# Clustering of Temporal Profiles Using a Bayesian Logistic Mixture Model: Analyzing Groundwater Level Data to Understand the Characteristics of Urban Groundwater Recharge

Yongsung Joo [*], Babette Brumback [†], Keunbaik Lee [‡], Seong-Taek Yun [§], Kyoung-Ho Kim [¶], and Chaeman Joo [‖]

August 19, 2008

[*]Assistant Professor, Department of Statistics, Dongguk University, Seoul 100-715, Korea. e-mail:yjoo@phhp.ufl.edu

[†]Associate Professor, Biostatistics, College of Public Health and Health Professions, University of Florida, Gainesville, FL 32611, USA.

[‡]Assistant Professor, Biostatistics, Louisiana State University-Health Science Center, New Orleans, LA 70122, USA.

[§]Professor, Department of Earth and Environmental Sciences and the Environmental Geosphere Research Lab (EGRL), Korea University, Seoul 136-701, Korea (South).

[¶]Graduate student, Department of Earth and Environmental Sciences and the Environmental Geosphere Research Lab (EGRL), Korea University, Seoul 136-701, Korea (South).

[‖]Graduate student, Department of Earth and Environmental Sciences and the Environmental Geosphere Research Lab (EGRL), Korea University, Seoul 136-701, Korea (South).

**Abstract**

The hydrogeological conditions of groundwater can be examined by carefully studying the patterns of fluctuations in groundwater levels. These fluctuations are spatially and temporally influenced by many complicated factors, including rainfall, topography, land use, and hydraulic properties of soils and bedrock (*i.e.*, aquifers). In this paper, we develop methodology based on a Bayesian logistic mixture model to simultaneously 1) cluster profiles of groundwater level changes over time and 2) estimate the relationship between the characteristics of each cluster and environmental variables. We apply the methodology to analyze groundwater level profiles from 37 monitoring wells in Seoul, South Korea, and we find four clusters of wells. Using the estimated relationship between the clusters and the environmental variables, we discern the hydrogeologic conditions of each cluster. Thus, we gain a better understanding of the recharge and subsurface flow of bedrock groundwater in an urban setting and the vulnerability of groundwater to the inflow of potential pollutants from ground surface.

**Key Words**: Model-based clustering; Clustering of time course data; Hydrogeology.

# 1 INTRODUCTION

Many environmental studies have emphasized the crucial importance of groundwater for a secure water supply in urban areas (Yang *et al.* 1999; Park *et al.* 2005). However, the hydrogeological condition and quality of urban groundwater have been significantly altered in many countries by over-usage (*e.g.*, over-pumping), decreased surface permeability due to paving and construction of subsurface facilities such as underground spaces and subways, inflow of surface pollutants, and leakage of water pipelines and sewage water. (Barrett *et al.* 1999; Yang *et al.* 1999; Park *et al.* 2005; Chae *et al.* 2008). To sustainably manage and utilize urban groundwater resources, a solid understanding is required of their hydrologic cycles, which consist of recharge from (infiltration of) rainwater and snowmelt, subsurface flow, and discharge into streams, lakes, and oceans (Lerner 2002).

A simplified conceptual model on the hydrologic cycle of urban groundwater is depicted in Figure 1. Potentially carrying diverse contaminants from the urban ground surface, the precipitated water *infiltrates* (*i.e.*, moves down almost vertically with relatively little lateral movement) from the ground surface to the water table (*i.e.*, the upper boundary of the water-saturated zone) through pores in soil or fracture networks in bedrock. Then, this infiltrated water *flows* along highly fractured bedrock zones, getting mixed in with infiltrated water from other locations. Once this groundwater reaches a monitoring well, the groundwater level rises around this well. The arrival of groundwater signals the entry of potential contaminants. As the groundwater flows away from the well to discharge zones (*e.g.*, streams, lakes, and rivers), the groundwater level falls back down.

[Figure 1 about here.]

This fluctuation of groundwater level varies depending on natural factors (*e.g.*, rainfall and hydraulic properties of aquifers) and anthropogenic factors (*e.g.*, land use, artificial

surface impermeability, over-pumping, and the like). Therefore, careful examination of the relationship may yield valuable information on the recharge characteristics and hydraulic properties of aquifers, and the vulnerability of groundwater to surface pollution. Groundwater level data have been analyzed by environmental scientists using spectral analysis (Larocque *et al.* 1998), the transfer function model (Lee and Lee 2000), principle component analysis (Moon *et al.* 2004), and a number of other methods. However, due to the complexity of the fluctuation patterns and of the environmental factors in urban areas, these methods have met with limited success.

In this paper, we propose a novel Bayesian logistic mixture model that simultaneously clusters temporal groundwater level profiles and examines the relationship between these clusters and environmental (geological and geographical) conditions. Analysis results can be easily interpreted and help environmental scientists and policy makers better understand recharge and subsurface flow of bedrock groundwater in the urban setting and identify locations with potentially high vulnerability to contaminants from the ground surface.

We apply the proposed methodology to analyze bedrock groundwater level data from Seoul, the capital city of South Korea. Seoul is one of the largest and most densely populated cities in the world, where approximately 10 million people reside in an area of 605 $km^2$ ($\approx$ 17000 people/$km^2$). As in many other major cities around the world (Yang *et al.* 1999; Zilberbrand *et al.* 2001; Vazquez-Sune *et al.* 2005), Seoul struggles with qualitative and quantitative maintenance of groundwater. Weekly averages of groundwater levels (unit:$cm$) were recorded at 37 monitoring wells in Seoul for 36 weeks (from March 1 to November 7, snow-free season) in 2001. This accounts for 37 temporal groundwater level profiles each with 36 observation times; see Figure 2. In addition, four environmental variables, discussed in Section 5.1.3, were measured on each of the wells. Wells-in-use

have not been considered in this study, to exclude the pumping effect on groundwater level change. Considered wells are designed to prohibit direct inflow of surface contaminants. These wells were drilled down to the fractured bedrock aquifer, with the average depth of about 35 $m$, and are cased with concrete down to the boundary between soil (or alluvium) and bedrock at depths of $<$20 $m$ from the land surface. Within the monitored period in 2001, a monsoon caused the major rainfalls between the 16th and the 24th weeks (gray-shaded vertical band in Figure 2), with the peak time occurring at the 20th week (thick black vertical line in Figure 2).

[Figure 2 about here.]

The plan of the paper is as follows. In Section 2, we review clustering methods based on mixture models and provide more motivation for clustering groundwater profiles. In Section 3, we describe the proposed clustering methodology in detail. Section 4 explains how to calculate the Deviance Information Criterion (DIC) to select the number of clusters for our model. We analyze the Seoul groundwater level data in Section 5. Additionally, we provide the hydrogeological background that enables us to interpret the results of our analysis, including information on the time lag between rainfall and a rise of groundwater level, the recharge mechanisms, and associated environmental variables. Section 6 concludes with a brief discussion.

# 2 MODELING BACKGROUND

## 2.1 Previous Use of the Mixture Model

Mixture models form the basis for the most common model-based clustering methods. The mixture model is

$$f(Y_i) = \sum_{k=1}^{K} p_k f_k(Y_i), \tag{1}$$

where $f_k(Y_i)$ is the probability density function of cluster $k$, $Y_i$ is a response variable of object $i$, $i = 1, \ldots, n$, $p_k$s are the mixing probabilities, and $\sum_{k=1}^{K} p_k = 1$. In the mixture model, the cluster membership of object $i$ is determined by the posterior membership probability:

$$\text{Prob[object } i \text{ belongs to cluster } k | Y_1, \ldots, Y_n] = \frac{p_k f_k(Y_i)}{\sum_{k'} p_{k'} f_{k'}(Y_i)}.$$

Compared to such model-free clustering methods as hierarchical clustering and k-means algorithm, the mixture model has two important advantages. First, the mixture model simultaneously provides the cluster membership probability of an object and estimates the distribution of each cluster, whereas model-free methods only provide the cluster membership with group labels and cannot offer detailed information on characteristics of clusters without additional data exploration. Second, methods based on mixture models can be easily customized to meet various research goals because they are model-based. A mixture of multivariate normals is commonly applied to cluster objects based on multivariate responses (Basford and McLachlan 1985). To cluster data points based on their different linear trends in each group, Turner (2000) studied the mixture of univariate

regressions, in which $f_k(Y_i) = f_k(Y_i|x_i)$ is the normal linear model with predictor $x_i$. Also, recent research has focused on clustering temporal profiles using mixture models (James and Sugar 2003; Luan and Li 2003; Ma *et al.* 2006).

Although the mixing probabilities are assumed to be constant parameters in all of the above mixture model references, these probabilities will vary with object-specific predictors ($W_i$) in many practical scenarios. To accommodate this, several researchers have applied mixture models with a logistic regression structure for the mixing probabilities ($p_k(W_i)$ instead of $p_k$). Thus, the posterior cluster membership probability of each object also depends on $W_i$. This type of model is called the mixtures-of-experts (Peng, Jacobs, and Tanner 1996; Jiang and Tanner 1999) or the *logistic mixture* model (Jeffries and Pfeiffer 2000; Wong and Li 2001; Pfeiffer *et al.* 2007; Joo *et al.* 2007). Peng *et al.* (1996) studied the logistic mixture of exponential family regression models for a speech recognition problem. Jiang and Tanner (1999) discussed theoretical aspects of this model. Jeffries and Pfeiffer (2000) and Pfeiffer *et al.* (2007) studied the logistic mixture of univariate log normal distributions and multivariate normal distributions. Wong and Li (2001) proposed the logistic mixture of autoregressive regression models. Joo *et al.* (2007) applied the logistic mixture of multivariate regressions to environmental pollution problems. To distinguish models with $p_k(W_i)$ and $p_k$, the model with a constant mixing probability $p_k$ will be called the *plain mixture* model.

For the purpose of analyzing our groundwater level data, we improve previous clustering methods for time course data (such as those proposed by James and Sugar 2003, Luan and Li 2003, and Ma *et al.* 2006) by incorporating logistic mixing probabilities, instead of constant parameters; thus our model can explain the relationship between cluster-specific groundwater level fluctuation patterns and environmental conditions.

## 2.2 Other Approaches for Groundwater Level Profiles

Let $\acute{Y}_{it}$ be the groundwater level (unit:$cm$) in well $i$ at time $t$, where $i = 1, \ldots, n$, and $t = 1, \ldots, T$. Here, $\acute{Y}_i = (\acute{Y}_{i1}, \ldots, \acute{Y}_{it}, \ldots, \acute{Y}_{iT})^T$ constructs the profile of well $i$ with $T$ time points. Let $W_i$ be a vector of well-specific environmental variables. To find the relationship between temporal changes in groundwater level and environmental conditions, one might consider the following approaches. As a naive approach, one may use a semi-parametric *additive model*

$$\acute{Y}_{it} = s(t) + g(W_i) + \acute{\epsilon}_{it}, \tag{2}$$

where $s(t)$ is a smooth function of $t$, $g(W_i)$ is a parametric function of environmental variables, and $\acute{\epsilon}_{it}$ is an independent normal error term with mean zero and variance $\acute{\sigma}^2$. In the naive model (2), $s(t)$ explains the common pattern of all profiles and $g(W_i)$ adjusts the intercepts of groundwater profiles with the environmental variable $W_i$. However, environmental conditions affect the shape of profiles as well as intercepts (See Section 5.1 for detailed hydrological reasonings). As another alternative, one might use a *non-additive model* with a smooth mean function of $t$ and $W_i$ to explain the changes of profile shapes with environmental conditions:

$$\acute{Y}_{it} = h(t, W_i) + \acute{\epsilon}_{it}, \tag{3}$$

where $h(t, W_i)$ is a non-additive function with smooth functions and interaction terms between spline bases of $t$ and $W_i$ (Ruppert *et al.* 2003). Unfortunately, this model is difficult to use when $n$ is relatively small compared to the number of knots in the smooth function, because $h(t, W_i)$ has a very large number of interaction terms even when $W_i$ contains only three or four environmental variables. Detailed discussions of

non-additive model (3) are given in Ma and Zhong (2008). Many studies found that patterns of groundwater level profiles tend to cluster (Larocque *et al.* 1998; Lee and Lee 2000; Moon *et al.* 2004). Thus, one might consider the following two-step analysis: 1) cluster the patterns of centered groundwater level profiles $Y_i = (Y_{i1}, \ldots, Y_{it}, \ldots, Y_{iT})^T$, where $Y_{it} = \acute{Y}_{it} - \sum_{t'} \acute{Y}_{it'}/T$, and 2) characterize each cluster based on environmental predictors related to infiltration and flow of the groundwater. Note that all monitoring wells have different average (or baseline) groundwater levels, which are strongly related to the altitude of the ground surface around the well (correlation = 0.96 in our data set). Therefore, groundwater levels need to be centered for a proper cluster analysis based on patterns of groundwater level profiles. This two-step analysis is logically simple and is easy to implement using a relatively small number of parameters. However, because the two steps are separated, it has disadvantages. First, the environmental factors are not reflected in determining clusters, even though their important relationships are well-known (Fetter 2000). Second, the uncertainty in the clustering analysis is not reflected in the second step. Our Bayesian logistic mixture model overcomes these disadvantages because it accomplishes the two objectives simultaneously.

# 3 OUR MODEL

Graphical visualization in Figure 2 showed that the profiles cannot readily be modeled with common parametric approaches, and furthermore that they are not periodic. We thus decided to use penalized splines to represent cluster-specific profile patterns as smooth semiparametric curves measured with error.

Recall that centered groundwater level profile $i$ is denoted $Y_i = (Y_{i1}, \ldots, Y_{it}, \ldots, Y_{iT})^T$. For the Seoul groundwater level data, $i = 1, \ldots, 37$ and $t = 1, \ldots, 36$. To develop our

method, we first suppose that a given profile $i$ belongs to a given cluster $k$, where $k \in \{1, \ldots, K\}$. We assume that the data in profile $i$ reflect a smooth underlying cluster-specific trend plus an additional measurement error:

$$Y_{it} = s_k(t) + \epsilon_{it},$$

where $s_k(t)$ is the smooth trend for cluster $k$, and $\epsilon_{it}$ is an independent normal mean zero error term with variance $\sigma_k^2$. To incorporate the assumption of smoothness into our analysis, we specified a Bayesian cubic smoothing spline for the $s_k$, as follows.

First, we selected a cubic B-spline basis with knots at inner time points $t = 2, \ldots, T-1$ and evaluated them at the observation times $t = 1, \ldots, T$. Let $B$ denote the resulting design matrix and let $B_t$ denote the $t^{th}$ row. The design matrix is identical for each profile because the observations occur at the same time points. Letting

$$s_k(t) = B_t \nu_k, \tag{4}$$

where $\nu_k$ is a parameter vector, we incorporate smoothness with an improper prior on $\nu_k$ (Brumback *et al.* 2007),

$$p(\nu_k) \propto \exp\left(-\frac{\eta_k}{2\sigma_k^2}\nu_k^T D \nu_k\right),$$

where $\eta_k$ is a tuning parameter controlling smoothness, and $D$ has $(s, r)^{th}$ entry

$$D_{sr} = \int_1^T B_{ts}^{(2)} B_{tr}^{(2)} dt,$$

where $B_{ts}^{(2)}$ is the second derivative of the $s^{th}$ B-spline basis function evaluated at $t$, and $(1, T)$ is the support of the B-spline basis. The prior $p(\nu_k)$ is constructed so that the

posterior mean of $s_k(t)$ will be a cubic smoothing spline.

An equivalent form of the model uses a spectral decomposition for $D$; denote it by $Q\Psi Q^T$, where $Q = (Q_1, Q_2)$ is an orthonormal matrix with the two columns in $Q_1$ corresponding to the zero eigenvalues of $D$ and the columns of $Q_2$ representing the other eigenvalues in decreasing order. $\Psi$ is the diagonal matrix of eigenvalues, which we set equal to the direct sum of $\Psi_1$ with all entries equal to zero and $\Psi_2$ with the nonzero eigenvalues on its diagonal. Using this decomposition, we can rewrite (4) as a mixed model:

$$s_k(t) = X_t\beta_k + Z_tU_k,$$

where $X_t\beta_k = B_tQ_1\beta_k^T$, $Z_t = B_tQ_2\Psi_2^{-1/2}$, $U_k$ is multivariate normal with mean zero and variance $\lambda_k^2 I_{T-2}$, $I_{T-2}$ is the $T-2$ by $T-2$ identity matrix, and $p(\beta_k) = 1$. It can be shown that a choice of $X_t = (1, t)$ is possible because $BQ_1$ spans the affine functions. Denote $X = (X_1^T, \ldots, X_t^T, \ldots, X_T^T)^T$, $Z_t = (Z_{t1}, \ldots, Z_{t\,T-2})$, and $Z = (Z_1^T, \ldots, Z_t^T, \ldots, Z_T^T)^T$. In this paper, we use a proper normal flat prior for $p(\beta_k)$ for easier implementation.

Next, we explain the logistic mixture structure, which allows the mixing probability $p_i$ in (1) to change with object-specific (well-specific) predictors. Let $W$ be the $n$ by $q+1$ design matrix in the logistic regression, which includes the intercept and $q$ predictors, $\alpha_k = (\alpha_{k0}, \ldots, \alpha_{kl}, \ldots, \alpha_{kq})^T$ be the vector of corresponding parameters for cluster $k$, and $\alpha = (\alpha_1, \ldots, \alpha_K)$. Also, let $W_i$ be the $i^{th}$ row of $W$, and $\theta = (\beta^T, \alpha^T, \sigma^2, \lambda^2)^T$ be a set of all parameters. Then, the likelihood function corresponding to our model is

$$f(Y|\theta, U) = \prod_{i=1}^{n} \sum_{k=1}^{K} \left[ p_k(W_i, \alpha) \prod_{t=1}^{T} \phi(Y_{it}|X_t\beta_k + Z_tU_{kt}, \sigma_k^2) \right], \tag{5}$$

where $\phi(\cdot)$ is the normal probability density function, $U = (U_1, \ldots, U_K)^T$, $U_k = (U_{k1}, \ldots, U_{k\,T-2})$,

$U_{kt} \sim N(0, \lambda_k^2)$, and

$$p_k(W_i, \alpha) = \frac{\exp(W_i \alpha_k)}{\sum_{k'=1}^{K} \exp(W_i \alpha_{k'})}.$$

We set $\alpha_1$ equal to a column vector of $q+1$ zeros because of the identification problem. Note that, because the profiles are the objects to be clustered in our model, the likelihood function $\prod_{t=1}^{T} \phi(Y_{it}|X_t\beta_k + Z_t U_{kt}, \sigma_k^2)$ replaces $f_k(Y_i)$ in (1).

For each regression parameter in vector $\beta_k$ and $\alpha_k$, a flat normal prior, $N(0, \tau^2)$, is used with a large variance $\tau^2$ $(= 10^7)$. For each variance parameter, $\sigma_k^2$ and $\lambda_k^2$, the inverse gamma prior, $IG(a, b)$, is used with small $a$ $(= 0.01)$ and $b$ $(= 0.01)$. Then, our model (5) is estimated using WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml). We also provided full conditional posterior distributions in Appendix A.

# 4 SELECTION OF THE NUMBER OF CLUSTERS USING THE DIC

We use the Deviance Information Criterion (DIC) (Spiegelhalter et al. 2002) to determine the number of clusters. Define $D(\theta) = -2 \log l(\theta|Y)$, $\bar{\theta} = E(\theta|Y)$, and $\bar{D}(\theta) = E[D(\theta)|Y]$, where $l(\theta|Y)$ is the observed data likelihood. The DIC is formally defined as $2\bar{D}(\theta) - D(\bar{\theta})$. The lowest DIC value indicates the most preferred model.

Celeux *et al.* (2006) compared different forms of DICs for the mixtures model and suggested that $\text{DIC}_3$ and $\text{DIC}_4$ were the most reliable of the DICs for the mixture models. The $\text{DIC}_3$ is based on the observed data likelihood, $l(\theta|Y)$, and the $\text{DIC}_4$ is based on the complete data likelihood, $l(\theta|Y, u, d)$, where $u$ is random effect, $d$ =

$(d_{11}, \ldots, d_{1K}, \ldots, d_{n1}, \ldots, d_{nK})$, and

$$d_{ik} = \begin{cases} 1, & \text{if object } i \text{ belongs to cluster } k; \\ 0, & \text{otherwise,} \end{cases}$$

Then $\mathrm{DIC}_3$ and $\mathrm{DIC}_4$ are given by

$$DIC_3 = -4E_\theta[\log l(\theta|Y)|Y] + 2\log \hat{f}(Y)$$

and

$$DIC_4 = -4E_{\theta,u,d}[\log l(\theta|Y,u,d)|Y] + 2E_{u,d}[\log l(E_\theta(\theta|Y,u,d)|Y)|Y],$$

where

$$\hat{f}(Y) = E_\theta\left[f(Y|\theta)|Y\right].$$

In this paper, we use $\mathrm{DIC}_4$ as the model selection criterion because $\mathrm{DIC}_3$ requires marginalized likelihood functions that are not available in a closed form for our proposed model. However $\mathrm{DIC}_4$ is easily calculated. The first and second terms of $\mathrm{DIC}_4$ for our proposed model are, respectively, approximated by MCMC algorithms as

$$\begin{aligned} & E_{\theta,u,d}[\log l(\theta|Y,u,d)|Y] \\ & \approx \frac{1}{R}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{R} P(d_{ik}=1|\theta^{(l)},U^{(l)},Y) \\ & \quad \times \left\{ \log p_k(W_i,\alpha^{(l)}) + \sum_{t=1}^{T} \log \phi(Y_{it}|X_t\beta_k^{(l)} + Z_tU_k^{(l)}, \sigma_k^{2(l)}) + \log \phi(U_k^{(l)}|0,\lambda_k^{2(l)}) \right\} \end{aligned}$$

and

$$
\begin{aligned}
&E_{u,d}[\log l(E_\theta(\theta|Y,u,d)|Y)|Y] \\
&\approx \frac{1}{R}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{R} P(d_{ik}=1|\hat{\theta},U^{(l)},Y) \\
&\quad\times\left\{\log p_k(W_i,\hat{\alpha})+\sum_{t=1}^{T}\log\phi(Y_{it}|X_i\hat{\beta}_k+Z_tU_k^{(l)},\hat{\sigma}_k^2)+\log\phi(U_k^{(l)}|0,\hat{\lambda}_k^2)\right\},
\end{aligned}
$$

where $R$ is the number of iterations, $\theta^{(l)} = ({\beta^{(l)}}^T, {\alpha^{(l)}}^T, {\sigma^2}^{(l)}, {\lambda^2}^{(l)})^T$ and $U^{(l)}$ are the simulated values at the $l^{th}$ MCMC iteration,

$$
P(d_{ik}=1|\theta,U,Y)=\frac{p_k(W_i,\alpha)\phi(U_k|0,\lambda_k^2)\prod_{t=1}^{T}\phi(Y_{it}|X_i\beta_k+Z_tU_k,\sigma_k^2)}{\sum_{k'=1}^{K}p_{k'}(W_i,\alpha)\phi(U_{k'}|0,{\lambda_{k'}}^2)\prod_{t=1}^{T}\phi(Y_{it}|X_i\beta_{k'}+Z_tU_{k'},\sigma_{k'}^2)},
$$

$\hat{\theta} = (\hat{\beta}^T, \hat{\alpha}^T, \hat{\sigma^2}, \hat{\lambda^2})^T$, and $(\hat{\beta}^T, \hat{\alpha}^T, \hat{\sigma^2}, \hat{\lambda^2}) = E_\theta(\beta^T, \alpha^T, \sigma^2, \lambda^2|Y, u, d)$.

# 5 APPLICATION

In this section, we provide more details on the necessary hydrological background in order to present and interpret the results of our data analysis.

## 5.1 Hydrogeological Background

### 5.1.1 Time Lag

Time lag between rainfall and a rise of groundwater level can be roughly estimated by comparing the time of peak rainfall and the time when groundwater level is the highest (Larocque *et al.* 1998). Because of unknown geological and hydrogeological features below the ground surface, the impact and time lag of rainfall is highly variable and thus is difficult

to estimate accurately. The time lag varies from a few hours to several months depending on environmental conditions such as the hydrogeological characteristics of aquifers and ground conditions. We may deduce vulnerability of groundwater to contaminants on the ground surface using the length of time lag. If homogeneous hydraulic conductivities can be assumed in the study area, we can also deduce the locality of infiltrated water. A short time lag indicates that the groundwater level of a monitoring well is raised by infiltrated water from neighboring ground surface or rapid flow through fractured zones with little dispersion or both. Particularly, the fracture flow may facilitate rapid and relatively long-range underground movement of contaminants.

### 5.1.2 Groundwater Recharge Mechanism and Subsurface Flow

Rainwater infiltrates almost vertically to the groundwater system. Then, the infiltrated water flows laterally following the hydraulic gradient (*i.e.*, change in groundwater levels over a unit distance). This lateral flow is often categorized into two types, local and regional, based on the relative distance between the point where rainwater infiltrates and the location of the monitoring well (Fetter 2000). The *local flow* is defined as lateral groundwater flow that is initiated by infiltration from the ground surface relatively close to the monitoring well, while the *regional flow* is initiated by infiltration from the ground surface relatively far from the well.

Generally, a watershed has multiple channels of local and regional flows. At locations that are not close to major discharge zones (*i.e.*, major rivers or oceans), local or regional flow often has its own path to a discharge zone without being mixed with others. When a well is located on that path, the groundwater profile will have a distinctive fluctuation pattern reflecting that flow. However, at locations that are close to major discharge zones, multiple local and regional flows usually get mixed with each other (Fetter 2000).

If infiltration occurs close to the monitoring well, most of the infiltrated water will flow laterally to the well, while some will disperse. Thus, the groundwater level will rise effectively with only a short time lag. If the infiltrated water travels a long distance to the monitoring well, most of it disperses and gets mixed with infiltrated water from other ground surface sources. Thus the immediate effect of rainfall will be sluggish and have a long time lag. In regional flow, groundwater level profiles can be even flat because the effect of rainfall is negligible and the time lag is long (sometimes longer than a year). Compared to other areas, the time lag is generally shorter in Seoul area, which is covered with relatively thin (less than $20m$) soil zones over crystalline bedrock.

### 5.1.3 Environmental Variables

Precipitation is one of the most important factors that influence groundwater level. Although the study area is a large metropolitan city, no spatial heterogeneity of precipitation was considered in our study; hence, precipitation was considered as an insignificant factor in clustering groundwater level fluctuation patterns. Instead, relationships between pattern in each cluster and precipitation were used for the interpretation of analysis results in Section 5.2. In this study, most wells are not located close to major natural and artificial underground structures, such as major faults, tunnels and subways. Also, they are far from the western coastal area. If wells were significantly affected by major underground structures or tides or storage-discharge, distinctive patterns (such as sudden drop or increase of groundwater level) of these effects must appear in the groundwater level profiles. But, we could not observe these patterns in our data.

As potentially important environmental (well-specific) predictors ($W_i$) that may affect groundwater level profiles, we consider four environmental variables: *hydraulic head*, *soil depth*, *percentage of permeable ground surface*, and *distance to river*. These data are

provided in Appendix B. The percentage of permeable ground surface was used as an indicator of runoff because infiltration (and runoff) is generally controlled by permeability of land surface. Hydraulic head, soil depth, and distance to river were used to explain the natural groundwater discharge mechanism. The four predictors are illustrated in Figure 3. Their detailed definitions and expected effects are as follows:

[Figure 3 about here.]

1) *Hydraulic head* is the annual average distance between groundwater level and sea level, while groundwater level is often defined as a relative measurement. If a sea level of zero is used in defining groundwater level, this groundwater level is the same as the hydraulic head. Because the interest of this paper is the change of groundwater level rather than hydraulic head, we constructed a clustering model for the centered groundwater level profiles. However, the shapes of the profiles may have meaningful relationships with hydraulic head. In mountain areas, hydraulic head is high and changes following the dramatic changes of topography; the higher the hydraulic head, generally the higher the hydraulic gradient. Because this high hydraulic gradient makes lateral groundwater flow fast (Bockgard 2004), we may often observe the effective rise of groundwater level after a rainy season at locations where hydraulic head is high. In basin areas, hydraulic gradient is usually low. Thus groundwater flow is slow and the effect of groundwater is sluggish. Hydraulic head is one of the most important factors that determine the effect of flow.

2) *Soil depth* is the thickness of the unconsolidated soil zone immediately below the ground surface. In general, if the soil zone is shallow, rainwater can infiltrate easily to the water table resulting in a rapid increase of groundwater level (Rodhe and Bockgard 2006).

3) *Percentage of permeable ground surface* is measured within a circle with $250m$ radius from a well. A low percentage of permeable ground surface increases surface runoff

of rainwater and reduces the infiltration and recharge of rainwater. For example, only a small amount of rainwater can infiltrate in densely paved areas.

4) *Distance to river* is the distance between a monitoring well and the main discharge zone of groundwater, which is the Han River in Seoul area. As explained in Section 5.1.2, a mixture of multiple local and regional flows will affect the groundwater level at locations close to the Han River, while profiles tend to take the distinctive pattern of either local or regional flow at locations not close to river (*i.e.*, close to mountainous areas) (Fetter 2000). Alternatively, we may consider the distance to the closest stream as an environmental variable. However, it has a couple of disadvantages. First, because there are too many small and curvy streams, it is difficult to quantify the distance to every small stream. Second, even when the closest stream is found, we should conduct an additional step of quantification on the characteristics (such as size and flow rate) of streams at the locations where groundwater is possibly discharged. These quantifications are difficult and beyond the scope of this paper.

## 5.2 Results

We considered six competing models: Bayesian logistic mixture models (5) with 2, 3, 4, and 5 clusters, the spline model without clustering (K=1 and $W_i = 1$ in (5)), and the plain mixture of 4 clusters (K=4 and $W_i = 1$ in (5)). The DIC values for these models were 20677.4, 20386.1, 19712.8, 19810.1, 21536.6 and 19794.4 respectively. Because the logistic mixture of 4 clusters had the lowest DIC value, we consider it the best for this data set among competing models. The spline model without clustering had a higher DIC value than any other considered mixture models. This indicated that clustering was useful in modeling groundwater level profiles. Also, the comparison between the logistic and plain mixture of 4 clusters models indicates that the mixture model was improved by

adding the logistic mixture structure in $p_k(W_i)$.

The likelihood of mixture models are invariant when labels of mixture components are switched (Diebolt and Robert, 1994; Richardson and Green, 1997). Thus, the posterior distributions have multimodal distributions. If components (clusters) are not well-separated, Markov Chains in posterior simulations can migrate easily from neighborhoods of one mode to another, switching labels. In this case, the means of the posterior distributions are practically meaningless in interpreting analysis results. As a solution to this problem, Stephens (2000) proposed to undo label switchings in Markov chains using his relabeling algorithm. Thus, each resulting Markov chain contains random numbers for only one of the components. This makes the resulting posterior distributions unimodal and posterior means easily interpretable. However, if the components are well-separated, Markov Chains may not migrate during a long period of simulations because the boundaries between components have very low probability densities. In our analysis, all simulated posterior distributions are unimodal (see Figure 4) and label switching was not observed during a reasonably large number (100,000) of iterations (see simulation history plots in Figure 5 as examples). We believe that, because the four clusters are distinctive, as shown in Figure 6, the Markov chains remained around a mode without migrating to neighborhoods of other modes. Therefore, we did not need to apply the relabeling algorithm to our data.

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

Sampling locations together with cluster memberships are presented in Figure 7. Also, the four clusters, sorted by the amplitudes of the mean temporal profiles, are shown in Figure 6. One can observe that Cluster 1 has the least variation over time and that Cluster 4 has the largest. Clusters 2 and 3 have similar amplitude. As described in Section 5.1, the logistic regression component in (6) is constructed with four environmental predictors: hydraulic head, soil depth, percentage of permeable ground surface, and distance to river. In Figure 8, the four clusters are compared in terms of these predictors. Posterior distributions of corresponding parameters are illustrated in Figure 4 and summarized in Table 1. Note that, because Cluster 1 is used as the reference in our model (5), the other clusters are described *relative* to it. The slope parameters for hydraulic head have posterior means equal to -0.05, -0.49, and -0.01 in Clusters 2, 3, and 4, respectively. Corresponding odds ratios are 0.95, 0.61, and 0.99, respectively. For example, odds of Cluster 3 is 0.61 times of odds of Cluster 1. Among three slope parameter estimates, only that for Cluster 3 has a 95% credible interval (CI) that does not include zero. Thus, the values for hydraulic head are similar among the wells in Clusters 1, 2, and 4. We can conclude that hydraulic head is relatively low in the wells of Cluster 3 and relatively high in the wells of Clusters 1, 2, and 4. Figure 8 also shows that wells in Cluster 3 have the lowest mean of hydraulic head and those in Cluster 1, 2, and 4 have similar means of hydraulic head. Similarly, we can conclude that Clusters 3 and 4 have wells with shallower soil zones and that Clusters 1 and 2 have wells with deeper soil zones. We found from Table 1 that the percentage of permeable ground surface within 250 $m$ radius does not distinguish the clusters in our data. Considering that Korean government regulation requires 500 $m$ radius around a residential well to be carefully protected, we also examined the same models using 500 $m$ and 1000 $m$ radius. Permeability was not significant in all cases. It is likely that the

surface permeability is not a significant factor because elevation of groundwater level in Seoul area is mainly affected by subsurface flow with more than a week of time lag, rather than by immediate infiltration with one or two days of time lag. Finally, we found that wells in Cluster 2 are located relatively far from the Han River, wells in Cluster 3 are close, and wells in Cluster 1 and 4 have medium distances.

[Table 1 about here.]

[Table 2 about here.]

Fractured crystalline bedrocks in Seoul consist of granites, gneisses and schist and they generally have the low and narrow-ranging hydraulic conductivity values from $2.3 \times 10^{-5}$ to $7.4 \times 10^{-3}$ $cm/sec$ (SMG, 1996). In interpreting data analysis results, we will assume hydraulic conductivity is moderately homogeneous in Seoul area and discuss the relationships between groundwater level profiles and origins of groundwater flow (regional or local flow). The relationships between clusters (Figure 6) and environmental characteristics (Table 1) can be interpreted as follows; for a summary refer to Table 2.

1) The profiles in Cluster 1 have a distinctively small or negligible degree of fluctuation. If homogeneous hydraulic conductivities are assumed in the study area, only regional flow can cause this distinctively flat pattern regardless of environmental conditions. Thus, we can expect that these wells are located on a path of regional flow. If hydraulic conductivities are distinctively low or unsaturated zone is distinctively thick around the monitoring wells in Cluster 1, even local flows may cause the flat pattern of groundwater level profiles. In both cases, infiltrated water approaches to the wells very slowly.

2) In Cluster 2, groundwater levels increase during the rainfall season and gradually decrease afterward, which is typically expected under natural conditions. In our data, the heavy rainfall season was between the 16th and the 24th weeks and the peak time was

around the 20th week. Peak groundwater level appears around the 23rd week. Therefore, we may characterize the fluctuation pattern of Cluster 2 as having a medium amplitude of fluctuation and a relatively short lag time between rainfall and a rise in groundwater level. Because the time lag is short, local flow seems dominant in Cluster 2. Table 1 shows that, for Cluster 2, hydraulic head is high, soil zone is deep, and the wells are located far from the Han River. Because these wells are not located closely to the river, we may expect a distinctive pattern of local flow. When homogeneous hydraulic conductivity is assumed, high hydraulic head can be another reason why the local flow can cause an effective response of the groundwater level rise. However, a deep soil zone seems to prevent the amplitudes of profiles in Cluster 2 from being as large as those in Cluster 4.

3) Groundwater levels in Cluster 3 increase slowly until the 26th week and decrease slowly afterward; thus the time lag of the rainfall effect in Cluster 3 is much longer than the lag in Clusters 2 and 4. In an area with homogeneous hydraulic conductivities, such a long time lag cannot be explained with local flow, which starts with infiltrated water from the ground surface adjacent to the monitoring well. We found that wells in Cluster 3 have a low hydraulic head and shallow soil zone, and are very close to the Han River; see Figure 7. Assuming homogeneous hydraulic conductivities, the slow increase and decrease of groundwater level can be explained as follows. First, because hydraulic head is low around wells in Cluster 3, groundwater flow will be slow. Thus, as compared to Cluster 2, the effect of local flow will be slower and weaker in Cluster 3. Second, because wells in Cluster 3 are located close to a major river, multiple local and regional flows mix with each other around these wells (Fetter 2000). If groundwater infiltrates at locations relatively close to a monitoring well, it will reach the well within a short time. However, if groundwater infiltrates at locations relatively far from the well, it will take longer to reach the well. Therefore, multiple groundwater flows will arrive at the monitoring well

with different time lags and generate a pattern of slow increase and slow decrease.

4) Profiles in Clusters 2 and 4 have similar fluctuation patterns, except for the larger amplitude of profiles in Cluster 4. Table 1 shows that wells in Cluster 4 appear to be subject to similar environmental conditions as those in Cluster 2, except for a shallow soil zone. The larger amplitude of profiles (stronger effect of local flow) in Cluster 4 can be explained by shallow soil depth, because this causes faster and more abundant infiltration to the water table.

Unless wells are located in groundwater recharge zone (i.e., high mountain area), the groundwater level tends to be affected by regional subsurface flow, which is recharged in high elevation areas, as well as local infiltration around monitoring wells. The significant change of groundwater level may be observed as a daily, weekly, or monthly base after a major rainfall event, depending on the hydrogeologic characteristics of aquifer. Because we used weekly average values for precipitation amount, our approach cannot explain daily variations. However, if rainfall elevates groundwater level in a day or two, these elevations will increase the average groundwater level of the first week. If this elevation is large enough, our method will detect the increase of groundwater level in the first week. After examining the rainfall peak time and the peak of groundwater levels (Larocque *et al.*, 1998) in Figure 6, we can find that most groundwater levels are affected by rainfall after more than a week.

# 6 CONCLUDING REMARKS

A Bayesian logistic mixture model was proposed to cluster groundwater level profiles and simultaneously discover relationships between clusters and environmental conditions around wells. We demonstrated that our model can explain these relationships adequately

using a manageable number of parameters, which is a strength relative to the alternative models: additive model (2) and non-additive model (3).

Successful application of our model helps us to better understand the pattern of groundwater level change with respect to the characteristics of recharge and subsurface flow. A short time lag between rainfall and a rise of groundwater level indicates that the groundwater recharges very rapidly around the monitoring well and thus can be vulnerable to the inflow of surface pollutants. Among the four clusters that we found, two of them, Clusters 2 and 4, have short time lags, apparently, being affected by local flow. Because profiles in Cluster 4 have large amplitudes, the groundwater is expected to be even more responsive to rainfall episodes around the monitoring wells in Cluster 4. Wells in Cluster 1 have a relatively stable profile pattern with very small or negligible fluctuation. Therefore, we expect effects of rainfall and pollution from ground surface to be more sluggish around these wells. Our results will be helpful in sustainably managing the groundwater resources on a local or regional scale. Particularly, in South Korea, our approach can be used to scientifically designate Groundwater Protection Areas (GPA) in an administrative district reinforced by current Groundwater Laws.

# ACKNOWLEDGEMENTS

Authors greatly appreciate associate editor's and reviewers useful comments that increased quality and readability of this paper. This work was partly supported by the Environmental Geosphere Research Lab (EGRL) of Korea University, which is funded by Korea Research Foundation. Dr. Gi-Tak Chae at the Korea Institute of Geoscience and Mineral Resources provided many suggestions and information that was extremely helpful in developing a framework and formulating conclusions for this study.

# References


Barrett, M.H., Hiscock, K.M., Pedley, S., Lerner, D.N., Tellam, J.H. and French, M.J. (1999) Marker species for identifying urban groundwater recharge sources: a review and case study in Nottingham, U.K. *Water Research*, **33**, 3083-3097.

Basford KE, McLachlan GJ. (1985) Likelihood estimation with normal mixture models. *Applied Statistics*, **34**, 282-289.

Bockgard, K. (2004) *Groundwater Recharge in Crystalline Bedrock: Processes, Estimation, and Modelling*, PhD Thesis, Uppsala University, Sweden.

Brumback BA, Brumback LC, and Lindstrom MJ. (2007) Penalized Spline Models for Longitudinal Data Book Chapter in *Advances in Longitudinal Data Analysis*, edited by Fitzmaurice G, Davidian M, Verbeke G, and Molenberghs G. London, Chapman and Hall. In press.

Celeux F, Forbes F, Robert CP, Titterington DM. (2006) Deviance information criterion for missing data models. *Bayesian Analysis* **1**, 651674.

Chae, G.T., Yun, S.T., Choi, B.Y., Yu, S.Y., Jo, H.Y., Mayer, B., Kim, Y.J., Lee, J.Y. (2008) Hydrochemistry of urban groundwater, Seoul, Korea: The impact of subway tunnels on groundwater quality. *Journal of Contaminant Hydrology*, doi:10.1016/j.jconhyd.2008.07.008.

Diebolt, J. and Robert, C.P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, series B*, 56, 363-375.

Fetter CW. (2000) *Applied Hydrogeology.* Prentice-Hall, Upper Sadle River, NJ.

Huntington, T. (2006) Evidence for intensification of the global water cycle: Review and synthesis. *Journal of Hydrology*, **319**, 83-95.

Jacobs, R., Jordan, M., Nowlan, S. and Hinton, G. (1991) Adaptative mixtures of local experts. *Neural Compuatation*, **3**, 79-87.

James, G.M and Sugar, C.A. Clustering for sparsely sampled functional data. (2003) *Journal of American Statistical Association*, **98**, 397-408.

Jeffries, N. and Pfeiffer, R. (2000) A mixture model for the probability distribution of rain rate. *Environmetrics*, **12**, 1-10.

Jiang, W. and Tanner, M. Hierarchical mixtures-of-experts for expnential family regression models: approximation and maximum likelihood estimation. (1999) *The Annals of Statistics*, **27**, 987-1011.

Joo Y, Lee K, Yun S, Min J. (2007) Logistic mixture of multivariate regressions for analysis of water quality impacted by agrochemicals. *Environmetrics*, **18**, 499-514.

Larocque, M., Mangin, A., Razack, M. and Banton, O. (1998) Contribution of correlation and spectral analyses to the regional study of a large karst aquifer (Charente, France). *Journal of Hydrology*, **205**, 217-231.

Lee, J.Y. and Lee, K.K. (2000) Use of hydrologic time series data for identification of recharge mechanism in a fractured bedrock aquifer system. *Journal of Hydrology*, **229**, 190-201.

Lerner, D.N. (2002) Identifying and quantifying urban recharge: a review. *Hydrogeology Journal*, **10**, 143-152.

Luan Y. and Li H. (2003) Clustering of time-course gene expression data using a mixed-effects model with B-splines. *Bioinformatics*, **19**, 474-482.

Ma, P., Castillo-Davis, C., Zhong, W. and Liu, J. (2006) A data-driven clustering method for time course gene expression. *Nucleic Acids Res*, **34**, 1261-69.

Ma, P., and Zhong, W. (2008) Penalized clustering of large scale functional data with multiple covariates. http://www.stat.uiuc.edu/∼pingma/research/fmclust-2.pdf

McLachlan, G.J., Bean, R.W. and Peel D. (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413-422

Moon, S.K., Woo, N.C. and Lee K.S. (2004) Statistical analysis of hydrographs and water-table fluctuation to estimate groundwater recharge. *Journal of Hydrology*, **292**, 198-209.

Park, S.S., Kim, S.O., Yun, S.T., Chae, G.T., Yu, S.Y., Kim, S. and Kim Y. (2005) Effects of land use on the spatial distribution of trace metals and volatile organic compounds in urban groundwater, Seoul, Korea. *Environmental Geology*, **48**, 1116-1131.

Peng, F., Jacobs, R. and Tanner, M. (1996) Bayesian inference in mixutres-of-experts and hierarchical mixutre of experts models with an application to speech recognition. *Journal of American Statistical Association*, **91**, 953-960.

Pfeiffer R.M., Carroll R.J., Wheeler W., Whitby D., and Mbulaiteye S. (2007) Combining assays for estimating prevalence of human herpesvirus 8 infection using multivariate mixture models. *Biostatistics*, in-press.

Richardson S., and Green P.J. (1997) On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, series B*, **59**, 731-792.

Rodhe, A. and Bockgard, N. (2006) Groundwater recharge in a hard rock aquifer-a conceptual model including surface loading effects. *Journal of Hydrology*, **330**, 389-401.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003) *Semiparametric Regression*, Cambridge University Press, Cambrigde, UK.

SMG (Seoul Metropolitan Government). (1996) *Master Plan of Groundwater Management of Seoul* (written in Korean). page 55.

Spiegelhalter DJ, Thomas A, Best NG, Lunn D. (2004) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, **64**: 583-640.

Stephens, M. (2000). Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society, Series B*, bf 62, 795-809.

Turner, T.R. (2000) Estimating the Propagation Rate of a Viral Infection of Potato Plants via Mixtures of Regressions,*Applied Statistics*, **49**, 371-384.

Vazquez-Sune, E., Sanchez-Vila, X. and Carrera, J. (2005) Introductory review of specific factors influencing uran groundwater, an emerging branch of hydrogeology, with reference to Barcelona, Spain. *Hydrogeology Journal*, **13**, 522-533.

Wong, C.S, and Li, W.k. (2001) On a logistic mixture autoregressive model. *Biometrika*, **88**, 833-846.

Yang, Y., Lerner, D.N., Barrett, M.H., and Tellam, J.H. (1999) Quantification of groundwater recharge in the city of Nottingham, UK. *Environmental Geology*, **38**, 183-198.

Zilberbrand, M., Rosenthal, E. and Shachnai, E. (2001) Impact of urbanization on hydrochemical evolution of groundwater and on unsaturated-zone gas composition in the coastal city of Tel Aviv, Israel. *Journal of Contaminant Hydrology*, **50**, 175-208.

# APPENDIX

## APPENDIX A: FULL CONDITIONAL POSTERIOR DISTRIBUTIONS

From the sampling distribution and prior distributions, we have the joint distribution given by

$$
\begin{aligned}
& f(Y, U, D, \beta, \alpha, \sigma^2, \lambda^2) \\
& \propto \prod_{i=1}^{N} \prod_{k=1}^{K} \left[ p_k(W_i, \alpha) \prod_{t=1}^{T} \phi(Y_{it} | X_t \beta_k + Z_t U_{kt}, \sigma_k^2) \right]^{D_{ik}} \\
& \times \prod_{k=1}^{K} \left\{ \prod_{t=2}^{T-1} \phi(U_{kt} | 0, \lambda_k^2 I) \cdot \phi(\beta_k | 0, \tau^2 I) \phi(\alpha_k | 0, \tau^2 I) \omega(\sigma_k^2 | a, b) \omega(\lambda_k^2 | a, b) \right\},
\end{aligned} \tag{6}
$$

where $\phi(\cdot)$ is the normal probability density function and $\omega(\cdot)$ is the inverse gamma probability density function. $D_{ik}$'s are unknown group membership indicator variables ($D_{ik} = 1$ if $Y_i$ belongs to group $k$; $D_{ik} = 0$ otherwise.). For the estimation of our model, Gibbs sampling is implemented. Let $\theta = (\beta, \alpha, \sigma^2, \lambda^2)$ and $\theta_{-a}$ be $\theta$ excluding $a$. Full

conditionals are given by

$$\begin{aligned} f(\beta_l|\theta_{-\beta_l}, U, D, Y) &\propto \prod_{j=1}^{N_l}\prod_{t=1}^{T} \phi(Y_{jt}|X_t\beta_l + Z_tU_{lt}, \sigma_l^2) \cdot \phi(\beta_l|0, \tau^2 I) \\ &\propto \phi(\beta_l|\mu(\beta_l), \Sigma(\beta_l)), \end{aligned}$$

where $N_l$ and $Y_{jt}$ are respectively the number of measurement and the response value such that $D_{il} = 1$ for $i = 1, \cdots, N$ and

$$\begin{aligned} \mu(\beta_l) &= \left( \frac{N_l \sum_{t=1}^{T} X_t^T X_t}{\sigma_l^2} + \frac{I}{\tau^2} \right)^{-1} \frac{1}{\sigma_l^2} \sum_{j=1}^{N_l}\sum_{t=1}^{T} \left(y_{jt} - Z_tU_{lt}\right) X_{it}^T, \\ \Sigma(\beta_l) &= \left( \frac{N_l \sum_{t=1}^{T} X_t^T X_t}{\sigma_l^2} + \frac{I}{\tau^2} \right). \end{aligned}$$

$$f(\alpha_l|\theta_{-\alpha_l}, U, D, Y) \propto \prod_{i=1}^{N}\prod_{k=1}^{K} \{p_k(W_i, \alpha)\}^{D_{ik}} \cdot \phi(\alpha_l|0, \tau^2 I),$$

where we may use a Metropolis-Hastings algorithm to generate $\alpha$.

$$\begin{aligned} f(\sigma_l^2|\theta_{-\sigma_l^2}, U, D, Y) &\propto \prod_{j=1}^{N_l}\prod_{t=1}^{T} \phi(Y_{jt}|X_t\beta_l + Z_tU_{lt}, \sigma_l^2) \cdot \omega(\sigma_l^2|a, b) \\ &\propto \omega(\sigma_l^2|a(\sigma_l^2), b(\sigma_l^2)), \end{aligned}$$

where $a(\sigma_l^2) = \frac{N_l T}{2} + a$ and $b(\sigma_l^2) = \frac{1}{2}\sum_{j=1}^{N_l}\sum_{t=1}^{T} \left(y_{jt} - X_t\beta_l - Z_tU_{lt}\right)^2 + b$.

$$\begin{aligned} f(\lambda_l^2|\theta_{\lambda_l^2}, U, D, Y) &\propto \prod_{t=2}^{T-1} \phi(U_{kt}|0, \lambda_l^2 I) \cdot \omega(\lambda_l^2|a, b) \\ &\propto \omega(\lambda_l^2|a(\lambda_l^2), b(\lambda_l^2)), \end{aligned}$$

where $a(\lambda_l^2) = \frac{T}{2} + a - 1$ and $b(\lambda_l^2) = \sum_{t=2}^{T-1} \frac{U_{lt}^2}{2} + b$.

$$\begin{aligned} f(D_{il} = 1|\theta, U, D_{-D_{il}}, Y) &\propto p_l(W_i, \alpha) \prod_{t=1}^{T} \phi(Y_{it}|X_t\beta_l + Z_t U_{lt}, \sigma_l^2) \\ &\propto \frac{p_l(W_i, \alpha) \prod_{t=1}^{T} \phi(Y_{it}|X_t\beta_l + Z_t U_{lt}, \sigma_l^2)}{\sum_{k=1}^{K} p_k(W_i, \alpha) \prod_{t=1}^{T} \phi(Y_{it}|X_t\beta_k + Z_t U_{kt}, \sigma_k^2)}. \end{aligned}$$

Thus, $D_i = (D_{i1}, \cdots, D_{iK})$ has a multinominal distribution $(P_1, \cdots, P_K)$ where $P_l = \frac{p_l(W_i,\alpha) \prod_{t=1}^{T} \phi(Y_{it}|X_t\beta_l + Z_t U_{lt}, \sigma_l^2)}{\sum_{k=1}^{K} p_k(W_i,\alpha) \prod_{t=1}^{T} \phi(Y_{it}|X_t\beta_k + Z_t U_{kt}, \sigma_k^2)}$ for $l = 1, \cdots, K$.

$$\begin{aligned} f(U_{lt}|\theta, U_{-U_{lt}}, D, Y) &\propto \prod_{j=1}^{N_l} \phi(Y_{jt}|X_t\beta_l + Z_t U_{lt}, \sigma_l^2) \cdot \phi(U_{lt}|0, \lambda_l^2 I) \\ &\propto \phi(U_{lt}|\mu(U_{lt}), \Sigma(U_{lt})), \end{aligned}$$

where

$$\begin{aligned} \mu(U_{lt}) &= \left( \frac{N_l Z_t^T Z_t}{\sigma_l^2} + \frac{I}{\lambda_l^2} \right)^{-1} \frac{1}{\sigma_l^2} \sum_{j=1}^{N_l} \left( y_{jt - X_t\beta_l} \right) Z_t^T, \\ \Sigma(U_{lt}) &= \left( \frac{N_l Z_t^T Z_t}{\sigma_l^2} + \frac{I}{\lambda_l^2} \right). \end{aligned}$$

In real data analysis, we used WinBugs (*http* : *//www.mrc* − *bsu.cam.ac.uk /bugs/winbugs/contents.shtml*) for posterior simulations.

# APPENDIX B: ENVIRONMENTAL VARIABLES

| Group ID Well | Hydraulic Head | Soil Depth | Percentage of Permeable Ground Surface | Distance to River |
|---|---|---|---|---|
| 1 | 25 | 23 | 45.16 | 4839 |
| 1 | 9 | 4 | 27.27 | 227 |
| 1 | 11 | 12 | 26.68 | 161 |
| 1 | 5 | 22 | 10.25 | 1953 |
| 1 | 32 | 19 | 22.38 | 182 |
| 1 | 12 | 11 | 0.00 | 447 |
| 1 | 4 | 15 | 43.03 | 124 |
| 1 | 28 | 15 | 6.76 | 1709 |
| 2 | 12 | 12 | 0.00 | 1060 |
| 2 | 11 | 20 | 12.17 | 1817 |
| 2 | 3 | 16 | 22.99 | 396 |
| 2 | 11 | 12 | 0.00 | 3392 |
| 2 | 5 | 21 | 37.74 | 212 |
| 2 | 5 | 19 | 0.00 | 693 |
| 2 | 13 | 13 | 4.19 | 4307 |
| 2 | 4 | 20 | 12.01 | 1858 |
| 2 | 10 | 12 | 0.17 | 2013 |
| 2 | 11 | 14 | 0.00 | 6227 |
| 2 | 54 | 8 | 43.17 | 5239 |
| 2 | 18 | 13 | 0.00 | 1311 |
| 2 | 27 | 11 | 42.19 | 6501 |
| 2 | 8 | 11 | 75.52 | 3365 |
| 3 | 8 | 19 | 28.42 | 386 |
| 3 | 5 | 2 | 13.03 | 1347 |
| 3 | 6 | 10 | 20.25 | 467 |
| 3 | 9 | 11 | 1.93 | 422 |
| 3 | 5 | 2 | 0.94 | 509 |
| 4 | 70 | 7 | 66.86 | 986 |
| 4 | 13 | 6 | 18.34 | 3345 |
| 4 | 11 | 13 | 100.00 | 2259 |
| 4 | 11 | 8 | 0.00 | 1086 |
| 4 | 61 | 12 | 63.41 | 3851 |
| 4 | 13 | 12 | 0.00 | 357 |
| 4 | 8 | 11 | 8.13 | 186 |
| 4 | 4 | 8 | 0.79 | 2629 |
| 4 | 3 | 25 | 0.00 | 2818 |
| 4 | 2 | 10 | 11.45 | 1774 |

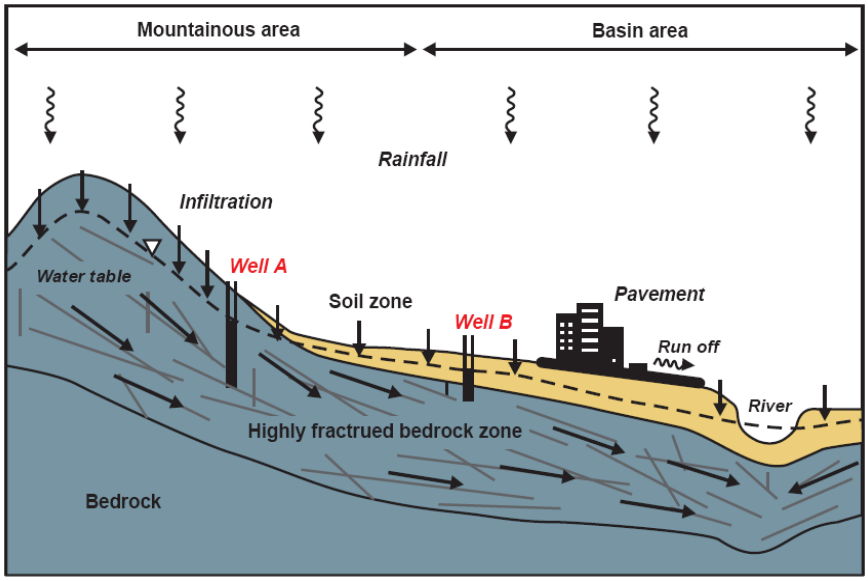


Figure 1: Diagram of infiltration and lateral groundwater flow in Seoul, Korea.

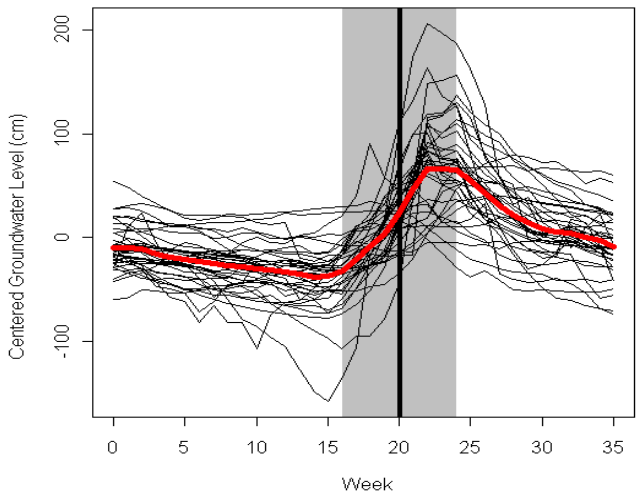


Figure 2: Temporal profiles of centered groundwater levels measured in 37 groundwater wells (unit: *cm*). As a reference line, the BLUP (Best Linear Unbiased Prediction) is drawn with a thick gray line. Gray-shaded vertical band and thick black vertical line indicate the rainy season of year 2001 and the peak time of this season, respectively.

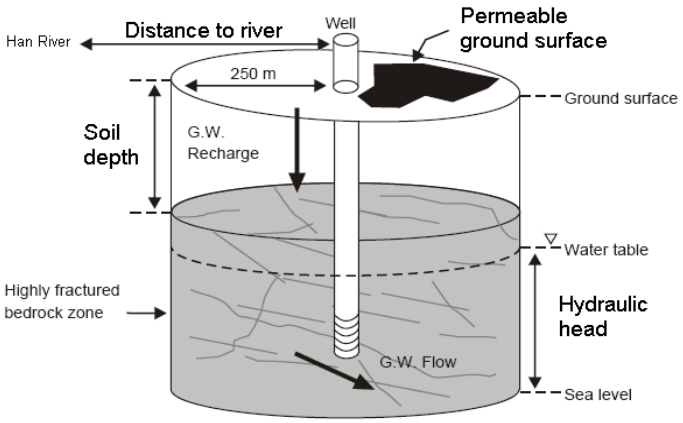


Figure 3: Environmental variables.

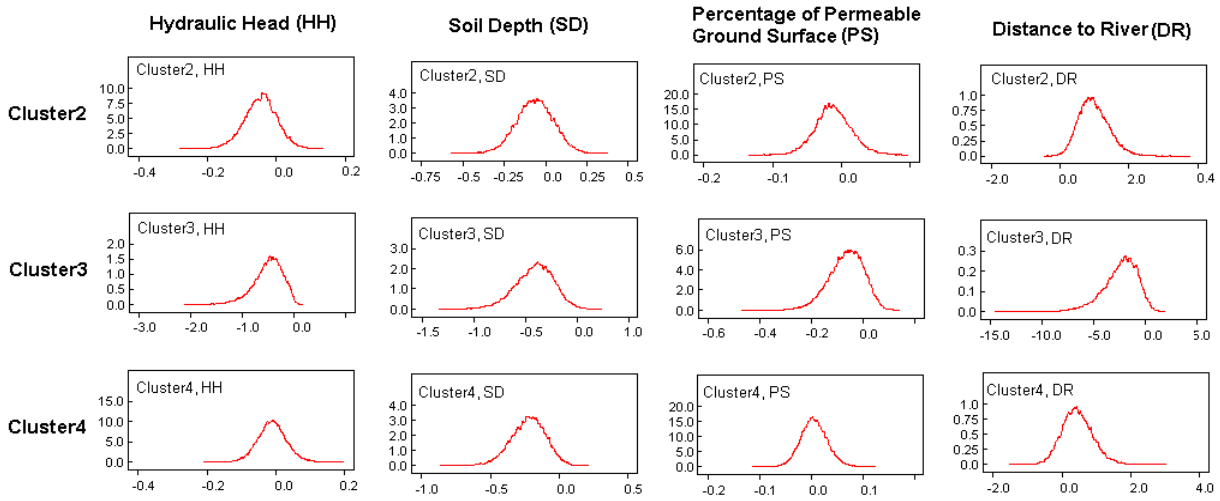


Figure 4: Posterior density distributions of slope parameters in the logistic regression component.

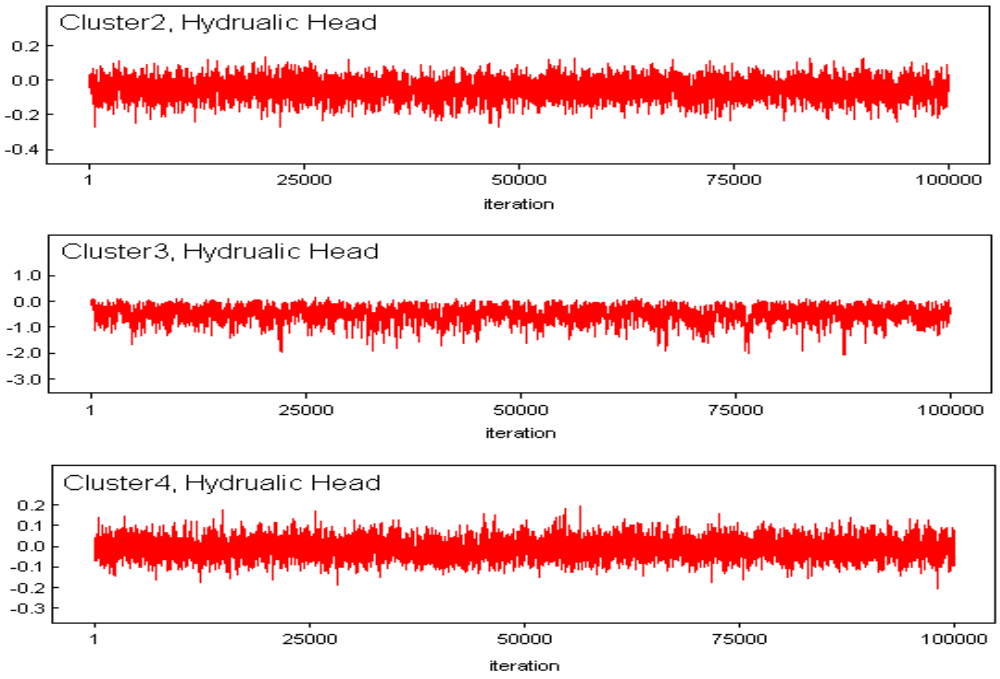


Figure 5: Simulations from the posterior distributions of the slope parameters for hydraulic head variable in the logistic regression component.

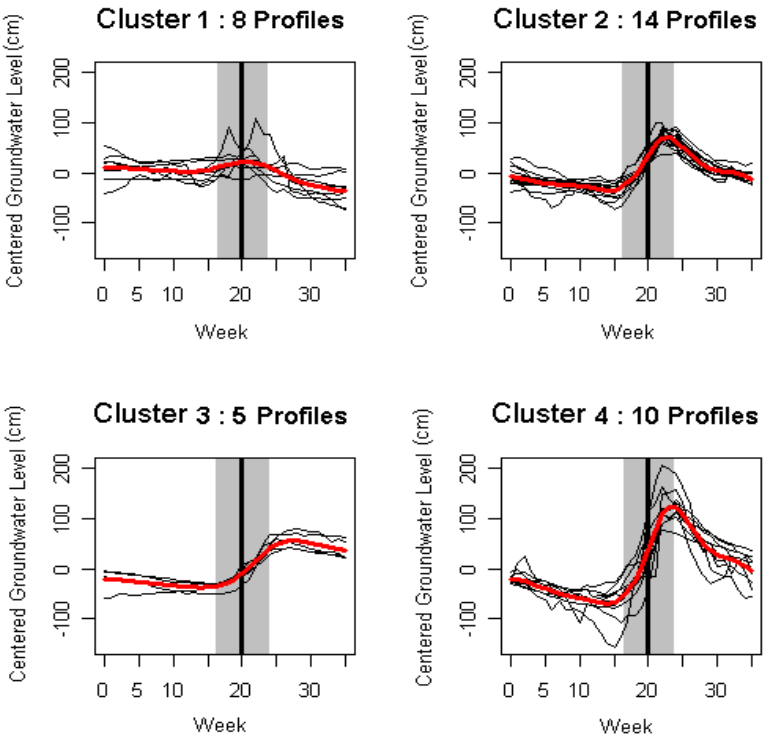


Figure 6: Clustered temporal profiles of centered groundwater levels (unit:*cm*). As reference lines, the BLUP (Best Linear Unbiased Prediction) profile for each cluster is drawn with a thick line. The gray-shaded vertical band and the thick black vertical line indicate the rainy season of year 2001 and the peak time of this season, respectively.

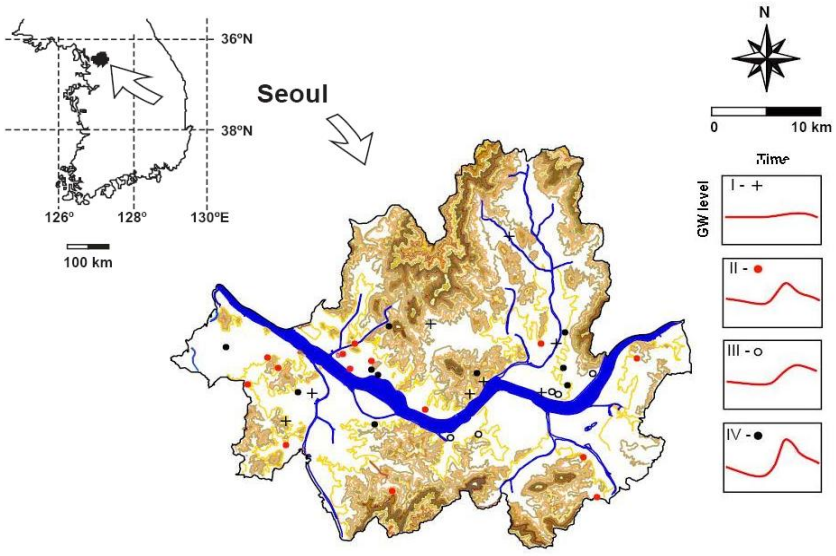


Figure 7: Well locations with cluster memberships.

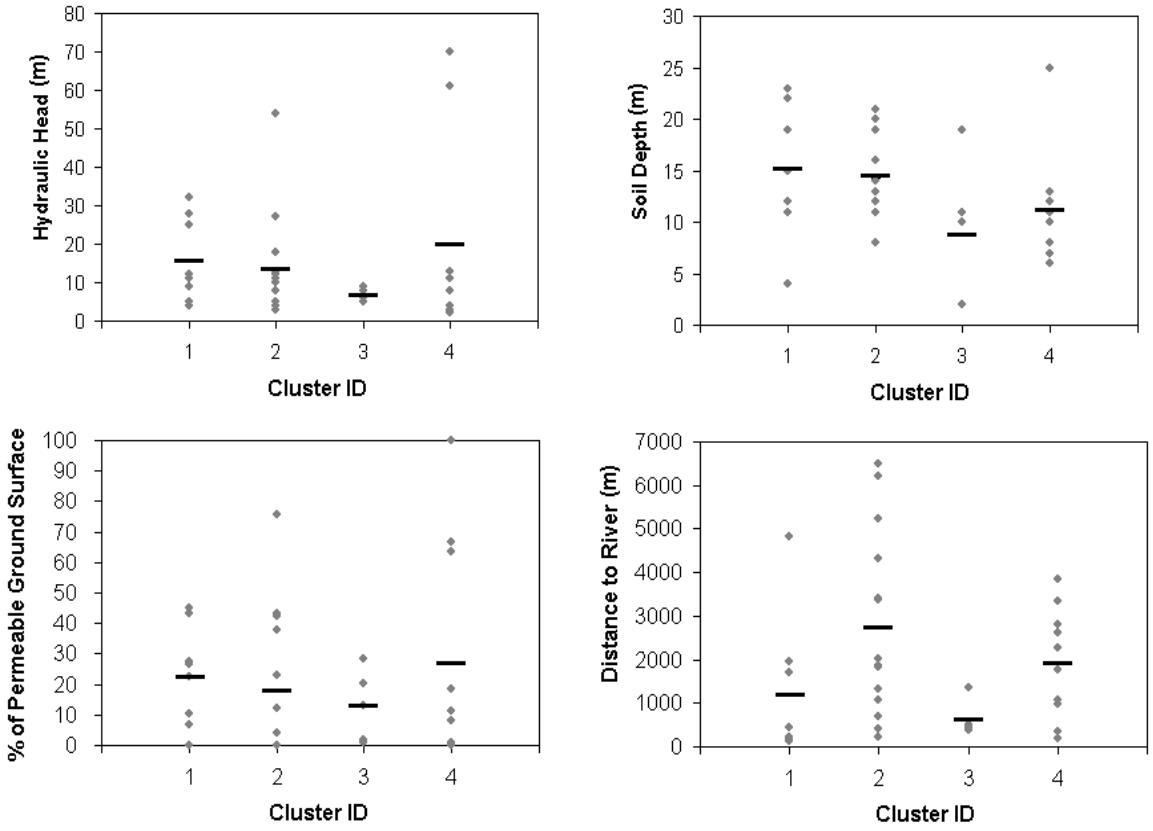


Figure 8: Scatterplots of environmental variables and group ID. For each well, the values of environmental variables are marked with gray circles. Means of each group are marked with dark lines.

Table 1: Environmental characteristics of four clusters and Bayesian estimates of slope parameters in the logistic regression component. Posterior means and corresponding posterior odds are reported in parentheses.

| | Hydraulic head | Soil depth | Percentage of permeable ground surface | Distance to river |
|---|---|---|---|---|
| Cluster 1 | high | deep | no difference | medium |
| Cluster 2 | high(-0.05, 0.95) | deep(-0.06, 0.94) | no difference (0.00, 1.00) | far(1.00**, 2.72) |
| Cluster 3 | low(-0.49**, 0.61) | shallow(-0.41**, 0.66) | no difference(-0.07, 0.93) | close(-2.35*, 0.10) |
| Cluster 4 | high(-0.01, 0.99) | shallow(-0.22*, 0.80) | no difference (0.00, 1.00) | medium (0.45, 1.57) |

* denotes parameters for which the 90 percent credible interval does not include zero.

** denotes parameters for which the 95 percent credible interval does not include zero.

Table 2: Profile patterns and groundwater flow.

| | Profile pattern | | Cause for profile patterns | |
|---|---|---|---|---|
| | Amplitude | Time lag | Local flow | Regional flow |
| Cluster 1 | negligible | N/A | weak | strong |
| Cluster 2 | medium | short | strong | weak |
| Cluster 3 | medium | long | medium | medium |
| Cluster 4 | large | short | very strong | weak |