CBMS Lecture 5

Alan E. Gelfand
Duke University

Univariate point-level modeling

- Basic Model:

$$Y(\mathbf{s}) = \mathbf{x}^T(\mathbf{s})\beta + w(\mathbf{s}) + \epsilon(\mathbf{s})$$

  The residual is partitioned into two pieces: one spatial, $w(\mathbf{s})$, and one non-spatial, $\epsilon(\mathbf{s})$. $w(\mathbf{s})$ is a stationary Gaussian process, introducing the partial sill (σ^2) and range (ϕ) parameters. $\epsilon(\mathbf{s})$ adds the nugget (τ^2) effect.

- Interpretations attached to $\epsilon(\mathbf{s})$:
  - pure error term; model is not perfectly spatial; τ^2, σ^2 are **variance components**;
  - measurement error or replication variability causing discontinuity in spatial surface $Y(\mathbf{s})$;
  - microscale uncertainty; distances smaller than the smallest inter-location distance, indep assumed.

More

- Suppose we have data $Y(\mathbf{s}_i), i = 1, \ldots, n$, and let $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^T$.
- Gaussian kriging models are special cases of of the general linear model, with a particular specification of the dispersion matrix

$$\Sigma = \sigma^2 H(\phi) + \tau^2 I.$$

- $H_{ij} = \rho(\mathbf{s}_i - \mathbf{s}_j; \phi)$, where ρ is a valid (and typically isotropic) correlation function.
- Setting $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \tau^2, \phi)^T$ (not a high dim problem), we require a prior $p(\boldsymbol{\theta})$, so the posterior is:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$$

Likelihood and priors

- The likelihood is given by:

$$\mathbf{Y}|\boldsymbol{\theta} \sim N(X\boldsymbol{\beta}, \sigma^2 H(\phi) + \tau^2 I)$$

- Typically, independent priors are chosen for the parameters:

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\beta})p(\sigma^2)p(\tau^2)p(\phi)$$

  Useful candidates are red multivariate normal for $\boldsymbol{\beta}$, and inverse gamma for σ^2 and τ^2.
- Specification of $p(\phi)$ depends upon choice of ρ function; a uniform or discrete prior is usually selected.

Priors cont.

- Informativeness: $p(\boldsymbol{\beta})$ can be "flat" (improper)
- Without nugget, τ^2, can't *identify* both σ^2 and ϕ (Zhang, 2004). With Matérn, can identify the product, $\sigma^2\phi^{2\nu}$.
- So an informative prior on at least one of these parameters
- With τ^2, ϕ and at least one of σ^2 and τ^2 require informative priors.
- If the prior on $\boldsymbol{\beta}, \sigma^2, \phi$ is of the form $\frac{\pi(\phi)}{\sigma^2}^{a+1}$ with $\pi(\cdot)$ proper, then, improper posterior if $a = 0$
- Shows the problem with using $IG(\epsilon, \epsilon)$ priors for σ^2 - $a + 1 = 1 + \epsilon$, "nearly" improper. Safer is $IG(a, b)$ with $a \geq 1$

Hierarchical modeling

- Foregoing is really a hierarchical setup by considering a conditional likelihood on the spatial random effects $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))$.
- **First stage**:

$$\mathbf{Y}|\boldsymbol{\theta}, \mathbf{w} \sim N(X\boldsymbol{\beta} + \mathbf{w}, \tau^2 I)$$

  The $Y(\mathbf{s}_i)$ are conditionally independent given the $w(\mathbf{s}_i)$'s.
- **Second stage:**

$$\mathbf{w}|\sigma^2, \phi \sim N(\mathbf{0}, \sigma^2 H(\phi))$$

- **Third stage:** priors on $(\boldsymbol{\beta}, \tau^2, \sigma^2, \phi)$

Computing the posterior

- We seek the marginal posterior $p(\boldsymbol{\theta}|\mathbf{y})$, which is the same under the original and hierarchical settings
- Choice: Fit as $f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ or as $f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{w})p(\mathbf{w}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.
- Fitting the marginal model is computationally more stable: lower dimensional sampler (no $\mathbf{w}$'s); $\sigma^2 H(\phi) + \tau^2 I$ more stable than $\sigma^2 H(\phi)$
- BUT the conditional model allows conjugate full conditionals for σ^2, τ^2 (inverse gamma), $\boldsymbol{\beta}$, and $\mathbf{w}$ (Gaussian) – easy updates!
- Marginalized model will need Metropolis updates for σ^2, τ^2, and ϕ. But these usually work well and often converge faster than the full Gibbs updates.

Where are the **w**'s?

- Interest often lies in the spatial surface $\mathbf{w}|\mathbf{y}$ (pattern of spatial adjustment)
- Have we lost the **w**'s with the marginalized sampling?
- **No**: They are easily recovered via composition sampling:

$$p(\mathbf{w}|\mathbf{y}) = \int p(\mathbf{w}|\boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$$

- Note that

$$p(\mathbf{w}|\boldsymbol{\theta}, \mathbf{y}) \propto f(\mathbf{y}|\mathbf{w}, \boldsymbol{\beta}, \tau^2) p(\mathbf{w}|\sigma^2, \phi)$$

  is a multivariate normal distribution, resulting in easy composition sampling, in fact 1-1 with posterior samples of $\boldsymbol{\theta}$

Spatial prediction (Bayesian kriging)

- Prediction of $Y(\mathbf{s}_0)$ at a new site $\mathbf{s}_0$ with associated covariates $\mathbf{x}_0 \equiv \mathbf{x}(\mathbf{s}_0)$.
- Predictive distribution:

$$p(y(\mathbf{s}_0)|\mathbf{y}, X, \mathbf{x}_0) = \int p(y(\mathbf{s}_0), \boldsymbol{\theta}|\mathbf{y}, X, \mathbf{x}_0) d\boldsymbol{\theta}$$
$$= \int p(y(\mathbf{s}_0)|\mathbf{y}, \boldsymbol{\theta}, X, \mathbf{x}_0) p(\boldsymbol{\theta}|\mathbf{y}, X) d\boldsymbol{\theta}$$

- $p(y(\mathbf{s}_0)|\mathbf{y}, \boldsymbol{\theta}, X, \mathbf{x}_0)$ is normal since $p(y(\mathbf{s}_0), \mathbf{y}|\boldsymbol{\theta}, X, \mathbf{x}_0)$ is!
- $\Rightarrow$ easy Monte Carlo estimate using composition with Gibbs draws $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(G)}$: For each $\boldsymbol{\theta}^{(g)}$ drawn from $p(\boldsymbol{\theta}|\mathbf{y}, X)$, draw $Y(\mathbf{s}_0)^{(g)}$ from $p(y(\mathbf{s}_0)|\mathbf{y}, \boldsymbol{\theta}^{(g)}, X, \mathbf{x}_0)$.

Joint prediction

- Suppose we want to predict at a set of m sites, say $S_0 = \{\mathbf{s}_{01}, \ldots, \mathbf{s}_{0m}\}$.
- We could individually predict each site "independently" using method of the previous frame
- BUT joint prediction may be of interest, e.g., bivariate predictive distributions to reveal pairwise dependence, to reflect posterior associations in the realized surface:
- Form the unobserved vector $\mathbf{Y}_0 = (\mathbf{Y}(\mathbf{s}_{01}), \ldots, \mathbf{Y}(\mathbf{s}_{0m}))$, with X_0 as covariate matrix for S_0, and compute

$$p(\mathbf{y}_0|\mathbf{y}, X, X_0) = \int p(\mathbf{y}_0|\mathbf{y}, \theta, X, X_0) p(\theta|\mathbf{y}, X)$$

- Again, posterior sampling using composition sampling.

Spatial Generalized Linear Models

- Some data sets preclude Gaussian modeling; $Y(\mathbf{s})$ need not be continuous
- Example: $Y(\mathbf{s})$ is a binary or count variable
  - Presence/absence of a species at a location; abundance of a species at a location
  - precipitation or deposition was measurable or not
  - number of insurance claims by residents of a single family home at $\mathbf{s}$
  - Land use classification at a location (not ordinal)
- $\Rightarrow$ replace Gaussian likelihood by an appropriate exponential family member if possible
- See Diggle Tawn and Moyeed (1998)

Spatial GLM (cont'd)

- **First stage:** $Y(\mathbf{s}_i)$ are conditionally independent given $\boldsymbol{\beta}$ and $w(\mathbf{s}_i)$ with $f(y(\mathbf{s}_i)|\boldsymbol{\beta}, w(\mathbf{s}_i), \gamma)$ an appropriate non-Gaussian likelihood such that

$$g(E(Y(\mathbf{s}_i))) = \eta(\mathbf{s}_i) = \mathbf{x}^T(\mathbf{s}_i)\boldsymbol{\beta} + w(\mathbf{s}_i) \ ,$$

  where η is a canonical link function (such as a log or logit) and γ is a dispersion parameter.
- **Second stage:** Model $w(\mathbf{s})$ as a Gaussian process:

$$\mathbf{w} \sim N(\mathbf{0}, \sigma^2 H(\phi))$$

- **Third stage:** Priors and hyperpriors.
- Lose conjugacy between first and second stage; not sensible to add a pure error term

Spatial GLM: comments

- Spatial random effects in the transformed mean with continuous covariates encourages the means of spatial variables at proximate locations to be close to each other
- Marginal spatial dependence is induced between, say, $Y(\mathbf{s})$ and $Y(\mathbf{s}')$, but the observed $Y(\mathbf{s})$ and $Y(\mathbf{s}')$ need not be close to each other. No smoothness in $(\mathbf{s})$ surface
- Our second stage modeling is attractive for spatial explanation in the *mean*
- First stage modeling is better for encouraging proximate observations to be close.
- Note that this approach offers a valid joint distribution for the $Y(\mathbf{s}_i)$, but not a spatial process model; we need not achieve a consistent stochastic process for the uncountable collection of $Y(\mathbf{s})$ values

Univariate areal data modeling

- In spatial epidemiology, interest in disease mapping, where we have data

  $Y_i =$ observed number of cases of disease in county i

  $E_i =$ expected number of cases of disease in county i

- Y_i are random, but the E_i are thought of as fixed and known functions of n_i, e.g.,

$$E_i = n_i \bar{r} \equiv n_i \left(\frac{\sum_i y_i}{\sum_i n_i} \right) ,$$

  what we expect under a constant disease rate across i

- This process is called internal standardization since it centers the disease rates, but uses the observed data to do so.

External standardization

- Internal standardization is "cheating" (or at least "empirical Bayes") in that we estimate the grand rate r from our current data, but do not account for this
- Need observed counts to obtain expected counts (more below)
- Better approach: use an existing standard table of age-adjusted rates for the disease.
- For example, after stratifying the population by age group, the E_i emerge as

$$E_i = \sum_j n_{ij} r_j \ ,$$

  where n_{ij} is the person-years at risk in area i for age group j, and r_j is the disease rate in age group j (taken from the standard table).
- This process is called external standardization

Traditional models and methods

- If E_i are not too large (disease is rare or regions i are small), we often assume

$$Y_i|\eta_i \sim Po(E_i\eta_i) \ ,$$

where η_i is the true relative risk of disease in region i.

- The maximum likelihood estimate (MLE) of η_i is

$$\hat{\eta}_i \equiv SMR_i = \frac{Y_i}{E_i} \ ,$$

the standardized morbidity (or mortality) ratio (SMR), i.e., the ratio of observed to expected disease cases (or deaths).

Traditional models and methods (cont'd)

- Note that $Var(SMR_i) = Var(Y_i)/E_i^2 = \eta_i/E_i$, and so we might take $\widehat{Var}(SMR_i) = \hat{\eta}_i/E_i = Y_i/E_i^2$...
- To find a confidence interval for η_i, easiest to assume that log SMR_i is roughly normally distributed. Using the delta method (Taylor series expansion),

$$Var[\log(SMR_i)] \approx \frac{1}{SMR_i^2} Var(SMR_i) = \frac{E_i^2}{Y_i^2} \times \frac{Y_i}{E_i^2} = \frac{1}{Y_i} .$$

- An approximate 95% CI for $\log \eta_i$ is thus $\log SMR_i \pm 1.96/\sqrt{Y_i}$, and so (transforming back) an approximate 95% CI for η_i is

$$\left(SMR_i \exp(-1.96/\sqrt{Y_i}) \, , \; SMR_i \exp(1.96/\sqrt{Y_i}) \right) .$$

Traditional models and methods (cont'd)

- Now suppose we wish to test whether the true relative risk in county i is elevated or not, i.e.,

$$H_0 : \eta_i = 1 \text{ versus } H_A : \eta_i > 1 \, .$$

- Under the null hypothesis, $Y_i \sim Po(E_i)$, so the p-value for this test is

$$Pr(X \geq Y_i | E_i) = 1 - Pr(X < Y_i | E_i) = 1 - \sum_{x=0}^{Y_i - 1} \frac{\exp(-E_i) E_i^x}{x!} \, .$$

- This is the (one-sided) p-value; if it is less than 0.05 the traditional approach would reject H_0, and conclude that there is a statistically significant excess risk in county i.

Hierarchical Bayesian methods

- Now think of the true underlying relative risks η_i as random effects, to allow "borrowing of strength" across regions
- Appropriate if we want to estimate and map the underlying risk surface
- The random effects here can be high dimensional, and are couched in a Poisson likelihood...
  $\Rightarrow$ most naturally handled using hierarchical Bayesian modeling!

Poisson-gamma model

- A standard hierarchical model is:

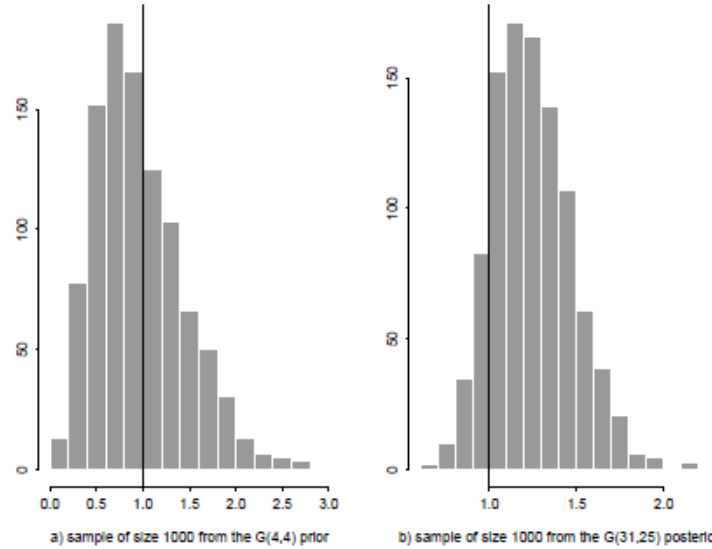
$$Y_i \mid \eta_i \overset{ind}{\sim} Po(E_i\eta_i)\ , i = 1, \ldots, I,$$
$$\text{and } \eta_i \overset{iid}{\sim} G(a, b)\ ,$$

  where $G(a, b)$ denotes the gamma distribution with mean $\mu = a/b$ and variance $\sigma^2 = a/b^2$ (this is the WinBUGS parametrization of the gamma)
- Solving these two equations for a and b we get

$$a = \mu^2/\sigma^2 \text{ and } b = \mu/\sigma^2\ .$$

- Setting $\mu = 1$ (the "null" value) and $\sigma^2 = (0.5)^2$, panel (a) of the next frame shows a sample of size 1000 from the resulting (fairly vague) $G(4, 4)$ prior...

Gamma prior and posterior



Note vertical reference line at $\eta_i = \mu = 1$ (the “null" value)

Gamma prior and posterior

- Thanks to the conjugacy of the gamma prior with the Poisson likelihood, the posterior distribution for η_i is again a gamma, namely a $G(y_i + a, E_i + b)$
- Point estimate of η_i: the posterior mean, $E(\eta_i|\mathbf{y})=$

$$E(\eta_i|y_i) = \frac{y_i + a}{E_i + b} = \frac{y_i + \frac{\mu^2}{\sigma^2}}{E_i + \frac{\mu}{\sigma^2}}$$

$$= \frac{E_i\left(\frac{y_i}{E_i}\right)}{E_i + \frac{\mu}{\sigma^2}} + \frac{\left(\frac{\mu}{\sigma^2}\right)\mu}{E_i + \frac{\mu}{\sigma^2}} = w_i\, SMR_i + (1 - w_i)\mu\ ,$$

where $w_i = E_i/[E_i + (\mu/\sigma^2)]$, so that $0 \leq w_i \leq 1$.

- Thus our point estimate is a **weighted average** of the data-based SMR for region i and the prior mean μ.

Poisson-gamma data example

- Suppose in county i we observe $y_i = 27$ disease cases, when we expected only $E_i = 21$
- Under our $G(4, 4)$ prior we obtain a $G(27 + 4, 21 + 4) = G(31, 25)$ posterior distribution; panel (b) of the figure (two slides ago) shows a sample of size 1000 drawn from this distribution.
- This distribution has mean $31/25 = 1.24$ (consistent with the figure), indicating slightly elevated risk (24%).
- However, the posterior probability that the true risk is bigger than 1 is $P(\eta_i > 1|y_i) = .863$

Poisson-gamma data example (cont'd)

- If we desired a $100 \times (1-\alpha)\%$ confidence interval for η_i, the easiest approach would be to simply take the upper and lower $\alpha/2$-points of the $G(31, 25)$ posterior.
- In our case, taking $\alpha = .05$ we obtain this 95% equal-tail credible interval as $(\eta_i^{(L)}, \eta_i^{(U)}) = (.842, 1.713)$, again indicating no "significant" elevation in risk for this county.
- To summarize I (instead of 1) posterior distributions (one for each county), we might use a choropleth map of the posterior means or 95% CI interval widths

Poisson-lognormal (spatial) model

- The gamma prior is very convenient computationally, but fails to allow for spatial correlation among the η_i
- Could contemplate a multivariate version of the gamma distribution, but instead we place a multivariate normal distribution on the $\psi_i \equiv \log \eta_i$, the log relative risks.
- Specifically, we augment our basic model to

$$Y_i \mid \psi_i \overset{ind}{\sim} Po\left(E_i\, e^{\psi_i}\right) ,$$

$$\text{where } \psi_i = \mathbf{x}_i'\boldsymbol{\beta} + \theta_i + \phi_i \text{ using}$$

- fixed effects $\boldsymbol{\beta}$ (for spatial covariates $\mathbf{x}_i$)
- heterogeneity random effects $\theta_i \overset{iid}{\sim} N(0\,,\, 1/\tau_h)$
- spatial clustering random effects $\boldsymbol{\phi} \sim CAR(\tau_c)$

Comments

- Identifiability - $\theta_i + \phi_i$
- Impropriety in CAR requires an identifiability constraint, e.g., $\sum \phi_i = 0$
- Specifying prior on pure error variance, equivalently on the precision τ_h, specifying prior on CAR "variance", equivalently, on the precision τ_c
- Reparametrization - "centering" to $\psi_i = \theta_i + \phi_i$ and θ_i
- Perhaps just focus on the spatial story
- Other first stage specifications - general linear areal data modeling

Back to internal standardization

- Again,much of the literature computes expected disease counts via internal standardization.
- Again, this places the data on both sides of the model, i.e., the counts are on the left side but they are also used to obtain the expected counts on the right side.
- So, these internally standardized models are incoherent and not generative because one cannot obtain E_i before Y_i's are realized.
- Probabilistically, they could not produce the data we observe.

A different story

- Instead, adopt the direct generative model for disease counts.
- Model disease incidence instead of relative risks, using a generalized logistic regression.
- Extract the relative risks post model fitting.

cont.

- Again, we observe disease counts Y_i as well as a set of region-specific covariates $\mathbf{X}_i$.
- Let p_i be the true incidence for region i and $\bar{p}$ be the overall disease rate across the entire study domain.
- The goal of disease mapping is to estimate the relative risk of the disease, $r_i = p_i/\bar{p}$, for each region.
- Usually we assume the number of individuals at risk in region i, n_i, is fixed and known
- Therefore, $\bar{p} = \frac{\sum_i n_i p_i}{\sum_i n_i}$.
- For rare diseases, it is reasonable to use the Poisson approximation to the binomial distribution.

cont.

- As above, the standard model in the literature:

$$\begin{aligned} Y_i \mid r_i &\overset{ind}{\sim} Po(E_i r_i), \\ \log(r_i) &= \mathbf{X}_i'\beta + \phi_i. \end{aligned}$$

- We propose a generative Poisson model for disease mapping with a specification for p_i rather than r_i:

$$\begin{aligned} Y_i \mid p_i &\overset{ind}{\sim} Po(n_i p_i), \\ F^{-1}(p_i) &= \mathbf{X}_i'\beta + \phi_i, \end{aligned}$$

  where $F(\cdot)$ is a cdf (e.g., the logit from the logit link) and the ϕ_i's, follow a CAR distribution.
- Same priors for the β and τ
- If we define $\tilde{r}_i = n_i p_i / E_i$ with E_i as above, can recover $\tilde{r}_i$'s from posterior samples of p_i, post model fitting
- The 'true" $r_i = \frac{p_i}{\bar{p}}$ where, again, $\bar{p} = \frac{\sum_i n_i p_i}{\sum_i n_i}$. So, posterior samples of the p_i will provide posterior samples of the r_i

Comparison of geostatistical vs. areal modeling

- Comparing point-referenced and areal data models
- Process vs. n-dimensional distribution
- Gaussian process vs. CAR (Markov random field)
- Model $\Sigma_{\mathbf{Y}}$ vs. $\Sigma_{\mathbf{Y}}^{-1}$
- Prediction vs explanation
- Likelihood evaluation

Misalignment

- Problems with a single variable
- Let's use the terminology "points" and "blocks"
- The variable is observed at some points but inference is desired at other points - Kriging
- The variable is observed at the point level but inference is desired at block levels
- "Block" average, block kriging

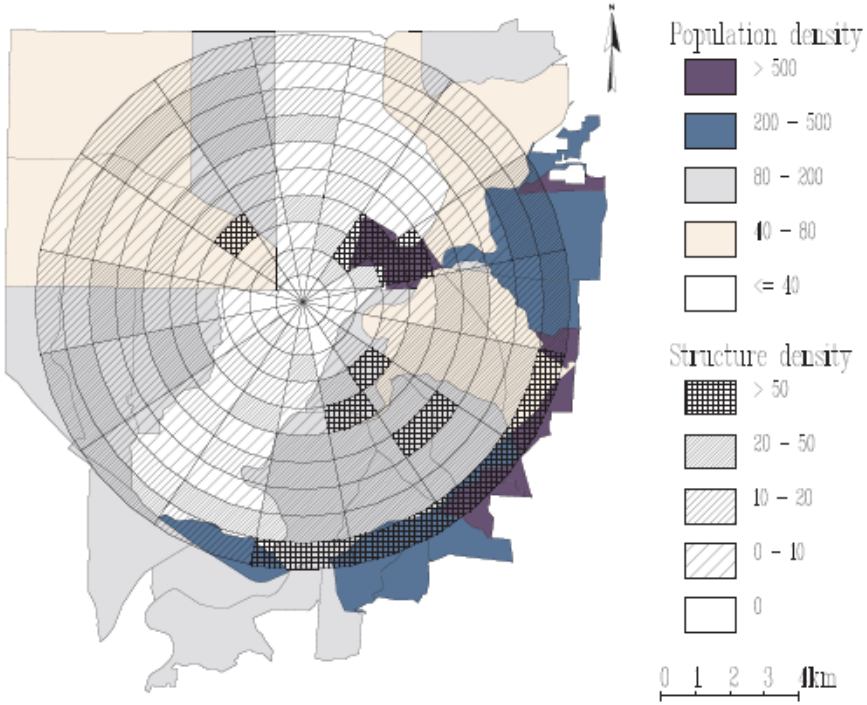
$$Y(A) = \frac{1}{|A|} \int_A y(\mathbf{s}) ds$$

- Stochastic integral, Monte Carlo approximation

$$\hat{Y}(A) = \frac{1}{L} \sum_l y(\mathbf{s}_l)$$

Block-Block Misalignment

Population by census tract; residential structures by "cell":

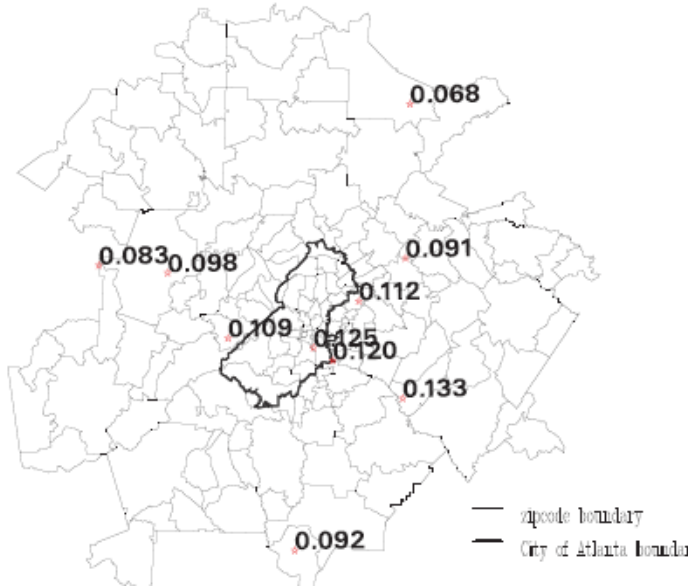


Misalignment cont.

- The variable is observed at block level with inference desired at other blocks
- Modified areal unit problem (MAUP)
- Alternatives to "areal allocation"
- The variable is observed at block level but inference is sought at point level
- Does this make sense? e.g., average rainfall for A vs. number of cases in A

Bivariate misalignment

Ozone measurements at fixed sites; counts of pediatric asthma cases by zip code in Atlanta, GA:



Misalignment cont.

- Problems with several variables; interest in regression
- X at point level, Y at other points
- X at point level, Y at block level
- X at block level, Y at point level
- X at block level, Y at block level
- Bring X's to the scale of the Y's
- With more than two variables, bring all the variables to a common scale. Highest resolution is obviously preferred.

The ecological fallacy

- Suppose for a set of regions A_i, $i = 1, ..., I$ partitioning A we have the population at risk, N_i, and observe the counts of cases Y_i.
- For simplicity, assume a univariate exposure surface denoted by $X(\mathbf{s})$ at location $\mathbf{s}$ in A.
- Within A, the exposure data $X(\mathbf{s}_k)$ are available from a set of monitoring stations at locations $\mathbf{s}_k$, $k = 1, ..., K$.
- A naive disease mapping model:

$$\begin{aligned} Y_i &\overset{ind}{\sim} \text{Poisson}(N_i p_i), \\ \text{logit}(p_i) &= \beta_0^* + \beta_1^* \bar{X}_i, \end{aligned}$$

where p_i is the disease incidence for region A_i, and $\bar{X}_i$ is the *mean* exposure within A_i.

cont.

- The exposure data are only observable at sparsely located monitoring stations but to obtain an average exposure for an A_i, we need a model for the entire exposure surface.
- A geostatistical approach specifies a model for the exposure surface $X(\mathbf{s})$ for $\mathbf{s} \in A$ employing say, a stationary Gaussian process.
- Then, the block average

$$\bar{X}_i = \int_{A_i} X(\mathbf{s}) f_i(\mathbf{s}) d\mathbf{s},$$

is formed, where $f_i(\mathbf{s})$ is the population density at location $\mathbf{s}$ in region A_i, i.e., $\int_{A_i} f_i(\mathbf{s}) d\mathbf{s} = 1$.

The ecological fallacy

- Leads to the *ecological fallacy*, an ecological bias arising from assumption that associations at the block level are the same as those for individuals within the blocks.
- To illustrate, let Y_{ij} denote a Bernoulli disease indicator for individual j in region A_i with individual level model

$$\begin{aligned} Y_{ij} &\overset{ind}{\sim} \text{Bernoulli}(p_{ij}), \\ \log(p_{ij}) &= \beta_0 + \beta_1 X_{ij}. \end{aligned}$$

Letting $Y_i = \sum_{j=1}^{N_i} Y_{ij}$, we have

$$\mathbb{E}(Y_i) = N_i q_i, \quad \text{with} \quad q_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \exp(\beta_0 + \beta_1 X_{ij}).$$

- q_i is the average disease incidence of individuals in region A_i. Clearly, $q_i \neq p_i$ in the above
- So, $\beta_1 \neq \beta_1^*$. Bias arises by summation of nonlinear (log) terms. Mean of the logs is **NOT** equal to the log of the mean.