0.968 0.934 0.169</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temp</td>
<td>-0.078-0.620 0.426 0.402 0.307 0.094 0.392 0.427 0.004 0.394 1</td>
<td>0.123 0.333 0.418 0.393 0.405 0.374-0.011</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EC</td>
<td>0.023-0.128 0.106 0.134 0.078 0.228 0.033 0.046-0.067 0.069 0.123 1</td>
<td>-0.010 0.001 0.010 0.057 0.059 0.268</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F-Coli</td>
<td>-0.109-0.335 0.653 0.505 0.366 0.262 0.345 0.230 0.287 0.587 0.333-0.010 1</td>
<td>0.709 0.342 0.559 0.461 0.104</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-Coli</td>
<td>-0.206-0.435 0.576 0.532 0.448 0.159 0.413 0.335 0.302 0.849 0.418 0.001 0.709 1</td>
<td>0.417 0.828 0.814 0.106</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DTN</td>
<td>-0.272-0.316 0.629 0.729 0.687 0.197 0.996 0.902 0.422 0.414 0.393 0.010 0.342 0.417 1</td>
<td>0.346 0.288 0.205</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DTP</td>
<td>-0.188-0.417 0.515 0.422 0.295 0.155 0.347 0.222 0.400 0.968 0.405 0.057 0.559 0.828 0.346 1</td>
<td>0.964 0.055</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PO₄-P</td>
<td>-0.198-0.406 0.437 0.374 0.269 0.136 0.286 0.207 0.298 0.934 0.374 0.059 0.461 0.814 0.288 0.964 1</td>
<td>0.050</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chl</td>
<td>0.067-0.031 0.330 0.553 0.461 0.225 0.212 0.235 0.039 0.169-0.011 0.268 0.104 0.106 0.205 0.055 0.050 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figures in bold are higher than 0.5.

### Table 4. Four major sources of B and C group in dendrogram.

<table>
<thead>
<tr>
<th>Cluster B</th>
<th>Cluster C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Station</td>
<td>Name</td>
</tr>
<tr>
<td>1017A50</td>
<td>Bokhacheon 2</td>
</tr>
<tr>
<td>1018A24</td>
<td>Wangsookcheon 3</td>
</tr>
<tr>
<td>1018A26</td>
<td>Wangsookcheon 4</td>
</tr>
<tr>
<td>1018A40</td>
<td>Tancheon 1</td>
</tr>
<tr>
<td>1018A42</td>
<td>Tancheon 2</td>
</tr>
<tr>
<td>1018A50</td>
<td>Tancheon 5</td>
</tr>
<tr>
<td>1018A72</td>
<td>Anyangcheon 1</td>
</tr>
<tr>
<td>1018A74</td>
<td>Anyangcheon 1-1</td>
</tr>
<tr>
<td>1018A75</td>
<td>Anyangcheon 2</td>
</tr>
<tr>
<td>1018A78</td>
<td>Mokgancheon</td>
</tr>
<tr>
<td>1018A80</td>
<td>Anyangcheon 4</td>
</tr>
<tr>
<td>1018A84</td>
<td>Anyangcheon 5</td>
</tr>
</tbody>
</table>

2) CA

The results of CA based on station scores of PCA for the three regions are shown in a dendrogram (Fig. 4), showing clustering of the sampling stations. The dissimilarity defined by Euclidean distance and the combination of cluster is based on Ward method. The dendrogram, displays all monitoring sites to be grouped in three statistically significant clusters; A (low pollution), B and C (high pollution).

Stations in cluster B and C on the dendrogram are listed in Table 4. Higher number of stations dominantly, belongs to cluster B than cluster C. Four
3) Biplot

In order to detect the homogeneities in the data set and to recognize the potential outliers in all of the stations under study, the PCA was performed (Fig. 5). In the PCA biplot that included the 18 water quality variables, the first principal component explained 4.37% of the variation, respectively and was negatively correlated with the two variables (DO and pH) and positively with other 16 variables. The second principal component explained 14.05% of the variation.

It is readily seen that a good homogeneity of the water quality for all sites is found. The station 1019A40 and 1018A30, 1019A35, 1022A50, 1022A55 are separated from other stations as extreme values. The distribution of stations over the region shows similar trend in the degree of relatedness as for the variable plot but behaves individually.

Station loading scores exhibited a fairly distinct spatial pattern along gradient with sampling stations, showing two major clusters (upper right and lower right) a dispersion from center to lower right (DTN, T-N, NH₃-N, NO₂-N, Chl, COD, SS) and upper right (F-Coli, T-Coli, T-P, DTP, PO₄-P and Temp). F-Coli and T-Coli, however, had a high correlation with DTP, T-P and PO₄-P. Moreover, plots of PC1 and PC2 axes show a good separation among subgroups, namely, group 1 (PO₄-P, DTP and T-P), 2 (TOC, BOD and COD), 3 (DTN, T-N and NH₃-N) and 4 (DO and pH) and 5 group (T-Coli, F-Coli).

IV. Discussion

The best model obtained as described in a Results section was a GA-MLR equation containing 5 descriptors considering 198 (N) data points (Table 2). This model explained 92.42% of the variance (adjusted coefficient of variation, $r_{adj}^2$) and a predicted 91.45% of the variance (LOO predicted
variance, $Q^2$). Validation of the model was performed using separation of the data into two independent sets, Y-randomization, cross-validation and bootstrap. The results indicated that the constructed GA–MLR model is a valid model with high statistical quality and low prediction errors. External predicted variance for this model was 88.17% ($r^2_{\text{pred}}$). In this work external predictivity was used to select the best model. The other statistics satisfied statistical interpretation. Scatter plots of observed vs. calculated or predicted values of the training and test set compounds are observed (Fig. 2.). The stations determined as potential outliers (1018A26, 1018A80 and 1022A55) and extreme points (1023A25 and 1022A50), i.e., outside of the AD, were confirmed in a comparative AD assessment by using the leverage approach. The proposed model could identify and provide an insight into some suggestions for the further design of sampling stations.

A strong correlation was found between COD and TOC with ($r^2$=0.79), indicating that the COD test is useful for monitoring and process control in the Han River. Water quality in the Keum River, comparable to Han River, was evaluated, where the COD in the dry season was mainly attributed to Chl and phytoplankton contains most of TP.

When the values of $pH$ are between 6.4 - 8, the nutrients are in their ionic and dissolved forms. Thus ammoniac nitrogen (NH$_4^+$) is in ionic form dissolves non-toxic, favorable to the development of algae, also the phosphorus as PO$_4$-P. In general, the highest concentrations reach at the station of the downstream in a dry season. DO levels less than 3 mg/L are stressful to most aquatic organisms. All streams in this study showed high dissolved oxygen concentrations (greater than 8 mg/L) are considered healthy streams except one station. Higher conductivity results from the presence of various ions. A value, 444 ìS/cm, corresponds to the limit of mineralization, under which river is weakly mineralized. The average value of the conductivity is about 761 μS/cm.

The highest fecal/total coliform concentrations exceeding the prescribed limits (T-Coli/F-Coli effluent limit: <3,000 MPN/100 mL) were observed at Tancheon (1018A40, 1018A42, 1018A44 and 1018A46), Gulpocheon (1019A40), Jungangcheon (1018A64), Shincheon (1022A50, 1022A54) and Osancheon (1101A57, 1101A50, 1101A47) sampling stations. In urban watersheds, fecal indicator bacteria are significantly correlated with human density. All members of the total coliform group can occur in animal or human feces. There was a positive correlation between TN and TP, but neither correlated ($r^2=0.22$) in this study. High T-P/T-N ratios (>120) were observed at Paldang-dam (1017A10), Paldang (1018A02), Deonkpung (1018A06), Dogok (1018A08), Wangsookcheon (1018A201018A22, 1018A24), Guri (1018A28), Amsa (1018A32) Guey (1018A34), Cheonggye (1018A59, 1018A60, 1018A63), Anyangecheon stations (1018A72, 1018A74, 1018A75, 1018A76, 1018A80 and 1018A84) Han part sites (1022A07, 1022A10, 1022A37, 1022A40), indicating that the effluent from industrial activity, household sewage or sewage treatment plants have the potential ability to increase the T-P concentration. Agricultural activities do not seem to be relevant to the high ratio. As far as T-N and T-P loadings are regulated by streamow regimes, TP is more likely to be retained under low ow conditions in dry winter season. Hence, there is an urgent need for efficient water treatment, enhanced water use efficiency and waste water recycling to decrease concentrations.

As might be expected from nutrient limitation theory, in lakes or rivers with very high TN:TP ratios (> 50), TP does indeed become more closely related to Chl-a concentrations than TN. Though in the short run a continued reduction in nitrogen loads appears to be the more promising approach of eutrophication management, we recommend enhanced efforts to diminish both N and P emissions. Chlorophyll-a (Chl) quantification allows the determination of algal blooms in waters body. The concentration levels reflect a wealth of micro-algal biomass through a chlorophyll activity relatively dense at 1022A55 (Sincheon), 1019A25 (Haengju), 1201A30 (Aracheon) and 1022A35 (Yeongpyeongcheon 3 of Han part) areas with the greatest levels. However, low correlation between T-N, T-P and Chl were observed ($r^2$=0.057 for T-N/Chl ratio and $r^2$=0.042 for T-P/Chl). Phytoplankton density was higher during low-flow seasons. It seemed that hydraulic residence time was the major factor controlling phytoplankton. As was indicated on the biplot, 1018A40 have

http://www.kseh.org/
strong influence on phosphorus nutrients loading of anthropogenic activity rather than a geographical (natural). The dendrogram indicated that Tancheon, Anyangcheon and Wangsookcheon were the most contaminated sites. Point sources such as Anyangcheon, Tancheon and Bokhacheon involve problems of high population density, industrial discharge, livestock farming, and treatment effluent. Waste from industries, animal Farms and household wastewater are leading sources of pollution in Anyangcheon, Tancheon, Shincheon Wangsookcheon and Munsancheon generating high levels of BOD, COD, T-P, T-N, NH$_4^+$-N, NO$_3^-$-N and F-Coli. Tannery and dyeing effluent discharged from Dongducheon and Yangju is considered to be a major source of the high EC, COD and NO$_3^-$-N in Shincheon.

The CA technique applied to the water quality data was able to relate the water quality variables with the sources of contamination. However, in cluster analysis, the clustering and the number of existing clusters are only a qualitative statement. Therefore, as suggested earlier, cluster analysis should be confirmed in an additional step such as applying the principal component analysis (PCA) technique.

V. Conclusion

Nutrient contents and nutrient ratios such as TN, TP and Chl indicated that phosphorus was a potential limiting element for algal growth in many upstream stations in January, and indicated that upstream of the Han River still needed a control measure to reduce nutrient concentrations.

The hybrid genetic algorithm-multiple linear regression (GA-MLR) was implemented to build the best MLR (BMLR) COD models to select and compare the models. Many 1018A stations had greater levels of water quality data as major pollution sources, as commonly shown in PCA and GA-MLR.

Results showed that point (municipal and industrial effluents) and nonpoint sources (agricultural runoff) in both upstream and downstream are the main contributors to organic and nutrient parameters. The upward trend of COD by recalcitrant organic matter in in the upper regions of the Han River is still problematic in specific areas.

This study presents a useful interpretation of water quality data sets with a view to meeting better information about water quality guideline for more effective management of water resources in river basins. A recent study proposed that: (1) upstream water treatment directly cuts off incoming pollutants, thereby presenting the smallest variation in its downstream effects on BOD or COD levels, (2) treatment is advantageous when reliability of water quality is a primary concern, (3) dam discharge is a flexible tool, and may be used strategically during a low-flow season.

The future research for the statistical interpretation using GA-MLR and multivariate methods is to detect hidden factors responsible for the data structure as well as to reveal discriminating chemical parameters. Further evaluation of the long-term data considering spatial and seasonal variation may also be beneficial to forecast water quality of the Han River.

Acknowledgements

This work received funding from research grant programme of the Dongduk Women’s University.

References


