

A Bayesian Analysis of Leukemia Incidence Data Surrounding an Inactive Hazardous Waste Site

Ronald C. Neath

Department of Statistics and CIS

Zicklin School of Business

Baruch College, City University of New York

June 18, 2009

Outline

1. Introduction: Description of data
2. Data: Graphical summaries
3. The model: A three-stage Bayesian hierarchical model
4. Selection of prior distributions
5. Analysis: R and WinBUGS
6. Conclusion: Final remarks

Introduction

Data analyzed by Waller, Turnbull, Clark, Nasca (*Case Studies in Biometry*, 1994)

Methodology of Wakefield and Morris (*JASA*, 2001)

Data:

Exposed population and leukemia cases by census block

Five-year period from 1978–1982

Area surrounding GE Auburn hazardous waste site, Cayuga County NY

Introduction (cont.)

Our goal: to quantify the increased risk of leukemia associated with proximity to putative point source

Methodology: Hierarchical Bayesian modeling, making use of R and WinBUGS software

The data

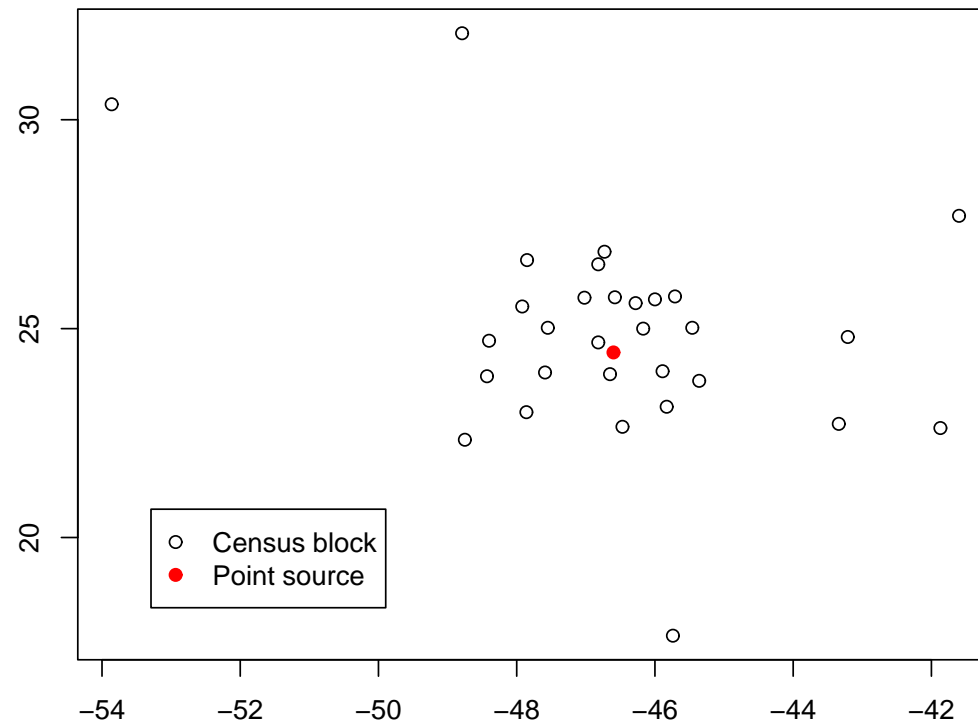
$n = 30$ regions (census blocks)

(x, y) -coordinates of geographic centroid of region, population and leukemia cases

x	y	pop	Y
-47.92	25.53	2422	0.66
-48.40	24.71	1019	0.28
-53.86	30.37	1618	1.28
-48.79	32.07	77	0.01
-41.60	27.70	1644	0.92
-47.85	26.64	2006	1.13

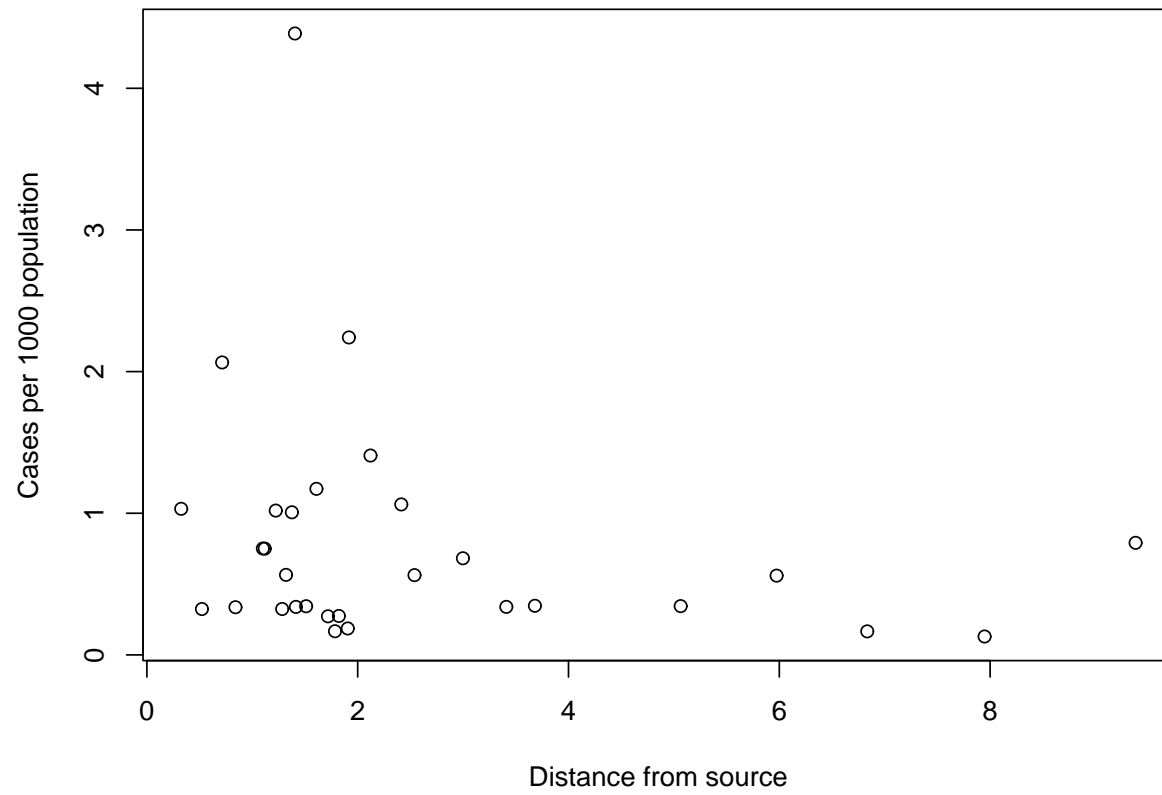
The data (cont.)

Geographic centroids of census blocks



The data (cont.)

Figure 1: Disease rate versus distance from point source



The model

Wakefield and Morris (*JASA*, 2001) analyzed incidence of stomach cancer relative to location of solid waste incinerator in north-east England

Y_i = observed disease count in area i

E_i = expected disease count in area i

X_i = log(population) in area i

d_i = distance from point source to centroid of area i

d_{ij} = distance between centroids of areas i and j

The model (cont.)

Assume

$$Y_i \sim \text{indep Poisson}(E_i \lambda_i)$$

where

$$\log \lambda_i = \log f(d_i; \alpha, \beta) + \eta_0 + \eta_1 X_i + U_i + V_i$$

where

$$f(d; \alpha, \beta) = 1 + \alpha \exp \left\{ (-d/\beta)^2 \right\}$$

is called the *location-risk function*

The model (cont.)

Random effects:

$$\mathbf{U} \sim N_n(\mathbf{0}, \sigma_u^2 \mathbf{I})$$

independent of

$$\mathbf{V} \sim N_n(\mathbf{0}, \sigma_v^2 \mathbf{H}(\varphi))$$

where

$$\mathbf{H}(\varphi)_{ij} = \exp\{-\varphi d_{ij}\}$$

Priors

Must assign prior distributions to:

location-risk parameters α and β ,

Poisson regression coefficients η_0 and η_1 ,

variance components σ_u^2 and σ_v^2 , and

spatial correlation parameter φ .

Priors (cont.)

Priors matter!

We took:

Regression coefficients $\eta_0, \eta_1 \sim \text{iid } N(0, 1000)$

Let $a \sim N(0, 1)$ and $\alpha = e^a - 1$, and let $\beta \sim \text{gamma}(2.5, 1)$

Precisions $\sigma_u^{-2}, \sigma_v^{-2} \sim \text{iid } \text{gamma}(.5, .0005)$

Spatial correlation parameter $\varphi \sim \text{gamma}(3, 1)$

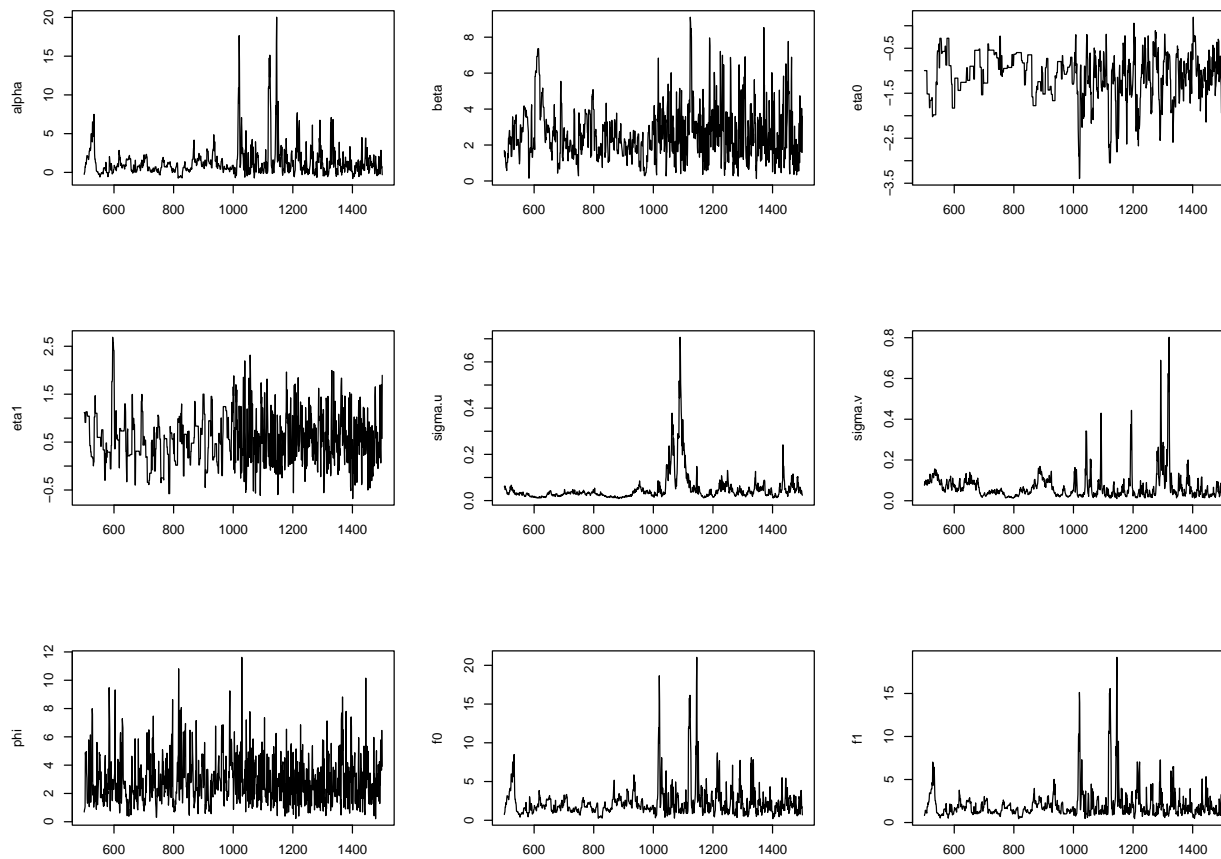
Analysis: Markov chain Monte Carlo

Simulate ergodic Markov chain whose unique stationary distribution is given by the posterior $\pi(\alpha, \beta, \eta_0, \eta_1, \sigma_u, \sigma_v, \varphi|y)$.

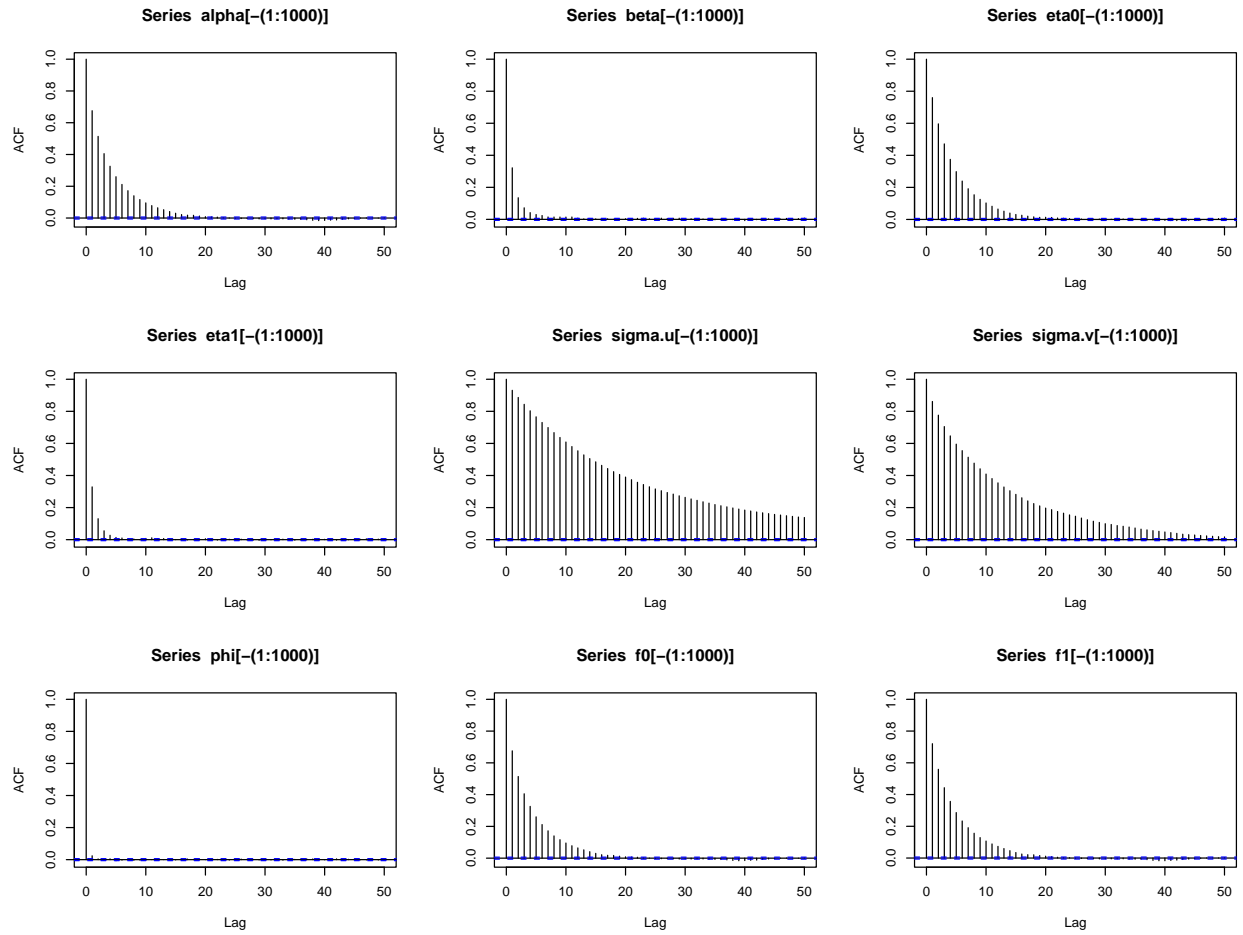
Use WinBUGS.

We ran the chain for 10^6 updates, and saved every 10th draw, so total MCMC sample size is 10^5 .

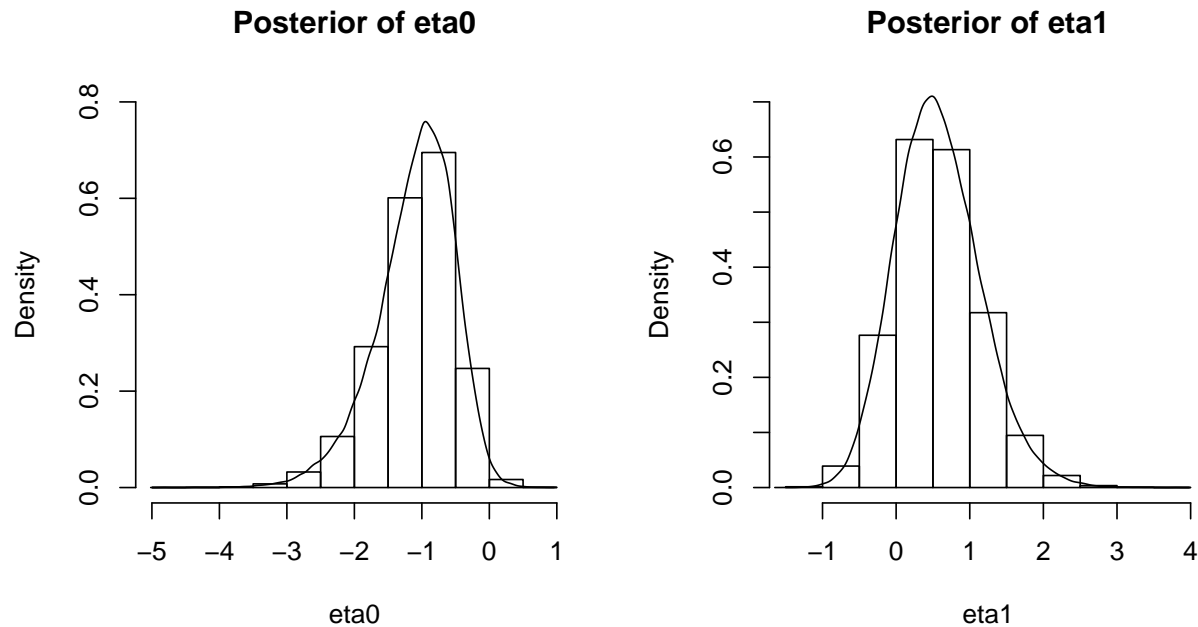
MCMC (cont.): Marginal trace plots



MCMC (cont.): Autocorrelation functions

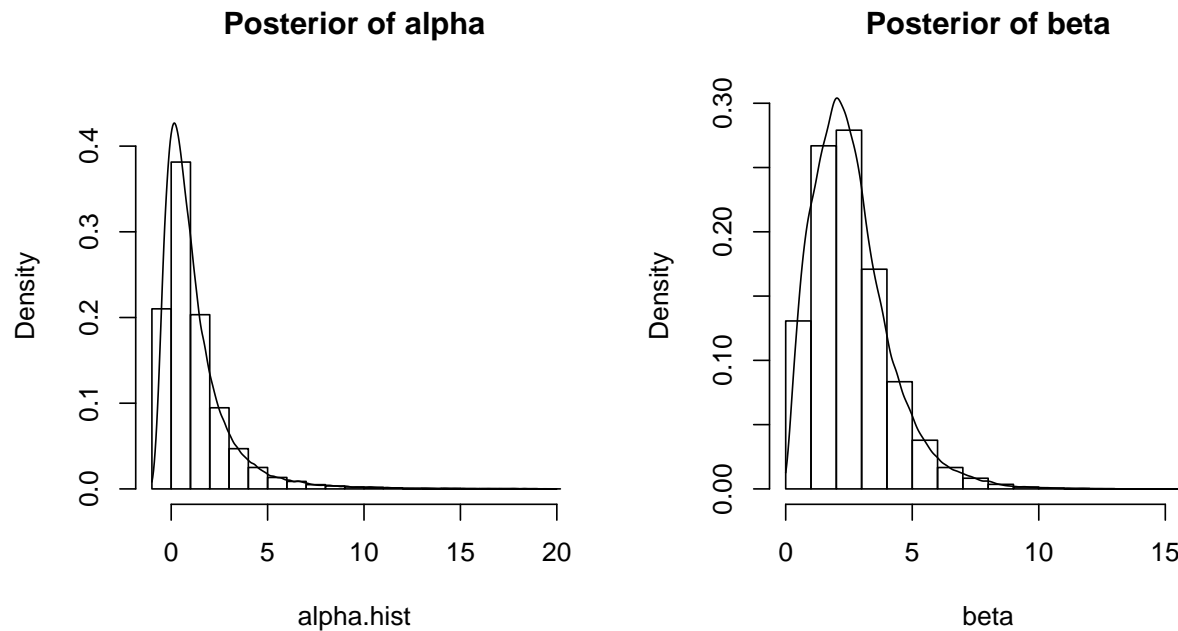


MCMC approximations to marginal posteriors: η_0 and η_1



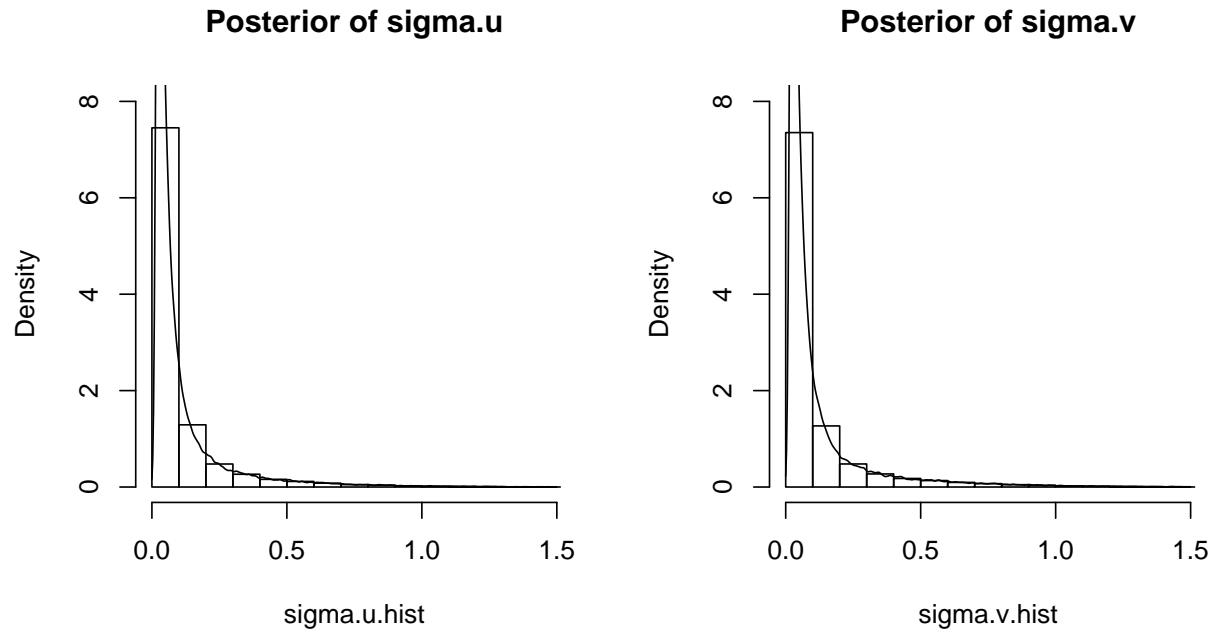
	Mean	MCSE	Std Dev	5%ile	Median	95%ile
η_0	-1.10	.005	0.58	-2.16	-1.03	-0.28
η_1	0.57	.003	0.58	-0.32	0.54	1.57

MCMC approximations to marginal posteriors: α and β



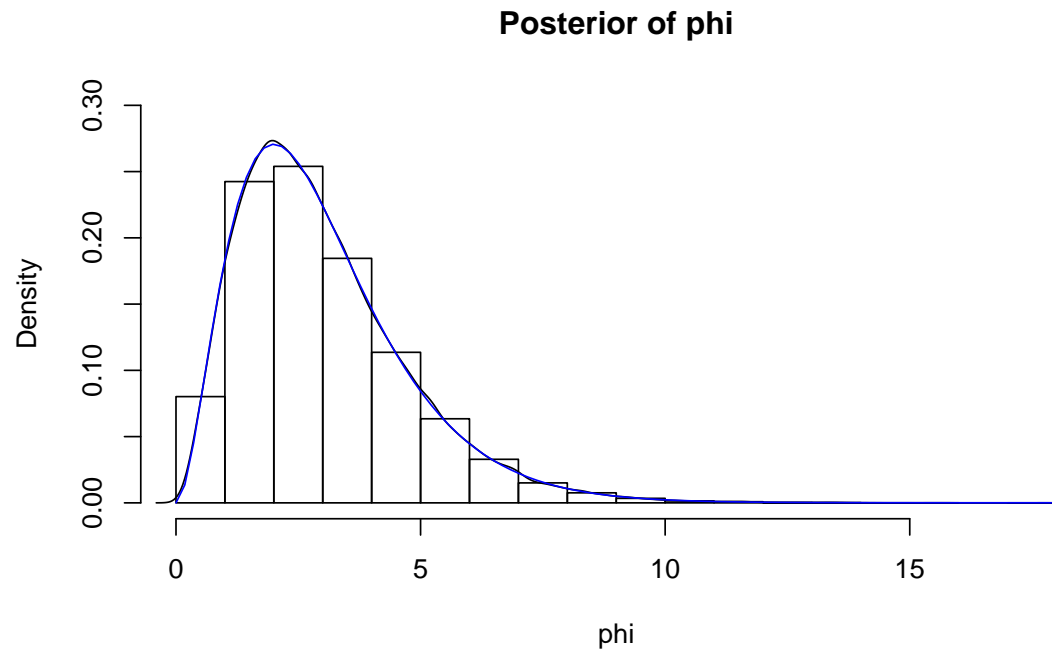
	Mean	MCSE	Std Dev	5%ile	Median	95%ile
α	1.21	.016	1.89	-0.46	0.72	4.48
β	2.58	.008	1.52	0.59	2.34	5.40

MCMC approximations to marginal posteriors: σ_u and σ_v



	Mean	MCSE	Std Dev	5%ile	Median	95%ile
σ_u	0.10	.003	0.16	0.02	0.05	0.41
σ_v	0.11	.003	0.18	0.02	0.05	0.48

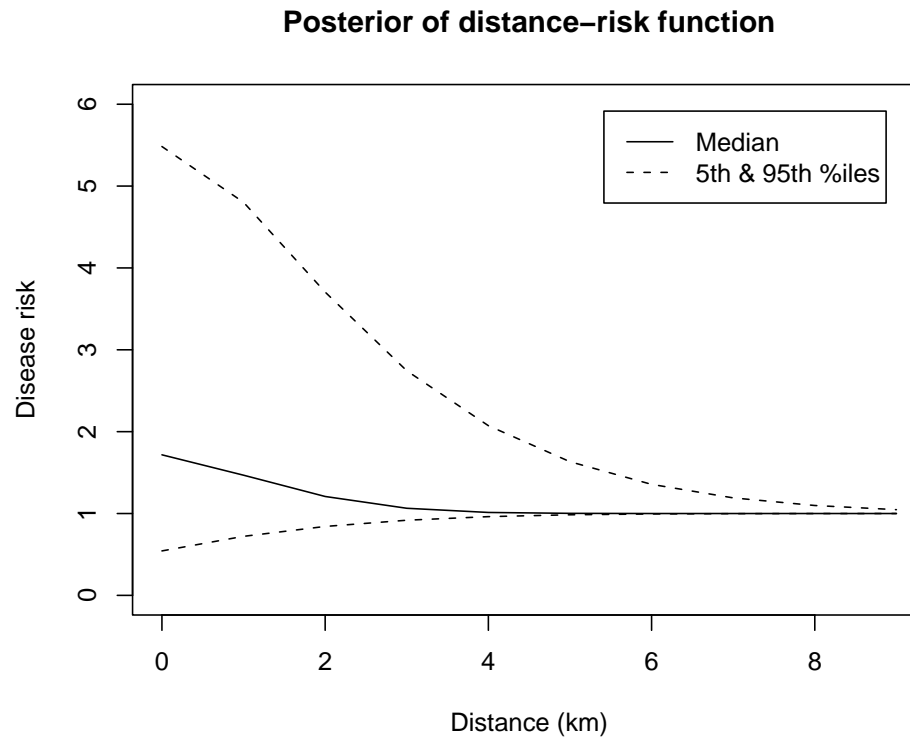
MCMC approximations to marginal posteriors: φ



	Mean	MCSE	Std Dev	5%ile	Median	95%ile
φ	3.00	.006	1.73	0.82	2.68	6.29

Analysis: Location-risk function

Location-risk function $f(d; \alpha, \beta) = 1 + \alpha \exp\{-(d/\beta)^2\}$



Analysis: a simplified model

The inclusion of random effects is controversial.

There is concern that random effects dilute the effect of the parameters of greatest interest (location-risk parameters α and β).

Is there compelling evidence of extra-Poisson variability in the data?

Fit equivalent model without random effects, i.e., assume $\sigma_u^2 = \sigma_v^2 = 0$.

Analysis: a simplified model (cont.)

Use R2WinBUGS (Sturtz, Ligges, and Gelman, 2005)

Inference for Bugs model at "C:/DOCUME~1/RNeath/MY...

5 chains, each with 1e+05 iterations (first 50000 discarded),

n.thin = 250

n.sims = 1000 iterations saved

	mean	sd	2.5%	25%	50%	75%	97.5%	Rhat	n.eff
eta0	-1.1	0.5	-2.2	-1.4	-1.0	-0.7	-0.2	1	1000
eta1	0.6	0.6	-0.4	0.2	0.6	0.9	1.7	1	960
alpha	1.2	1.6	-0.6	0.1	0.7	1.8	5.6	1	1000
beta	2.6	1.5	0.4	1.5	2.5	3.4	6.0	1	1000
deviance	64.4	2.2	61.5	62.7	63.9	65.5	69.8	1	1000

Final remarks

1. Prediction of random effects: Take means of $U_i + V_i$ from MCMC output.
2. Here we considered only a subset of the data collected by Waller et al. In fact there were 11 hazardous waste sites in a larger region. In future work I might consider the model

$$\log \lambda_i = \sum_{k=1}^{11} \log f(d_{ik}; \alpha_k, \beta_k) + \eta_0 + \eta_1 X_i + U_i + V_i$$

3. The use of R and WinBUGS was essential to this project's success.