ENVIRONMETRICS
Environmetrics 2005; **16**: 305–318
Published online 2 February 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/env.707

A mixture model approach to analyzing major element chemistry data of the Changjiang (Yangtze River)

Lin Xue[1], James C. Fu[2], Feiyue Wang[3] and Liqun Wang[2,*,†]

[1]*Cancer Care Manitoba, 675 McDermot Avenue, Winnipeg, Manitoba, R3E OV9, Canada*
[2]*Department of Statistics, University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada*
[3]*Environmental Science Program and Department of Chemistry, University of Manitoba, Winnipeg, Manitoba, R3T 2N2, Canada*

SUMMARY

In this article we study the statistical distributions of major chemical compositions (HCO_3, Ca; charges are neglected for simplicity) and the total dissolved solid (TDS) concentration in the river water of the Changjiang (Yangtze River) of China. We propose a Bayesian finite mixture model with an unknown number of components for the multi-year averages of continuously monitored data over the period 1958–1990 at 191 stations in the drainage basin. A discretization-based Monte Carlo sampling approach is used to estimate the posterior distributions of the parameters in the model. Two sub-populations are identified for the levels of TDS, HCO_3 and Ca, and observations from the 191 stations are classified into two groups using the posterior classification probabilities. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: stochastic modeling; environmental geochemistry; physical processes; Bayesian mixture model; discretization-based Monte Carlo sampling; classification

1. INTRODUCTION

The Changjiang (Yangtze River) is the third largest among the rivers in the world in terms of length (6300 km) and the fourth largest in terms of water discharge (900 km^3/year). The river originates from the Qinghai–Tibet Plateau in western China and flows through the entire central region before it empties into the Pacific Ocean on the east coast. Joined by a large number of tributaries, the Changjiang drainage basin covers an area of $1.8 \times 10^6\,km^2$ and is home to about 400 million people. The geological, geographic and social–economic setting of the river basin is very complex (Chen *et al.*, 2002).

Despite its global significance, the major element chemistry of the Changjiang were not well studied until recently by Chen *et al.* (2002). Based on monthly monitored chemical data at 191 stations in the drainage basin for the period 1958–1990, Chen *et al.* (2002) thoroughly studied the chemical compositions of the river water throughout the basin, their long-term trend of change, and the underlying natural and anthropogenic processes contributed to such changes.

In this article, we use a Bayesian finite model to further study the statistical distributions of major chemical elements in the Changjiang basin, using the multi-year data sets reported in Chen *et al.*

*Correspondence to: L. Wang, Department of Statistics, University of Manitoba, Winnipeg, Manitoba R3T 2N2, Canada.
†E-mail: liqun_wang@umanitoba.ca

Received 3 May 2004
Accepted 20 September 2004

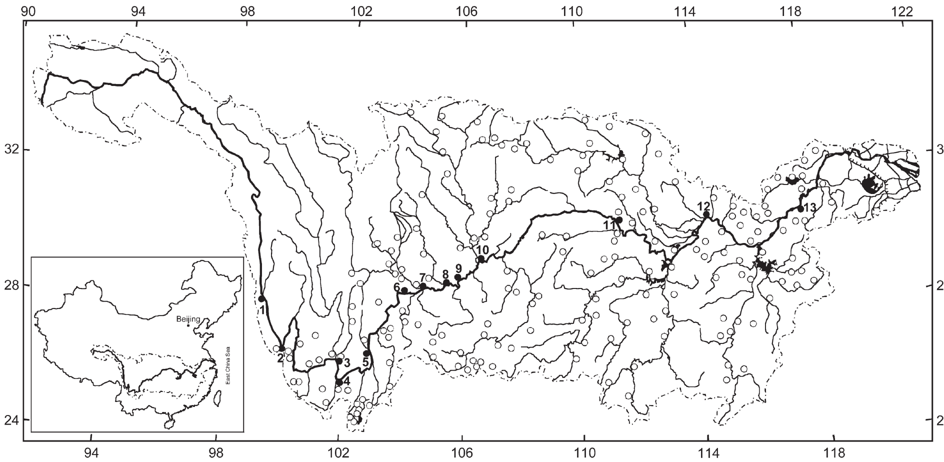

Figure 1. The Changjiang drainage basin and 191 sampling stations. The filled circles are stations along the main channel and the open circles are stations on tributaries

(2002). Figure 1 shows the Changjiang drainage basin and the 191 stations where major element concentrations were monitored almost monthly for the period 1958–1990. In particular, we focus on three major chemical composition variables: the total dissolved solid (TDS) concentration in milligrams per liter (mg/l); the bicarbonate (HCO_3; the charge is neglected for simplicity) concentration in mg/l, which constitutes on average 64% of TDS; and the calcium (Ca) concentration in mg/l, which constitutes on average 16% of the TDS. Histograms of these three variables, as illustrated in Figure 2 for TDS, show similar patterns of a mixture of several overlapping right-skewed distributions. To clarify the mixture pattern and to symmetrize the components, we apply the logarithmic transformation on the original data.

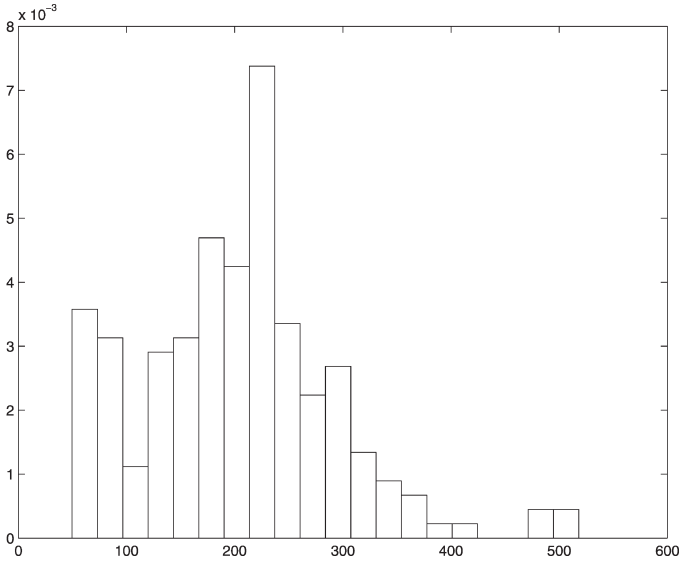

Figure 2. Histogram of TDS

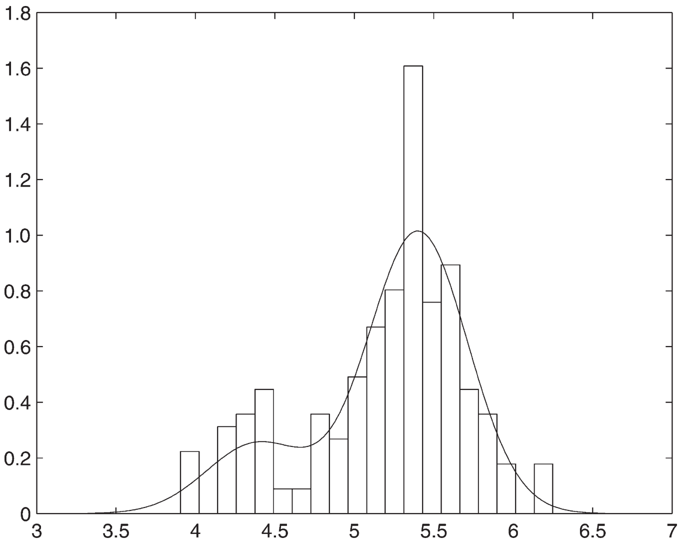

Figure 3. Histogram of log (TDS) and the predictive density with $k = 2$

Histograms of the log-transformed data are shown in Figures 3–5 (superimposed with predictive densities described in Section 3). They clearly show that the underlying distribution of each data set is compatible with a mixture of several symmetric distributions, indicating that a finite mixture model may be appropriate to describe the data. The general form of a finite mixture distribution is

$$f(x) = \sum_{j=1}^{k} w_j f_j(x)$$

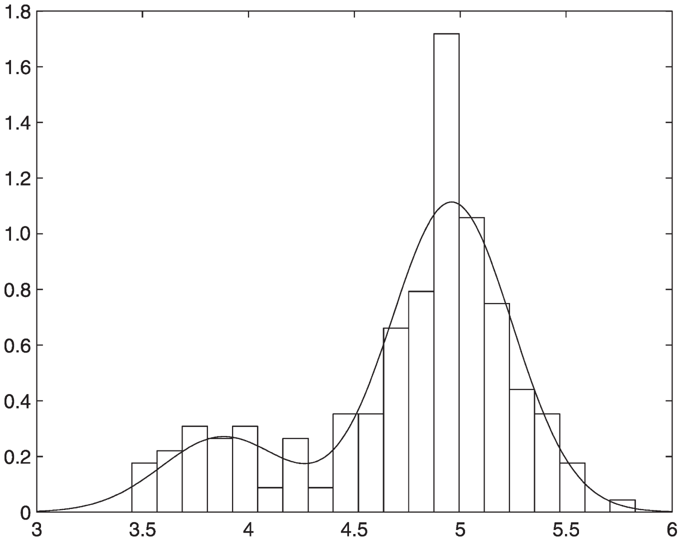

Figure 4. Histogram of log (HCO_3) and the predictive density with $k = 2$

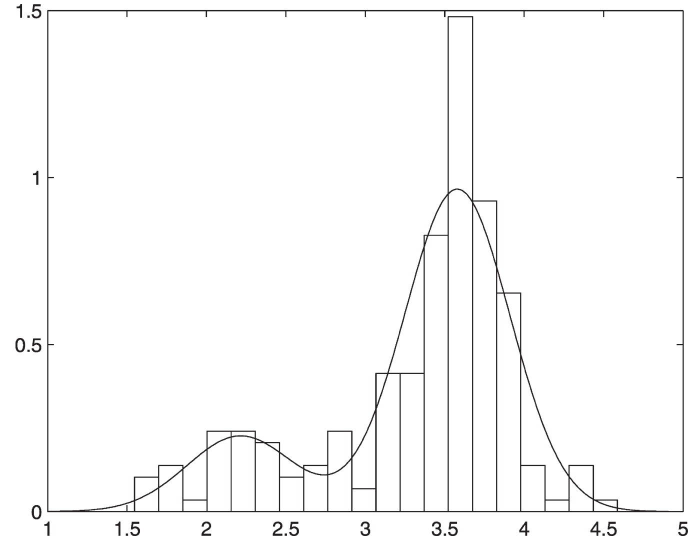

Figure 5. Histogram of log (Ca) and the predictive density with $k = 2$

where $f_j(x), j = 1, 2, \ldots, k$, are the component densities and $w_j, j = 1, 2, \ldots, k$, are the mixing weights satisfying $w_j \geq 0$ and $\sum w_j = 1$. Each weight w_j represents the probability that an observation of X comes from the sub-population $f_j(x)$. In practice, mixtures of parametric distributions such as normal densities are usually used. The classical maximum likelihood analysis of finite mixture models requires the specification of the number of components k in the mixture distribution, which is usually done heuristically. For the Changjiang data in Figures 3–5, it is difficult to tell from the histograms alone how many components there are in each mixture. It is also difficult to specify the number of components from other sources, because of the lack of theoretical support. Therefore it is realistic and desirable to have a more flexible model in which the number of components is allowed to vary. Consequently, we propose a Bayesian finite mixture model where the number of components is treated as an unknown parameter and is estimated from data along with other model parameters.

2. THE BAYESIAN FINITE MIXTURE MODEL

In this Section, we specify a common statistical model for the three variables TDS, HCO_3 and Ca, because their histograms in Figures 3–5 show similar patterns except for possibly different numbers of components. Therefore, we denote each of these variables generically as X. Since all components in the histograms look more or less symmetric, we use a mixture of normal densities. Normal mixture models are frequently used for both theoretical and practical reasons (e.g. Titterington *et al.*, 1985; Gelman *et al.*, 1995; Leonard and Hsu, 1999). The probability density of a k-component normal mixture for X is

$$f\left(x \mid \boldsymbol{\mu}_k, \sigma_k^2, \mathbf{w}_k, k\right) = \sum_{j=1}^{k} \frac{w_{kj}}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{\left(x - \mu_{kj}\right)^2}{2\sigma_k^2}\right], -\infty < x < \infty \tag{1}$$

where $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2}, \ldots, \mu_{kk})$ are the component means, σ_k^2 is the common variance for all components and $\mathbf{w}_k = (w_{k1}, w_{k2}, \ldots, w_{kk})$ are the mixing weights satisfying $w_{kj} \geq 0$ and $\sum w_{kj} = 1$. The component mean μ_{kj} is the center of the *j*th sub-population. In finite mixture modeling, it is common to assume that all normal components have the same variance for both theoretical and practical convenience. As mentioned earlier, in our Bayesian model the number of components is treated as a possible outcome of a random variable K. Thus, for different values of K, there are different sets of parameters. For example, conditional on $K = 1$, there are parameters μ_{11}, σ_1^2 and $w_{11} = 1$; conditional on $K = 2$, there are parameters μ_{21}, μ_{22}, σ_2^2 and w_{21}, w_{22} $(w_{21} + w_{22} = 1)$; and so on.

Suppose $\{x_i, i = 1, 2, \ldots, n\}$ is an independent and identically distributed random sample of X. Then the likelihood function is, for $K = k$,

$$\prod_{i=1}^{n} f\left(x_i \mid \boldsymbol{\mu}_k, \sigma_k^2, \mathbf{w}_k, k\right) = \prod_{i=1}^{n} \sum_{j=1}^{k} \frac{w_{kj}}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{(x_i - \mu_{kj})^2}{2\sigma_k^2}\right] \tag{2}$$

Now we specify the prior distributions for parameters $\boldsymbol{\mu}_k$, σ_k^2, $\mathbf{w}_k$ and K. Since all other parameters depend on the value of K, our model has a hierarchical structure. Firstly, we assume that, given $K = k$, $\boldsymbol{\mu}_k$, σ_k^2 and $\mathbf{w}_k$ are conditionally independent, so that the full posterior distribution for $K = k$ is proportional to

$$\prod_{i=1}^{N} f\left(x_i \mid \boldsymbol{\mu}_k, \sigma_k^2, \mathbf{w}_k, k\right) p(\boldsymbol{\mu}_k \mid k) p\left(\sigma_k^2 \mid k\right) p(\mathbf{w}_k \mid k) p(k) \tag{3}$$

In practice, the component means $\mu_{k1}, \mu_{k2}, \ldots, \mu_{kk}$ are usually treated as being randomly drawn from a common normal distribution $\mathrm{N}(\mu_0, \sigma_0^2)$ (Leonard and Hsu, 1999; Gelman *et al.*, 1995). Note that an issue of identifiability arises here, because the likelihood function in (2) remains the same for all permutations of the components. A usual remedy for this problem, as we assume in this article, is to impose the order restriction $\mu_{k1} < \mu_{k2} < \cdots < \mu_{kk}$ on the means. Thus, conditional on $K = k$, the prior distribution for the component means is

$$p(\boldsymbol{\mu}_k \mid k) = k! \prod_{j=1}^{k} \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{(\mu_{kj} - \mu_0)^2}{2\sigma_0^2}\right], \mu_{k1} < \mu_{k2} < \cdots < \mu_{kk} \tag{4}$$

A commonly used prior for σ_k^2 is the inverse-gamma distribution (Ibrahim *et al.*, 2002; Escobar and West, 1995), though sometimes an inverse-χ^2 distribution is also used (Belisle *et al.*, 2002; Gelman *et al.*, 1995). In this article, we use the inverse-gamma distribution for the component variance σ_k^2, whose density is

$$p\left(\sigma_k^2 \mid k\right) = \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\sigma_k^2\right)^{-\alpha-1} e^{-\beta/\sigma_k^2}, \sigma_k^2 > 0 \tag{5}$$

For the weights $\mathbf{w}_k$, a natural choice of prior distribution is the Dirichlet distribution (Diebolt and Robert, 1994; Stephens, 2000) with density

$$p(\mathbf{w}_k \mid k) = \frac{\Gamma(k\gamma)}{\Gamma(\gamma)} \prod_{j=1}^{k-1} w_{kj}^{\gamma-1} \left(1 - \sum_{j=1}^{k-1} w_{kj}\right)^{\gamma-1}, \ 0 \leq \sum_{j=1}^{k-1} w_{kj} \leq 1 \tag{6}$$

Finally, for the number of components K we assign the discrete uniform distribution over the set $\{1, 2, \ldots, k_{\max}\}$, so that

$$p(k) = 1/k_{\max}, \ k = 1, 2, \ldots, k_{\max} \tag{7}$$

This is a non-informative prior which provides equal support for k between 1 and $k_{\max}$.

Prior distributions (4)–(7) depend on the so-called hyper-parameters μ_0, σ_0^2, α, β, γ and $k_{\max}$. In Bayesian hierarchical modeling, prior distributions for the hyper-parameters are specified. To simplify analysis, in this article we use the following simple strategy to assign a degenerate (single point) prior for all hyper-parameters.

Since μ_0 and σ_0^2 are the mean and the variance of μ_{kj}, one can extract information about them from the data. Let M_x and R_x denote the midrange and the range of the data, respectively. Then, it is reasonable to set $\mu_0 = M_x$ and $\sigma_0^2 = R_x^2$. For the inverse-gamma prior, we choose $\alpha = 2$ and $\beta = 1$ to prevent the values of σ_k^2 from getting too close to zero (Escobar and West, 1995). In the Dirichlet distribution, we set γ=1. This corresponds to an uniform distribution over the range of values of the weights. Finally, we choose $k_{\max} = 3$, because one can see from the histograms in Figures 3–5 that the underlying mixture distribution is unlikely to have more than three components. This configuration results in a 13-dimensional posterior distribution.

The above strategy of assigning hyper-parameter values was used by Richardson and Green (1997) and Stephens (2000). Note that the choice of values for hyper-parameters will have an influence on the posterior distribution and subsequent inference from it. Such influence can be examined via sensitivity analysis. Richardson and Green (1997) have carried out a sensitivity analysis for their Bayesian model and they found that the posterior estimates are not very sensitive to the values of hyper-parameters chosen according to this strategy.

3. COMPUTATIONAL ALGORITHM

While the Bayesian mixture model provides a flexible and practical description of data coming from a population with a varying number of sub-populations, statistical inference based on this model remains a challenging task because of its mathematical complexity. As in many Bayesian analysis, the inference is usually based on a random sample generated from the posterior distribution. Whereas the expectation–maximization (EM) algorithms are commonly used for the classical maximum likelihood analysis, a general tool for Bayesian computation is the Markov chain Monte Carlo (MCMC) method (Gilks *et al.*, 1998). For a mixture model with an unknown number of components, a reversible jump MCMC algorithm was proposed by Green (1995) and an alternative algorithm was developed by Stephens (2000). In practice, however, the actual implementation of these procedures is quite involved (e.g. Besag and Green, 1993; Brooks *et al.*, 2003).

In this article, we use a direct and simple algorithm developed by Fu and Wang (2002) to estimate the model. This algorithm is non-iterative, easy to implement and overcomes some difficulties associated with MCMC procedures. It allows one to quickly generate a large sample from a given posterior density up to a multiplicative constant, which can be used to estimate the marginal posterior densities of all parameters. In the following, we briefly introduce this algorithm. More details may be found in Fu and Wang (2002).

Suppose we are given a d-dimensional posterior density $p(\theta)$ up to a multiplicative constant, where θ is the vector of all unknown parameters in the model. Basically, the algorithm consists of the following steps. The first step is called discretization. In this step, first an initial compact set $C(p) = [a, b]^d$ $(-\infty < a < b < \infty)$ containing the significant region of $p(\theta)$ is determined based on its properties. If $p(\theta)$ has a bounded support $S(p)$, then $C(p) = S(p)$. Then a discrete set $S_N(p) = \{\theta_j \in C(p), j = 1, 2, \ldots, N\}$ is generated using either a deterministic (such as low discrepancy sequence) or stochastic (such as independent and uniformly distributed random numbers) sequence. The second step is called contourization, in which a discrete probability distribution $\{P_M(i), i = 1, 2, \ldots, M\}$ over a finite partition of $S_N(p)$ is constructed. This distribution approximates the original density $p(\theta)$. The third step is re-sampling. In this step, a random sample is generated from $S_N(f)$ according to the discrete distribution $\{P_M(i)\}$. This sample of θ can be used to visualize the marginal distributions of $p(\theta)$. The histograms from this sample will show whether the initial compact set $C(p)$ is appropriate. If $C(p)$ is either too large or too small, then it is modified and the whole procedure is repeated. The sample obtained in the final stage of this procedure can be regarded as an i.i.d. random sample from $p(\theta)$.

As a by-product, this algorithm also finds the posterior modes of $p(\theta)$, which is practical in real applications. Moreover, applying this algorithm to the likelihood function $\ell(\theta)$ defined on the parameter space $S(\ell)$, the modes give the approximate maximum likelihood estimates.

Another object of interest in Bayesian inference is the predictive distribution, which specifies, given the present sample, what values a future observation of X may take and what probabilities are associated with them. Let $\mathbf{X}_n$ denote the present sample of size n. Then the predictive density of a future observation $X = x$ is defined as

$$p(x \mid \mathbf{X}_n) = \int f(x \mid \theta) p(\theta \mid \mathbf{X}_n)\, \mathrm{d}\theta$$

where $f(x \mid \theta)$ is the density of X and $p(\theta \mid \mathbf{X}_n)$ is the posterior density of θ. Once a random sample $\theta_1, \theta_2, \ldots, \theta_N$ is generated from the posterior distribution $p(\theta \mid \mathbf{X}_n)$, then the predictive density can be estimated as

$$\hat{p}(x \mid \mathbf{X}_n) = \frac{1}{N} \sum_{i=1}^{N} f(x \mid \theta_i)$$

This formula is used to calculate the predictive densities for the Changjiang data in the next section.

4. ANALYSIS OF THE CHANGJIANG DATA

In this section we present the analytical results of the Changjiang data using the Bayesian mixture model and the computational algorithm described in Sections 2 and 3. All numerical computations are

Table 1. Prior and posterior distributions of K, for log (TDS)

K	1	2	3
Prior	1/3	1/3	1/3
Posterior	0	0.6337	0.3663

carried out using the computer package MATLAB in a Unix environment. The MATLAB code is available from the corresponding author upon request. The following numerical results are based on a sample of size 3000 drawn from each joint posterior distribution.

4.1. Distribution of the TDS

Using the strategy of Section 2, for this data set the hyper-parameters in the prior distribution of the component means are assigned the values $\mu_0 = 5.0781$ and $\sigma_0^2 = 5.4951$. The marginal posterior distribution (mpd) of K is shown in Table 1. It clearly favors two components, though there is clearly some uncertainty about the number of components. The mpds for other parameters are shown in Figures 6–8, whereas the numerical summaries of posterior means and standard deviations are given in Tables 2 and 3.

Figure 6 shows the mpds of the sub-population means and weights for $k = 2$. The distributions of μ_{21} and μ_{22} are roughly symmetric. The center of μ_{21} is near 4.4, and its values range from about 4.2 to 4.6. The center of μ_{22} is near 5.4 with range from about 5.3 to 5.5. The mpds of the weights are slightly skewed. w_{21} is centered around 0.2, ranging from about 0.1 to 0.3, whereas w_{22} is centered around 0.8 with range from about 0.7 to 0.9. The mpd of the population variance σ_2^2 is shown in Figure 8. It has a shape of an inverse-gamma distribution and its values range from about 0.06 to 0.15 with mode around 0.1. The predictive density for a future observation of log (TDS) with $k = 2$ is shown in Figure 3. It has

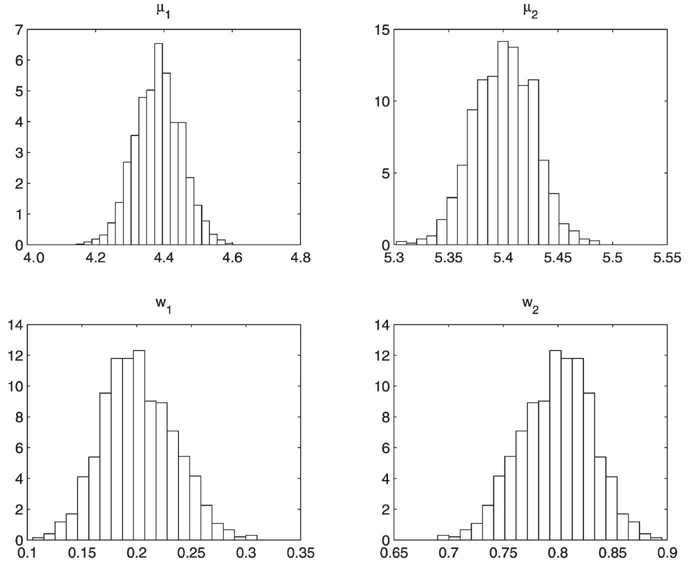

Figure 6. Marginal posterior distributions of $\boldsymbol{\mu}_k$ and $\mathbf{w}_k$ ($k = 2$), for log (TDS)

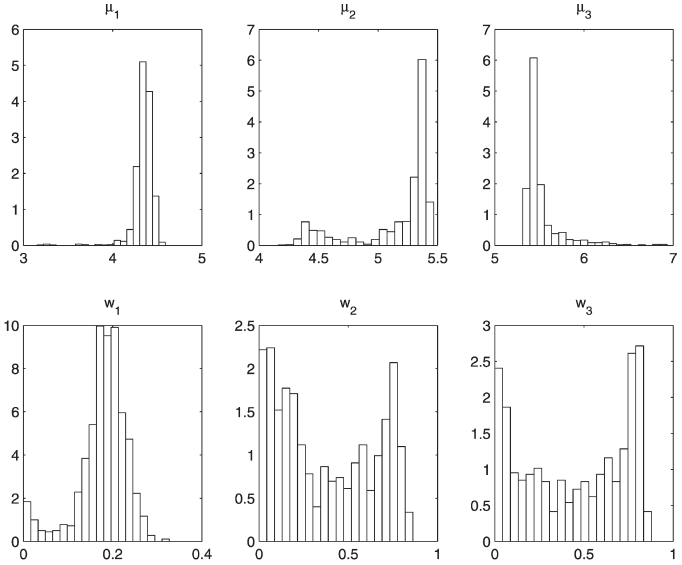

Figure 7. Marginal posterior distributions of $\boldsymbol{\mu}_k$ and $\mathbf{w}_k$ $(k = 3)$, for $\log(\mathrm{TDS})$

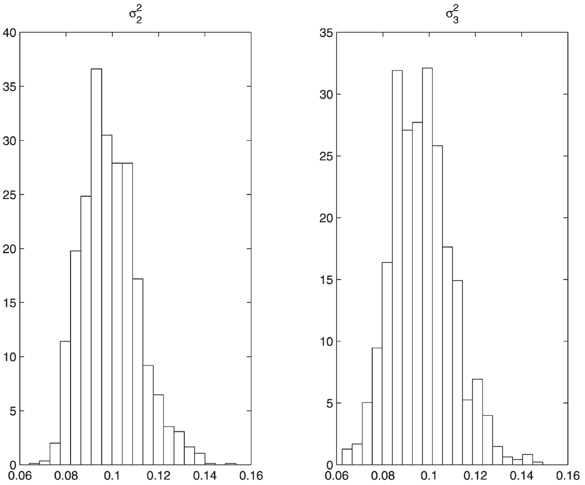

Figure 8. Marginal posterior distributions of σ_k^2 $(k = 2, 3)$, for $\log(\mathrm{TDS})$

Table 2. Posterior means and standard deviations of $\boldsymbol{\mu}_k$, $\mathbf{w}_k$ and σ_k^2 $(k = 2)$, for $\log(\mathrm{TDS})$

	$\boldsymbol{\mu}_k$		$\mathbf{w}_k$		σ_k^2
Mean	4.3823	5.4006	0.2015	0.7985	0.0990
SD	0.0674	0.0275	0.0330	0.0323	0.0120

Table 3. Posterior means and standard deviations of $\boldsymbol{\mu}_k$, $\mathbf{w}_k$ and σ_k^2 ($k = 3$), for log (TDS)

	$\boldsymbol{\mu}_k$			$\mathbf{w}_k$			σ_k^2
Mean	4.3457	5.1617	5.5178	0.1780	0.3754	0.4466	0.0969
SD	0.1215	0.3356	0.2154	0.0553	0.2715	0.2929	0.0130

Table 4. AMLE of σ_k^2, $\boldsymbol{\mu}_k$ and $\mathbf{w}_k$ ($k = 2, 3$), for log (TDS)

	$k = 2$		$k = 3$		
$\boldsymbol{\mu}_k$	4.3854	5.4022	4.3419	5.3463	5.7898
$\mathbf{w}_k$	0.2009	0.7991	0.1926	0.7192	0.0883
σ_k^2	0.0856		0.0736		

two modes at around 4.4 and 5.4 and fits the overall shape of the histogram quite well, though it seems to have a little lower peak. If prediction is a primary goal of a study, one might improve this by choosing other component densities rather than normal, such as double-exponential densities.

The mpds of parameters for $k = 3$ are similarly interpreted. It should be noted, however, that the mpd of μ_{32} in Figure 7 has two modes at about 4.6 and 5.4, which are the centers of μ_{31} and μ_{33}, respectively. This is a consequence of forcing a third component by the model which is poorly supported by the data. Therefore it is another indication besides the mpd of K that the data support two-component model.

For comparison, we also present in Table 4 the approximate maximum likelihood estimates (AMLE) of $\boldsymbol{\mu}_k$, $\mathbf{w}_k$ and σ_k^2 for $k = 2$ and 3, respectively. These estimates are computed automatically by the algorithm of Fu and Wang (2002). They are similar to the posterior mean estimates of the corresponding parameters, except the AMLE for the weights $\mathbf{w}_3$.

4.2. Distributions of HCO$_3$ and Ca

Again using the strategy of Section 2, the hyper-parameters are computed as $\mu_0 = 4.6381$, $\sigma_0^2 = 5.6468$ for HCO$_3$ and $\mu_0 = 3.0673$, $\sigma_0^2 = 9.2382$ for Ca.

The marginal posterior distributions for both variables are similar to the corresponding distributions for TDS. The numerical summaries of the distributions for HCO$_3$ are given in Tables 5–8, whereas those for Ca are given in Tables 9–12. Because the interpretations of these distributions are analogous to that for the TDS, they are not repeated here.

Table 5. Prior and posterior distributions of K, for log (HCO$_3$)

k	1	2	3
Prior	1/3	1/3	1/3
Posterior	0	0.6263	0.3737

Table 6. Posterior means and standard deviations of $\boldsymbol{\mu}_k$, $\mathbf{w}_k$ and σ_k^2 ($k = 2$), for $\log(\mathrm{HCO_3})$

	$\boldsymbol{\mu}_k$		$\mathbf{w}_k$		σ_k^2
Mean	3.8773	4.9594	0.1976	0.8024	0.0828
SD	0.0578	0.0254	0.0307	0.0307	0.0097

Table 7. Posterior means and standard deviations of $\boldsymbol{\mu}_k$, $\mathbf{w}_k$ and σ_k^2 ($k = 3$), for $\log(\mathrm{HCO_3})$

	$\boldsymbol{\mu}_k$			$\mathbf{w}_k$			σ_k^2
Mean	3.8527	4.6545	5.0273	0.1740	0.3059	0.5201	0.0811
S.D.	0.0953	0.3870	0.1467	0.0568	0.2715	0.2778	0.0095

Table 8. AMLE of $\boldsymbol{\mu}_k$, $\mathbf{w}_k$ and σ_k^2 ($k = 2, 3$), for $\log(\mathrm{HCO_3})$

	$k = 2$		$k = 3$		
σ_k^2	0.0715		0.0667		
$\boldsymbol{\mu}_k$	3.8697	4.9693	3.8857	4.9456	5.1335
$\mathbf{w}_k$	0.2032	0.7968	0.2014	0.7360	0.0625

Table 9. Prior and posterior distributions of K, for $\log(\mathrm{Ca})$

k	1	2	3
Prior	1/3	1/3	1/3
Posterior	0	0.6790	0.3210

Table 10. Posterior means and standard deviations of $\boldsymbol{\mu}_k$, $\mathbf{w}_k$ and σ_k^2 ($k = 2$), for $\log(\mathrm{Ca})$

	$\boldsymbol{\mu}_k$		$\mathbf{w}_k$		σ_k^2
Mean	2.2104	3.5787	0.1923	0.8077	0.1116
SD	0.0661	0.0291	0.0308	0.0308	0.0127

Table 11. Posterior means and standard deviations of $\boldsymbol{\mu}_k$, $\mathbf{w}_k$ and σ_k^2 ($k = 3$), for $\log(\mathrm{Ca})$

	$\boldsymbol{\mu}_k$			$\mathbf{w}_k$			σ_k^2
Mean	2.1477	3.0570	3.6862	0.1569	0.2864	0.5567	0.1079
SD	0.1438	0.5300	0.2442	0.0616	0.2754	0.2982	0.0142

Table 12. AMLE of $\boldsymbol{\mu}_k$, $\mathbf{w}_k$ and σ_k^2 $(k = 2, 3)$, for log (Ca)

	$k = 2$		$k = 3$		
σ_k^2	0.1006		0.0831		
$\boldsymbol{\mu}_k$	2.2227	3.5804	2.0492	2.5783	3.5950
$\mathbf{w}_k$	0.1834	0.8166	0.1494	0.0840	0.7666

4.3. Classification of the Changjiang basin

From the computational results in last section, the posterior distribution of K favors two components for all three variables. This means that most likely observations of each variable comes from a population with two sub-populations. An interesting question is how these two sub-populations are composed. In this section, we use the estimated posterior distribution to classify all observations of each variable into two groups.

We first introduce the general notation of classification. Let $f(x \mid \mu_{kj}, \sigma_k^2)$ be the jth component density in the mixture and $\hat{\mu}_{kj}$, $\hat{w}_{kj}$ and $\hat{\sigma}_k^2$ are the posterior mean estimates. For each $1 \leq j \leq k$, the classification probability of an observation $X = x$ belonging to jth sub-population can be defined as

$$P_j(x) = \frac{\hat{w}_{kj} f(x \mid \hat{\mu}_{kj}, \hat{\sigma}_k^2)}{\sum_{\ell=1}^{k} \hat{w}_{k\ell} f(x \mid \hat{\mu}_{k\ell}, \hat{\sigma}_k^2)} \tag{8}$$

Then we can classify each observation x_i in the present sample to sub-population j with the highest classification probability $P_j(x_i)$. For more discussion of classification, see, for example, Everitt *et al.* (2001).

Since there are two sub-populations for the Changjiang data sets, classification of the sample point x_i according to (8) is equivalent to comparing $P_1(x_i)$ and $P_2(x_i)$. In other words, the observation x_i is classified into sub-population 1, if $P_1(x_i) > P_2(x_i)$, and into sub-population 2 otherwise. The classification results of three variables are shown in Figures 9–11.

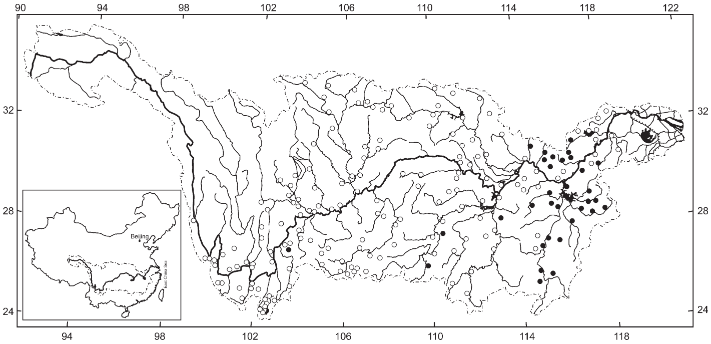

Figure 9. Classification of 191 observations of log (TDS). The filled circles represent sub-population 1 and the open circles represent sub-population 2

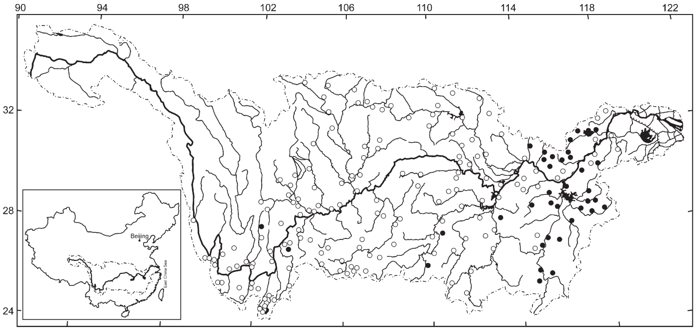

Figure 10. Classification of 191 observations of log (HCO_3). The filled circles represent sub-population 1 and the open circles represent sub-population 2

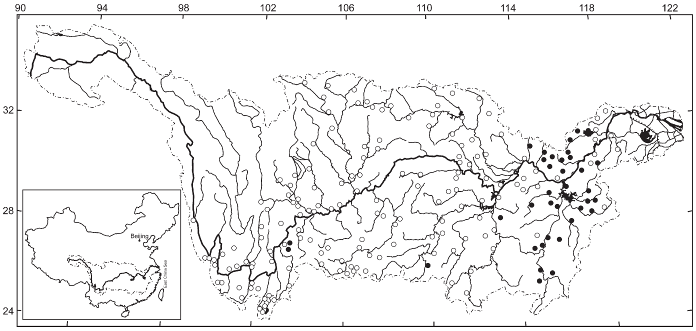

Figure 11. Classification of 191 observations of log (Ca). The filled circles represent sub-population 1 and the open circles represent sub-population 2

5. CONCLUSIONS

We have studied the statistical distributions of two major chemical elements (Ca and HCO_3) and the total dissolved solid concentration in the Changjiang, using a Bayesian finite mixture model and a novel computational algorithm. The results obtained provide better understanding of the distributions of chemical elements in the Changjing basin. Another major contribution is the identification of two sub-populations for each variable studied, which shows how the geological locations are associated with the distributions.

According to the classification results, most of the group 1 stations are located in the lower reaches of the Changjiang river and the stations of group 2 are located in the upper and middle reaches of the river area. The major chemical elements of the Changjiang are mainly controlled by chemical weathering, atmospheric precipitation and other natural processes as well as human activities (Chen *et al.*, 2002). In the lower reaches of the river basin, annual average precipitation is higher than in the middle and upper reaches of the river. Carbonate rocks, the weathering of which produces Ca and HCO_3 in the river water, are also less abundant in the lower reaches (Chen *et al.*, 2002). We think that these are the main causes as to why most of the stations in the lower reaches have a lower level of the

Table 13. Estimated mean levels (mg/l) of TDS, HCO_3 and Ca in two sub-populations

	TDS	HCO_3	Ca
Sub-population 1	79.84	48.42	9.12
Sub-population 2	221.41	142.59	35.87

chemical elements studied. The analytical results are based on the log-transformed data. The estimated sub-population means in the original scale can be easily calculated, and they are given in Table 13.

This work represents the first systematic study of the statistical distributions of chemical elements in the Changjiang basin. The proposed model is flexible enough to account for the unknown number of components in the mixture distribution. The algorithm used provides a simple and effective approach to the computation of these types of models, which is usually a challenging task in Bayesian computation.

ACKNOWLEDGEMENT

We would like to thank the Editor, Abdel El-Shaarawi, for helpful comments and encouragement. Financial support by Science and Engineering Research Canada (NSERC) is gratefully acknowledged.

REFERENCES

Belisle P, Joseph L, Wolfson DB, Zhou X. 2002. Bayesian estimation of cognitive decline in patients with Alzheimer's disease. *The Canadian Journal of Statistics* **30**: 1–176.

Besag J, Green PJ. 1993. Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, B* **55**: 25–37.

Brooks SP, Giudici P, Roberts GO. 2003. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society, B* **65**: 1–37.

Chen J, Wang F, Xie X, Zhang L. 2002. Major element chemistry of the Changjiang (Yangtze River). *Chemical Geology* **187**: 231–255.

Diebolt J, Robert CP. 1994. Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, B* **56**: 363–375.

Escobar MD, West M. 1995. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**: 577–588.

Everitt BS, Landau S, Leese M. 2001. *Cluster Analysis* (4th edn). Arnold: London.

Fu J, Wang L. 2002. A random-discretization based Monte Carlo sampling method and its application. *Methodology and Computing in Applied Probability* **4**: 5–25.

Gelman A, Carlin JB, Stern HS, Rubin DB. 1995. *Bayesian Data Analysis*. Chapman & Hall: London.

Gilks WR, Richardson S, Spiegelhalter DJ. 1998. *Markov Chain Monte Carlo in Practice*. Chapman & Hall: London.

Green PJ. 1995. Reversible jump MCMC computation and Bayesian model determination. *Biometrica* **82**: 711–732.

Ibrahim JG, Chen MH, Gray RJ. 2002. Bayesian models for gene expression with DNA microarray data. *Journal of the American Statistical Association* **97**: 88–99.

Leonard T, Hsu JSJ. 1999. *Bayesian Methods (An Analysis for Statisticians and Interdisciplinary Researchers)*, Cambridge, UK.

Richardson S, Green P. 1997. On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, B* **59**: 731–792.

Stephens M. 2000. Bayesian analysis of mixture models with an unknown number of components–an alternative to reversible jump methods. *Annals of Statistics* **28**: 40–74.

Titterington DM, Smith AFM, Makov UE. 1985. *Statistical Analysis of Finite Mixture Distribution*. John Wiley & Sons Ltd: UK.