Methodology

- Probability model with covariates
  - lognormal distribution with mean of ln's a function of
    - region,
    - source type (sw vs. gw), and
    - size of utility (population served)
  - constant variance of ln's
- Bayesian methodology
  - prior distribution for model parameters
  - posterior distribution computed using Markov Chain Monte Carlo (MCMC)
    - necessitated by model complexity and BDL data

Arsenic Occurrence Databases

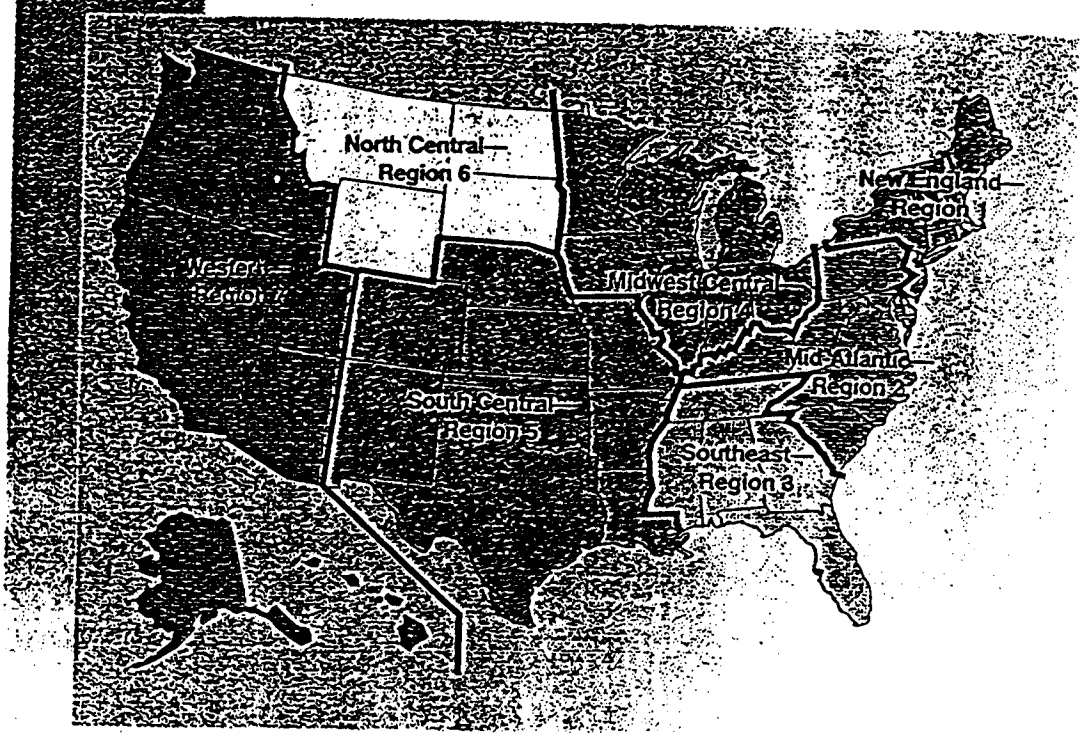
Database	Sample Locations	Source Type	Number of Sites	Detection Limit, µg/L	Percentage Below Detection Limit
WITAF National Arsenic Occurrence Survey (NAOS)	Raw water	Surface and groundwater	441	0.5	37%
EPA National Inorganics and Radionuclides Survey (NIRS)	Finished water	Groundwater	982	5	93%
Association of California Water Agencies	Raw water and finished water	Surface and groundwater	1,542	1	37%
12 State Data Reported to EPA	Raw water and finished water	Surface and groundwater	>11,000	Varies	65%

MODEL SPECIFICATION

$$Y_{ij} = \mu_i + \beta x_{ij} + \gamma g_{ij} + \epsilon_{ij}$$

- Y_{ij} is the natural logarithm of arsenic concentration in μg/L at j^{th} source in i^{th} region
- μ_i is a constant for i^{th} region, where i ranges over the seven geographical regions specified in NAOS
- x_{ij} is the natural logarithm of the population served by j^{th} source in i^{th} region (an indicator of the size and flow rate of the utility source)
- g_{ij} is 0 if j^{th} source in i^{th} region is a surface water source and 1 if it is a ground water source
- ϵ_{ij} represents those sources of random variation present at the jth source in i^{th} region but not captured by the covariates in the model.

FIGURE 2 US geographic regions based on arsenic NOFs



Source: Frey and Edwards, 1997.

DISTRIBUTIONAL ASSUMPTIONS

In the model

$$Y_{ij} = \mu_i + \beta x_{ij} + \gamma g_{ij} + \epsilon_{ij}$$

it is assumed that

$$\epsilon_{ij} \sim N(0, \sigma_\epsilon^2) \quad \forall i, j$$

$$\mu_i \sim N(\psi, \tau^2) \quad i = 1, \ldots, 7$$

That is, μ_i are sampled from a parent normal distribution (hierarchical model).

The normality assumption of ϵ_{ij} implies that conditional on all parameters,

$$Y_{ij} \sim N(\mu_i + \beta x_{ij} + \gamma g_{ij}, \sigma_\epsilon^2)$$

BAYESIAN METHODOLOGY

- Probability model:

$$X \sim f_{X|\Theta}(x|\theta)$$

where Θ are parameters $\Theta \sim f_{\Theta}(\theta)$

- Begin with prior distribution $f^0_{\Theta}(\theta)$
- Observe sample $X = \vec{x}_s$
- Compute posterior distribution

$$f_{\Theta|X}(\theta|\vec{x}_s) = \frac{f_{X|\Theta}(\vec{x}_s|\theta) f^0_{\Theta}(\theta)}{\int_{\Theta} f_{X|\Theta}(\vec{x}_s|\theta) f^0_{\Theta}(\theta) d\theta}$$

PRIOR DISTRIBUTIONS

Without substantive prior knowledge about parameters of hierarchical model, our priors were diffuse:

$$
\begin{aligned}
\psi &\sim N(0, 3^2) \\
\beta &\sim N(0, 10^2) \\
\gamma &\sim N(0, 10^2) \\
log(\sigma^2) &\sim N(0, 10^2) \\
log(\tau^2) &\sim N(0, 10^2)
\end{aligned}
$$

These parameters are assumed independent *a priori*, but are dependent in the posterior.

POSTERIOR ESTIMATES

Parameter	P.M.	P.S.D.
μ_1	-3.13	0.65
μ_2	-3.50	0.61
μ_3	-3.62	0.61
μ_4	-1.76	0.57
μ_5	-1.84	0.59
μ_6	-1.04	0.66
μ_7	-1.41	0.62
σ^2	2.23	0.21
ψ	-2.27	0.74
τ^2	1.76	1.76
β	0.21	0.05
γ	0.14	0.19

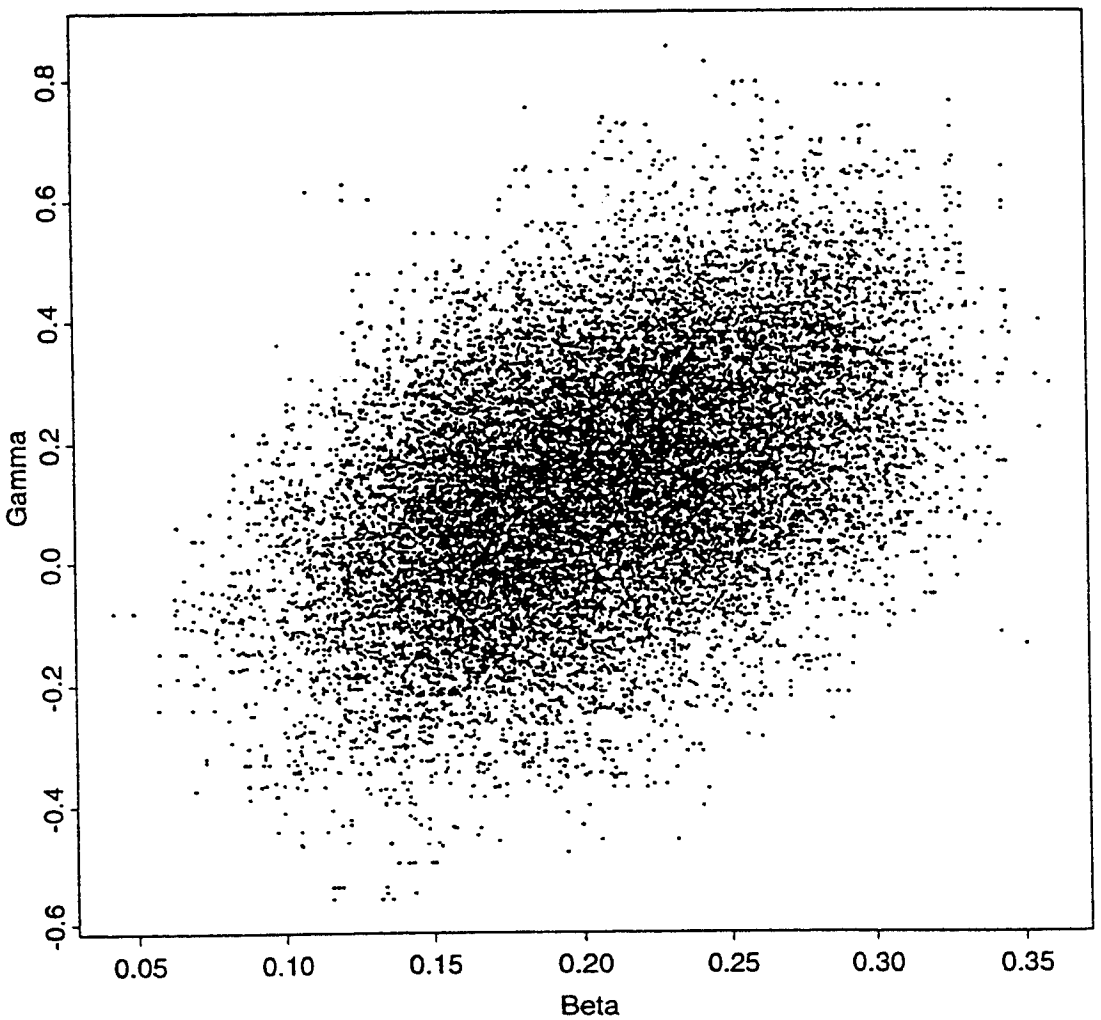


Figure 2: Scatterplot of γ versus β from a sample of size 30000 from the joint posterior distribution.