Logistic mixture of multivariate regressions for analysis of water quality impacted by agrochemicals

Yongsung Joo1*, Keunbaik Lee2, Joong-Hyuk Min3, Seong-Taek Yun4 and Trevor Park2

1 Biostatistics, College of Public Health and Health Professions, University of Florida, Gainesville, FL 32611, U.S.A.
2 Department of Statistics, University of Florida, Gainesville, FL 32611, U.S.A.
3 Department of Environmental Engineering Sciences, University of Florida, Gainesville, FL 32611, U.S.A.
4 Department of Earth and Environmental Sciences, Korea University, Seoul 136-701, Korea

SUMMARY

In this paper, we study the impacts of two representative agricultural activities, fertilizers and lime application, on water quality. Because of heavy usage of nitrogen fertilizers, nitrate (NO\textsubscript{3}⁻) concentration in water is considered as one of the best indicators for agricultural pollution. The mixture of normal distributions has been widely applied with (NO\textsubscript{3}⁻) concentrations to cluster water samples into two environmentally interested groups (water impacted by agrochemicals and natural background water groups). However, this method fails to yield satisfying results because it cannot distinguish low-level fertilizer impact and natural background noise. To improve performance of cluster analysis, we introduce the logistic mixture of multivariate regressions model (LMMR). In this approach, water samples are clustered based on the relationships between major element concentrations and physicochemical variables, which are different in impacted water and natural background water. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: mixture of regressions; model-based clustering; mixture of normal distributions; ECM; water quality; agricultural pollution

1. INTRODUCTION

Ground and surface water quality has been continuously threatened by the direct input of agrochemicals, such as fertilizers and pesticides, and the associated biogeochemical reactions (Hamilton and Helsel, 1995; Kraft et al., 1999). To understand the impact of agricultural activities on water quality, investigations on the major chemical compositions in alluvial and bedrock groundwater and surface water were conducted in alluvial areas of Nakdong River watershed, which is the second largest in South Korea (Figure 1). Year-round vigorous agricultural activities are the typical land-use pattern in these areas. Large amounts of synthetic nitrogen-fertilizers and composite fertilizers (N-P-K = 21-17-17) are applied to the farmland. Also, farmers commonly apply lime (CaO) to neutralize soils that are acidified through heavy agricultural activities.
This paper has two objectives: (1) clustering waters (water samples) of the Nakdong River watershed into two environmentally meaningful groups (water impacted by agrochemicals vs. natural background water), and (2) estimating the relationship between agrochemical-originated element concentrations and physicochemical variables. To cluster waters based on major element concentrations that can be originated from the agrochemicals, many statisticians (Xue et al., 2005) and environmental scientists (Chen et al., 2002; Park et al., 2005) applied the mixture of univariate normal distributions (MUN):

$$f(y_i) = \sum_{j=1}^{J} p_j \phi(y_i | \mu_j, \sigma_j^2)$$  \hspace{1cm} (1)

Figure 1. Nakdong River watershed and sampling locations
where $y_i$ is a major element concentration in a log scale, $p_j$ is the mixing probability for $j$th group, and $\phi(y_i; \mu_j, \sigma_j^2)$ is the probability density of normal distribution with mean $\mu_j$ and variance $\sigma_j^2$. However, this method does not utilize the fact that concentrations of agrochemical-originated elements change significantly with physicochemistry of water, which can be explained by well-known biochemical equations. Our proposed logistic mixture of multivariate regressions (LMMR) model achieves these objectives simultaneously.

Since the EM algorithm was developed by Dempster et al. (1977), it has been regarded as one of the most popular approaches in estimating mixture models and has accelerated usage of these models. Everitt (1984) compared estimation algorithms for the mixture of two univariate normal densities and concluded that Newton and EM algorithms provide very satisfactory results. Often, random variables of interest are correlated with each other. To incorporate these correlations into the mixture model, Basford and McLachlan (1985) proposed the mixture of multivariate normals. Xue et al. (2005) exploited Bayesian mixtures of univariate normals to cluster waters in the Yangtze River watershed, their goals being similar to ours.

Naturally, these models evolved into a regression framework to cluster observations based on the relationships between a response variable and predictors. Turner (2000) studied the mixture of univariate regressions to cluster observations based on the relationship between a response variable and predictors. Jeffries and Pfeiffer (2000) suggested a mixture of univariate normals model that has a logistic regression structure for mixing probabilities. Thus, the mixing probability can vary with the covariates. This paper extends their frameworks to the LMMR. As the name implies, this model includes all these conventional mixture models as special cases and contains advantages of each method:

1. It explains the relationship between a response variable and predictors as in the mixture of univariate regressions (Turner, 2000; Razzaghi, 2002).
2. It uses the covariance structure of response variables as in the mixture of multivariate normals (Basford and McLachlan, 1985).
3. It allows the mixing probabilities to depend on predictors as in the logistic mixture model (Jeffries and Pfeiffer, 2000).
4. It conducts clustering and estimation of parameters simultaneously as in any other mixture model.

Even though the proposed model has the possibility of overparameterization as a disadvantage, this is not observed in our case study with Nakdong River watershed data.

In Section 2, Nakdong River watershed data and water chemistry related to agrochemicals are described. In Section 3, the LMMRs and the estimation (ECM) algorithm are introduced. In Section 4, Nakdong River watershed data are analyzed with the mixture models. Section 5 concludes the paper.

2. NAKDONG RIVER WATERSHED DATA AND WATER CHEMISTRY

It is well known by water chemists that concentrations of major agrochemical-originated elements and physicochemical variables of waters have different relationships in impacted and natural background waters. The main idea of our proposed model is to use these relationships in clustering waters into two groups.

2.1. Data collection

Water samples were collected at 93 locations from October 1999 to August 2000. The sampling locations are widely scattered to catch the impact of agrochemicals on the entire study area. These
consist of 72 shallow alluvial groundwater (61 irrigation wells for agricultural activities and 11 domestics wells), 15 deep alluvial groundwater, and 6 surface water sampling locations at three small-scale rural watersheds (Wolha, Daesan, and Yongdang) in the middle to lower reaches of the Nakdong River (Figure 1). Shallow irrigation and domestic well waters are collected mostly in the vicinity of the water table with median depth 8 m (4–37 m), while deep groundwaters are sampled at median depth 120 m (80–198 m). These samples cover various types of waters from surface waters to deep groundwaters, and from unpolluted waters to waters highly polluted by agricultural activity.

We assume that major element concentrations at the 93 sampling locations are independent for the following reasons. First, crop fields in the study areas of the Nakdong River watershed had typically even and random applications of agrochemicals because similar crops were grown. This study area did not have any locally massive pollution, which may spread out and cause dependency among sampling locations. Second, concentrations of agrochemical-related major elements change randomly with various factors in the underground environment. In general, physicochemical properties of soil porous media (i.e., porosity, hydraulic conductivity, temperature, redox, organic matter content, and so on) and underground geological structures (i.e., joints and fractures) vary spatially, controlling groundwater flow path, flow rate, and water chemistry. Onsoy et al. (2005) discusses heterogeneity of alluvial groundwater system with agrochemical-induced pollutants. These heterogeneous characteristics of subsurface systems make the samples independent, even between horizontally adjacent groundwater samples.

Annual averages of agrochemical-related major element concentrations (NO\textsubscript{3}, Ca\textsuperscript{2+}, K\textsuperscript{+}, and Cl\textsuperscript{-}) and physicochemistry of water samples (alkalinity and pH) are measured at each location. In particular, NO\textsubscript{3} and Ca\textsuperscript{2+} are two major elements in fertilizer and lime. In this study, the alkalinity corresponds to HCO\textsubscript{3} concentration of the water samples because natural waters in the Nakdong River watershed are within neutral pH range (Drever, 1997). See Min et al. (2002) for details.

2.2. Water chemistry

Numerous environmental studies have shown that deep groundwaters in rural watersheds are less susceptible to pollution by agricultural activities than shallow groundwaters, because they are farther from the ground surface where agrochemicals are applied (Hendry et al., 1983). Less permeable geologic units prevent the infiltration of impacted shallow groundwater into the deeper zone. In addition, pollutants from agrochemicals tend to get retarded or degraded as they move to the deeper zone (Hallberg and Keeney, 1993; Spalding and Exner, 1993; Stites and Kraft, 2001; Kraft and Stites, 2003). All irrigation wells in this study are shallow and located within crop fields. Therefore, we can assume that groundwaters in irrigation wells are impacted, while deep groundwaters are not impacted and have characteristics of natural background waters. With regard to domestic well and surface waters, no assumption can be made because their impact status depends on adjacency to farmland and unknown underground geology.

There has been much research in water chemistry on the relationships between major agrochemical-originated elements and physicochemical variables (Min et al., 2002; Chae et al., 2004). In general, the natural background water chemistry is primarily controlled by water–rock interactions. In particular, the dissolution of carbonate mineral (CaCO\textsubscript{3}) plays a major role in regulating the water chemistry due to its high solubility, which is described by

\[
\text{CaCO}_3 + \text{CO}_2 + \text{H}_2\text{O} \rightleftharpoons \text{Ca}^{2+} + 2\text{HCO}_3^-
\]  

(2)
In other words, even though log concentrations of Ca$^{2+}$ and alkalinity (HCO$_3^-$) may vary from location to location depending on severity of weathering processes, they will have a positive relationship. For example, Chen et al. (2002) report that there is a very strong positive relationship between log(Ca$^{2+}$) and log(alkalinity) in Yangtze river water. However, our Nakdong River watershed data has more complicated trends than the Yangtze River watershed. When only natural background waters (deep groundwater, which is denoted with D in Figure 3(a)) are considered, a positive trend is observed. When only impacted waters (irrigation well groundwater, which is denoted with * in Figure 3(a)) are considered, a negative trend is observed. This can be explained by the following two processes associated with agricultural activities.

First, nitrification of nitrogen fertilizers (NH$_4^+$), which is described by Equation (3), raises NO$_3^-$ concentration and lowers pH.

$$\text{NH}_4^+ + 2\text{O}_2 \leftrightarrow \text{NO}_3^- + 2\text{H}^+ + \text{H}_2\text{O} \quad (3)$$

Negative relationships between log(NO$_3^-$) and pH appear in Figure 3(c). As a consequence of nitrification, lower pH reduces alkalinity of water (as measured by level of HCO$_3^-$, Min et al., 2002). In other words, the nitrification process makes a negative trend between log(NO$_3^-$) and log(alkalinity) in impacted water (* in Figure 3(b)). This process is absent or very mild in natural background water (D in Figure 3(b)). Second, lime (CaO) is applied to neutralize farmland soils where nitrogen fertilizers are used. This makes log(Ca$^{2+}$) and log(NO$_3^-$) have a positive relationship only among impacted waters (* in Figure 3(d)). Natural background waters do not have any strong trend (D in Figure 3(d)). These two processes make log(Ca$^{2+}$) and log(alkalinity) have a negative trend in impacted water (* in Figure 3(a)).

3. LOGISTIC MIXTURE OF MULTIVARIATE REGRESSIONS MODEL AND ESTIMATION

In this section, we propose the LMMRs model and an ECM (Meng and Rubin, 1993) algorithm to find the MLE of this model.

3.1. Proposed model

Suppose that we are interested in the mixture of J groups. Also, let $y_i$ be a column vector of M response variables for the $i$th observation (annual average of measurements at $i$th sampling location), $x_i$ be a predictor vector for the mean functions $\mu_j(x_i)$ where $j = 1, \ldots, J$, and $z_i$ be a predictor vector for the logistic functions for mixing probability $p_j(z_i)$. The proposed model is given by

$$f(y_i | x_i, z_i) = \sum_{j=1}^{J} p_j(z_i) \cdot \phi(y_i | \mu_j(x_i), \Sigma_j) \quad (4)$$

where $\phi(y_i | \mu_j(x_i), \Sigma_j)$ is a multivariate normal probability density function of $y_i$ with mean vector $\mu_j(x_i) = X_i \beta_j$ and covariance matrix $\Sigma_j$,

$$y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iM} \end{pmatrix}, \quad \mu_j(x_i) = \begin{pmatrix} \mu_{j1}(x_i) \\ \vdots \\ \mu_{jM}(x_i) \end{pmatrix}, \quad X_i = I_{M \times M} \otimes x_i \quad \beta_j = \begin{pmatrix} \beta_{j1} \\ \vdots \\ \beta_{jM} \end{pmatrix}$$
3.2. Parameter estimation

Let \( w_i = (w_{i1}, \cdots, w_{ij}) \) be an indicator vector for group memberships: If \( y_i \) belongs to group \( j \), \( w_{ij} = 1 \) and \( w_{ij'} = 0 \) for all other \( j' \neq j \). Also, let \( \theta = (\beta_1, \cdots, \beta_j, \Sigma_1, \cdots, \Sigma_J, \gamma_1, \cdots, \gamma_{J-1}) \). Parameters in the LMMRs model are estimated using the ECM algorithm (Meng and Rubin, 1993) as follows (see Appendix for details):

**Step 1.** Provide initial values, \( w^{(0)} = (w_{1}^{(0)}, \cdots, w_{n}^{(0)})^T \) and

\[
\theta^{(0)} = \left( \beta_1^{(0)}, \cdots, \beta_j^{(0)}, \Sigma_1^{(0)}, \cdots, \Sigma_J^{(0)}, \gamma_1^{(0)}, \cdots, \gamma_{J-1}^{(0)} \right)
\]

**Step 2.** Let

\[
P_{ij}^{(k_1, k_2, k_3)} = \frac{P_j^{(k)}(z_i) \phi \left( y_i \mid \mu_j^{(k)}(x_i), \Sigma_j^{(k)} \right)}{\sum_{l=1}^L P_l^{(k)}(z_i) \phi \left( y_i \mid \mu_l^{(k)}(x_i), \Sigma_l^{(k)} \right)}
\]

and calculate \( P_{ij}^{(k-1, k-1, k-1)} \). Note that, after the algorithm converges, this probability can be interpreted as

\[
\text{Prob}[w_{ij} = 1 \text{ and } w_{ij'} = 0 \text{ for all } j' \neq j | y, x, z]
\]

**Step 3.** Calculate

\[
\beta_j^{(k)} = \left\{ \sum_{i=1}^N X_i^T \Sigma_j^{(k-1)} X_i P_{ij}^{(k-1, k-1, k-1)} \right\}^{-1} \left\{ \sum_{i=1}^N X_i^T \Sigma_j^{(k-1)} y_i P_{ij}^{(k-1, k-1, k-1)} \right\}
\]

and update \( P_{ij}^{(k, k-1, k-1)} \) with \( \beta_j^{(k)} \).
Step 4. Calculate

$$
\Sigma_j^{(k)} = \frac{1}{\sum_{i=1}^{N} p_{ij}^{(k,k-1)}} \sum_{i=1}^{N} \{y_i - \mu_j^{(k)}(x_i)\} \{y_i - \mu_j^{(k)}(x_i)\}^T P_{ij}^{(k,k-1)}
$$

and update $P_{ij}^{(k,k-1)}$ with $\Sigma_j^{(k)}$.

Step 5. Set $l = 1$. Also, let $\lambda^{(0)} = \gamma^{(k-1)}$ and

$$
\lambda^{(l)} = \lambda^{(l-1)} - \left[ \frac{\partial h(\lambda^{(l-1)})}{\partial \lambda^{(l-1)}} \right]^{-1} h(\lambda^{(l-1)}) \quad (6)
$$

where

$$
\begin{align*}
h(\lambda) &= (h_1(\lambda), \ldots, h_{J-1}(\lambda))^T \\
h_a(\lambda) &= \sum_{i=1}^{N} \left\{ \left( P_{ia}^{(k,k-1)} - P_a^{(k)}(z_i) \right) z_i \right\} \\
\frac{\partial h(\lambda)}{\partial \lambda} &= \left[ \frac{\partial h_a(\lambda)}{\partial \lambda_a} \right]_{ab} \\
\frac{\partial h_a(\lambda)}{\partial \lambda_b} &= \left\{ \begin{array}{ll}
- \sum_{i=1}^{N} p_a^{(k)}(z_i) \left( 1 - p_a^{(k)}(z_i) \right) z_i z_i^T, & \text{if } a = b \\
\sum_{i=1}^{N} p_b^{(k)}(z_i) p_a^{(k)}(z_i) z_i z_i^T, & \text{if } a \neq b
\end{array} \right.
\end{align*}
$$

for $a, b = 1, \ldots, J-1$. Iterate Equation (6) until $\lambda^{(l)}$ converges. Then set $\lambda^{(k)} = \lambda^{(l)}$.

Step 6. Set $k = k + 1$ and repeat Steps 2–5 until convergence.

3.3. Standard errors of estimates

A well-known disadvantage of the EM algorithm is the difficulty in calculating standard errors for MLEs. There have been many studies, including Louis (1982), Meng and Rubin (1991), and Oaks (1999). However, their methods are difficult to implement for the LMMR because the model does not have an exponential family density, and analytic calculations of second derivatives are very difficult. Basford et al. (1997) proposed using two simpler alternatives for mixture of normals: an information-based method and a bootstrap method. Although the bootstrap method is computationally intensive, it provides more reliable estimates when the sample size is relatively small or means in the mixture model are not well-separated (Basford et al., 1997). In this paper we use the bootstrap method to find the standard errors of the MLE because our ECM algorithm converges fast enough.

4. ANALYSIS OF WATER CHEMISTRY DATA IN NAKDONG RIVER WATERSHED, KOREA

Hamilton and Helsel (1995) and Min et al. (2002) reported that NO$_3^-$ and Ca$^{2+}$ concentrations were significantly elevated in groundwater beneath agricultural areas. In particular, NO$_3^-$ concentration in
water has been considered as the most important variable to describe the environmental impacts of agricultural activities on water quality (Böhlke and Denver, 1995; Canter, 1997; Min et al., 2003; Chae et al., 2004).

While many other types of clustering methods, such as k-means and hierarchical clustering (Johnson and Wichern, 2002), provide actual assignment of each observation to a group, the mixture models calculate classification probabilities:

\[
\text{Prob} \{ w_{ij} = 1 \} \text{ and } w_{ij} = 0 \text{ for all } j' \neq j | y, x, z \} = \frac{\hat{P}_j(z_i)\phi(y_i | \mu_j(x_i), \Sigma_j)}{\sum_{l=1}^{J} \hat{P}_l(z_i)\phi(y_i | \mu_l(x_i), \Sigma_l)}
\]

(7)

Then the \(i\)th water is assigned to the \(j\)th group (normal distribution component) that has the biggest classification probability. This approach helps analysts understand the relationship between an observation and groups (clusters) in terms of the classification probability. In this section, we apply the MUN and the LMMR to cluster the same Nakdong River watershed data (93 water sampling locations). All measurements except \(pH\) are used in log scales.

4.1. Mixture of univariate normal distributions

The MUN approach (Eq. 1) has been applied very widely in environmental sciences (Jeffries and Pfeiffer, 2000; Borgmann et al., 2005; Park et al., 2005). In this subsection, the MUN is applied to cluster the whole set of Nakdong River watershed observations (93 waters) into two groups based on each of four elements (\(NO_3^-\), \(Cl^-\), \(K^+\), and \(Ca^{2+}\)).

When we assume that higher concentrations of these elements indicate higher agricultural impact, the left side group (normal curve) in Figure 2 is for natural background waters and the right side group is for impacted waters. Concentrations of \(NO_3^-\) in Figure 2(a) have a very large group on the right side and a very small one on the left side, which shows that the major portion of the waters are impacted by agricultural activity. However, \(Ca^{2+}\), \(K^+\), and \(Cl^-\) concentrations in Figure 2(b), (c), and (d) show that most waters are not impacted. Even though concentrations of the four major elements are measured from the same water samples, results of these clustering analyses contradict each other. This implies that the MUN is not a proper clustering method for these waters.

As discussed in Section 2.2, we may assume that, among the 93 waters, the 61 irrigation well waters are impacted and the 15 deep groundwaters are from natural background. The performance of a clustering method can be evaluated by checking how often it finds the right group memberships of the 76 (=61 + 15) impacted or natural background waters. After the MUNs are applied to the whole data (93 observations), classification probabilities (Eq. 7) are calculated. Each water is classified into a group that has classification probability greater than 1/2. Then two successful classification rates and their average are calculated (Table 1): the proportion of irrigation well groundwaters that are classified to the impacted water group, and the proportion of deep groundwaters that are classified to the natural background water group. Note that the 17 domestic or surface waters are used for model estimation, but are not considered in calculating successful classification rates because their group memberships cannot be assumed. When a good clustering method is used, all or most of the irrigation well waters should be classified to the impacted water group that has higher concentrations of agrochemical-originated elements and all or most of the deep groundwaters to the natural background water group that has lower concentrations. This makes both successful classification rates high. When a clustering model tends to assign too many observations (waters) to one of the groups, one of the successful
classification rates will be very high and the other will be very low. If the random variable of the model contains poor information to identify targeting clusters, both rates will be low. A better clustering method will have both rates high, and therefore the average will be high. In this paper, the average of rates is used as one of methods in comparing clustering models.

Because of heavy usage of nitrogen fertilizers, the concentration of NO$_3^-$ is usually regarded as the best indicator for agricultural pollutants. However, it does not perform as expected with the Nakdong River watershed data. The MUN with NO$_3^-$ assigns many natural background waters (10 out of 15 natural background waters) to the impacted water group, in addition to assigning all impacted waters (61 out of 61 impacted waters) to the impacted water group. In other words, too many waters are assigned to the impacted water group (Figure 2(a)). As a result, the impacted waters have successful classification rate 1.000 ($=61/61$) and natural background waters do only 0.333 ($=5/15$). The average of these two rates is 0.667. When the MUNs are used with K$^+$ and Cl$^-$ concentrations, the successful classification rates have similar tendencies, but in the opposite direction. The MUNs with K$^+$ and Cl$^-$ have rates 0.148 and 0.066 for impacted waters and 0.933 and 0.800 for natural background waters. Averages of these rates are 0.541 and 0.433. Performance of the clustering model based on Ca$^{2+}$ is even worse. Impacted waters have 0.164, natural background waters have 0.267, and the average of the rates

![Logistic Mixture of Multivariate Regressions](https://example.com/image.png)

Figure 2. Histograms of log(NO$_3^-$), log(Ca$^{2+}$), log(K$^+$), and log(Cl$^-$) in waters of the Nakdong River watershed and the estimated mixture of normal distributions in solid lines.
Table 1. Successful classification rate using the MUNs (water impacted by agrochemicals vs. natural background water)

<table>
<thead>
<tr>
<th>Agrochemical elements*</th>
<th>NO₃⁻</th>
<th>Ca²⁺</th>
<th>K⁺</th>
<th>Cl⁻</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impacted natural average</td>
<td>1.000</td>
<td>0.164</td>
<td>0.148</td>
<td>0.066</td>
</tr>
<tr>
<td>Natural</td>
<td>0.333</td>
<td>0.267</td>
<td>0.933</td>
<td>0.800</td>
</tr>
<tr>
<td>Average</td>
<td>0.667</td>
<td>0.216</td>
<td>0.541</td>
<td>0.433</td>
</tr>
</tbody>
</table>

*Used in log scales.

Figure 3. Scatter plots of log(Ca²⁺), log(NO₃⁻), log(alkalinity), and pH in Nakdong River watershed: Triangles are deep groundwater (natural background water group), circles are shallow irrigation water (impacted water group), and solid dots are surface water or domestic well groundwaters (unknown group membership)
is 0.216. Even though the performance of the clustering model based on NO$_3^-$ is not good, it is still the best among these four MUNs in terms of average rate. These results are summarized in Table 1.

4.2. Logistic mixture of multivariate regressions

Chemical relationships among major agrochemical-originated elements (NO$_3^-$ and Ca$^{+}$) and physicochemical variables (alkalinity and pH) are described in Section 2.2. To utilize those relationships, the LMMR is applied using major agrochemical-originated elements as response variables and physicochemical variables as predictors in four mean functions (2 groups $\times$ 2 response variables) and the logistic function for the mixing probability. Mean or logistic functions may have one of four possible sets of predictors—{intercept}, {intercept, alkalinity}, {intercept, pH}, and {intercept, alkalinity, pH}, assuming nonzero intercepts. We compare six competing models in Table 2 and examine details of the best among these six models later on.

4.2.1. Model comparisons. All competing models in Table 2 have NO$_3^-$ and Ca$^{+}$ as response variables. Their common predictors of four mean functions ($x$), predictors of the logistic function for mixing probability ($z$), the likelihood and successful classification rates are described in Table 2. The full model, Model #6, is given by

$$
\mu_{jm}(x) = \beta_{jm0} + \beta_{jm1} \cdot \text{alkalinity} + \beta_{jm2} \cdot \text{pH}
$$

and

$$
\text{logit}(p_1(z)) = \gamma_0 + \gamma_1 \cdot \text{alkalinity} + \gamma_2 \cdot \text{pH}
$$

where $j = 1$ for impacted water group, $j = 2$ for natural background water group, $m = 1$ for response Ca$^{+}$, $m = 2$ for response NO$_3^-$, and $\mu_{jm}(x)$ is the mean function of the $m$th response variable in the $j$th group. Because the LMMRs in Table 2 are the mixture of two groups, the $j = 2$ group is used as the baseline of the logistic function. For example, Model #1 has $\beta_{111} = \beta_{112} = \beta_{121} = \beta_{122} = \beta_{211} = \beta_{212} = \beta_{221} = \beta_{222} = \gamma_2 = 0$. Each water is classified into a group that has classification probability (Eq. 7) greater than 1/2.

Comparisons of results in Tables 1 and 2 show that, overall, the LMMRs have much higher averages of successful classification rates than the MUNs. However, comparisons between the LMMRs based on these rates are not obvious as the overall comparison between the LMMRs and the MUNs, because most LMMRs in Table 2 have similarly high rates. As an alternative approach, we conducted bootstrap

<table>
<thead>
<tr>
<th>Model</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
<th>#6</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$ (mean)</td>
<td>Intercept only</td>
<td>Alkalinity</td>
<td>Intercept only</td>
<td>Alkalinity</td>
<td>Intercept only</td>
<td>Alkalinity, pH</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>-161.6</td>
<td>-166.3</td>
<td>-139.8</td>
<td>-159.9</td>
<td>-148.0</td>
<td>-116.3</td>
</tr>
<tr>
<td>Impacted</td>
<td>1.000</td>
<td>0.984</td>
<td>0.934</td>
<td>1.000</td>
<td>0.869</td>
<td>0.787</td>
</tr>
<tr>
<td>Natural</td>
<td>0.933</td>
<td>0.533</td>
<td>0.933</td>
<td>0.933</td>
<td>0.800</td>
<td>1.000</td>
</tr>
<tr>
<td>Average</td>
<td>0.967</td>
<td>0.759</td>
<td>0.934</td>
<td>0.967</td>
<td>0.835</td>
<td>0.894</td>
</tr>
</tbody>
</table>

*Common predictors of the mean functions.
*Except pH, predictors and response variables are used in log scales.
*Predictors of the logistic function.
likelihood ratio tests to compare the LMMRs. Note that likelihood ratio tests cannot be used to compare a LMMR and a MUN or to compare between the MUNs, because these models are not constructed based on the same set of random (response) variables.

Because of an identification (label-switching) problem, the usual likelihood ratio test (Wilks, 1938) does not work with the mixture model. Following McLachlan’s (1987) framework, likelihood ratios are bootstrapped to compare each of Models #1–#5 with Model #6. Test results show that Model #6 is significantly better than any other ($p$-value < 0.01 in each test). In other words, Model #6 is the best model based on likelihood ratio tests.

In terms of average of successful classification rates, Models #1 and #4 have the highest rate 0.967, while Model #6 has 0.894. However, considering uncertainty of these successful classification rates, it is difficult to determine if this difference is large enough. Also, Models #1 and #4 do not reflect that the relationships between major agrochemical-originated elements ($\text{NO}_3^-$ and $\text{Ca}^{2+}$) and physicochemical variables (alkalinity and pH) are different in impacted water and natural background water groups, as shown by the likelihood ratio tests and well-studied by environmental scientists (Min et al., 2002; Chae et al., 2004). Therefore, we suggest Model #6 as the best model for the Nakdong River watershed data.

We may also consider using likelihood-based model selection criteria, AIC or BIC, which have been widely used for various types of selection problems. However, derivations of these criteria are invalid for mixture models because of the identification problem (Hagiwara et al., 2000; Miloslavsky and Van der Laan, 2003). Nevertheless, because of the computational simplicity, they are still considered as popular model selection criteria for the selection of the number of groups (Miloslavsky and Van der Laan, 2003).

Based on Model #6, all of 6 surface waters and 5 out of 11 domestic well groundwaters are classified as natural background groupwaters.

4.2.2. Relationships of major elements in impacted and natural background waters. In addition to clustering, the LMMR estimates the relationships among major elements and physicochemical variables, which differ in the two groups of waters. Even though chemical equations such as Equations (2) and (3) can describe these relationships roughly, there are also many other biological and chemical reactions that cannot be fully detected or described. One statistical challenge is that it is difficult to predetermine whether some waters belong to the impacted or natural background water group. For example, the group memberships of domestic well groundwaters and surface waters cannot be predetermined or assumed in the Nakdong River watershed data. The LMMR calculates the classification probability (Eq. 7) of each water and finds the relationships within each group simultaneously. To show the relationships among major elements that play important roles in nitrification, estimates of parameters in Model #6 are reported in Table 3. Also, their standard errors and $p$-values are calculated using bootstrapping. Significant linear relationships are observed between log(alkalinity) and log($\text{Ca}^{2+}$) in both impacted ($\hat{\beta}_{111} = -0.49$) and natural background water ($\hat{\beta}_{211} = 0.33$) groups. The table also shows that log($\text{NO}_3^-$) has a strong negative linear relationship with log(alkalinity) in the impacted group ($\hat{\beta}_{121} = -0.81$). However, log($\text{NO}_3^-$) and log(alkalinity) do not have a significant linear relationship in natural background water. Significant negative relationships are found between pH and log($\text{NO}_3^-$) in both groups ($\hat{\beta}_{122} = -0.46$ and $\hat{\beta}_{222} = -2.54$). But pH is not significantly related with log($\text{Ca}^{2+}$). In impacted water, a strong correlation (0.92) is found between two response variables, log($\text{Ca}^{2+}$) and log($\text{NO}_3^-$). This shows that nitrogen fertilizers and lime (CaO) have been applied together on crop fields. All significant relationships are consistent with research in water chemistry (Section 2.2).
Table 3. Parameter estimates for Model #6

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Intercept</th>
<th>Alkalinity</th>
<th>pH</th>
<th>Std dev.</th>
<th>Intercept</th>
<th>Alkalinity</th>
<th>pH</th>
<th>Std dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impacted: j = 1</td>
<td>5.52 (0.45)</td>
<td>-0.49 (0.13)</td>
<td>-0.05 (0.09)</td>
<td>0.37</td>
<td>9.72 (0.70)</td>
<td>-0.81 (0.21)</td>
<td>-0.46 (0.14)</td>
<td>0.60</td>
</tr>
<tr>
<td>Natural: j = 2</td>
<td>1.08 (0.65)</td>
<td>0.33 (0.12)</td>
<td>0.19 (0.12)</td>
<td>0.34</td>
<td>14.69 (2.90)</td>
<td>1.04 (0.57)</td>
<td>-2.54 (0.60)</td>
<td>1.66</td>
</tr>
</tbody>
</table>

Logistic function for mixing probability

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Intercept</th>
<th>Alkalinity</th>
<th>pH</th>
<th>Std dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impacted: j = 1</td>
<td>19.00 (6.41)</td>
<td>-7.84 (2.68)</td>
<td>1.99 (0.95)</td>
<td></td>
</tr>
</tbody>
</table>

*p-value < 0.01.

Standard errors of mean function parameter estimates are in parentheses.
Because Model #6 uses the same predictors (alkalinity and pH) five times, for one logistic function and four mean functions (2 groups × 2 response variables), the existence of collinearity might be a concern among the five slope parameters of each predictor. The correlation matrix for the five slope parameters of alkalinity is estimated using Monte Carlo simulations as follows:

\[
\text{Corr} \begin{pmatrix}
\hat{\beta}_{111} \\
\hat{\beta}_{211} \\
\hat{\beta}_{121} \\
\hat{\beta}_{221} \\
\hat{\gamma}_1
\end{pmatrix} = \begin{pmatrix}
1.00 & 0.12 & 0.91 & 0.05 & 0.13 \\
0.12 & 1.00 & 0.07 & 0.06 & -0.10 \\
0.91 & 0.07 & 1.00 & 0.02 & 0.16 \\
0.05 & 0.06 & 0.02 & 1.00 & 0.03 \\
0.13 & -0.10 & 0.16 & 0.03 & 1.00
\end{pmatrix}
\]

As defined in Equation (8), \(\hat{\beta}_{111}\) and \(\hat{\beta}_{121}\) are the slope parameters of \textit{alkalinity} in the two mean functions (response variables) for the impacted water group, \(\hat{\beta}_{211}\) and \(\hat{\beta}_{221}\) are the slope parameters for the natural background water group, and \(\hat{\gamma}_1\) is the slope parameter in the logistic function. Except for only one correlation (\(\text{corr}(\hat{\beta}_{111}, \hat{\beta}_{121}) = 0.91\)), magnitudes of all other correlations are smaller than 0.20. A high correlation between \(\hat{\beta}_{111}\) and \(\hat{\beta}_{121}\) can be explained as follows. First, response variables \(\text{Ca}^{2+}\) and \(\text{NO}_3^-\) have a high correlation within the impacted water group (0.92 in Table 3). Second, \textit{alkalinity} is used as a common predictor in the two mean functions within the impacted water group. Therefore, their slope parameters (\(\hat{\beta}_{111}\) and \(\hat{\beta}_{121}\)) should have a high correlation within the impacted water group. Note that the slope parameters have a low correlation within the natural background water group (\(\text{corr}(\hat{\beta}_{211}, \hat{\beta}_{221}) = 0.06\)), where the correlation between two response variables is low (0.04 in Table 3). Similar patterns are observed with the five slope parameters of pH as follows:

\[
\text{Corr} \begin{pmatrix}
\hat{\beta}_{112} \\
\hat{\beta}_{212} \\
\hat{\beta}_{122} \\
\hat{\beta}_{222} \\
\hat{\gamma}_2
\end{pmatrix} = \begin{pmatrix}
1.00 & -0.06 & 0.92 & -0.08 & 0.16 \\
-0.16 & 1.00 & -0.07 & 0.01 & 0.11 \\
0.92 & -0.07 & 1.00 & -0.09 & 0.18 \\
-0.08 & 0.01 & -0.09 & 1.00 & -0.09 \\
0.16 & 0.11 & 0.18 & -0.09 & 1.00
\end{pmatrix}
\]

Except for only one correlation (\(\text{corr}(\hat{\beta}_{122}, \hat{\beta}_{122}) = 0.92\)), magnitudes of all correlations are smaller than 0.20. Again, this can be explained with a high correlation between the two response variables. Overall, we did not observe any evidence of multicollinearity that is caused by using the same predictors for the logistic function and the mean functions.

5. CONCLUSION

Shallow groundwater and surface water systems in rural areas are vulnerable to contamination by the excessive application of agrochemicals, such as fertilizer, pesticides, and lime. Even though nitrogen fertilizer is one of the biggest agrochemical pollutant sources, our results show that waters cannot be clustered properly into environmentally meaningful groups based only on the concentration of \(\text{NO}_3^-\). This paper proposes the LMMR to cluster waters and estimate the relationship between the major agrochemical-originated elements and physicochemical variables. Based on the likelihood ratio test, the best model is found with log(\(\text{NO}_3^-\)) and log(\(\text{Ca}^{2+}\)) as responses and log(alkalinity) and pH as predictors. Parameter estimates of the best model are consistent with research in environmental sciences.
REFERENCES


Here is detailed description of the expectation-conditional maximization (ECM), Meng and Rubin, 1993) algorithm for the logistic mixture of multivariate normals.

**E-step:** Calculate the expectation of log complete data likelihood.

When \( w_i = (w_{i1}, \ldots, w_{ij}) \) is considered as the missing variable, the log complete data likelihood of \( y \) and \( w \) is

\[
\log f(y, w|\theta) = \sum_{i=1}^{N} \sum_{j=1}^{J} w_{ij} \cdot \log \{ \phi(y_i|\mu_j(x_i), \Sigma_j) \cdot p_j(z_i) \}
\]

Then, the expectation of log complete data likelihood is

\[
Q(\beta, \Sigma, \gamma | P^{(k-1,k-1)}_{ij}, y) = E_w \{ \log f(y, w|\beta, \Sigma, \gamma) \}
\]

\[
= \sum_{i=1}^{N} \sum_{j=1}^{J} P^{(k-1,k-1)}_{ij} \cdot \log \{ \phi(y_i|\mu_j(x_i), \Sigma_j) \cdot p_j(z_i) \}
\]

**CM-step:** Maximize \( Q(\cdot) \) with respect to \( \beta = (\beta_1, \ldots, \beta_J), \Sigma = (\Sigma_1, \ldots, \Sigma_J) \) and \( \gamma = (\gamma_1, \ldots, \gamma_{J-1}) \).

To calculate \( \beta^{(k)}_j \),

\[
\frac{\partial Q(\beta, \Sigma, \gamma | P^{(k-1,k-1)}_{ij}, y)}{\partial \beta_j} \bigg|_{\beta=\beta^{(k-1)},\gamma=\gamma^{(k-1)}} = 0
\]

for \( J = 1, \ldots, J \). Then, this leads to Step 3.

To calculate \( \Sigma^{(k)}_j \),

\[
\frac{\partial Q(\beta, \Sigma, \gamma | P^{(k-1,k-1)}_{ij}, y)}{\partial \Sigma_j} \bigg|_{\beta=\beta^{(k-1)},\gamma=\gamma^{(k-1)}} = 0
\]

for \( J = 1, \ldots, J \). Then, this leads to Step 4.

To calculate \( \gamma^{(k)}_j \),

\[
\frac{\partial Q(\beta, \Sigma, \gamma | P^{(k-1,k-1)}_{ij}, y)}{\partial \gamma_j} \bigg|_{\beta=\beta^{(k)}, \Sigma=\Sigma^{(k-1)}} = \sum_{i=1}^{N} \left\{ \left( P^{(k,k-1)}_{ij} - P^{(k)}_{ij}(z_i) \right) z_i \right\} = 0
\]

for \( J = 1, \ldots, J \). Since the solution is not available in the closed form, we use a Newton–Raphson algorithm as in Step 5.