Bayesian Data Analysis: Final Project

Meuse Dataset

Madhura Baxi and Elizabeth Spencer

December 10, 2019

Introduction

Lead toxicity is a major problem impacting the ecosystem and human health. Lead can be added to soil through various sources such as discharge of waste streams to nearby water bodies. Increased amount of lead in soil can cause a multitude of detrimental effects in plants and animals including decrease in the rate of growth and reproduction. Consumption of high amounts of lead can also lead to neurological effects in vertebrates. Once consumed into the body, lead gets absorbed in the blood and distributed throughout the body hampering the oxygen carrying capacity of the blood as well as gets accumulated in the bones. In this way, lead exposure can adversely affect kidney function, immune and cardiovascular system system, reproduction as well as nervous system. Especially, infants and children in their developmental phase are most sensitive to lead exposure which can adversely affect their brain development causing them to suffer from various mental health problems in the future. Therefore, it is essential to study and quantify lead concentration in soil and understand how and to what degree is this lead concentration in soil affected by various sources of lead and soil type. Hence, in this study we aimed to model the lead concentration found in a variety of sites as a function of distance to the river and soil type. We choose to use a hierarchical linear model such that the sites are grouped by soil type.

Method

Dataset Description

The meuse data set from sp package in R software, comprised of concentrations of four heavy metals (lead, zinc, copper, cadmium) measured in the soil at a number of site locations in a flood plain along the river Meuse. Dataset also contained information on other measurements including the distance to river from each site, (x,y) coordinates of each site and soil type found at each site. Figure 1 shows the distances between sites.

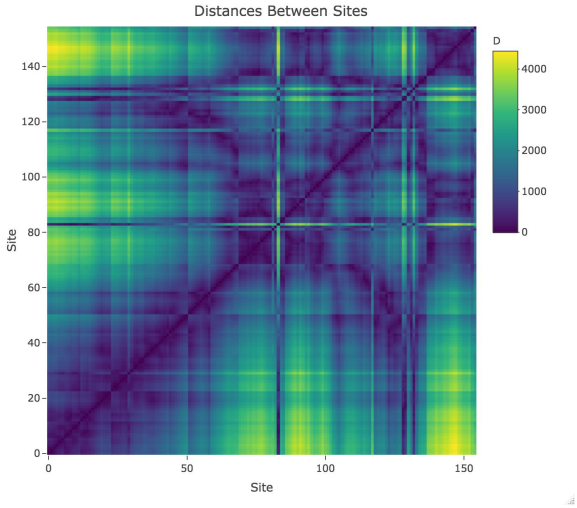

Figure 1: Euclidean distance (meters) between each pair of sites.

Model: Likelihood

The Meuse dataset contained lead concentrations from n sites. The aim of this study was to investigate if the distance from the Meuse river and soil type impact the measurement of lead. Figure 2 (left), demonstrates a plot of log of lead concentration as a function of distance to the river in meters and it is visually evident that there is some negative trend showing that as the sites go farther from the river, there is less lead. However, dataset contained three different kinds of soil types and each may have different properties that impact the amount of lead concentration. We explored this further in Figure 2 (right), by plotting the same relationship, but now divided by soil types 1, 2 and 3. We see here that there is still a negative trend between distance and lead concentration; however, the slopes and intercepts of each line vary slightly around the slope and intercept in the pooled data in Figure 2 (right). It is important to note that by dividing the data into three groups we lose data points and thus statistical power, especially for soil type three which has only 12 site measurements.

We hypothesize that the farther from the river, lead concentration decreases and that this relationship varies by soil type. We choose to implement a hierarchical model as a compromise between entirely pooling the data, as in Figure 2 (right), which ignores the potential confound of soil type completely, or between treating the data separately by soil type, as in Figure 2 (left), which does not let us leverage the whole dataset. This final point is important because we have few measurements for soil type three as shown in Figure 2 (right). In our implementation of the hierarchical model, we will effectively fit three separate regressions for each soil type separately, but we will treat each of the fit intercepts and slopes as random effects from fixed mean for the slopes and the intercepts.

If we consider our data vector y of measurements of the log of lead concentration from our n sites, the likelihood is defined as follows:

$$y|\ \beta,\ \sigma^2, \tau^2\ \sim\ N(X\beta, \sigma^2 I_n\ +\ \tau^2 H(\phi))$$

Our design matrix, X, is of size n x 6 covariates. We have three covariates that are indicator vectors for the three soil types, and three covariates that are the soil type indicator vectors * the distance to the river. This effectively estimates three intercept terms and three slope terms modelling log of lead concentration as a function of distance to the river for each soil type separately. We assume that there is some variance in the data due to error, σ^2, as well as some spatial covariance, $\tau^2 H(\phi)$, where sites that are nearby may have similar amounts of lead. In Figure 3, we plot $H(\phi)$ for three different values of ϕ and observe that as ϕ increases, there is more spatial correlation. To make posterior derivation easier, we expand the likelihood as follows:

$$y|\ \beta,\ \sigma^2, \tau^2\ \sim\ N(X\beta\ +\ \delta, \sigma^2 I_n),$$

and set the prior for delta such that:

$$\delta \sim N(0,\ \tau^2 H(\phi)).$$

Next, we place a prior distribution on β as follows:

$$\beta|\ \alpha,\ \kappa^2\ \sim N(X_\beta \alpha,\ \kappa^2 I_6),$$

where $X_\beta = [1\ 1\ 1\ 0\ 0\ 0;\ 0\ 0\ 0\ 1\ 1\ 1\ ]^T$ and α is a 2 x 1 vector containing the fixed effects estimates for slope and intercept.

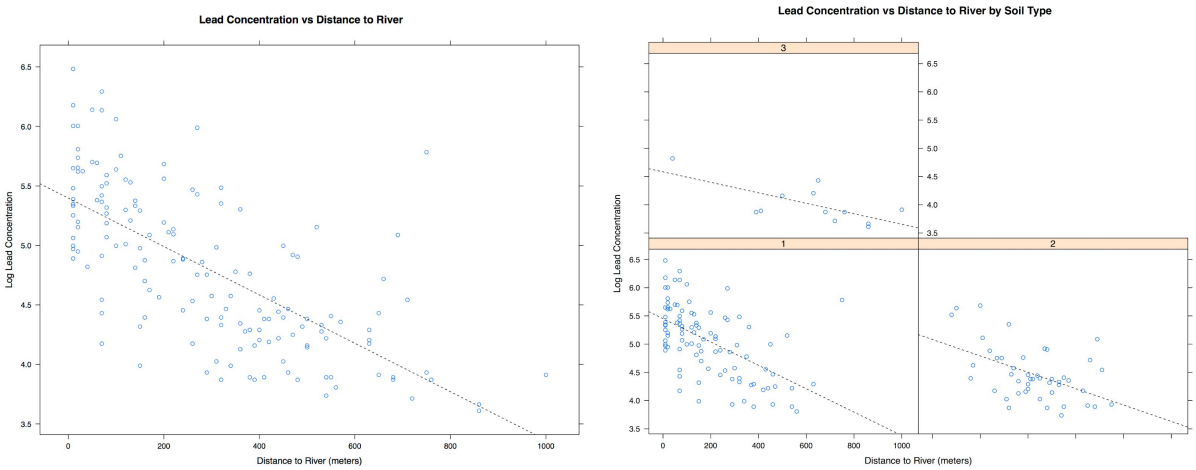

Figure 2: **(Left)** Linear fit showing the effect of distance to river on lead concentration with data from all soil types pooled together. The estimates for the fit line are 5.039 for the intercept and -0.0015 for the slope. **(Right)** Linear fit showing the effect of distance to river on lead concentration separately for each soil type.

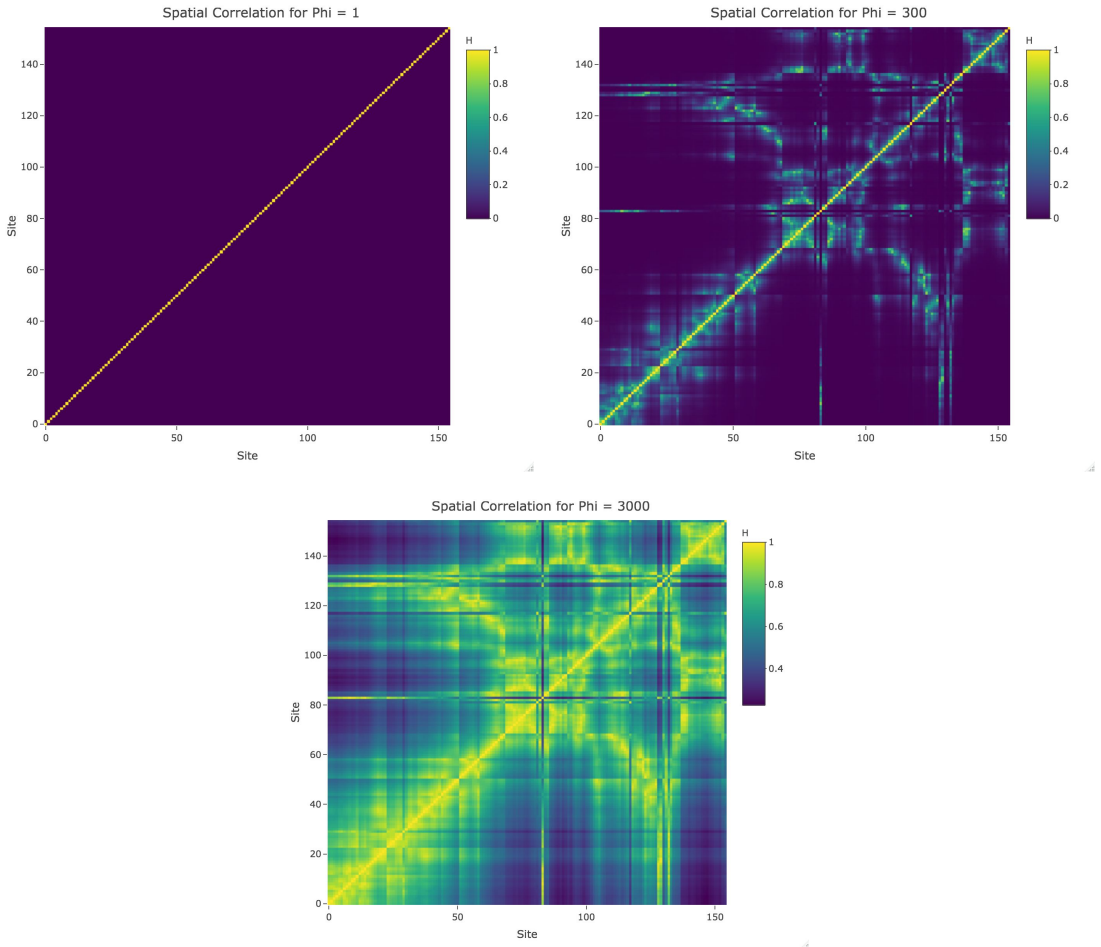

Figure 3: Spatial correlation matrices for different values of phi. As phi increases, the range at which nearby site are correlated increases. **(Top Left)** Phi=1 **(Top Right)** Phi=300 **(Bottom)** Phi = 3000.

Model: Priors

For the priors on α and σ^2, we chose flat uniform priors, and for the priors on κ^2, τ^2, and ϕ, we chose an $Inverse-Gamma$ distribution:

1. $P(\alpha) \propto 1$: flat uniform prior
2. $P(\sigma^2) \propto 1/\sigma^2$: flat uniform prior in $log(\sigma^2)$
3. $P(\kappa^2) \propto Inverse-Gamma(\alpha_\kappa, \beta_\kappa)$
4. $P(\tau^2) \propto Inverse-Gamma(\alpha_\tau, \beta_\tau)$
5. $P(\phi) \propto Inverse-Gamma(\alpha_\phi, \alpha_\phi)$

We initialized starting values for β using MLE estimates for $log(lead) \sim f(distance\ to\ river)$ (linear fit), $\widehat{\beta}$, separating by soil type. In order to determine the

starting values for σ^2 and τ^2, the variance for the estimation error and spatial component in estimating β from y, we computed the deviance from the above mentioned linear fit, $log(lead) \sim f(distance\ to\ river)$. We then divided the deviance by n, to get the deviance per observation. Then assuming equal weight on the variance for the estimation error and spatial component, we divided the deviance per observation by 2 and set this value as our initial values τ^2 and σ^2. We set the parameters for the prior on τ^2, $Inverse-Gamma(\alpha_\tau = 10, \beta_\tau = 4)$, so that the expectation of τ^2, $\beta_\tau/(\alpha_\tau - 1)$, would be close to our initialization value for τ^2.

To initialize κ^2, we did a linear fit $\widehat{\beta} \sim X_\beta * \alpha$, where $\widehat{\beta}$ is our MLE estimates computed for the β starting values. The parameters α_κ and β_κ in the $Inverse-Gamma$ for sampling κ^2 were set to 10 and 4 respectively to match those of τ^2.

Lastly, we set α_ϕ to the value of 10, upon trying various different values such as 0.001, 0.1, 1, 10. Values of α_ϕ lower than 0.1 led to extremely high values of ϕ implying extremely high spatial correlation. We chose to set α_ϕ to be 10 because then expectation of ϕ is approximately 1 and $H(\phi)$ is approximately the identity matrix, thus giving us a noninformative prior.

Full joint posterior:

Assuming independence of parameters, we can write the joint distribution as follows:

$$
\begin{aligned}
P(\beta,\ \alpha,\ \kappa^2, \sigma^2, \tau^2, \phi\ , \delta, | y) &= P(y \mid \beta,\ \alpha,\ \kappa^2, \sigma^2, \tau^2, \phi,\ \delta\ ) * P(\beta,\ \alpha,\ \kappa^2, \sigma^2, \tau^2, \phi\ , \delta) \\
&= |\sigma^2 I_n|^{-1/2} exp\{-\tfrac{1}{2}\ (y\ -\ (X\beta + \delta))^T (\sigma^2 I_n)^{-1} (y\ -\ (X\beta + \delta))\}\ *\ ... \\
&\quad |\kappa^2 I_n|^{-1/2} * exp\{-\tfrac{1}{2}(\beta - X_\beta \alpha)^T (\kappa^2 I_n)^{-1} (\beta - X_\beta \alpha)\}\ * 1 * ... \\
&\quad (\kappa^2)^{(\alpha_\kappa + 1)} exp\{-\beta_\kappa/\kappa^2\}\ * 1/\sigma^2\ *\ (\tau^2)^{(\alpha_\tau + 1)} exp\{-\beta_\tau/\tau^2\}\ * ... \\
&\quad (\phi\ )^{(\alpha_\phi + 1)} exp\{-\alpha_\phi/\phi\}\ * ... \\
&\quad |\tau^2 H(\phi)|^{-1/2}\ exp\{-\tfrac{1}{2}\delta^T (\tau^2 H(\phi))^{-1} \delta\}
\end{aligned}
$$

Conditional posterior derivation:

To find the conditional posterior for each parameter, we isolate the terms containing the parameter of interest from the full joint posterior above.

1. $P(\sigma^2 |\ \beta,\ \alpha,\ \kappa^2, \tau^2, \phi\ , \delta, y\ ) \propto$

$$
\begin{aligned}
&1/\sigma^2\ *\ |\sigma^2 I_n|^{-1/2} exp\{-\tfrac{1}{2}\ *(y\ -\ (X\beta + \delta))^T * (\sigma^2 I_n)^{-1} * (y\ -\ (X\beta + \delta))\} \\
&= 1/\sigma^2\ *\ 1/(\sigma^2)^{n/2} * exp\{-\tfrac{1}{2\sigma^2} * (y\ -\ (X\beta + \delta))^T * (y\ -\ (X\beta + \delta))\} \\
&= 1/(\sigma^2)^{1+n/2}\ *\ \ exp\{-\tfrac{1}{2\sigma^2} * ((y\ -\ (X\beta + \delta))^T * (y\ -\ (X\beta + \delta))) * n/n\} \\
&= Scaled-Inv-Chi^2(n, \tfrac{1}{n} *\ [y\ -\ (X\beta + \delta)]^T * [y\ -\ (X\beta + \delta)])
\end{aligned}
$$

2. $P(\tau^2|\ \beta,\ \alpha,\ \kappa^2, \sigma^2, \phi\ , \delta, y\ ) \propto$
$\tau^{2(\alpha_\tau+1)} * exp\{-\beta_\tau/\tau^2\}\ *\ |\tau^2 H(\phi)|^{-1/2} exp\{-(\frac{\beta_\tau}{\tau^2}+\frac{1}{2}\ * \delta^T * (\tau^2 H(\phi))^{-1} * \delta)\}$

$= \tau^{2(\alpha_\tau+1)} * \tau^{2(-n/2)} *\ exp\{-\beta_\tau/\tau^2\}\ *\ |H(\phi)|^{-1/2} exp\{-(\frac{\beta_\tau}{\tau^2}+\frac{1}{2}\ * \delta^T * (\tau^2 H(\phi))^{-1} * \delta)\}$
$= \tau^{2(\frac{2\alpha_\tau+n}{2}+1)}\ * |H(\phi)|^{-1/2}\ exp\{-\frac{1}{2\tau^2}(2\beta_\tau + \delta^T * H(\phi)^{-1} * \delta)\}$
$\propto \tau^{2(\frac{2\alpha_\tau+n}{2}+1)} *\ exp\{-\frac{1}{2\tau^2}(2\beta_\tau + \delta^T * H(\phi)^{-1} * \delta)\ *\ \frac{2\alpha_\tau+n}{2\alpha_\tau+n})\}$
$= Scaled-Inv-Chi^2(2\alpha_\tau + n, \frac{1}{2\alpha_\tau+n} * [2\beta_\tau *\ \delta^{\ T} * H(\phi)^{-1} * \delta])$

3. Similarly,
$P(\kappa^2|\ \beta,\ \alpha,\ \tau^2, \sigma^2, \phi\ , \delta, y\ ) \propto Scaled-Inv-Chi^2(n+2\alpha_\kappa, \frac{1}{n+2\alpha_\kappa} * 2\beta_\kappa * [\beta - X_\beta \alpha]^T [\beta - X_\beta \alpha])$,
where n is the length of β, in this case $n = 6$.

All of the above parameters will be updated using Gibbs sampling. The spatial range parameter, ϕ, will be updated using the Metropolis-Hastings random walk step (a normal distribution with mean as last updated phi and standard deviation 0.1), and β and α will be estimated using expectation maximization.

Results

In order to estimate our parameters, we implemented our Metropolis-Gibbs sampling procedure using four chains, each with 10,000 simulations and a warm-up period of 2,000. Table 1 shows the quantiles of the estimates for each variable. For the beta variables, the last two subscripts indicate soil type (A=1,B=2,C=3) and intercept or slope (0 or 1 respectively). For example, BetaA0 and BetaA1 represent the intercept and slope term respectively for the model fit on the soil type 1 data. Comparing the 50% quantiles for all the intercept terms for the separate and pooled parts of the model, (BetaA0, BetaB0, BetaC0, and Alpha0), we see the estimates were not identical though approximately around 5. Similarly, the slope terms, BetaA1, BetaB1, BetaC1, and Alpha1, were all negative and in the same order of magnitude. The Alpha0 and Alpha1 estimates were close to the fit estimates on the simple linear model $log(lead\ concentration) \sim f(distance\ to\ river)$, ignoring soil type, which were 5.039 for the intercept and -0.0015 for the slope as depicted in Figure 2 (left).

The estimate for ϕ which represents the spatial range on the covariance matrix is approximately one, implying very low covariance between neighboring regions, and a variance structure of approximately the identity matrix as shown in Figure 3 (top left).

Table 1: Statistics of Sampled Variable from their Posterior Distributions

	2.5%	25%	50%	75%	97.5%
BetaA0	5.31	5.40	5.45	5.50	5.59
BetaB0	4.70	4.95	5.08	5.21	5.46
BetaC0	4.06	4.49	4.72	4.95	5.38
BetaA1	-0.0036	-0.0022	-0.0021	-0.0018	-0.0015
BetaB1	-0.0023	-0.0017	-0.0014	-0.0012	-0.00062
BetaC1	-0.002	-0.001	-0.0018	-0.0008	-0.0001
Alpha1	-0.73	-0.24	-0.002	0.24	0.73
Alpha0	4.30	4.83	5.08	5.34	5.87
Kappa2	0.18	0.28	0.37	0.50	0.94
Sigma2	0.18	0.21	0.23	0.25	0.29
Tau2	0.03	0.04	0.05	0.05	0.06
Phi	0.58	0.84	1.04	1.29	2.01

Model Checking

Bulk_ESS and Tail_ESS were computed for each variable in our model, which gave us crude measures of effective sample size for bulk and tail quantities respectively. An ESS > 100 per chain is considered good, and Rhat is the potential scale reduction factor on rank normalized split chains (at convergence, Rhat <= 1.05). As seen in Table 2, we observed all variables to have ESS > 100. ESS for ϕ was the least of all variables but still > 100. Rhat for all variables was also observed to be <=1.05.

Table 2: Model Inference Statistics.

	Mean	SD	Rhat	Bulk_ESS	Tail_ESS
BetaA0	5.4	0.1	1.00	15993	15926
BetaB0	5.1	0.2	1.00	14511	15650
BetaC0	4.7	0.3	1.00	12649	14940
BetaA1	0.0	0.0	1.00	15780	16019
BetaB1	0.0	0.0	1.00	14463	15642
BetaC1	0.0	0.0	1.00	13058	15084
Alpha1	0.0	0.4	1.00	15196	14803
Alpha0	5.1	0.4	1.00	11991	14473
Kappa2	0.4	0.2	1.00	13720	14634
Sigma2	0.2	0.0	1.00	13645	14894
Tau2	0.0	0.0	1.00	15340	16081
Phi	1.1	0.4	1.01	267	193

Next, we used two main tools to assess the convergence of our variables using (1) trace plots (Figure 4), (2) autocorrelation plots (Figure 5) and (3) density plots (Figure 6). First, we analyze the trace plots and observe excellent mixing of the four chains for all variables. Second, we use autocorrelation plots to assess the stationarity of each variable. We observe low autocorrelation for all variables at all nonzero lags, except for ϕ, which had strong correlation at all lags. Third, we compare the posterior density plots for each variable (Figure 6) and observe that the posterior distributions for all variables match well except for ϕ, where we observe some differences in the posterior distributions from 4 chains.

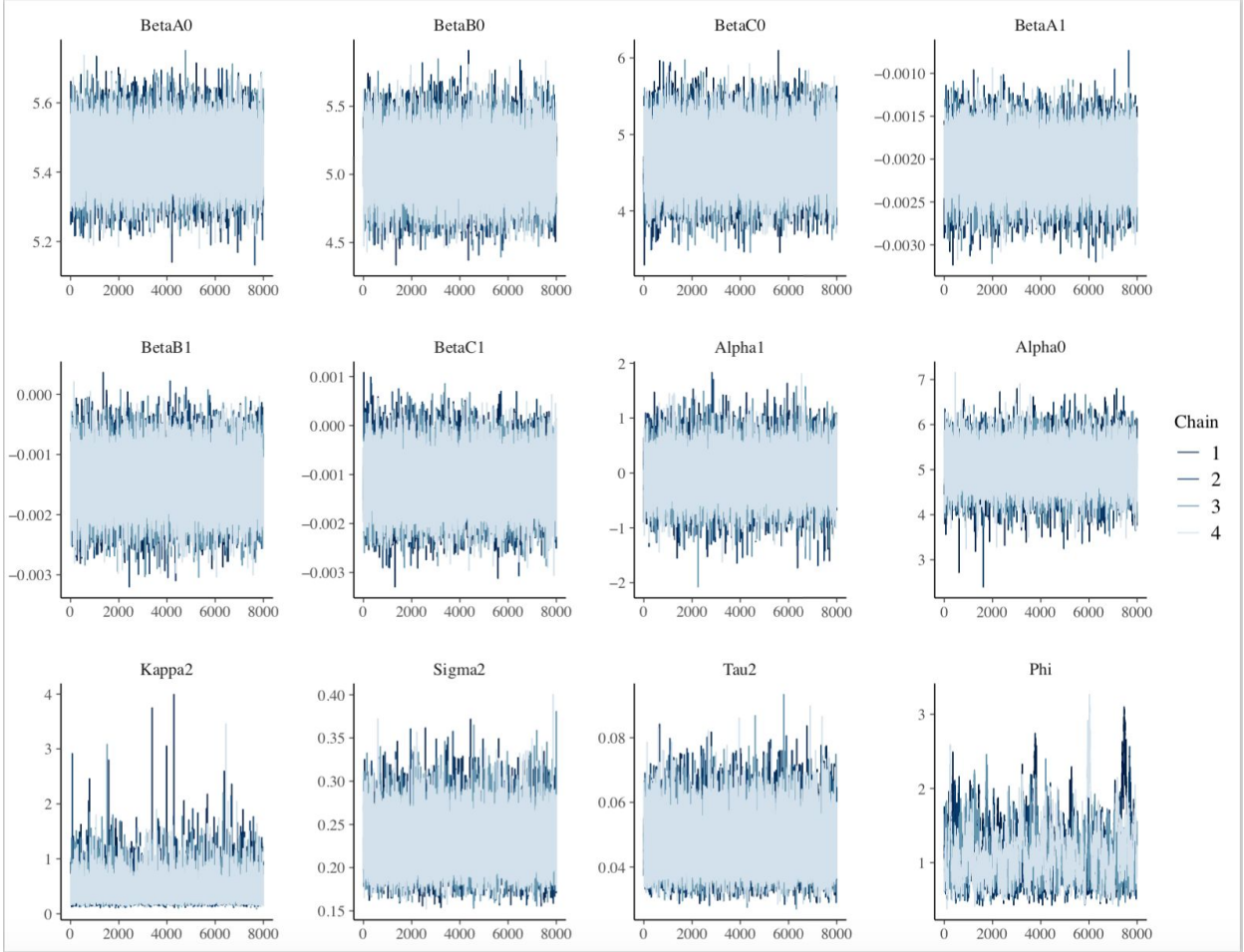

Figure 4: Trace Plots for each sampled variable

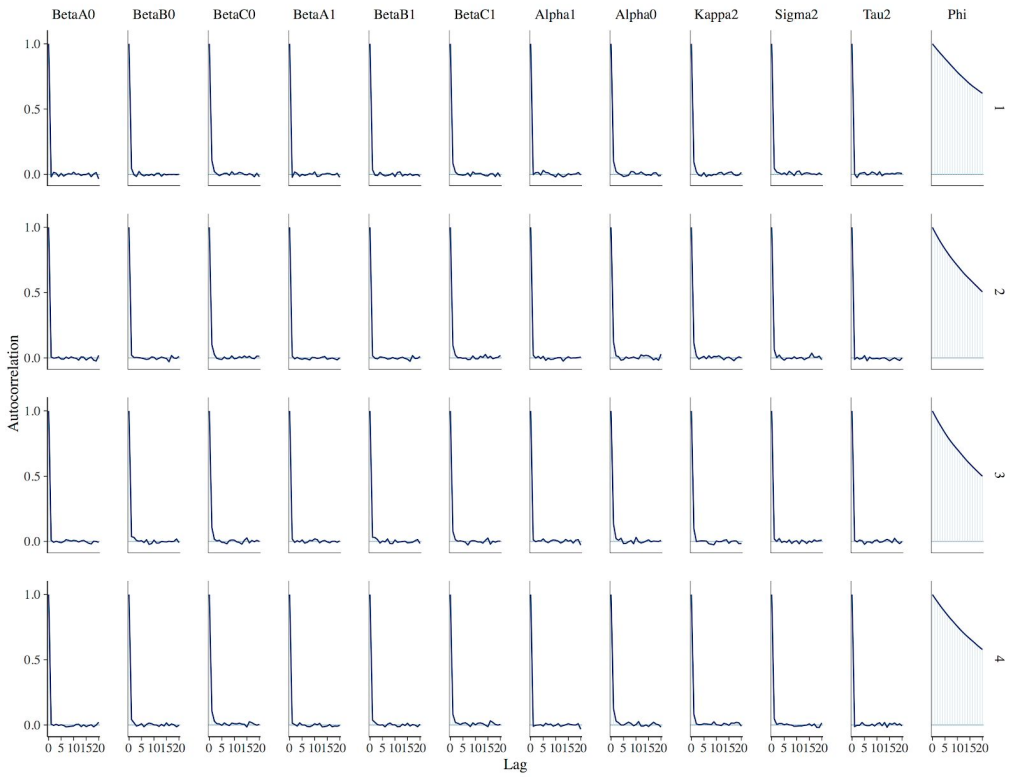

Figure 5: Autocorrelation Plots for each variable

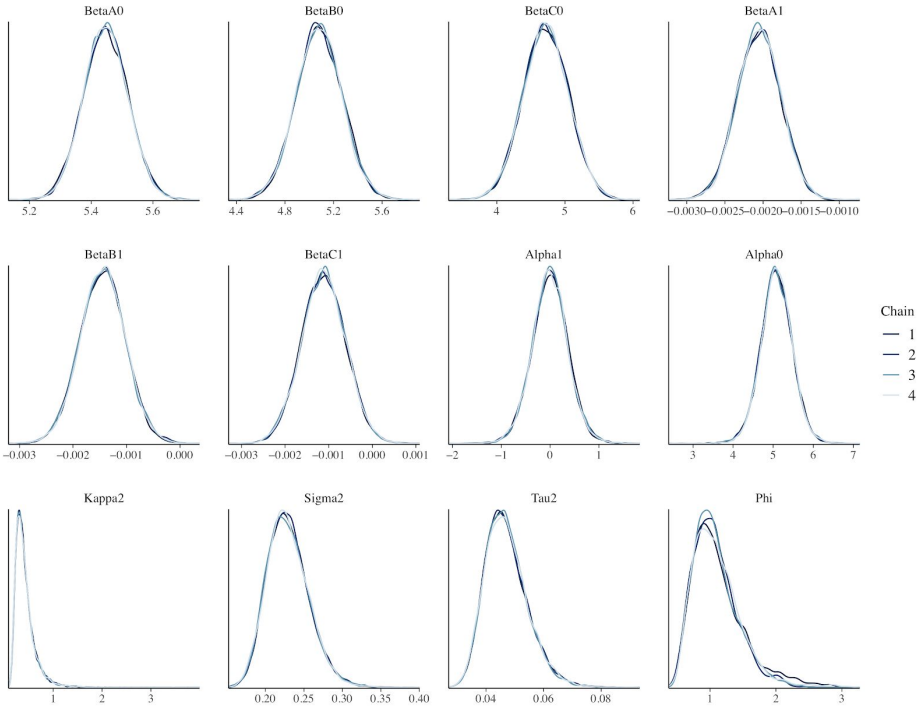

Figure 6: Density Plot for each Variable

Discussion & Conclusion

By implementing a hierarchical linear model, we were able to assess the relationship between lead concentration, and distance to the Meuse river and soil type. We determined there is a negative relationship between lead concentration and distance to the river, implying that the closer the site is to the river, the higher the lead concentration. Furthermore, we observed different relationships by soil type. Soil type 1 has the strongest negative relationship and then soil type 2 and 3 follow in that order; however, the amount of observations are far fewer for soil type 3 so we cannot make any conclusive arguments about the impact of soil type on lead concentration.

Most of our variables converge and achieve stationarity implying a good model fit. The Rhat and ESS values for all parameters are within the acceptable range for convergence. The trace plots, autocorrelation plots and density plots all show convergence of all parameters, except for ϕ. The trace plots for the four chains mix well, the autocorrelation plots indicate that the distribution of the variables have reached steady state and are not dependent on the past, and the posterior density plots match well for all the chains. There is also low standard deviation on most estimated variables indicating good fit. In summary, we achieve an overall good fit of our hierarchical linear model with the potential exception of ϕ which does not pass all of our model tests.

The following limitations of this model should be kept in mind while interpreting the results. To start, we made a strong assumption that the variance would be the same for the intercept and the slope of our fit line. Additionally, we made the assumption that the prior on the spatial range parameter ϕ is an $Inverse-Gamma$ with parameters $\alpha_\phi = \beta_\phi$. This is a strong assumption because it is encouraging our estimate of ϕ to remain approximately 1. In Table 1, we observed that the 50% quantile of ϕ is approximately 1 corresponding to a spatial covariance matrix that is approximately the identity (see Figure 3 top left). This implies that the sites do not necessarily have similar concentrations of lead. This could be due to the fact that there is true low spatial correlation which is consistent with the observation that site measurements are measured far apart (Figure 1), or it is also possible that there is true spatial correlation but our model failed to represent it. This could be reflected in the fact that ϕ did not pass all of our goodness of fit assessments. Further work can be done to better pick priors for ϕ and τ^2 so that ϕ may converge more strongly.

This study suggests that a possibly high concentration of lead is being dumped into the Meuse river from some source such as waste from nearby industrial plants. The Meuse river may be carrying the polluted sediment which eventually leads to lead deposits in the soil close to the river bank. The actual source of this polluted sediment that is dumped into the river should be further investigated because lead is extremely

harmful to human and animal populations of nearby regions. Planes near the river are usually fertile and can be used to grow plants and crops. Lead deposition in the river and eventually in the nearby fields, could have lasting detrimental effects on the mental and physical health of the people living in the close-by towns.