Arsenic in Public Water Systems – A Bayesian Approach

Crystal Dong

Background

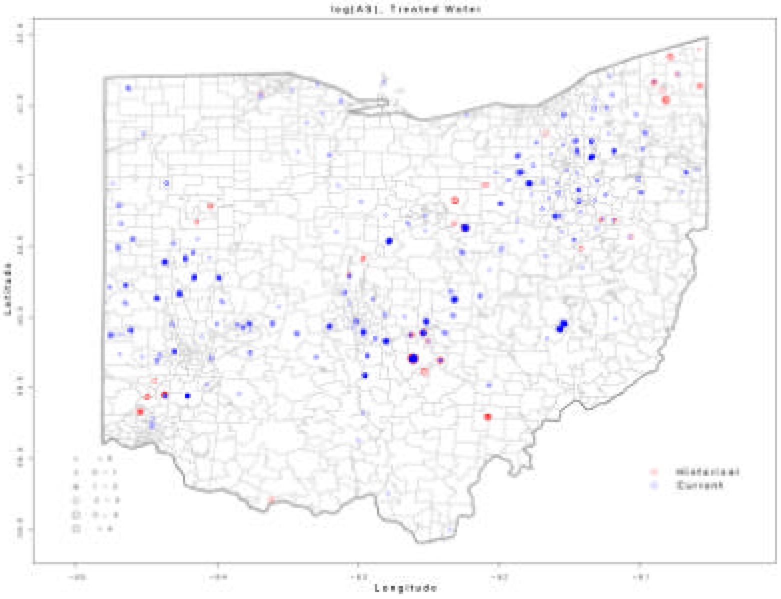
- STAR Grant Project
- Source to Biomarker (STB)
- First stage: Source to Aerial
- The goal:
  - a map of metal concentration
  - at the scale of county (or census track).
  - soil, water, air, and food
  - feed into later stages

Science

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

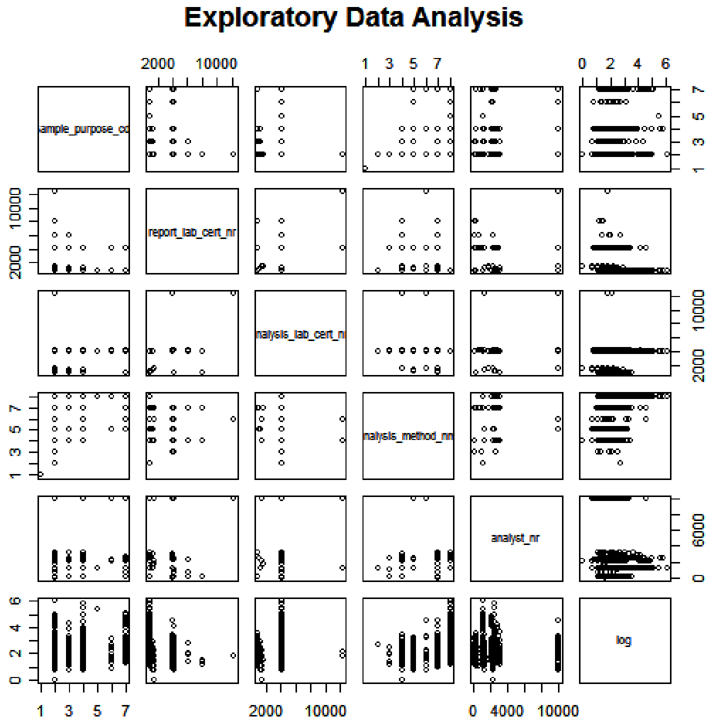
Getting Data

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



Choose variables

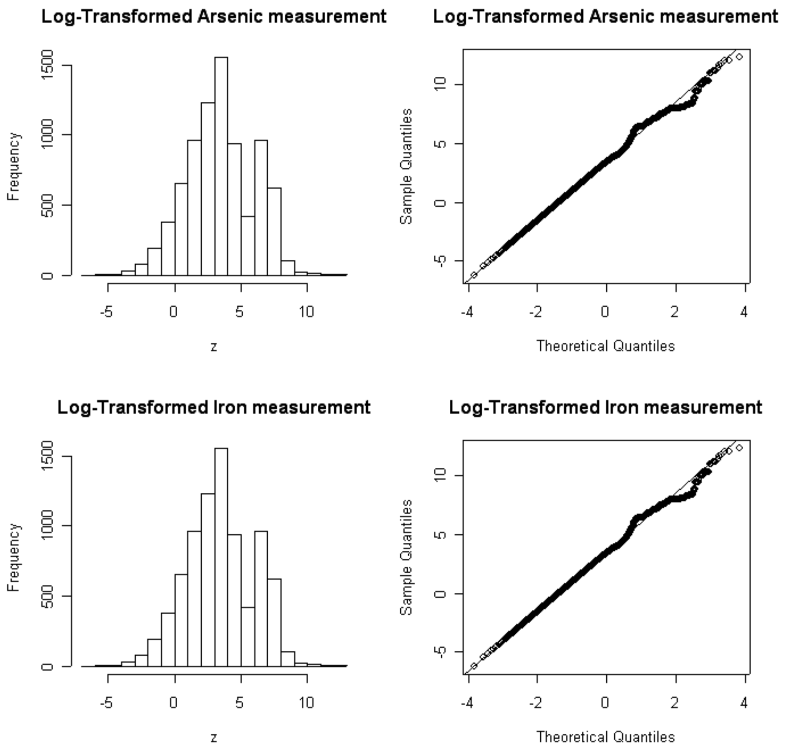
- # connection highly correlated with population
- Source (GW, SW, PSW, PGW)
- Iron level
- Others?



Getting values for non-detects

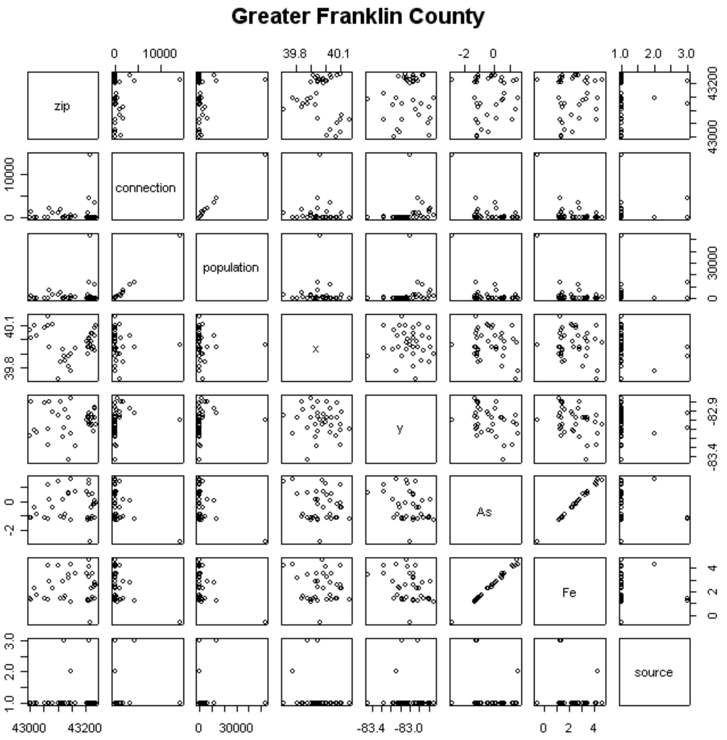
- Quantile Method
  - Assume normal
  - Fit straight line

```
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

- **As | μ**, s2, **S** ~ MVN(**μ**, s2 **S**) (1)
- E(μ[i] | a0, a1, a2, a3) = a0 + a1 population[i] + a2 Fe[i] + a3 source[i] (2)
- a0~ N(0.0,1.0E-6) (3)

  a1~ N(0.0,1.0E-6)

  a2~ N(0.0,1.0E-6)

  a3~ N(0.0,1.0E-6)

  t ~ Gamma(0.001, 0.001)

  s2 = 1/ t

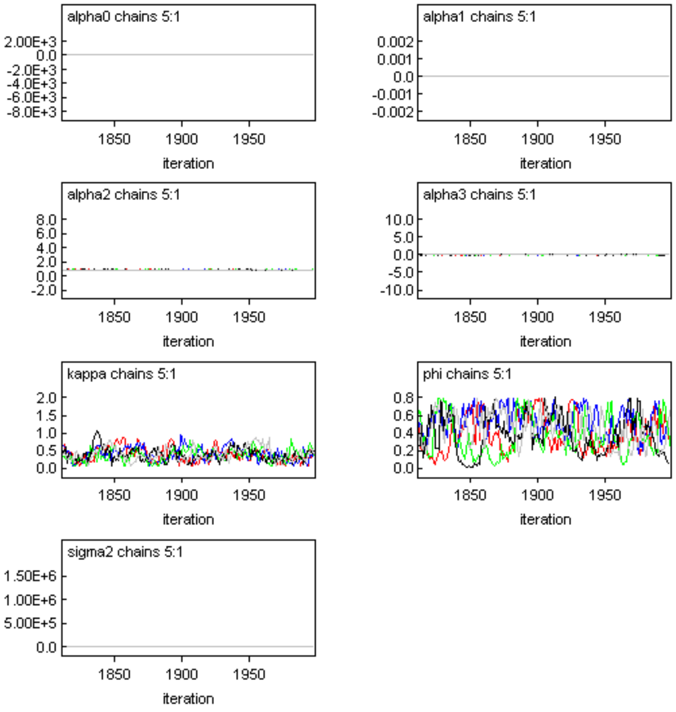
  f ~ U(0.001, 0.8)

  ? ~ U(0.05,1.95)

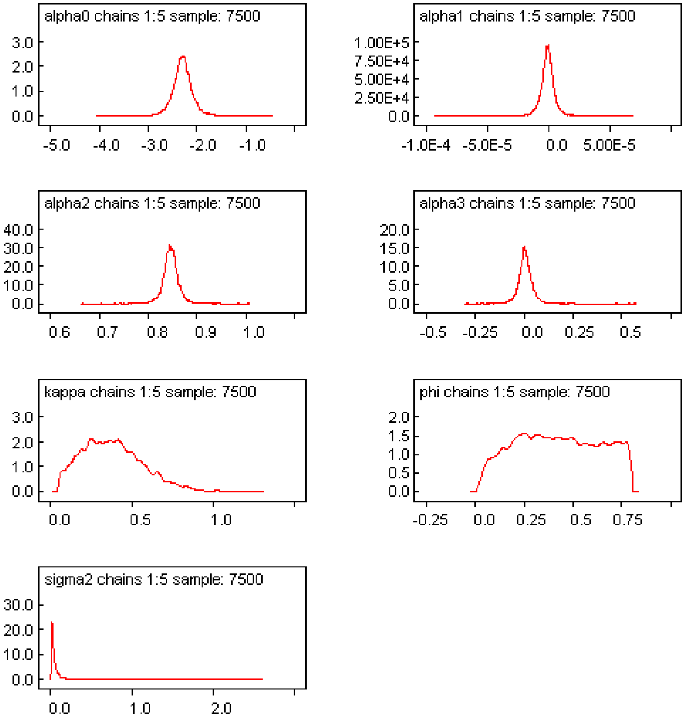
Spatial Part

- 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 re smoothed. Large values of ? lead to greater smoothing.

WinBUGS



alpha0 chains 5:1
2.00E+3
0.0
-2.0E+3
-4.0E+3
-6.0E+3
-8.0E+3
1850
1900
1950
iteration
alpha1 chains 5:1
0.002
0.001
0.0
-0.001
-0.002
1850
1900
1950
iteration
alpha2 chains 5:1
8.0
6.0
4.0
2.0
0.0
-2.0
1850
1900
1950
iteration
alpha3 chains 5:1
10.0
5.0
0.0
-5.0
-10.0
1850
1900
1950
iteration
kappa chains 5:1
2.0
1.5
1.0
0.5
0.0
1850
1900
1950
iteration
phi chains 5:1
0.8
0.6
0.4
0.2
0.0
1850
1900
1950
iteration
sigma2 chains 5:1
1.50E+6
1.00E+6
5.00E+5
0.0
1850
1900
1950
iteration



alpha0 chains 1:5 sample: 7500
3.0
2.0
1.0
0.0
-5.0
-4.0
-3.0
-2.0
-1.0
alpha1 chains 1:5 sample: 7500
1.00E+5
7.50E+4
5.00E+4
2.50E+4
0.0
-1.0E-4
-5.0E-5
0.0
5.00E-5
alpha2 chains 1:5 sample: 7500
40.0
30.0
20.0
10.0
0.0
0.6
0.7
0.8
0.9
1.0
alpha3 chains 1:5 sample: 7500
20.0
15.0
10.0
5.0
0.0
-0.5
-0.25
0.0
0.25
0.5
kappa chains 1:5 sample: 7500
3.0
2.0
1.0
0.0
0.0
0.5
1.0
phi chains 1:5 sample: 7500
2.0
1.5
1.0
0.5
0.0
-0.25
0.0
0.25
0.5
0.75
sigma2 chains 1:5 sample: 7500
30.0
20.0
10.0
0.0
0.0
1.0
2.0

MCMC results

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

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