Applying a Multivariate Statistical Analysis Model to Evaluate the Water Quality of a Watershed

Edward Ming-Yang Wu¹, Shu-Lung Kuo²*

ABSTRACT: Multivariate statistics have been applied to evaluate the water quality data collected at six monitoring stations in the Feitsui Reservoir watershed of Taipei, Taiwan. The objective is to evaluate the mutual correlations among the various water quality parameters to reveal the primary factors that affect reservoir water quality, and the differences among the various water quality parameters in the watershed. In this study, using water quality samples collected over a period of two and a half years will effectively raise the efficacy and reliability of the factor analysis results. This will be a valuable reference for managing water pollution in the watershed. Additionally, results obtained using the proposed theory and method to analyze and interpret statistical data must be examined to verify their similarity to field data collected on the stream geographical and geological characteristics, the physical and chemical phenomena of stream self-purification, and the stream hydrological phenomena. In this research, the water quality data has been collected over two and a half years so that sufficient sets of water quality data are available to increase the stability, effectiveness, and reliability of the final factor analysis results. These data sets can be valuable references for managing, regulating, and remediating water pollution in a reservoir watershed. Water Environ. Res., 84, 2075 (2012).

KEYWORDS: multivariate statistical analysis, watershed, factor analysis, pollution factor, discriminant analysis.

doi: 10.2175/106143012X13415215906979

Introduction

Multivariate analysis, which is a branch of statistics, is capable of simultaneously analyzing the observation data concerning two or more variables. It is a useful tool to find the regularity among the complicated factors in an environmental system so that complex phenomena can be simplified to excerpt important relevant information for analyzing and discriminating test results in order to propose effective, responsive management and control policies. Some multivariate statistical techniques, such as cluster analysis, factor analysis, principal component analysis (PCA), and discriminant analysis have been widely used as unbiased methods in the analysis of water quality data, air pollution in urban environments, and waste recycling methods for obtaining meaningful information (Bengrain and Marhawa, 2003; Brown et al., 1996; Helena et al., 2000; Huang, 1988; Lin et al., 2010; Liu et al., 2003; Reghunath et al., 2002; Simeonov et al., 2003; Vega et al., 1998; Voncina et al., 2002; Wunderlin et al., 2001). The multivariate treatment of data is widely used to characterize and evaluate surface and freshwater quality, and it is useful for revealing temporal and spatial variations caused by natural and anthropogenic factors linked to seasonality (Helena et al., 2000; Reisenhofer et al., 1998; Vega et al., 1998). Arslan (2009) explored the potential capabilities of geographic-information-system-based joint multivariate statistical analyses for water quality assessment of the Porsuk River in Turkey.

In addition, the various multivariate statistical techniques, for example, cluster analysis, discriminant analysis, and PCA/factor analysis, can be implemented for interpreting complex databases to offer a better understanding of water quality in the study region (Singh et al., 2005). The multivariate treatment of environmental data, which is commonly used to characterize and evaluate surface and groundwater quality, is useful for detecting temporal and spatial variations caused by natural and anthropogenic factors. These techniques also permit identification of the possible factors that influence the water systems and cause variations in water quality. Hence, they are valuable tools for developing appropriate strategies for effective management of water resources (Farnham et al., 2002; Lambrakis et al., 2004; Mellinger, 1987; Mendiguchia et al., 2004; Simeonov et al., 2003; Simeonova et al., 2003; Singh et al., 2004).

Functions of reservoir watersheds are typically divided into four main types: (1) source of water treatment plant; (2) water used to maintain sustainable development of a water body; (3) recreation and leisure, including direct contact (swimming) and indirect contact (boating), in use of water; and (4) agricultural and other industrial use. At present, because of improper development and use of land, some reservoir watersheds in Taiwan have encountered the issue of eutrophication. This has a huge effect on water use and flooding prevention, as well as shortening the life of reservoirs.

The multivariate statistics in this study are able to identify the pollution characteristics of water quality in watersheds, as well as the distribution characteristics of each cluster. It can also validate data collected from field investigation. After the statistical model was investigated and verified, this study found that multivariate statistics are able to evaluate the effectiveness of a water management program. This study also incorporates the River Pollution Index (RPI) with multivariate statistics and provides a discussion concerning clusters that satisfy the objectives of the indicator system and management. From this, the study determines the information needed to make effective decisions about managing reservoir watersheds.

1 Department of Civil and Ecological Engineering, I-Shou University, Kaohsiung City, Taiwan.
2 Kelee Environmental Consultant Corporation, Kaohsiung City, Taiwan.
* Kelee Environmental Consultant Corporation, 6F.-2, No. 288–8, Sinya Road, Kaohsiung City 806, Taiwan; Tel: 886-7-8150815, ext. 161; Fax: 886-7-8150816; e-mail: singsuey@ms28.hinet.net.
Methodology

Categorizing Water Quality Parameters and Sampling Time. Water samples collected at the six monitoring stations in the Feitsui Reservoir (Taipei, Taiwan) watershed between May 2008 and December 2010 were analyzed using factor analysis, cluster analysis, and discriminant analysis to understand the correlation among the major water quality variables and the tendencies of pollution. Additionally, nine water quality parameters—water pH, temperature, dissolved oxygen, biochemical oxygen demand (BOD), suspended solids, anionic surfactant (MBAS), ammonia nitrogen (NH$_3$-N), total phosphorous, and chlorophyll—were selected in this study to manifest the completeness and diversity of water quality information for the watershed. All statistical analyses were carried out with SPSS for Windows, version 16.0.

The Application of RPI. In this paper, the water quality data collected at six water quality monitoring stations in the Feitsui Reservoir watershed was analyzed using factor analysis to reveal the common factors in water quality for these monitoring stations, and to investigate the mutual correlations between the nine selected water quality parameters. Additionally, the RPI implemented by the Environmental Protection Administration of Taiwan was combined with the water quality data to be analyzed using multivariate statistical analysis for categorizing the water quality monitoring stations into appropriate clusters. The results are expected to reveal the characteristics and extent of pollution for each cluster so that the spatial distribution of water pollution in the watershed can be evaluated to reflect the real difference of water quality for monitoring stations, and to develop an evaluation model suitable for accessing the characteristics of stream pollution in Taiwan, which will be a valuable reference for managing stream pollution. The aforementioned RPI can be used to determine stream pollution based on the four water quality parameters—suspended solids, BOD, dissolved oxygen, and NH$_3$-N—shown in Table 1. Based on RPI results, the stream pollution can be categorized into four classes: “unpolluted or slightly polluted”, “lightly polluted”, “moderately polluted”, and “seriously polluted”. Although RPI classification is a rapid and simple method for classifying stream pollution, this study takes into consideration that the stream pollution index is only applicable to determining the pollution levels of a stream section; it is not suitable for reservoir water quality measurement. The latter is affected by many factors, such as the upstream land use for growing tea and highland cold weather vegetables and applications of fertilizer and pesticides that will increase concentrations of total phosphorous or MBAS and chlorophyll (an important eutrophication index). In addition, pH and temperature are also important factors for evaluating the water quality of streams and reservoirs. Therefore, the above reasoning clearly shows that the water quality for the whole watershed of the reservoir, especially the reservoir’s downstream water quality, cannot be evaluated based on only the aforementioned four RPI parameters. Hence, one of the tasks pioneered in this study is the use of the nine water quality parameters for effectively evaluating the water quality of the whole watershed. The RPI established by the Taiwan Environmental Protection Administration is relatively applicable to evaluating streams not connected with reservoirs.

Study Area. This study analyzes the time series water quality data collected at the six water quality monitoring stations located in the Feitsui Reservoir watershed in northern Taiwan. The watershed, with a total area of 303 km$^2$, which covers all of the Pinglin District and portions of the Shuangxi District, Shiding District, and Xindian District of New Taipei City, is about 30 km from Taipei. The Feitsui Reservoir is the second largest reservoir in Taiwan. It is located in the upper region of the Shixi stream, which flows into the Xindian stream, and receives the effluents from six tributary streams flowing through the watershed (Hsieh et al., 2010). This reservoir has been developed with a single objective: to provide water to residents of the Taipei municipality. It is also the only reservoir in Taiwan with protected water sources. To lessen the burden of increasing development and recreational use, as well as deteriorating water quality, in 2001 the Feitsui Reservoir Administration prohibited water recreational activities in the watershed. Figure 1 shows the geographic location of the Feitsui Reservoir in northern Taiwan (water quality monitoring sites, w1 to w6).

Application of Multivariate Analysis. Factor Analysis. Factor analysis is a useful tool for extracting latent information, such as relationships between variables that are not directly observable (Einax et al., 1998). The original data matrix is decomposed into the product of a matrix of factor loadings and a matrix of factor scores plus a residual matrix. In general, by applying the eigenvalue-one criterion, the number of extracted factors is less than the number of measured characteristics. Thus the dimensionality of the original data space can be decreased by means of factor analysis. After rotation of the factor loading matrix (e.g., by varimax rotation, which involves scaling the loadings by dividing them by the corresponding communality), the factors can often be interpreted as origins or common sources. There is some confusion between PCA and factor analysis in the literature. PCA is a data reduction technique that aims to explain most of the variance in the data while reducing the number of variables to a few uncorrelated components. In contrast, the aim of factor analysis is to identify underlying factors that are responsible for the correlation among the variables. Both methods enable the identification of groups of variables or individuals.

Table 1—Classification of stream pollution.

<table>
<thead>
<tr>
<th></th>
<th>Unpolluted or slightly polluted</th>
<th>Lightly polluted</th>
<th>Moderately polluted</th>
<th>Seriously polluted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dissolved oxygen (mg/L)</td>
<td>&gt;6.5</td>
<td>4.6–6.5</td>
<td>2.0–4.5</td>
<td>&lt;2.0</td>
</tr>
<tr>
<td>BOD (mg/L)</td>
<td>&lt;3.0</td>
<td>3.0–4.9</td>
<td>5.0–15</td>
<td>&gt;15</td>
</tr>
<tr>
<td>Suspended solids (mg/L)</td>
<td>&lt;2.0</td>
<td>20–49</td>
<td>50–100</td>
<td>&gt;100</td>
</tr>
<tr>
<td>NH$_3$-N (mg/L)</td>
<td>&lt;0.5</td>
<td>0.5–0.99</td>
<td>1.0–3.0</td>
<td>&gt;3.0</td>
</tr>
<tr>
<td>Point record</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>Average</td>
<td>&lt;2.0</td>
<td>2.0–3.0</td>
<td>3.1–6.0</td>
<td>&gt;6.0</td>
</tr>
</tbody>
</table>
In factor analysis, each variable, that is, $x_1$–$x_p$, in a set of $p$ variables is decomposed into $q$ common factors, that is, $f_1$–$f_p$ ($q \leq p$), which are linearly combined with special factors $\xi_i$.

The model of factor analysis can be represented by

\[
X_1 = \mu_1 + \ell_{11} f_1 + \ell_{12} f_2 + \ldots + \ell_{1q} f_q + \xi_1 \\
X_2 = \mu_2 + \ell_{21} f_1 + \ell_{22} f_2 + \ldots + \ell_{2q} f_q + \xi_2 \\
\vdots \\
X_i = \mu_i + \ell_{i1} f_1 + \ell_{i2} f_2 + \ldots + \ell_{iq} f_q + \xi_i \\
\vdots \\
X_p = \mu_p + \ell_{p1} f_1 + \ell_{p2} f_2 + \ldots + \ell_{pq} f_q + \xi_p
\]

(1)

Where $f_1$,...,$f_q$ are common factors contained in each variable $x_i$; $\xi_i$ is a special factor contained only in the $i$th variable ($x_i$); and $\ell_{ij}$ is the loading of $i$th factor to the $j$th common factor ($f_j$).

Because of considerations of efficiency, this study selects eigenvalue $>1$. The researchers in this study set up the percentage threshold of cumulative variance selected as higher than 60%.

Cluster Analysis. Cluster analysis (Vialle et al., 2011) was used to search for natural groupings among objects to discover latent structures present in the data set. Sun et al. (2011) attempted to develop a genetic algorithm aided stepwise cluster analysis method to describe the nonlinear relationships between the selected state variables and the carbon:nitrogen ratio in food waste composting. The parameters analyzed were sorted into groups, or clusters, so that the degree of association is strong between members of different clusters. Prior to cluster analysis, the descriptor variables were block-standardized by range to avoid effects of scale or units on the distance measurements.

Hierarchical agglomerative cluster analysis was then performed on the normalized data set with the Ward’s method, using Euclidean distances as a measure of similarity. Each cluster thus describes, in terms of the data collected, the class to which its members belong, and this description may be abstracted by changing it from the particular to the general class or type. Hierarchical agglomerative clustering is the most common approach to provide intuitive similarity relationships between any sample in the cluster and the entire data set, as is typically illustrated by a dendrogram (tree diagram) (Einax et al., 1998; McKenna, 2003). The dendrogram provides a visual summary of the clustering processes by presenting a picture of the groups and their proximity with a dramatic reduction in dimensionality of the original data. The root-mean-square standard deviation (RMSSTD) method that is the integrated standard deviations of all variables is used in this study as the evaluation index. The RMSSTD is defined as

\[
\text{RMSSTD} = \sqrt{\frac{\sum_{i=1}^{p} s_i^2}{p}}
\]

(2)

where $s_i$ is the sum of all standard deviations for the $i$th variable in the various clusters and $p$ is the number of variables. A smaller RMSSTD value means higher similarity for members in a cluster.

Discriminant Analysis. Discriminant analysis undertakes the same task as multiple linear regression by predicting an outcome. However, multiple linear regression is limited to cases where the dependent variable on the $Y$ axis is an interval variable.
so that the combination of predictors will, through the regression equation, produce estimated mean population numerical Y values for given values of weighted combinations of X values. But many interesting variables are categorical, such as political party voting intention; migrant/nonmigrant status; making a profit or not; holding a particular credit card; owning, renting, or paying a mortgage for a house; employed/unemployed; satisfied versus dissatisfied employees; which customers are likely to buy a product or not buy; whether a person is a credit risk or not; etc.

Discriminant analysis involves the determination of a linear equation using regression analysis to predict which group the case belongs to. The equation or function is defined as

$$D = v_1X_1 + v_2X_2 + v_3X_3 + \ldots + v_iX_i + a$$

(3)

where $D$ is the discriminate function, $v$ is the discriminant coefficient or weight for that variable, $X$ is the respondent's score for that variable, $a$ is a constant, and $i$ is the number of predictor variables.

This function is similar to a regression equation or function. The $v$'s, which are unstandardized discriminant coefficients analogous to the $b$'s in the regression equation, maximize the distance between the means of the criterion (dependent) variable. Standardized discriminant coefficients can also be used like a “beta weight” in regression analyses. This equation is applied to maximize the distance among the categories so that an equation that has strong discriminatory power among groups can be obtained. After an existing data set is used to calculate the discriminant function to classify any new cases that can then be classified. The number of discriminant functions is one less the number of groups. There is only one function for the basic two group discriminant analysis. The flow chart of the main structure and analysis relevant to this study is shown in Figure 2.

Results and Discussion

Selection and Processing of Water Quality Data. Daily data collected at six automatic water quality monitoring stations in the Feitsui Reservoir watershed from May 2008 to December 2010 were used in this study. During that period, if the value of water quality monitoring data was smaller than that of the maximum detection limit, that entry was discarded. Additionally, incomplete water quality data sets caused by occasional instrument failures at some water quality monitoring stations were deleted. In short, 281 effective entries were acquired.

Before the introduction of multivariate statistics to the analysis of parameters of water quality, because of seasonal or periodical characteristics of some water quality parameters, this study first standardizes the initial data with this equation:

$$Z_{ij} = \frac{X_{ij} - \bar{X}_i}{S_i}$$

(4)

Where $X_{ij}$ refers to variance $i$, the measured value of sample $j$; $\bar{X}_i$ is the mean of variance $i$; and $S_i$ is the standard deviation of variance $i$.

Results of Implementing Factor Analysis. In the factor analyses implemented in this study, the varimax rotation is used to carry out orthogonal rotation for explaining the number characteristics of factors. Results of analyses show that four factors have characteristics values greater than “1” (see Table 2). Hence, these four factors are selected for illustrating the major factors that affect the water quality of the watershed. Furthermore, a Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy suggested by Kaiser (1974) is assigned to a data set. The size of the KMO value has no statistically critical point, but, according to empirical experience, the larger the KMO value, the more common factors suitable for factor analysis there are. If the KMO value is greater than 0.8, this indicates that the data set is fit for factor analysis, but if the KMO value is smaller than 0.5, factor analysis is not suitable. As indicated in Table 3, the KMO value of this study is 0.814, suitable for factor analysis. Chi-square distribution ($\chi^2$) of Bartlett's Test of Sphericity, however, is 2187.568, reaching the significant level and presenting the existence of a common factor between relevant matrixes of the parent population, also suitable for factor analysis.

Determination of Factors. The number of factors with characteristics values exceeding “1” can be used to determine the number of major factors. The number of major factors is then subject to orthogonal rotation to yield component matrices (Table 4) that can be used for selecting the parameters among various factors, explaining the characteristics of the four factors, describing the major factors that affect the water quality in the watershed, and expressing the differences among these four factors.

Selection of Factors. Table 4 shows the four primary factors that significantly affect the variation of important water quality parameters in the watershed. Additionally, the three-dimensional distribution of factors that affect the water quality in the watershed shows four axes from Figure 3. In Figure 3, BOD, dissolved oxygen, and NH$_3$-N fall on the first axis, and they belong to the first factor class; total phosphorous and MBAS are on the second, and they belong to the second factor class; water temperature alone is on the third axis, and it belongs to the third factor class; suspended solids and pH are on the fourth axis, and they belong to the fourth factor class.

The First Factor. The first factor consists of BOD, dissolved oxygen, and NH$_3$-N, with a total variation of 22.440% as shown in Table 2. In Table 4, BOD has the highest loading extent of 0.771. BOD, which is the quantity of oxygen consumed microbiologically to decompose organic matter under conditions of a constant temperature within a certain period of time, is used to index the quantity of microbiologically degradable organic matter and the organic pollution. Because the microbial degradation of organic matter is a slow process, a complete decomposition of all organic matter is temperature-dependent and may take more than 20 days. BOD is a significant parameter to index the water pollution caused by organic matter, and it is an important parameter to be based on and specified in limiting wastewater discharges classifying water body, evaluating pollution degree of stream, inspecting environmental quality, and enforcing environmental laws and regulations. The watershed in question is located in a remote region in northern Taiwan. The extent of BOD pollution is not high, but nonpoint sources of fertilizer and pesticide discharged from tea plantations in the surrounding mountain regions may cause high BOD and NH$_3$-N in the watershed. The analysis results show that BOD contributes the major pollution in this factor.

Table 4 also shows that dissolved oxygen has a high pollution loading extent of 0.667 in the first factor. Among all the water quality analyses, dissolved oxygen is the best index for indicating environmental quality because all aquatic life depends on
dissolved oxygen to metabolize food for producing energy to maintain life and growth. Hence, the dissolved oxygen level in a water body is of great importance to all aerobic aquatic life; higher levels of dissolved oxygen will maintain the biological diversity. Any changes caused by natural or anthropogenic sources will be reflected by variations of the measurable dissolved oxygen in the water body. The seasonal variation of dissolved oxygen in the stream downstream of the watershed has more water quality variations than the stream upstream, so that the extent of dissolved oxygen variations is not as obvious for upstream as for downstream (Chen, 2006). Hence, the magnitude of dissolved oxygen in a water body is used in this study for evaluating the current extent of pollution loading to the water body.

In Table 4, the loading extent of NH3-N in the first factor is 0.636. Among the numerous sources of NH3-N found in the water body, the nitrogen released by the decomposition of organic matter and fertilizer released from agricultural applications are the major contributors. The ammonia contained in the water gives it a bad odor and make the water corrosive to metal and buildings, toxic to aquatic life, and unfit for drinking purposes. As mentioned previously, the many tea plantations in the watershed contribute nonpoint sources of pollution that have a higher percentage of ammonia than in pesticide and fertilizer. Discharge of ammonia-rich sources into a water body will cause its eutrophication and lower the stream’s recreational

---

**Table 2—Results of factor analyses and the variance explained.**

<table>
<thead>
<tr>
<th>Components</th>
<th>Initial eigenvalues</th>
<th>% of total variance</th>
<th>Cumulative variance %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.016</td>
<td>22.440</td>
<td>22.240</td>
</tr>
<tr>
<td>2</td>
<td>1.275</td>
<td>14.171</td>
<td>36.571</td>
</tr>
<tr>
<td>3</td>
<td>1.151</td>
<td>12.794</td>
<td>49.365</td>
</tr>
<tr>
<td>4</td>
<td>1.016</td>
<td>11.291</td>
<td>60.656</td>
</tr>
<tr>
<td>5</td>
<td>0.883</td>
<td>9.807</td>
<td>70.464</td>
</tr>
<tr>
<td>6</td>
<td>0.795</td>
<td>8.830</td>
<td>79.293</td>
</tr>
<tr>
<td>7</td>
<td>0.768</td>
<td>8.532</td>
<td>87.826</td>
</tr>
<tr>
<td>8</td>
<td>0.621</td>
<td>6.903</td>
<td>94.729</td>
</tr>
<tr>
<td>9</td>
<td>0.471</td>
<td>5.271</td>
<td>100.0000</td>
</tr>
</tbody>
</table>

---

**Table 3—KMO measures and Bartlett Test of Sphericity.**

<table>
<thead>
<tr>
<th>Kaiser-Meyer-Olkin measure of sampling adequacy</th>
<th>0.814</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bartlett Test of Sphericity Chi-square distribution</td>
<td>2187.568</td>
</tr>
<tr>
<td>Degree of freedom</td>
<td>21</td>
</tr>
<tr>
<td>Significance</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Table 4—Matrix of water quality factor loadings for the watershed.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Factors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>BOD</td>
<td>0.771</td>
</tr>
<tr>
<td>Dissolved oxygen</td>
<td>0.667</td>
</tr>
<tr>
<td>NH₃-N</td>
<td>0.636</td>
</tr>
<tr>
<td>Chlorophyll</td>
<td>0.413</td>
</tr>
<tr>
<td>Total phosphorus</td>
<td>0.057</td>
</tr>
<tr>
<td>MBAS</td>
<td>0.147</td>
</tr>
<tr>
<td>Temperature</td>
<td>0.035</td>
</tr>
<tr>
<td>Suspended solids</td>
<td>−0.002</td>
</tr>
<tr>
<td>pH</td>
<td>0.175</td>
</tr>
</tbody>
</table>

and ornamental value. Additionally, organic nitrogen will become oxidized into nitrites and nitrates; both species adversely impact aquatic life and even human health. Hence, NH₃-N is also used as an important water quality index to show its influence with respect to the extent of pollution that can be sustained by the stream.

Integration of the above analysis results show that the major pollution in the watershed, as seen in concentrations of BOD, dissolved oxygen, and NH₃-N in the water body, is closely related to the production and decomposition of waterborne organic matter that comes mainly from wastewater discharges and nonpoint sources. These organic matters are easily oxidized to cause concentrations of waterborne BOD, dissolved oxygen, and NH₃-N, so that the first factor may be termed the organic pollution factor.

The Second Factor. The second factor is mainly composed of total phosphorus and MBAS; its total variation of 14.171% is seen in Table 2. Table 4 shows that the total phosphorus has the relative high loading extent of 0.752 in the second factor. The higher total phosphorus in this watershed may be caused by the surface runoff from storms, which carries the residual fertilizer from upstream tea plantations. The mud and sand particles suspended in the surface runoff contain high quantities of total phosphorus, which cause an increase in the stream’s pollution loading. Discharges of surfeit nutrients into the neighboring stream water bodies cause occasional eutrophication that leads to prolific algal growth. Decomposition of dead algae brings about excessive consumption of dissolved oxygen, which results in oxygen depletion and causes the water quality to rapidly deteriorate (Wu, 2009). Additionally, the increasing local recreational bed-and-breakfast housing development will greatly increase the pollution load to local streams. Hence, the total phosphorus will be an important index for indicating the stream pollution.

As shown in Table 4, the loading extent of MBAS in the second factor is 0.724. MBAS originates mainly from domestic wastewater as a result of the use of household detergents such as branched alkylbenzene sulfonate. It is easily accumulated at the interface between the solid phase and liquid phase in the environment, in river (or lake) bottom sludge, polluted water bodies, and soil environments, to cause serious damage to the environmental ecological system. Through bio-accumulation, MBAS may cause a direct threat to human health. Because MBAS cannot exist naturally in the environment, MBAS pollution in the watershed is directly related to the agricultural spraying of pesticide in upstream regions. Additionally, the organic elements contained in MBAS may also cause eutrophication in a water body. Eutrophication in this area has increasingly become serious, and the possible cause is the increase of point and nonpoint source pollution loadings. In particular, nonpoint source pollution is the main factor contributing to eutrophication in reservoir watersheds. Possible methods of controlling nonpoint pollution include control of upstream fertilizer use, change of cropping patterns, and land utilization management.

The aforementioned analysis results are synthesized to conclude that the NH₃-N and MBAS pollution is closely related to the eutrophication phenomenon in the watershed, and hence the second factor can be called the eutrophication pollution factor.

The Third Factor. The third factor consists of only water temperature; its total variation is 12.794%, as shown in Table 2. Table 4 shows that the loading extent of water temperature is as high as 0.882. Like the water pH value, the water temperature significantly influences aquatic life because it is, for the most part, very sensitive to changes in water temperature and can survive only in a narrow temperature range. In addition, the water’s physical properties, such as density, viscosity, vapor pressure, and surface tension, and chemical properties, such as biochemical reaction rate, are significantly affected by changes in water temperature. Hence, the temperature of a water body is an important parameter for determining the quality of a water body. Excessively high or low water temperature beyond the optimum range will adversely influence aquatic life and make it lose its ability to find food or move around. Hence, maintaining the water temperature within a certain optimum range is an important link in the stream and reservoir management strategy. Rising stream water temperature increases the biological metabolic rate to enhance the growth rate of vegetation but...
may adversely weaken the resistance capability of aquatic life. The stream temperature may vary depending on seasons and locations. The upstream section of the watershed, which has dense forest coverage with rapidly flowing waters, is lower in water temperature and higher in stream dissolved oxygen, with more algal growth, than the downstream section that is flatter and has more anthropogenic development.

Therefore, the magnitude of a stream's temperature directly affects its ecology, the survival of aquatic life, and even the concentration of stream dissolved oxygen, so that stream temperature can be an important parameter for indexing water quality. Because it varies seasonally, the third factor can be regarded as a seasonal factor.

The Fourth Factor. The fourth factor is composed of suspended solids and pH value; its total variation of 11.291% is shown in Table 2. Table 4 shows that, in the fourth factor, the loading extents are −0.673 for suspended solids, and 0.653 for pH. When a stream has a relatively high level of suspended solids, its self-purification capacity will be lowered, so that water quality deteriorates. Suspended solids will interfere with the aesthetic appearance of the water body in addition to consuming dissolved oxygen by suspended living and organic particles. The waterborne suspended particles in a stream consist of organic and inorganic particles such as sand and clay; inorganic suspended particles may settle, whereas organic particles will consume dissolved oxygen. The stream pH will affect vegetative growth, the settling and dissolving of suspended particles, and the treatment of water and wastewater. Under conditions of high stream pH levels, suspended particles with sizes ranging from fine flocs to coarse particles are extremely susceptible to chemical coagulation, so that they settle and cause a reduction of the stream suspended solids. This phenomenon confirms the mutual correlation between suspended solids and pH in this factor.

The above analysis results reveal that the stream suspended solids and pH are closely correlated, and hence the fourth factor is called the settling pollution factor.

Cluster Analysis of Water Pollution Characteristics in the Watershed. A normalized component score coefficient matrix is generated prior to conducting cluster analysis. The matrix is subject to linear combination by calculating the “Factor Score”, or the mathematical product of the component score coefficient and the original variable. Using the two-stage method, the results are first roughly clustered and then the k-means method is used to test the number of different clusters. In this study, five clusters are selected for determining the difference among various water quality parameters and conducting further analyses on various clusters. The relationship among various clusters and factors is shown in Figure 3. Here, these five clusters are reverted back to the original items of the nine water quality parameters and are compared with the current RPI so that the unique characteristics of the water quality in this watershed can be more clearly understood; the results are shown in Table 5. The six water quality monitoring stations are distributed in upstream and downstream regions of the Feitsui Reservoir watershed. Anthropogenic land use and crop harvest, as well as natural soil erosion and pollution in the neighboring regions of the watershed, have caused some differences in the groundwater hydrology near the watershed. However, all six water quality monitoring stations have very similar geographical and geological features (Chen, 2006; Yu, 2006). Hence, the following five clusters are used to fully investigate the water quality characteristics and degree of pollution in this watershed.

First Cluster. As shown in Figure 4, this cluster has the highest score in the fourth factor (sedimentation factor) and the next highest score in the third factor (seasonal factor). In Table 5, this cluster has the lowest pH of between 6.0 to 6.8, and the highest suspended solids of 57 to 75.5 mg/L among all clusters. The waterborne suspended particles in a stream with low pH will not easily be affected by chemical coagulation reaction and produce precipitates. Hence, the concentration of suspended solids in the stream obviously increases. This phenomenon can be confirmed by the negative correlation coefficient between pH and suspended solids contained in the water quality factor loading matrix. Certainly, the upstream of the receiving water body may carry increasing concentrations of suspended solids carried over by the mountain stormwater (Chen, 2006). However, the monitoring results on suspended solids shown in Table 5 indicate that upstream soil conservation has been well-maintained in this watershed. In addition, the upstream region has not been subject to inappropriate development. Hence, the stream suspended solid concentration has not shown any abnormal increases. Additionally, Table 4 reveals that the next highest factor in this cluster is the water temperature, indicating that most water quality data in this cluster were collected during the period when autumn was changing to winter. Further, the water temperature during this period is generally between 17.2 and 23.6 °C; the difference between the highest and the lowest temperatures is relatively low, so that seasons do not greatly affect the water quality in this watershed. Overall, if the RPI is used to determine the classification of stream pollution, stream pollution in this region belongs to moderate pollution. If based on a single water quality item, stream pollution can be considered “medium pollution” because of relatively high suspended solids.

The suspended solids in this cluster obviously have a wider range of distribution than those in the other four clusters, mainly because surface erosion during storms boosts the stream suspended solids levels within a short period. This can be confirmed by the observation that suspended solids with high concentrations in this cluster occur mostly during the wet season from June to September each year in northern Taiwan caused by the plum rains and monsoons. Stormwater may also dilute the stream water so that concentrations of other water quality parameters in this cluster are lower than those in other clusters, as shown in Table 5. In this cluster, the influence of suspended solids on RPI is relatively high and is primarily affected by the settling factor. Hence, the water has the characteristics of relatively low pH and high suspended solids, so that this cluster can be regarded as having “high settleable pollution water quality”.

Second Cluster. As shown in Figure 4, the factor scores for this cluster in four factors are generally not high. For example, the factor score in the first factor is “4”, and it is the lowest in the fourth factor (seasonal factor). Additionally, this cluster shows average concentration for any water quality parameter. If the three water quality items in the first factor that most obviously affect the water quality are used as examples, Table 5 shows that a dissolved oxygen of 6.73 mg/L average concentration is the third lowest in all five clusters; a BOD of 0.93 mg/L average concentration is the third highest in all five clusters. NH3-N, whose average concentration is 0.066 mg/L, ranks second in all
five clusters; its concentration in this cluster is almost the same as in the third and fifth clusters.

Overall, the RPI in this cluster belongs to the nonpolluted or slightly polluted category. The characteristic of this cluster is that four water quality parameters in this cluster, if judged based on water pollution classification, may also be considered to belong in the nonpolluted or slightly polluted category. Further, all other water quality items do not tend to have high concentration, which indicates that this cluster in the watershed belongs to the good water quality category. The aforementioned analysis results reveal that concentrations of water quality parameters in this cluster have small ranges of variation and distribute evenly over the various seasons of the year. This cluster has more of the total 228 sets of data collected in this study than all other four clusters, so that this cluster may be termed unpolluted, indicating that the water in this watershed has no pollution.

**Third Cluster.** This cluster is shown in Figure 4 to have the highest score for the eutrophication pollution factor, with the next highest for the organic pollution factor that is also seen in Table 5. After reversion, the average concentration of total phosphorus in the eutrophication factor of 0.180 mg/L is the highest among all clusters and is much higher than the total phosphorus concentrations in all other clusters. However, the concentration of total phosphorus has not been considered as a parameter to index stream pollution; therefore, the influence of total phosphorus on stream water quality in this watershed cannot be evaluated based on the total phosphorus concentration. As shown in the results of Chang (2009), tea gardens, farmland, woodlands, marketplaces, and grasslands contributed 72% of nonpoint pollution sources to reservoir eutrophication whereas recreational use of camping grounds, animal husbandry, and the discharge from water treatment plants contributed the other 28%. This shows that among the eutrophication pollution factors of this cluster, the major pollution source is man-made nonpoint pollution. In addition, the same report (Chang, 2009) also pointed out the trend of increasing total phosphorus concentration in the Feitsui Reservoir watershed since 1996, which is related closely to the commencement of Taipei-Ilan Highway project. The effect of large-scale excavation and water and soil conservation is also one of the factors influencing total phosphorus concentration in this cluster.

When the data in this cluster is used for evaluating stream pollution, the influence of total phosphorus concentration on stream water quality cannot be evaluated. In addition to high total phosphorus concentration, the eutrophication factor in this

---

**Table 5—The average and extreme values of water quality parameters in the five clusters.**

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Clusters</th>
<th>Clusters</th>
<th>Clusters</th>
<th>Clusters</th>
<th>Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>6.43; 6.0–6.8</td>
<td>7.18; 6.1–8.4</td>
<td>7.26; 6.6–8.0</td>
<td>7.26; 6.4–8.1</td>
<td>7.26; 6.3–8.1</td>
</tr>
<tr>
<td>Temperature (℃)</td>
<td>22.83; 17.2–23.6</td>
<td>23.86; 20.1–30.3</td>
<td>25.97; 21.5–29.1</td>
<td>25.97; 20.3–32.2</td>
<td>16.78; 11.4–20.8</td>
</tr>
<tr>
<td>Dissolved oxygen (mg/L)</td>
<td>6.80; 6.3–7.4</td>
<td>6.73; 3.4–8.7</td>
<td>6.04; 5.1–7.8</td>
<td>6.00; 5.0–8.3</td>
<td>7.01; 5.3–8.2</td>
</tr>
<tr>
<td>BOD (mg/L)</td>
<td>0.65; 0.5–0.7</td>
<td>0.39; 0.2–7.2</td>
<td>0.72; 0.2–1.3</td>
<td>1.32; 0.2–5.4</td>
<td>0.96; 0.2–4.1</td>
</tr>
<tr>
<td>Suspended solids (mg/L)</td>
<td>65.83; 57–75.5</td>
<td>2.22; 0.1–6.6</td>
<td>25.11; 17–41</td>
<td>5.99; 1.6–15</td>
<td>3.08; 0.2–8.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 4—Factor scores of the five clusters.**
cluster also includes an MBAS that has the second highest concentration among all five clusters. Hence, the water quality associated with this cluster is more obviously subject to eutrophication than other clusters. More than normal algal growth also leads to the lowest dissolved oxygen among all five of the clusters. However, the relatively higher eutrophication level does not obviously affect the organic pollution factor. This is possibly because the nutrients that cause eutrophication do not contain much organic matter. Hence, this cluster affects the settling factor more than the eutrophication factor, and the algal growth will also affect the formation of sedimentation in the water body. Hence, this cluster has the second highest score in the “settling pollution factor”.

Overall, the water quality in this cluster belongs to the lightly polluted category. With respect to the pollution classification of water quality parameters, dissolved oxygen and suspended solids in the four water quality parameters are classified as lightly polluted. The aforementioned relatively high concentrations of total phosphorus and MBAS indicate that this cluster is mostly distributed in regions with frequent agricultural activities, and these regions are subject to the use of fertilizers and pesticides. Additionally, this cluster is evenly distributed among the six water quality parameters, and hence the agricultural activities in this cluster are not limited to specific sections. The conclusion of the above analyses is that this cluster generally has relatively higher nutrient levels, but the RPI result only shows it to be lightly polluted. As such, this cluster can be considered to have “light eutrophication pollution”.

Fourth Cluster. As shown in Figure 4, the organic factor in this cluster has the highest score and the eutrophication factor has the second highest score, indicating that this cluster seems to have a relatively high level of pollution. However, based on the results presented in Table 5, with respect to the reverted organic pollution factor of this cluster, the average concentration of dissolved oxygen is 6.00 mg/L, the lowest among all clusters. The BOD concentration of 1.25 mg/L is the highest among all clusters, and the average NH$_3$-N concentration of 0.088 mg/L is also the highest among all clusters. However, this cluster has low scores in the third and fourth factors. In addition, the organic matter contained in the water body strongly affects the monitoring stations in this cluster, thus leading to higher concentrations of organic-matter-related water quality parameters. In terms of cluster distribution interval, tourists and local residents generate more organic pollutants during weekends and consecutive holidays. Kuo and Lee (2004) suggested that, in addition to the construction site of the Taipei-Ilan Highway in this watershed, the recent increase in the number of tourists, as well as agricultural pollution such as that generated by tea gardens, resulted in an increase of organic loading in rivers and water bodies, and also greatly affected the water quality of this watershed. Additionally, this cluster has the lowest factor score in the seasonal factor category; this indicates that the water quality for all monitoring stations in this cluster is less affected by seasonal change than the water quality in other clusters.

Overall, Table 4 shows that this cluster has a higher NH$_3$-N concentration than the highest NH$_3$-N concentration of 0.45 mg/L for other clusters. If evaluated with the pollution classification for each water quality parameter, NH$_3$-N is either in the nonpolluted or slightly polluted parameter. Among the other four water quality parameters, dissolved oxygen belongs to the slightly polluted parameter, whereas BOD and suspended solids are nonpolluted parameters. Additionally, the chlorophyll concentration is the highest among all clusters; the total phosphorous concentration is second but still is obviously lower than those in other clusters. Thus, the water in this cluster is not subject to eutrophication, as that in the third cluster is. Concerning distribution, this cluster has a higher organic pollution concentration than other clusters, and this pollution is mainly distributed in the summer. Thus, this cluster can be called *water quality with lightly organic pollution*.

Fifth Cluster. Figure 4 shows that this cluster has the highest factor score in the seasonal factor. It is much higher than those in all other clusters. The water temperature of 16.78 °C in this cluster is lower than the water temperature in other clusters, as shown in Table 5. Thus, this cluster is generally distributed in the winter season; the number of samples in this cluster is only one-quarter of the total number of all samples in this cluster, indicating that this cluster is closely related to seasonal distribution. In a natural environment, lower water temperature will lead to increased water viscosity and dissolved oxygen concentration. The data in Table 5 confirm this phenomenon and show that dissolved oxygen in this cluster is the highest among all five clusters. However, this watershed is located in northern Taiwan's mountain region, where the background stream water quality is low in pollution, so that concentrations of BOD and dissolved oxygen in this cluster are lower than those in other clusters. Hence, both BOD and dissolved oxygen in this cluster are not high when compared with those in other clusters. Additionally, winter is a dry season in this cluster, with a relatively low frequency of storms, so that the suspended solid concentration in the settling factor for this cluster is the lowest among all five clusters. Its concentration is almost the same as that in unpolluted water.

In general, the water quality in this cluster belongs to the unpolluted or slightly polluted category. If the water quality classification is assessed, the four water quality items can be classified as unpolluted. In addition, this cluster is mainly distributed in the winter season. This cluster is only subject to nonpoint source pollution without obvious point source pollution caused by anthropogenic or animal husbandry activities. Hence, the water quality of the monitoring stations in this region shows no abnormality, without deviant aberrant high or low levels. This cluster can be called *unpolluted seasonal water quality*.

**Discriminant Analysis.** Table 6 lists the results of discriminant analysis. The similarities (i.e., percentage of correct discrimination) between the “discriminant cluster” obtained using the discriminant functions as shown in Table 5, and the “actual cluster” obtained using the discriminant analyses, are high for all clusters. They are 100% for the first cluster, 92.43% for the second, 100% for the third, 93.02% for the fourth, and 98.55% for the fifth. In these discriminant analyses, the discriminant accuracies for the first and the third clusters are 100%, indicating that the percentage of accurate analyses is high for discriminating water quality with high sedimentation and eutrophication. In other words, factors that have a low influence on the stream's overall quality have a better probability of discrimination.

**Conclusions**

The factor analysis method in multivariate statistical analysis has been applied to explain the principal factors of the nine...
water quality parameters for the Feitsui Reservoir located in northern Taiwan. This study uses factor analysis to identify characteristics and changes of this watershed and provides the local water authority with information concerning the real-time organic, eutrophication, seasonal, and settling pollution factors of water quality. Second, cluster analyses help to sufficiently determine factor features and pollution levels, for example, the increase of organic substances generated after weekends and consecutive holidays; potential pollution from eutrophication related to land utilization and agricultural development in the watershed; seasonal factors such as drought periods in winter, with lower percentages of heavy rain in mountain areas and man-made pollution; and settling pollution factors such as the influence of summer typhoons and the rainy season.

Factor analyses in this study fail to take into account the pollution characteristics and levels in the watersheds (such as unpolluted, lightly polluted, moderately polluted, and seriously polluted indicated by RPI) whereas cluster analyses help researchers understand the above information. RPI scores can also be used to determine the pollution level of each cluster. In the end, this study uses SPSS, as well as results of cluster analyses, to reach 95.37% discriminant accuracy and prove the efficiency of cluster analysis.

Because overall water quality of the reservoir and pollution in the watershed and reservoir are complicated by large-scale development and pollution (e.g., upstream land use and the application of pesticides and fertilizers), this research uses nine water quality parameters to effectively assess the overall water quality of the watershed. The proposed implementation of this testing will obviously make up for the deficiency of the RPI, which is based on only four water quality parameters, to evaluate the classification of stream pollution. Additionally, the principles and methodology of multivariate statistical analysis that are used in this research for establishing a method based on statistical quantitative analysis using variations of water quality in a selected watershed can be referenced and implemented in future studies. The results obtained using statistical analysis and interpretation must be verified using on-site data on the physical and chemical reactions in stream self-purification as collected from sites that have similar geological and geographic characteristics. The validity and applicability of statistical results that have been confirmed by these comparisons can be applied in implementing resource protection strategies by authorities responsible for managing streams and reservoirs. This is the most significant contribution of this research.

It is trusted that the use of multivariate statistics and the incorporation of applications and discussion of the RPI among clusters meet the objectives of indicator management systems in execution. This management information is able to reflect the needs of reservoir watersheds and serve as a valuable tool for decision makers.

Acknowledgments

The authors sincerely acknowledge the financial support under Grant NSC 2011-2221-E214-011 from Taiwan's National Science Council.

Submitted for publication April 5, 2012; accepted for publication July 6, 2012.

References


Table 6—Results of Discriminant Analyses.

<table>
<thead>
<tr>
<th>Discriminated cluster</th>
<th>Percentage of accurate discrimination (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual cluster</td>
<td>Cluster 1</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>35</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>0</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>0</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>0</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>36</td>
</tr>
</tbody>
</table>


