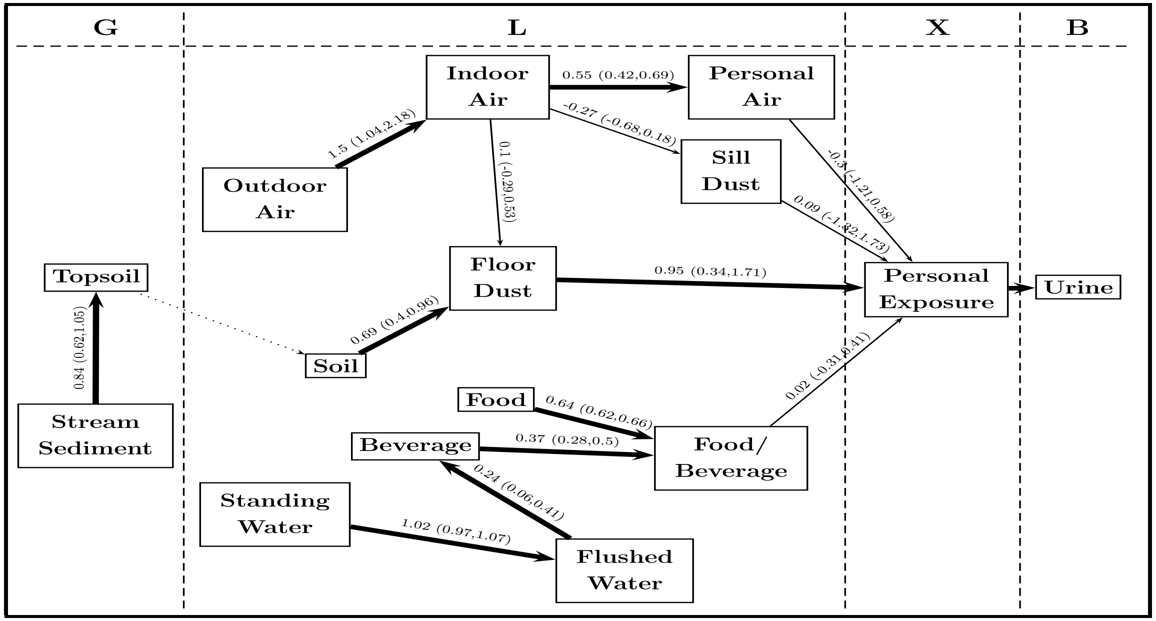
Tutorial on Hierarchical Bayesian Modeling for Exposure to Arsenic

The preparation of this webpage was supported by the American Chemistry Council's (ACC) Long-Range Research Initiative under an agreement (2004-2007) between the ACC and the Ohio State University Research Foundation.

Purpose

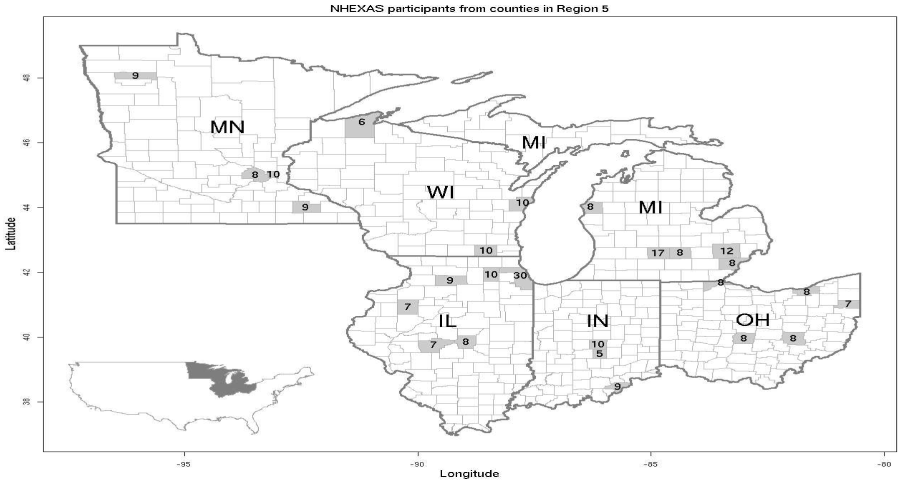
A multi-scale (individual level, county level) **hierarchical Bayesian model (HBM)** was used to investigate the pathways of human exposure to Arsenic.

The HBM has explicit stages for pollutant sources, global and local environmental levels, personal exposures, and biomarkers. Analysis is on the **log** scale.



Purpose, ctd.

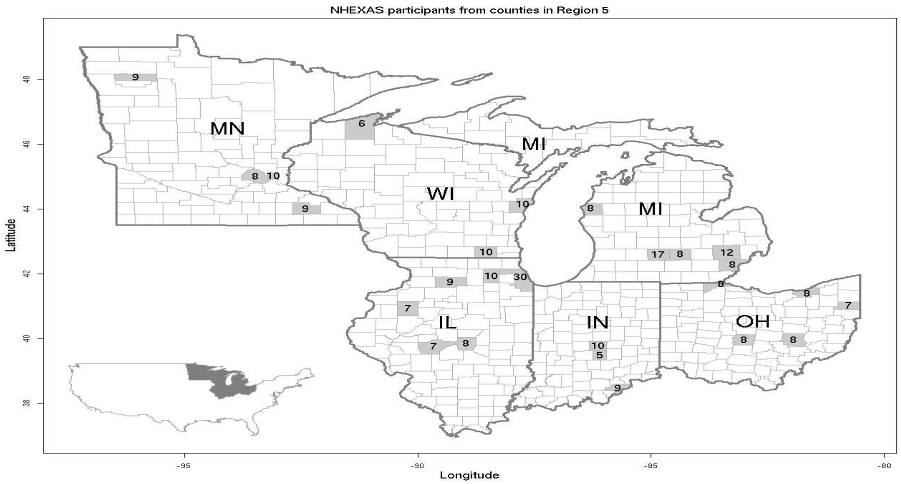
The focus is on **Arsenic** and its air, soil, water, and food pathways of exposure for individuals in the U.S. Environmental Protection Agency’s Region 5 (**Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin**):



The metal Arsenic (from among Arsenic, Lead, Chromium, and Cadmium) was chosen for a prototype HBM because the U.S. Environmental Protection Agency (EPA) is interested in revisiting environmental standards for levels of Arsenic (As).

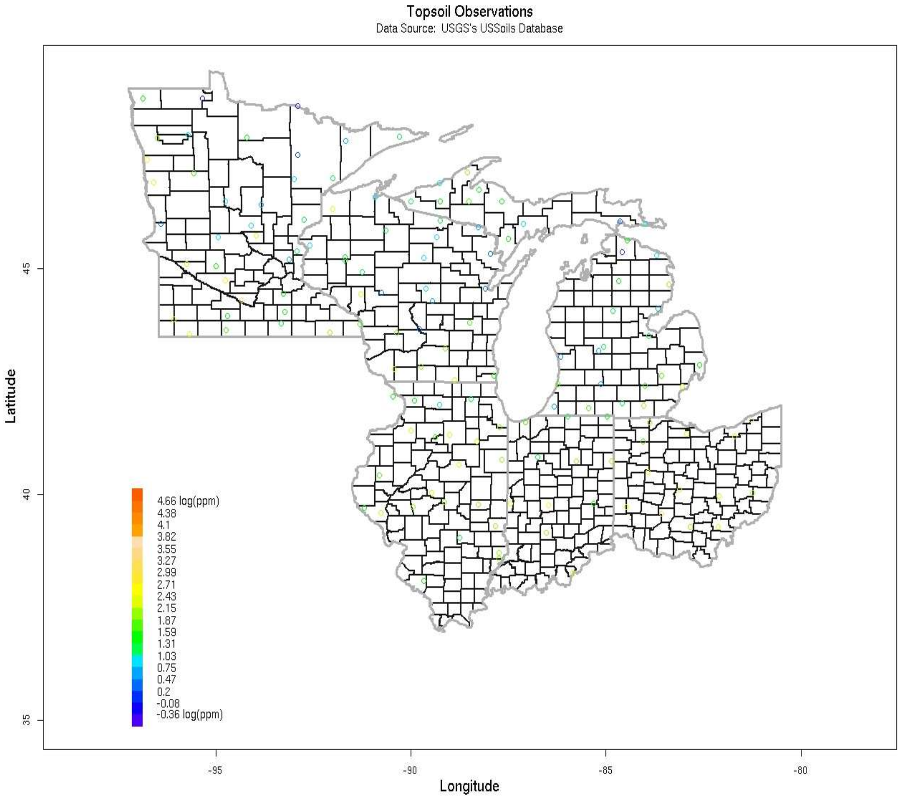
Data

- Primary source of data was **NHEXAS** (National Human Exposure Assessment Survey) - a U.S. EPA, CDC, FDA study for surveillance of exposure to toxic agents - conducted 1995-1998 (supplemental NHEXAS demographic information from questionnaires is not included in prototype models but is planned for use in future models). The figure shows county of residence for the 249 NHEXAS participants sampled. Residential exposure and biomarker data (blood and urine) were collected from participants.



Data, ctd.

- USDA-NRCS Soil Geochemistry Spatial Database: Arsenic concentration measurements in topsoil - background contaminant (**global-environment level**)

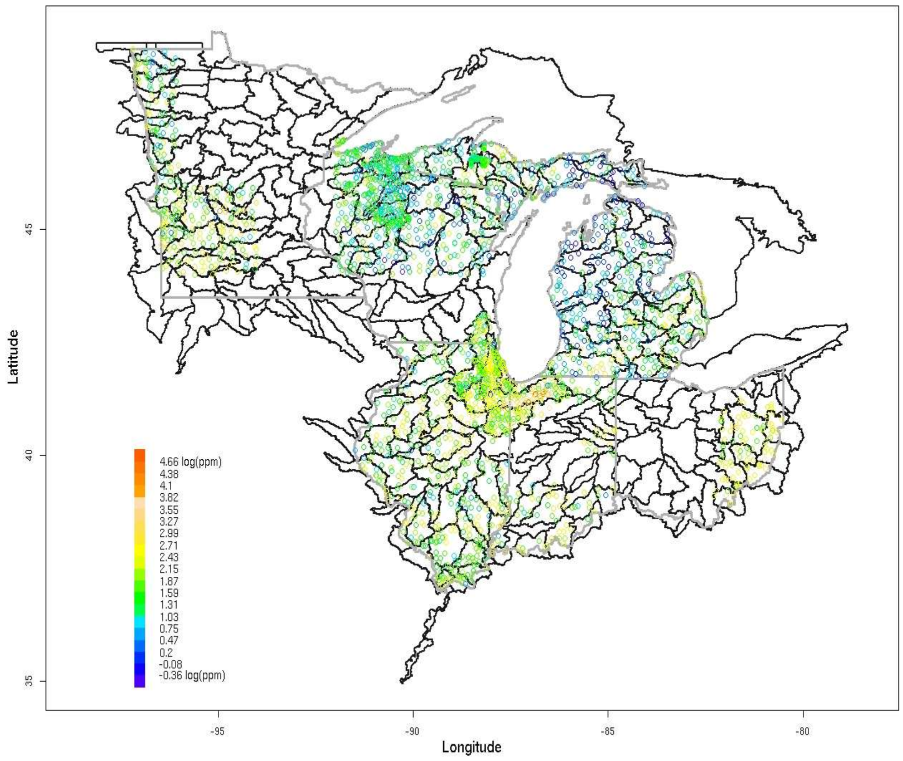


Data, ctd.

- USGS National Geochemical Survey: Arsenic concentration measurements in stream sediments - background contaminant (**global-environment level**)

Stream-Sediment Observations

Data Source: USGS's National Geochemical Survey (NGS) Database



Hierarchical Bayesian Models (HBMs)

- We define the statistical model using a series of hierarchical steps:

  $$\text{Data Model} \longleftarrow \text{Process Model} \longleftarrow \text{Parameter Model}$$

- HBMs provide a mechanism to **merge science with data**.
- We use Bayes' Theorem to derive the posterior distribution of "the processes and parameters" given "the data".
- Posterior distribution is usually **not available in closed form**.

A Standard Model Structure

- Generic notation: $\mathbf{Y} \equiv$ data, $\mathbf{X} \equiv$ process, $\boldsymbol{\theta} \equiv$ parameter.
- Let $[A]$ denote the joint distribution for the random quantity A, and $[A|B]$ denote the conditional distribution for A conditional on B.
- We specify our Bayesian model in steps:

1. Data model: $[\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}_1]$.
2. Process model: $[\mathbf{X}|\boldsymbol{\theta}_2]$.
3. Parameter model (prior): $[\boldsymbol{\theta}]$, where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$.

- We then calculate the posterior distribution:

$$[\boldsymbol{\theta}, \mathbf{X}|\mathbf{Y}] \quad \propto \quad [\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}_1] \times [\mathbf{X}|\boldsymbol{\theta}_2] \times [\boldsymbol{\theta}].$$

In practice, the **proportionality constant** can be difficult to obtain. (We know the posterior has total mass $= 1$.)

Data Models

- Three types of data:
  - observed (above detection limit)
  - left-censored (below detection limit)
  - unobserved (missing)
- Illustrate with biomarker data:
  - $\mathcal{S}^{B}_{\mathcal{A}} \equiv$ NHEXAS participants with biomarker level $>$ **MDL**
  - $\mathcal{S}^{B}_{\mathcal{B}} \equiv$ NHEXAS participants with biomarker level $\leq$ **MDL**
  - $\mathcal{S}^{B}_{\mathcal{M}} \equiv$ NHEXAS participants with biomarker level **missing**

Data Models, ctd.

- Available biomarker data:

$$\mathbf{Y}^B \equiv \{Y_i^B : i \in \mathcal{S}_{\mathcal{A}}^B \cup \mathcal{S}_{\mathcal{B}}^B\},$$

  where if data are censored, define $Y_i^B \equiv M_i^B$, the MDL for $i \in \mathcal{S}_{\mathcal{B}}^B$.

- Assume $Y_i^B \sim N(B_i, 1/\omega^B)$; $i \in \mathcal{S}_{\mathcal{A}}^B$.
- Both censored and missing data are handled coherently in the Bayesian formulation. That is, $\{Y_i^{*B} : i \in \mathcal{S}_{\mathcal{B}}^B \cup \mathcal{S}_{\mathcal{M}}^B\}$ can be imputed.

Process Models

- **Biomarker given Personal Exposure**

$$[\mathbf{B}|\mathbf{X}, \mu^B, \beta^B, \tau^B] = \prod_{i=1}^{N^I} [B_i | X_i, \mu^B, \beta^B, \tau^B] \,,$$

where N^I is the number of NHEXAS individuals ($N^I = 249$), and

$$[B_i | X_i, \mu^B, \beta^B, \tau^B] \sim N\left(\mu^B + \sum_{j=1}^{N^X} \beta_j^B X_{ij}, 1/\tau_B\right) \,,$$

for N^X the number of personal exposure routes.

Process Models, ctd.

- **Personal Exposure given Local Environment**

$$[\mathbf{X}|\mathbf{L},\mu^X,\beta^X,\tau^X]=\prod_{i=1}^{N^I}\prod_{j=1}^{N^X}[X_{ij}|\{L_{ik}\}_{k\in\mathcal{S}^{X_j}},\mu_j^X,\beta_j^X,\tau_j^X]\,,$$

where $\mathcal{S}^{X_j}\subset\{1,\ldots,N^L\}$ is the selector set of local variables linked **causally** to the j-th personal exposure, and

$$[X_{ij}|\{L_{ik}\}_{k\in\mathcal{S}^{X_j}},\mu_j^X,\beta_j^X,\tau_j^X]\sim N\left(\mu_j^X+\sum_{k\in\mathcal{S}^{X_j}}\beta_{jk}^X L_{ik},1/\tau_j^X\right)\;.$$

Process Models, ctd.

- **Local Environment given Global Environment**

$$[\mathbf{L}|\mathbf{G}, \beta^L, \tau^L] = \prod_{i=1}^{N^I} \prod_{j=1}^{N^L} [L_{ij}|\{L_{ik}\}_{k \in \mathcal{S}^{L_j}}, G_j, \beta_j^L, \tau_j^L],$$

  where $\mathcal{S}^{L_j} \subset \{1, \ldots, N^L\}$ is the selector set describing **causal relations** within local variables, and

$$[L_{ij}|\{L_{ik}\}_{k \in \mathcal{S}^{L_j}}, G_j, \beta_j^L, \tau_j^L] \sim N(G_{c(i),j} + \sum_{k \in \mathcal{S}^{L_j}} \beta_{jk}^L L_{ik}, 1/\tau_j^L),$$

  for $c(i)$ the county in which the i-th individuals resides.

- **Global Environment**
  $[G_T|\mu^{G_T}, \beta^{G_T}, \tau^{G_T}]$ has independent components, where "T" stands for "topsoil"
  $[G_S|\mu^{G_S}, \Sigma^{G_S}]$ is a spatial CAR model based on watersheds, where "S" stands for "stream sediments".

Parameter Models

- Parameters are assumed mutually **independent**.
- Mean parameters and regression parameters are assumed **Gaussian** distributed.
- Precision parameters (1/variance) are assumed **gamma** distributed.
- Spatial CAR dependence parameter assumed **uniform** on its parameter space, $(1/\lambda_{\min}, 1/\lambda_{\max})$, where $\{\lambda_i\}$ are the eigenvalues of the CAR spatial-dependence matrix that has zeros down the diagonal.

Posterior Inference - MCMC based

- Posterior for the exposure-pathways parameters (that relate biomarker concentrations of toxic metals to local and global environmental exposures) is **not available in closed form**.
- Instead, we sample from the posterior distribution. Two common approaches are **MCMC** based:
  1. Gibbs sampling.
  2. Metropolis-Hastings algorithm.
- Inference is based on MCMC samples of parameter values from the posterior distribution.
  - We need to derive algorithms to sample “$\mathbf{X}$” and “$\boldsymbol{\theta}$”, conditional on the data “$\mathbf{Y}$”.
  - Then we need to code and implement the algorithms.

Gibbs Sampling

- Suppose we can decompose $\mathbf{X}$ into J components, and $\boldsymbol{\theta}$ into K components:

$$\begin{aligned}\mathbf{X} &= (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_J) \\ \boldsymbol{\theta} &= (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K).\end{aligned}$$

- Letting, for example, $\mathbf{X}_{-j}$ denote the $J - 1$ components of $\mathbf{X}$ **without** $\mathbf{X}_j$, we calculate the following conditional posterior distributions:

$$\begin{aligned}&[\mathbf{X}_j | \mathbf{Y}, \mathbf{X}_{-j}, \boldsymbol{\theta}]; \qquad j = 1, \ldots, J \\ &[\boldsymbol{\theta}_k | \mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}_{-k}]; \qquad k = 1, \ldots, K\,.\end{aligned}$$

- In Gibbs sampling, we **repeatedly sample** parameters from these conditional distributions (the sample step).
- We **always accept** these sampled values (i.e., we accept with probability 1).
- The samples converge in distribution to the posterior distribution (under certain regularity conditions).

Metropolis-Hastings (M-H)

(Simplest case: symmetric proposal distributions)

- As an example, consider $\boldsymbol{\theta}$ (it is similar for $\mathbf{X}$).
- Suppose our current parameter value is $\boldsymbol{\theta}^c$. In the sample step:
  - For some $j = 1, \ldots, J$, we sample a proposal $\boldsymbol{\theta}_j^*$ from a proposal distribution (e.g., a normal distribution).
- We accept the proposal $\boldsymbol{\theta}_j^*$ with probability equal to

$$\alpha \quad = \quad \min\left\{ \frac{[\boldsymbol{\theta}_j^* | \mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}_{-j}]}{[\boldsymbol{\theta}_j^c | \mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}_{-j}]}, \ 1 \right\} ;$$

  otherwise we stay at the current $\boldsymbol{\theta}_j^c$ (i.e., we **accept with probability $\boldsymbol{\alpha}$**).
- Repeat a large number of times.
  - The samples converge in distribution to the posterior distribution (again under certain regularity conditions).

Notes on M-H Algorithm

- This is a rejection-sampling algorithm.
- There are other versions of M-H that may have better convergence properties.
- We can think of Gibbs sampling as a special case of M-H with an acceptance probability of 1.
- Advantages:
  - We do not need to calculate full posterior or conditional posterior distributions.
  - In particular, we do not need to know the proportionality constant, $1/[\mathbf{Y}]$.
- Disadvantages:
  - Choice of proposal distributions and algorithm always requires careful thought and tuning.

Model Fitting

- Correct implementation of MCMC is a non-trivial task.
- Our strategy:

  1. A **model definition document** is created before the model is implemented in code.
     - The document defines all the notation and the sampling steps.
     - People (i.e., more than one!) check the document for errors.
  2. All code is stored in a **Subversion** (http://subversion.tigris.org/) archive.
     - This software keeps track of multiple versions (e.g., allows the user to “go back”).
     - It simplifies project management involving multiple users.
  3. Any **utility code** that will be needed (e.g., random number generators, functions to calculate the Cholesky decomposition of a matrix) are stored in a separate directory.
     - The utility code is checked independently before being used in the MCMC.

Model Fitting, ctd.

- Our strategy, ctd.:

  4. The routines to read in the data are coded up and checked before we implement and run the MCMC.
  5. We successively add parameters into the model.
     - Variables that are unsampled, are fixed.
     - We **check** that the output is correct **with simulated data**, before running on actual data.

- The same strategy applies when we make changes to the model.
  - We try to compare the output from previous and updated versions of the code.

Diagnostics

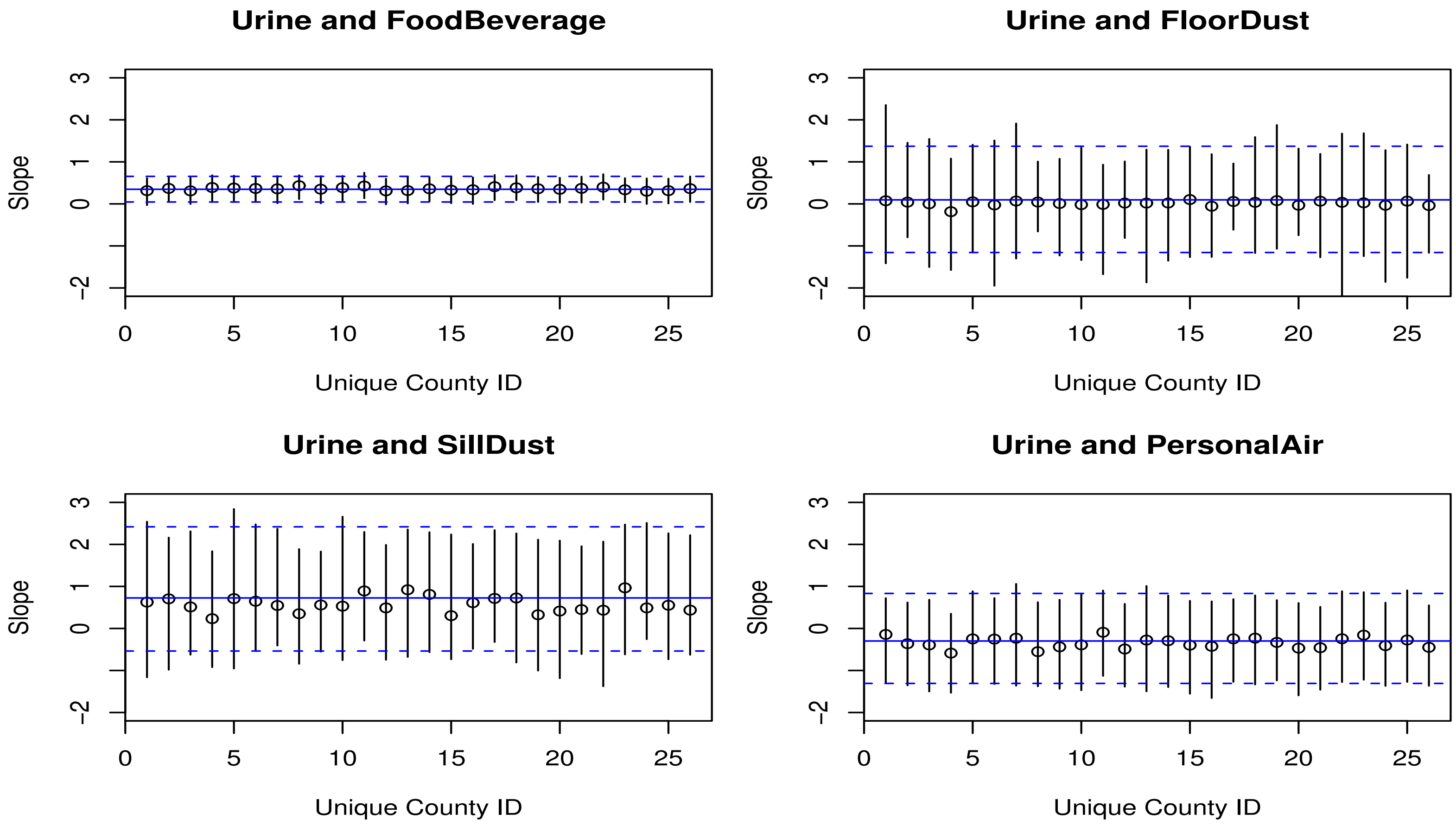
- Although the MCMC library is helpful, in practice one needs to investigate:
  - That the sampling and acceptance steps are **implemented correctly**.
  - Whether the model **mixes well**.
  - The **sensitivity** of the results to choice of **prior**.
  - The **sensitivity** of the results to the **data**.
- All these issues are important for our exposure-pathways models, because:
  1. There are a large number of pathways (processes).
  2. We incorporate different data sources.
     - Each dataset can have issues with **missingness** and **censoring** (minimum detection limits).
  3. The data have different spatial scales.
     - There can be a lack of spatial coverage or resolution.

Sensitivity to the Data

- Given that the data sources vary in quality and scope, we re-run the model to investigate **sensitivity** of the results to subsets of the **data**.
  Examples:
  - What is the effect of leaving out the global soil model, upon the exposure-pathways relationships?
  - How sensitive are the results to specific subsets of the NHEXAS data?

Example: Sensitivity To Leaving Out Cases in a County

(Arsenic pathways model fit to EPA Region 5)



Sensitivity to Choice of Prior

- Prior choice is important in our model.
- Examples:
  - Certain layers of the model (e.g., the personal-exposure level) contain no data – prior choice is critical here to ensure identifiability.
  - The prior on the spatial-dependence parameter in the CAR model for soil needs to ensure that the spatial covariance matrix is positive-definite.
- To understand how the prior affects the results, we re-run the model under different choices.

Bayesian Computation: Summary

- Large Bayesian hierarchical models need to be implemented carefully.
    - The model should be coded up and tested in bite-size pieces.
    - Since it can be hard to get the MCMC chains to mix, the choice of prior and parameter updates are critical.
- Datasets need to be managed, in a documented fashion.
- Sensitivity of the results to the data is often overlooked, but it is important.
    - For example, in the exposure pathways models that relate the Arsenic levels measured in urine, to local and global environmental measurements, there are no data in a county that heavily influences the model fit.
- Fast, reliable, MCMC computation with an extendable, modifiable library has been crucial.

From Sources to Biomarkers

- We have chosen to analyze one metal (Arsenic) in one biomarker (urine) for the six states that make up the U.S. Environmental Protection Agency's Region 5 (Illinois, Indiana, Michigan, Minnesota, Ohio, and Wisconsin). The analysis is on the **log** scale.

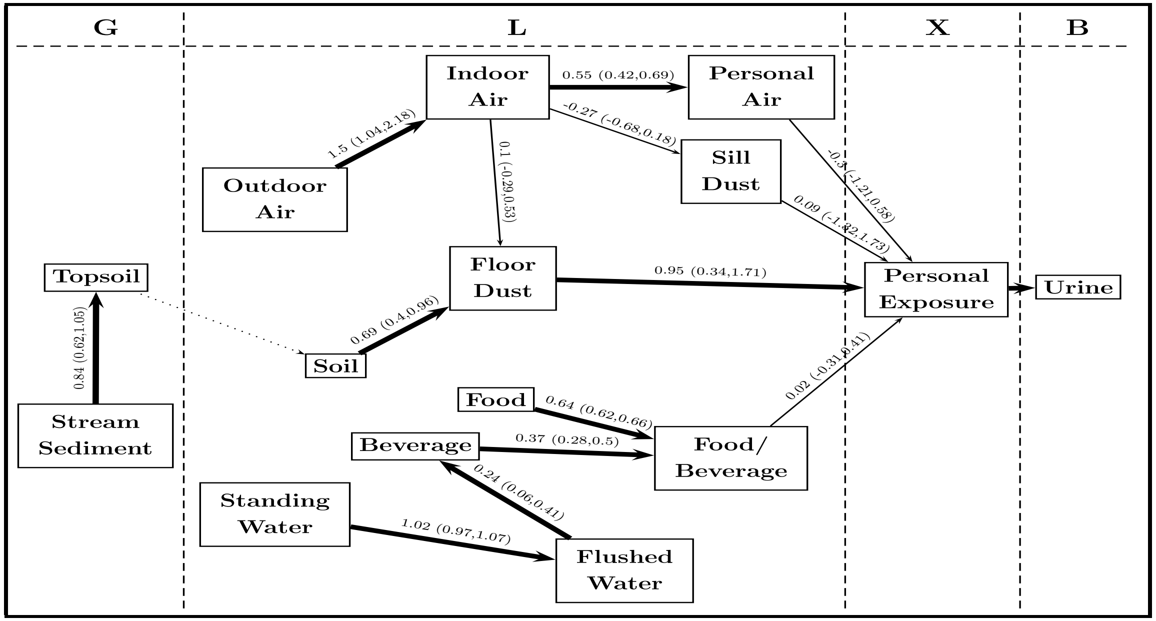
- Markov Chain Monte Carlo (**MCMC**) simulation methods were used to fit a HBM - the Gibbs sampler successively samples from the full conditional distributions; where necessary, the Metropolis-Hastings algorithm is used.

- Key parameters of interest are the **slope parameters** in the regression models, which link, for example, biomarkers to personal exposure, personal exposure to local environmental exposures, and local exposures to global exposures.

- Slope parameters are declared to **exhibit causality** if 95% posterior credible intervals **do not contain 0**.

From Sources to Biomarkers, ctd.

The figure shows a prototype sources-to-biomarkers (PSTB) model involving global variables G (sources such as soil, water, etc.), local variables L (personal environment), personal-exposure variables X (inhalation, ingestion, etc.), and biomarker variables B (in urine and blood) linked in a causal manner that can be visualized as an acyclic directed graph.

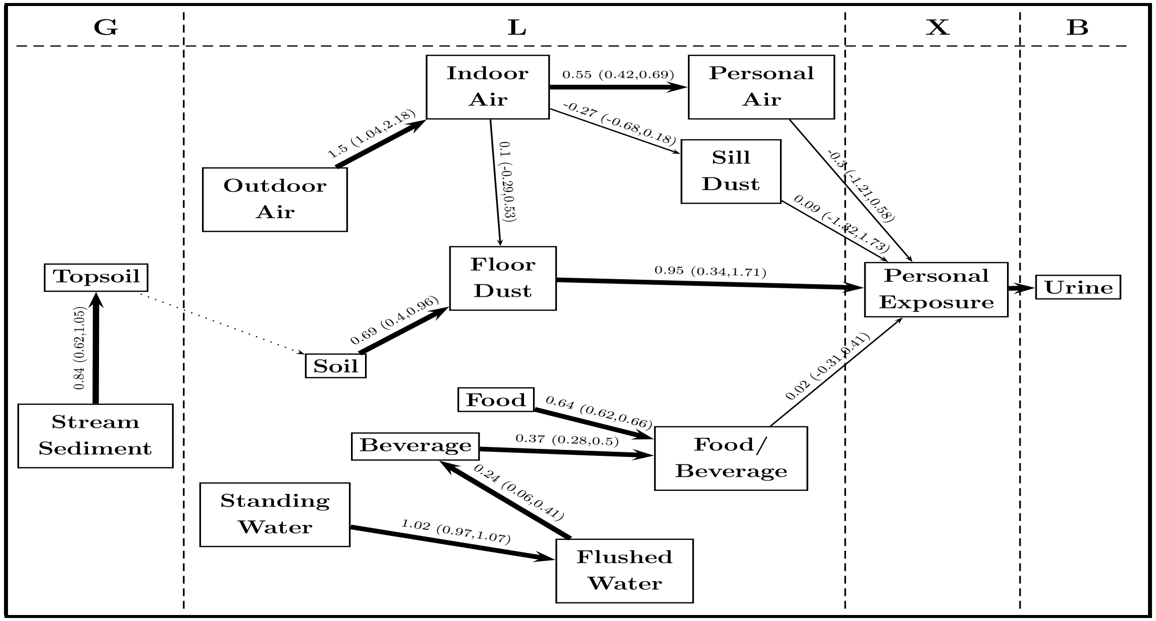


From Sources to Biomarkers, ctd.

For our PSTB model of exposure to Arsenic, we have one biomarker variable (urine); one latent personal-exposure variable; 11 local variables (outdoor air, indoor air, personal air, floor dust, soil, sill dust, standing water, flushed water, beverage, food, and food/beverage); and two global variables (topsoil in counties and stream sediments in watersheds).

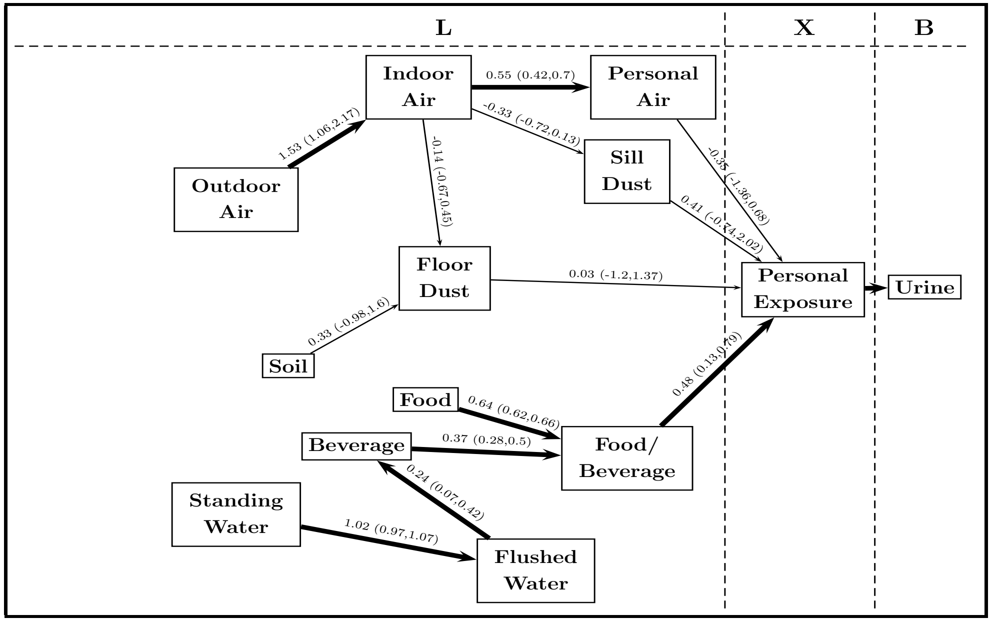
Results

Slope parameters represent **causative effects** in the model, so a value different from zero says something substantial about how the data in our study support the science from which we constructed the pathways. **The soil-to-floor-dust-to-personal-exposure route exhibits strong positive dependence.**



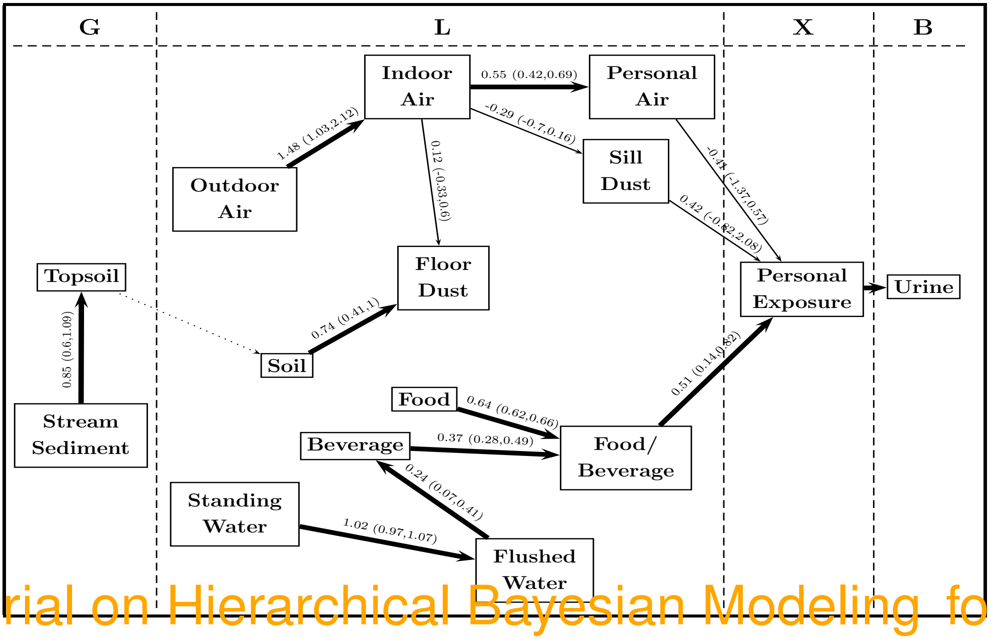
Results, ctd.

Now **take out the global-soil component** (i.e., local-environment components only). Then the soil-to-floor-dust-to-personal-exposure route exhibits only weak positive dependence, and the slope parameter for **the food/beverage-to-personal-exposure route exhibits strong positive dependence**. This is in line with the findings of Clayton et al. (2002), who did not use global-soil data (nor a HBM).



Results, ctd.

Now **leave in the global-soil component**, but **take out the floor-dust-to-personal-exposure route**. Once again, the food-beverage-to-personal-exposure route exhibits strong positive dependence. That is, when the global-soil and local-soil data are not able to inform about personal exposure, the food/beverage-to-personal-exposure route dominates. Lesson: Look for **partial** associations.



Conclusions

- Using Bayesian inference, we conclude from the data and the PSTB model that the **route from floor dust to personal exposure** is the strongest. In the presence of information about the source of Arsenic in soil, the route from food/beverage to personal exposure is considered to be weaker than that from floor dust.
- The findings of this research are to appear in 2007 in the *Journal of Statistical Planning and Inference*. For a preprint, see: **http://www.stat.osu.edu/∼sses/papers.html#2005**
- The HBM approach developed in this article is computationally intensive but shows considerable flexibility in modeling human exposure.