Arsenic in Public Water Systems – A Bayesian Approach

Crystal Dong
Background

- STAR Grant Project
- Source to Biomarker (STB)
- First stage: Source to Aerial
- The goal:
  - a map of metal concentration
  - at the scale of county (or census track).
  - soil, water, air, and food
  - feed into later stages
Science

- What is Arsenic?
- Natural?
- Harmful?
- EPA rule: 50 µg/L to 10 µg/L, 2006
- Ohio EPA insight: connection with iron
Getting Data

- PWS – public water system
- Why only Ohio
- Why only Franklin county
- Why is the map so scarce
Choose variables

- # connection highly correlated with population
- Source (GW, SW, PSW, PGW)
- Iron level
- Others?
Getting values for non-detects

- Quantile Method
  - Assume normal
  - Fit straight line

```r
MDL<-function(n1, n2, y2){
i<-seq(1,n1+n2)
z<-qnorm((i-.5)/(n1+n2))
line.fit<-lm(y2~z[(n1+1):(n1+n2)])
mu.hat<-line.fit$coeff[1]
sigma.hat<-line.fit$coeff[2]
y1<-mu.hat+sigma.hat*z[1:n1]
y1
}
```
Input Data

- Greater Franklin County
- 12 out of 48 missing
- 43015 outlier
- Population
- Source
- Iron level
Model Specification

- \( \text{As} | \mu, s^2, S \sim \text{MVN}(\mu, s^2 S) \) \hspace{1cm} (1)
- \( \text{E}(\mu[i] | a_0, a_1, a_2, a_3) = a_0 + a_1 \text{population}[i] + a_2 \text{Fe}[i] + a_3 \text{source}[i] \) \hspace{1cm} (2)
- \( a_0 \sim N(0.0,1.0E-6) \) \hspace{1cm} (3)
- \( a_1 \sim N(0.0,1.0E-6) \)
- \( a_2 \sim N(0.0,1.0E-6) \)
- \( a_3 \sim N(0.0,1.0E-6) \)
- \( t \sim \text{Gamma}(0.001, 0.001) \)
- \( s^2 = 1/ t \)
- \( f \sim U(0.001, 0.8) \)
- \( ? \sim U(0.05,1.95) \)
Spatial Part

- Between-area correlation matrix:
- \( S_{ij} | ? = f(d_{ij}; ?) \)
  - where \( d_{ij} = \) distance between area \( i \) and \( j \).

- powered exponential family
  - \( f(d_{ij}; f, ?) = \exp[-(f d_{ij})^? ] \) where \( f > 0 \) and \( ? \in (0, 2] \).
  - The larger \( f \) is, the more rapid the rate of decline of correlation with distance. The parameter \( ? \) controls the amount by which spatial variations in the data are smoothed. Large values of \( ? \) lead to greater smoothing.
WinBUGS
## MCMC results

<table>
<thead>
<tr>
<th>node</th>
<th>mean</th>
<th>sd</th>
<th>MC error</th>
<th>2.50%</th>
<th>median</th>
<th>97.50%</th>
<th>start</th>
<th>sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_0$</td>
<td>-2.297</td>
<td>0.2197</td>
<td>0.00288</td>
<td>-2.725</td>
<td>-2.298</td>
<td>-1.841</td>
<td>501</td>
<td>7500</td>
</tr>
<tr>
<td>$a_1$</td>
<td>-8.2E-07</td>
<td>6.2E-06</td>
<td>9.9E-08</td>
<td>-1.3E-05</td>
<td>-7.3E-07</td>
<td>1.1E-05</td>
<td>501</td>
<td>7500</td>
</tr>
<tr>
<td>$a_2$</td>
<td>0.8463</td>
<td>0.0173</td>
<td>2.73E-04</td>
<td>0.8125</td>
<td>0.846</td>
<td>0.881</td>
<td>501</td>
<td>7500</td>
</tr>
<tr>
<td>$a_3$</td>
<td>0.005848</td>
<td>0.03921</td>
<td>7.15E-04</td>
<td>-0.06941</td>
<td>0.004609</td>
<td>0.08326</td>
<td>501</td>
<td>7500</td>
</tr>
<tr>
<td>$?_0$</td>
<td>0.3791</td>
<td>0.1829</td>
<td>0.005395</td>
<td>0.0789</td>
<td>0.365</td>
<td>0.7787</td>
<td>501</td>
<td>7500</td>
</tr>
<tr>
<td>$f$</td>
<td>0.4281</td>
<td>0.2113</td>
<td>0.008089</td>
<td>0.06349</td>
<td>0.4227</td>
<td>0.7825</td>
<td>501</td>
<td>7500</td>
</tr>
<tr>
<td>$s^2$</td>
<td>0.05601</td>
<td>0.08785</td>
<td>0.003591</td>
<td>0.01043</td>
<td>0.03269</td>
<td>0.2527</td>
<td>501</td>
<td>7500</td>
</tr>
</tbody>
</table>
Future work

- C program
- Link between raw water and treated water
- Log-transformed normal assumption
- Iron dominates
  - Population?
  - Source?
  - Others?
- Re-examine data pre-processing