# Effects of spatial scale on spatial confounding bias and precision of spatial regression estimators

Chris Paciorek
Department of Biostatistics
Harvard School of Public Health
www.biostat.harvard.edu/˜paciorek

December 3, 2008

# Spatially-correlated Residuals

$$Y \sim \mathcal{N}(X\beta, \Sigma)$$

What do we know?

- Under known correlation structure:
  1. GLS is more efficient than OLS for estimating exposure effect, $\beta$.
  2. Standard OLS variance estimator is incorrect.
  3. Estimating the correlation structure complicates matters.

What don't we know?

- What happens if the residual is correlated with the exposure; what can we say about bias?
- How does the spatial scale of the residual affect bias, efficiency, and variance estimation?
- How does spatial scale in exposure affect matters?

# The Core Issue

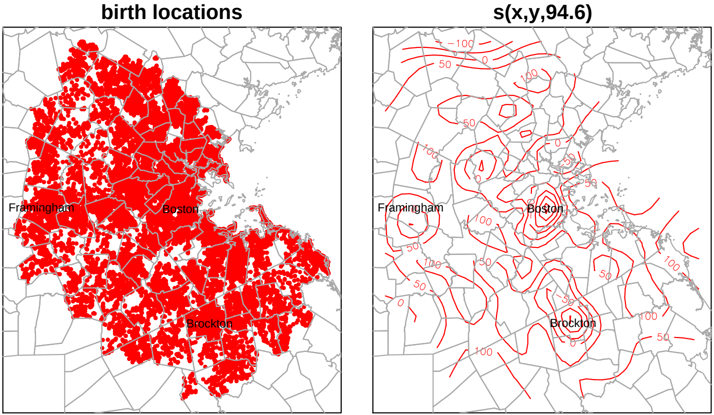
- Is the spatial residual structure correlated with the exposure?
    1. The spatial structure may be caused by unmeasured confounders.
    2. If exposure and residual have large-scale variation, dependence seems likely.
- If so, this correlation violates a key assumption of standard random effects models, including kriging models.

# Example of Air Pollution Epidemiology

- Estimates of chronic health effects of air pollution are identified from cross-sectional (i.e. spatial) variation in exposure.
- Large-scale spatial differences are easier to measure than small-scale differences in exposure.
- Hypothesis: large-scale variation is more likely to be confounded than smaller-scale variation.
  - regional variation in diet, exercise, cultural factors, socioeconomic status
- So if regions with lower income or less healthy diets are regions with higher pollution, you would expect spatial confounding bias.

# Birthweight and Traffic Pollution in Eastern Massachusetts

All births in eastern Massachusetts, 1996-2001



birth locations

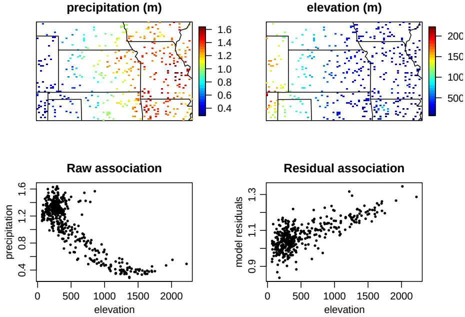
s(x,y,94.6)

For comparison, sex effect is ~130 g, black carbon estimate of ~7 g.

# Scale Matters

## How does elevation affect precipitation in the central United States?

- Large-scale negative association, but elevation is not the causal effect.



- A spatial model $y_i = \beta_0 + \beta_1 x_i + g(s_i) + \epsilon_i$ can isolate the elevation effect to the effect of elevation at small scales (positive association).

# A Simple Modeling Framework

Consider the linear model with correlated residuals:

$$Y \sim \mathcal{N}(\mathcal{X}\beta, \Sigma)$$

This can be obtained using a simple mixed model,

$$Y_i \sim \mathcal{N}(\beta_0 + \beta_x X(s_i) + g(s_i), \tau^2)$$

with spatially-correlated, normally-distributed random effects,

$$g \sim \mathcal{N}(0, \sigma_g^2 R(\theta_g)).$$

Marginalizing over $g$ gives

$$Y \sim \mathcal{N}(\beta_0 1 + \beta_x X, \sigma_g^2 R(\theta_g) + \tau^2 I).$$

$X$ is likely to be spatially correlated (e.g., if $X$ is generated by a Gaussian process, $X \sim \mathcal{N}(0, \sigma_x^2 R(\theta_x))$.

Note that this model is essentially equivalent to a universal kriging model.

# A Potential Problem

- What if $X$ and $g$ are dependent?
  - We have integrated over the marginal for $g$ (because the usual random effects model assumes the random effects are independent of the covariates) when we should have integrated over the conditional for $g|X$.
- Letting $\epsilon_i^* = g(s_i) + \epsilon_i$, we have the model $Y_i = \beta_0 + \beta_x X(s_i) + \epsilon_i^*$.
  - The usual regression model assumes the covariate and the residual are independent
  - Violating this assumption induces bias.

# Identifiability

- There is a fundamental non-identifiability in the model

$$Y_i = X(s_i)\beta + g(s_i) + \epsilon_i$$

which we could re-express as

$$Y_i = g^*(s_i) + \epsilon_i.$$

How do we separate the pollution effect from the spatial effect (spatial confounder) if the pollution effect is just another form of spatial effect?

# Constraints Provide Identifiability

- Constraints on $g$ provide identifiability: penalized likelihood, distribution on random effects (mixed effects model or Bayesian model)
- Such penalized models favor attribution to the fixed effect:
  - Penalty on smoothness of $g$
  - Random effects density (prior) for $g$
- Key question: Do such models reduce spatial confounding bias?
  - Potential mechanism for bias reduction: attribute variability from confounder to $g$.
- Conventional Wisdom?
  - Accounting for spatial correlation in the residual, $g$, can account for spatial confounding and reduce (eliminate?) bias.

# General Analytic Framework

Assume there is an unmeasured spatially-varying confounder, $Z(s)$. Let the data generating mechanism be

$$Y_i = \beta_0 + \beta_x X(s_i) + \beta_z Z(s_i) + \epsilon_i, \;\; \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau^2)$$

Assume that $X(s)$ and $Z(s)$ are Gaussian processes and that at a given location $\text{Corr}(X(s_i), Z(s_i)) = \rho$.

- $X$ and $Z$ could be considered deterministic, in which case, $\rho$ stands in for the empirical association of $X$ and $Z$,

$$\hat{\rho} = \frac{\sum (x_i - \bar{x})(z_i - \bar{z})}{s_x s_z}$$

# Bias Implications (1)

## Known parameters, single scale

- Suppose $X(s)$ and $Z(s)$ share the same range of spatial correlation, but may be scaled differently in magnitude, namely, $\text{Cov}(X) = \sigma_x^2 R(\theta_c)$ and $\text{Cov}(\beta_z Z) = \beta_z^2 \sigma_z^2 R(\theta_c)$, then

$$\begin{aligned} \text{E}(\hat{\beta}_x|X) &= \beta_x + \left[(\mathcal{X}^T\Sigma^{-1}\mathcal{X})^{-1}\mathcal{X}^T\Sigma^{-1}\text{E}(Z|X)\beta_z\right]_2 \\ &= \beta_x + \rho\frac{\sigma_z}{\sigma_x}\beta_z \end{aligned}$$

  because $E(Z|X) = \mu_x + \rho\sigma_z\sigma_x R(\theta_c)\sigma_x^{-2}R(\theta_c)^{-1}(X - \mu_x 1)$.

- The bias, $\rho\frac{\sigma_z}{\sigma_x}\beta_z$, is the same as if the covariates were not spatially structured.

- Heuristic: the model attributes variability from the confounder to the covariate of interest.

# Bias Implications (2)

## Known parameters, multi-scale

Let $X(s) = X_c(s) + X_u(s)$ with $\mathrm{Cov}(X) = \sigma_c^2 R(\theta_c) + \sigma_u^2 R(\theta_u)$.
Let $\mathrm{Cov}(Z) = \sigma_z^2 R(\theta_c)$ and $\mathrm{Cor}(X_c(s_i), Z(s_i)) = \rho$.

$$\begin{aligned} \mathrm{E}(\hat{\beta}_x|X) &= \beta_x + \left[(\mathcal{X}^T\Sigma^{*-1}\mathcal{X})^{-1}\mathcal{X}^T\Sigma^{*-1}M(X-\mu_x 1)\right]_2 p_c\rho\frac{\sigma_z}{\sigma_c}\beta_z \\ &= \beta_x + c(X)\rho\frac{\sigma_z}{\sigma_c}\beta_z \end{aligned}$$

where

$$\Sigma^* \equiv \frac{\beta_z^2\sigma_z^2 R(\theta_c) + \tau^2 I}{\beta_z^2\sigma_z^2 + \tau^2} = ((1-p_z)I + p_z R(\theta_c))$$

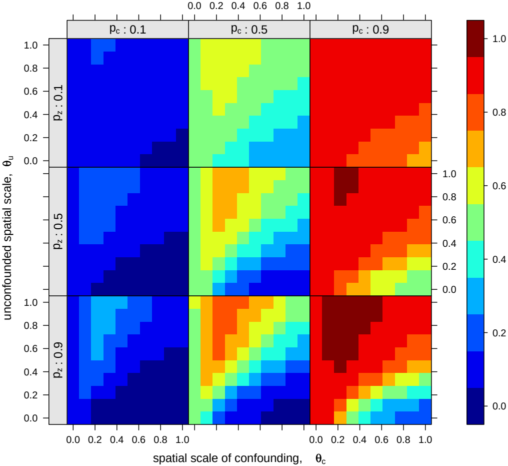
and

$$M \equiv (p_c I + (1-p_c)R(\theta_u)R(\theta_c)^{-1})^{-1}$$

and $p_z \equiv \beta_z^2\sigma_z^2/(\beta_z^2\sigma_z^2+\tau^2)$ and $p_c \equiv \sigma_c^2/(\sigma_c^2+\sigma_u^2)$

# Bias Implications (2)

## Known parameters, multi-scale



# Bias Implications (2)

## Known parameters, multi-scale

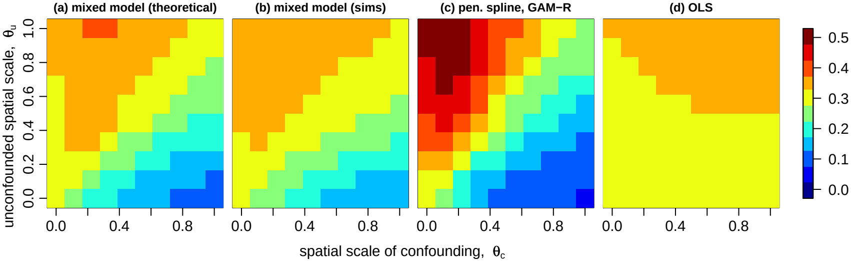
- Reducing bias requires the covariate of interest to have a spatial scale at which it is unconfounded, and that scale must be smaller than the scale at which confounding operates.
- We would like the covariate to have as much variation at the unconfounded scale and as little at the confounded scale as possible.

$$\begin{aligned} \mathrm{E}(\hat{\beta}_x|X) &= \beta_x + \left[(\mathcal{X}^T\Sigma^{*-1}\mathcal{X})^{-1}\mathcal{X}^T\Sigma^{*-1}M(X-\mu_x 1)\right]_2 p_c\rho\frac{\sigma_z}{\sigma_c}\beta_z \\ &= \beta_x + c(X)\rho\frac{\sigma_z}{\sigma_c}\beta_z \end{aligned}$$

- Other results are straightforward and match the non-spatial setting for confounding. We want:
  - the magnitude of variation in the confounder (or its effect on the outcome) be small.
  - the correlation between confounder and covariate to be small.

# Bias Implications (3)

## Unknown parameters, multi-scale: Simulation results

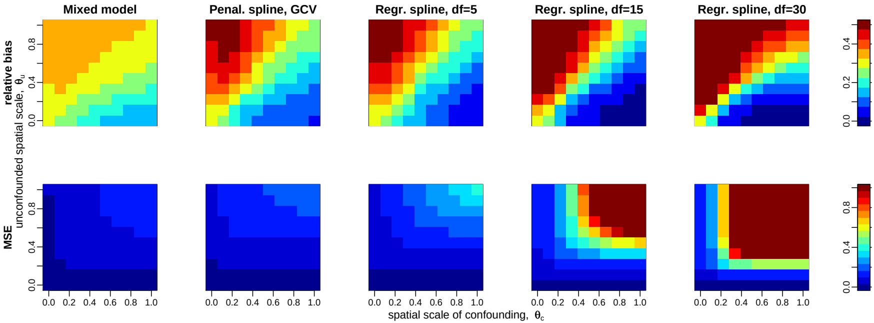


Further simulations indicate that bias is somewhat reduced by having unconfounded small-scale residual variability $(\beta_z Z + g + \epsilon)$.

- This increases the variation attributed to the spatial residual.
- This fits with the partial spline literature, which suggests undersmoothing to reduce bias.

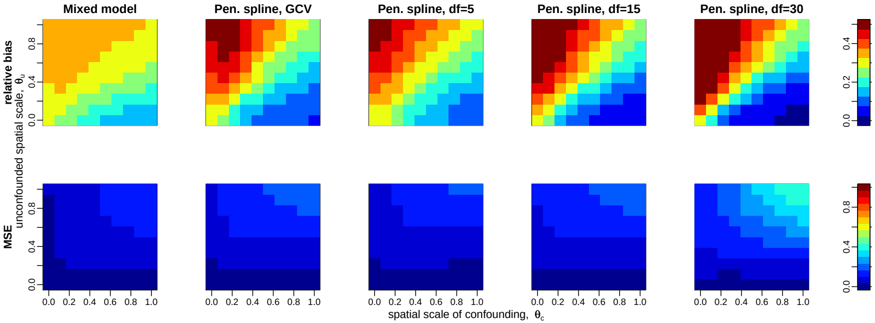
# Bias-Variance Tradeoff

Peng et al. (2006) and Zeger et al. (2007) suggest fixing the degrees of freedom and assessing sensitivity to different df values.
If there is unconfounded small-scale variation, choosing a df that captures the large-scale variation should reduce bias.



# Penalized vs. regression splines

Regression splines show less bias (but much higher variance) than penalized splines with equivalent df, presumably related to the fact that the penalized spline smoothing matrix is not a projection matrix.



# Birthweight Analysis

- Covariates: mother's age, mother's race, gestational age, mother's cigarette use, mother's health conditions, previous preterm birth, previous large birth, sex of baby, year of birth, index of prenatal care, maternal education, census tract income
- Exposure: 9-month black carbon as predicted from Gryparis et al. (2007) spatio-temporal/land use model
- Gryparis et al. (2008) found a black carbon effect of -7.27 $g$ (s.e. 3.78) per $\mu$g/m$^3$ black carbon

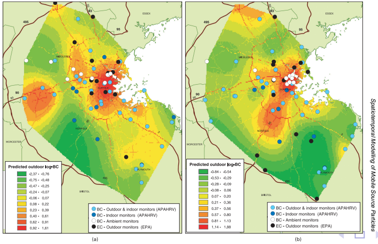
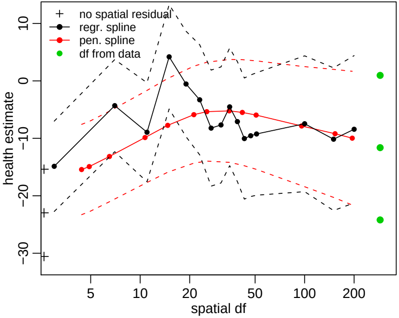


**Fig. 4.** Median predicted outdoor levels of BC for (a) winter and (b) summer: the winter predictions are for December 26th, 2002, and the summer predictions

## Naive Analysis

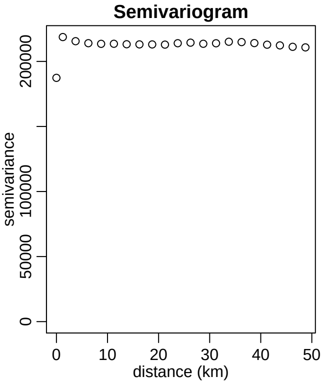
### Assume individual covariates largely unavailable

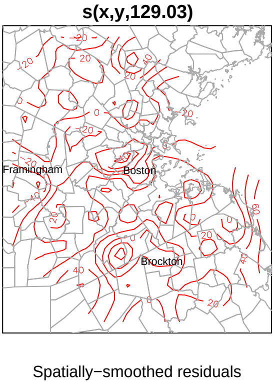
- Covariates: mother's age, gestational age, sex of baby, year of birth
- Exposure: 9-month black carbon as predicted from Gryparis et al. (2007) spatio-temporal/land use model
- Model: $y_i = \mathcal{X}_i^T \beta + g(s_i; \text{df}) + \epsilon_i$



# Residual Assessment in Full Model

Question: is there residual spatial correlation and does accounting for potential spatial confounding affect epidemiological results?

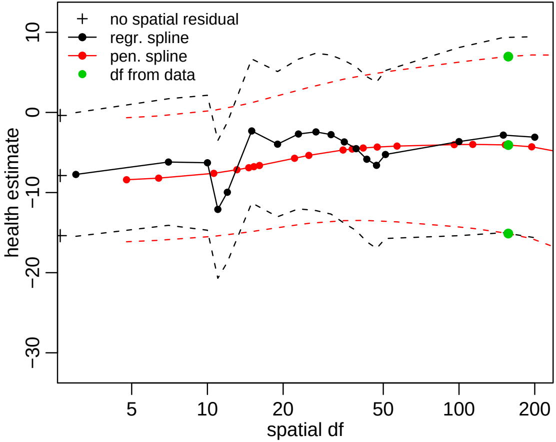




Variograms may fail to detect small magnitude spatial variation that can affect bias.

# Sensitivity Analysis

## Could previous results be affected by spatial confounding?

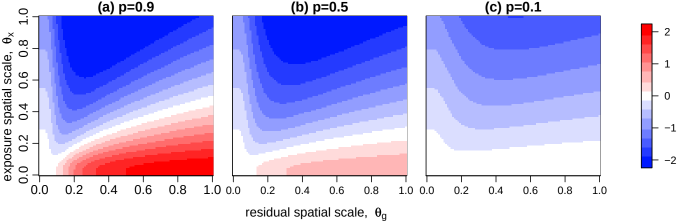


# Spatial Scales and Precision

## Does it help or hurt to have spatial variation in your data?

Relative to the equivalent amount of non-spatial variation, what is the precision of GLS estimation in the presence of residual spatial structure?

$$\log \frac{\mathrm{E}_X(\mathrm{Var}(\hat{\beta}_x)^{-1})\text{ with spatial data}}{\mathrm{E}_X(\mathrm{Var}(\hat{\beta}_x)^{-1})\text{ with non-spatial data}}$$



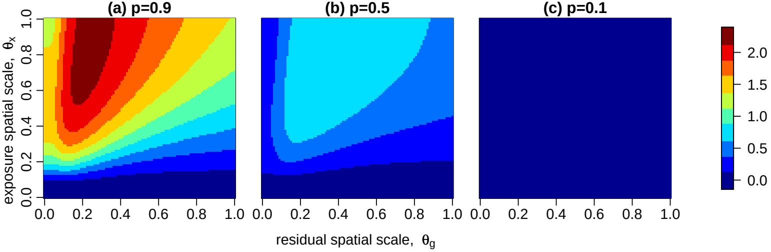
Intuition: Model treats spatial structure as a covariate that reduces residual variance, $Y_i = \mathcal{X}_i^T \beta + g(s_i) + \epsilon_i$.

# Spatial Scales and Relative Efficiency

## When does accounting for spatial variation increase efficiency?

What is the relative efficiency of GLS compared to OLS?

$$\log \mathrm{E}_X \frac{\mathrm{Var}(\hat{\beta}_x^{\mathrm{GLS}})^{-1}}{\mathrm{Var}(\hat{\beta}_x^{\mathrm{OLS}})^{-1}}$$



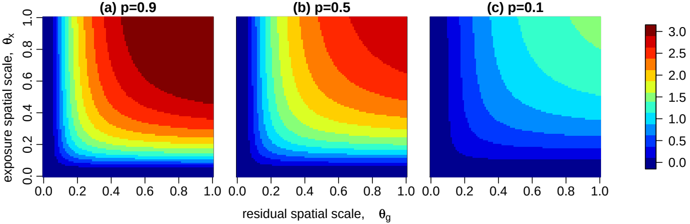
Take-home message: Benefits of GLS kick in primarily when spatial variation in exposure is moderate to large in scale.

# Spatial Scales and Uncertainty Estimation

## When is the naive OLS variance estimate OK?

What is the expected ratio of the naive and correct OLS variance estimators?

$$\log \mathrm{E}_X \frac{\mathrm{Var}_{\mathrm{correct}}(\hat{\beta}_x^{\mathrm{OLS}})}{\mathrm{Var}_{\mathrm{naive}}(\hat{\beta}_x^{\mathrm{OLS}})}$$



Take-home message: Using the naive variance estimator may be reasonable when either the exposure or residual spatial scales are small.

# Conclusions

Scale is critical: Assess the spatial scale of variation in the residuals and exposure.

- Bias:
  - Large-scale exposure variation only: little ability to reduce bias.
  - If small-scale variation in exposure exists, large-scale bias can be reduced.
    - Having small-scale variation in the residual does not reduce bias at that scale but can result in less smoothing and therefore reduced bias at larger scales.
  - Use fixed df spatial terms to assess bias-variance tradeoff in exposure estimates.
  - Measurement error in fine-scale exposure estimates may be a concern.
- Precision
  - Accounting for large-scale residual correlation is also critical for efficiency and uncertainty estimation.
  - Try to account for effect of spatial residual on uncertainty estimation, but if scale of residual is small, effect may be minor

# Implications for Areal Spatial Data

- Areal data by construction lack fine-scale variation in exposure.
- Standard areal spatial models (conditional auto-regression; CAR) vary at the scale of the areas.
- These results suggest models cannot account for bias at that scale.
- However, to the extent the CAR structure fits both small- and large-scale spatial patterns, standard CAR models may reduce bias from large-scale confounding.