Spatial Statistics and Spatial Scales in Environmental Health

Chris Paciorek
Department of Statistics; University of California, Berkeley
and
Department of Biostatistics; Harvard School of Public Health

www.biostat.harvard.edu/~paciorek

April 14, 2010

Spatial Statistics in Environmental Health

- Estimation of chronic health effects generally relies on cross-sectional variation in the exposure of interest.
- Often this variation is correlated over space, and we want to use this fact to help us estimate variation in exposure amongst individuals.
- Spatial statistical methods can help to
  1. Estimate exposure based on available data,
  2. Consider measurement error (exposure misclassification) arising from this estimation, and
  3. Account for spatial correlation in the health outcome data.
- Applications include air pollution, climate, built environment, infectious disease.

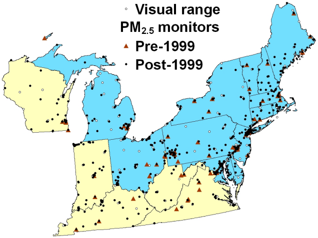
Exposure Estimation Methods for Air Pollution

Often researchers estimate ambient concentrations and use these as a proxy for exposure.

- Methods Using Monitoring Data
  - Nearest Monitor
  - Local Averaging
  - Inverse-Distance Weighted Averaging
  - Kriging
  - Land Use Regression
  - Spatio-temporal Statistical Models
- Other Sources of Information
  - Remote Sensing
  - Atmospheric Modeling
- *The Future: Atmospheric models that assimilate data and provide uncertainty estimation?*

Exposure Estimation and Spatial Scales

- We'd like to exploit as much of the true exposure variation as possible, at all scales.
  1. This can help improve precision in health analyses.
  2. Exposure at different scales may provide different information about health effects (e.g., PM components).
  3. Contrasts at different scales may be differently affected by unmeasured confounding.
- Example: estimate PM_{10} and $PM_{2.5}$ concentrations monthly at Nurses' Health Study residences.



Spatio-temporal Statistical Modeling

- A spatio-temporal statistical model (Yanosky et al. 2009; Paciorek et al. 2009):
  - First stage for monthly spatial variation:

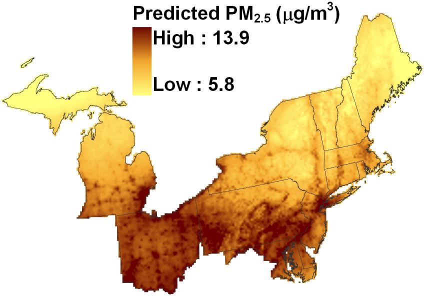
$$\log \mathrm{PM}_{it} = \mu_i + W_{it} B_W + p_t(s_i) + \epsilon_{it}$$

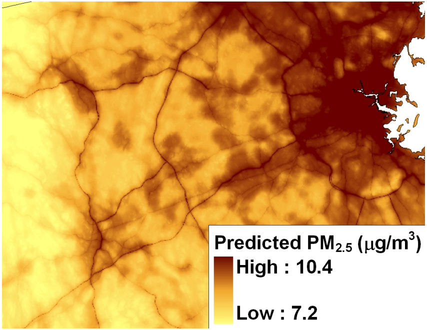
  - Second stage model for spatial-only effects:

$$\mu_i = Z_i B_Z + p_\mu(s_i) + \delta_i$$

- W's are temporally-varying predictors, while Z's vary only spatially. Either might provide fine-scale exposure information.
- Spatio-temporal ($p_t(s)$) and spatial ($p(s)$) terms act as in kriging (distance-weighted averaging).

PM Predictions (Ambient Exposure Estimates)





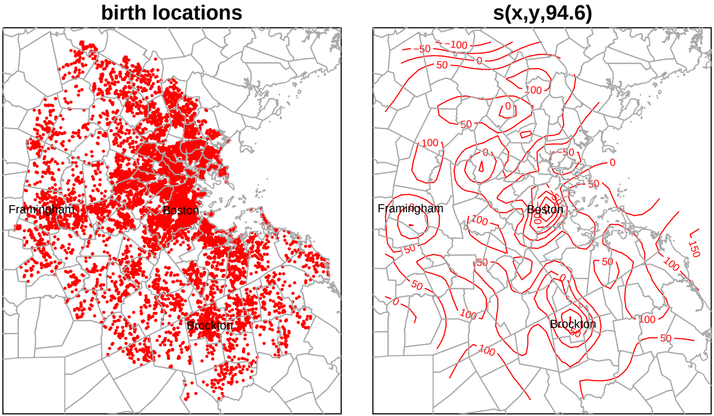
$PM_{2.5}$ predictions: northeast US (left) and greater Boston (right)

Spatial Confounding in Air Pollution Epidemiology

- Estimates of chronic health effects of air pollution are identified from cross-sectional (i.e. spatial) variation in exposure.
    - E.g., Puett et al. (2008, 2009) fit Cox survival models to estimate effects of PM exposure on mortality and coronary heart disease.
- Hypothesis: large-scale exposure variation is more prone to confounding than smaller-scale variation.
    - regional variation in diet, exercise, cultural factors, socioeconomic status
    - E.g., if regions with less healthy diets or lower income are regions with higher pollution, you would expect spatial confounding bias from unmeasured spatially-varying confounders.

Birthweight and Traffic Pollution in Eastern Massachusetts

All births in eastern Massachusetts, 1996-2001



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

Spatially-correlated Residuals

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

What do we know?

- Under known correlation structure:
  1. GLS ($\hat{\beta} = (\mathcal{X}^T\Sigma^{-1}\mathcal{X})^{-1}\mathcal{X}^T\Sigma^{-1}Y$) is more efficient than OLS for estimating exposure effect, β_x.
  2. Standard OLS variance estimator ($\hat{\sigma}^2(\mathcal{X}^T\mathcal{X})^{-1}$) is incorrect.
  3. Estimating the correlation structure complicates matters.

What don't we know?

- If the residual is correlated with the exposure (X), what can we say about bias in $\hat{\beta}_x$?
- How does the spatial scale of the residual affect bias, efficiency, and variance estimation?
- How does the spatial scale of the exposure affect matters?

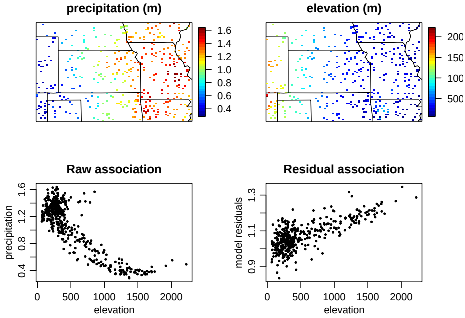
The Core Issue

- Is the spatial residual structure correlated with the exposure?
  1. The spatial structure may be caused by unmeasured confounders.
  2. Even without clear potential confounders, if exposure and residual have large-scale variation, dependence/concurvity seem likely.
- A typical approach would be to model the residual spatial variation, e.g., using spatial random effects.
- But the association violates a key assumption of standard random effects models, including kriging models.
- So can a spatial health model really help us?

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 $y_i = \beta_0 + \beta_x x_i + g(s_i) + \epsilon_i$ can (mostly) isolate the elevation effect to a positive effect of elevation at small scales.

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,

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

The mixed model is equivalent to the GLS approach (by marginalizing over g):

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

Our interest is in situations where X is also spatially correlated.

Spatial Confounding Bias

- What if X and g are dependent?
- 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 Xs and the error term are independent.
  - Violating this assumption induces bias.
- A different perspective is to consider the difficulty in separating the influence of the two spatial effects in

$$Y_i = \beta_0 + \beta_x X(s_i) + g(s_i) + \epsilon_i.$$

General Analytic Framework

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

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

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

- The value of ρ indexes the magnitude of the association (concurvity) between X and Z.

Bias Implications (1)

Known variance parameters, single scale

- Suppose $X(s)$ and $Z(s)$ share the same scale of spatial correlation, θ_c, then

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

- 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 $X(s) = X_c(s) + X_u(s)$ where only X_c is correlated with Z and has the same scale of spatial correlation, θ_c, while X_u is independent of Z and has spatial scale θ_u:

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

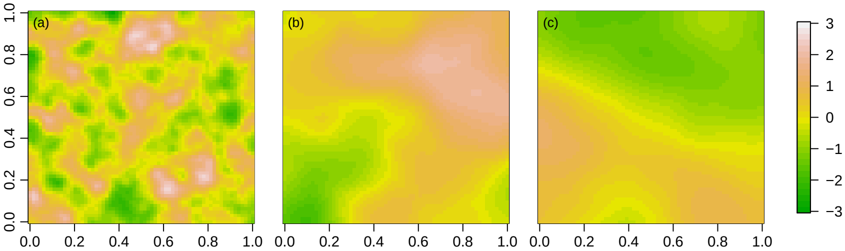
where

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

and

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

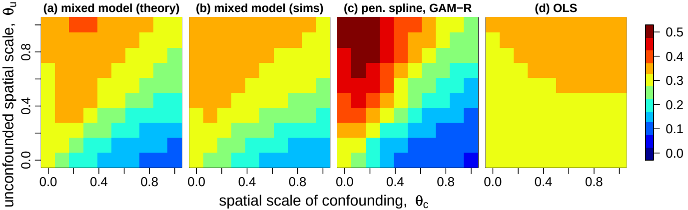
Detour: Spatial processes



(a) Fine-scale (high-frequency) variability ($\theta = 0.1$)
(b) Moderate-scale variability ($\theta = 0.5$)
(c) Large-scale (low-frequency) variability ($\theta = 0.9$).

Bias Implications (3): Simulation Results

- 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.



- Either (b) a mixed model/kriging/GLS approach or (c) using a spline term for the spatial term, $g(s)$, reduce but not eliminate bias.
- In all approaches, we must choose a parameter that determines how much smoothing we do in estimating $g(s)$.

Using Splines

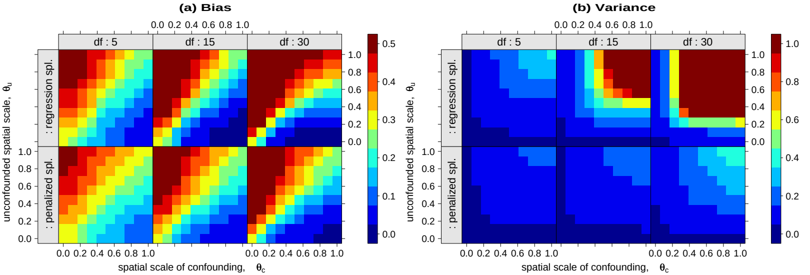
- Let's consider fitting

$$Y_i = \beta_0 + \beta_x X(s_i) + g(s_i) + \epsilon_i$$

  using a spline term to represent $g(s) := Bu$.
- A regression spline is just like ordinary regression but with spatial 'covariates', B.
- A penalized spline is like a regression spline but the magnitude of the u values is penalized, shrinking the $\hat{u}$ values toward 0. Mixed models are closely related to penalized splines.
- The effective degrees of freedom (df) quantify the complexity of $g(s)$ and thereby determine how much smoothing of the data is done.

Bias-Variance Tradeoff

- Peng et al. (2006) and Zeger et al. (2007) suggest fixing the df 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.
- Regression splines show less bias (but much higher variance) than penalized splines with equivalent df.
  - Why? Regression spline conditioning is as in OLS.



Birthweight Analysis

- Exposure: 9-month black carbon as predicted from Gryparis et al. (2007) spatio-temporal/land use model. (See also Bliznyuk et al.)
- 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.
- Gryparis et al. (2009) found a black carbon effect of -7.27 g (s.e. 3.78) per $\mu g/m^3$ black carbon.

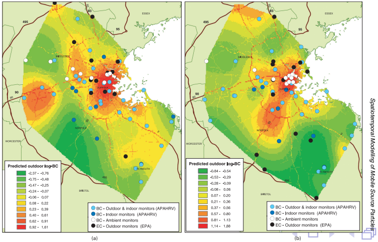
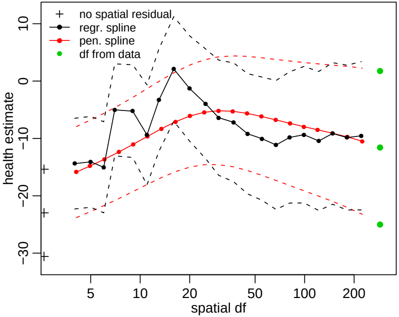


Fig. 4. Median predicted outdoor levels of BC for (a) winter and (b) summer: the winter predictions are for December 26th, 2002, and the summer predictions

Naive Analysis

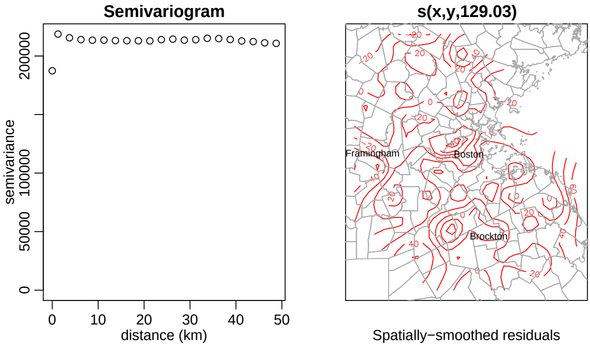
Assume individual covariates largely unavailable

- Exposure: 9-month black carbon as predicted from Gryparis et al. (2007) spatio-temporal/land use model.
- Covariates: mother's age, gestational age, sex of baby, year of birth.
- Model: $y_i = \mathcal{X}_i\beta + g(s_i; \mathrm{df}) + \epsilon_i$.



Residual Assessment in Full Model

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

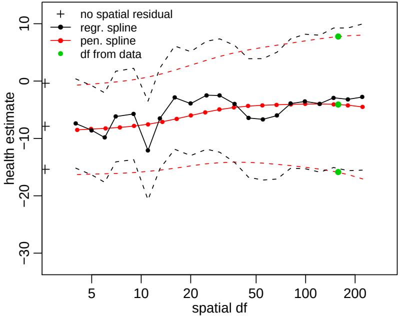


Spatially−smoothed residuals

Sensitivity Analysis

Could published results be affected by spatial confounding?

- Exposure: 9-month black carbon as predicted from Gryparis et al. (2007) spatio-temporal/land use model.
- Covariates: full set of covariates.
- Model: $y_i = \mathcal{X}_i\beta + g(s_i; \mathrm{df}) + \epsilon_i$.



Spatial Exposure Measurement Error

- Spatial exposure models can much more readily distinguish large-scale than small-scale variation, unless there are good predictors that vary at small scales.
- This suggests that in attempting to reduce large-scale confounding bias by relying on small-scale variation, we pay a price in terms of increased measurement error.
  - Also know as exposure misclassification in epidemiology/environmental health.
- What might be the effects of this?

Generic Measurement Error

- A basic regression model:

$$Y_i \sim \beta_0 + \beta_X X_i + \epsilon_i$$

  Regressing on $\hat{X}_i \neq X_i$ affects statistical properties of $\hat{\beta}_X$.
- Classical error:

$$W_i = X_i + U_i$$

  If you regress on W rather than X, $\hat{\beta}_W$ is biased, potentially badly.
- Berkson error:

$$X_i = W_i + V_i$$

  If you regress on W here, $\hat{\beta}_W$ is not biased but is more variable than the estimate $\hat{\beta}_X$ from regressing on X.
  With Berkson error, you miss components of the variation in the exposure.

Exposure Measurement Error and Scales

- We've shown that estimating exposure using methods such as land use regression and kriging are a form of regression calibration, which in principle leads to a Berkson-like formulation with limited health effects bias (Gryparis et al. 2009).
- However, uncertainty in the exposure model parameters can induce bias (Szpiro et al. 2009).
- Fine-scale variation is hard to estimate well.
    - We hypothesize that attempts to use exposure estimates of fine-scale variability may induce classical-like exposure error that could induce bias in health effects estimation.

Exposure Measurement Error Strategies

- Gryparis et al. (2009) also show that:
  - Bayesian approaches hold promise, but are often computationally expensive and potentially sensitive to model misspecification.
  - Basing health effects estimates on simulating multiple exposure estimates can be seriously biased.
- Ongoing work involves bootstrap methods to account for both Berkson-like and classical-like measurement error.
- Exposure error in multi-pollutant health analyses is a major open issue.

Conclusions: Scale is Critical

- Exposure Estimation
    - Spatial statistics methods provide a way to estimate larger-scale variation in exposure.
    - Leveraging fine-scale predictors and (hopefully) atmospheric models and remote sensing can help with finer-scale variation.
- Spatial Confounding Bias:
    - Large-scale exposure variation only: little ability to reduce bias.
    - Small-scale exposure variation present: large-scale confounding bias can be reduced.
    - Use fixed df spatial terms to assess the bias-variance tradeoff.
- Exposure Measurement Error
    - Reliance on small-scale exposure variation carries measurement error risks.
    - The impacts of such measurement error and methods for accounting for it are unsettled territory.

References

- Core material on spatial confounding:
  - Paciorek. 2010. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. In review, Statistical Science.
- Other references
  - Bliznyuk, Paciorek, and Coull. Spatio-temporal modeling of mobile source particles with temporal change of support. In preparation.
  - Gryparis, Paciorek, Zeka, Schwartz, and Coull. 2009. Measurement error caused by spatial misalignment in environmental epidemiology. Biostatistics 10:258-274.
  - Gryparis, Coull, Schwartz, and Suh. 2007. Semiparametric latent variable regression models for spatio-temporal modeling of mobile source particles in the greater Boston area. Applied Statistics 56: 183-209.
  - Paciorek, Yanosky, Puett, Laden, and Suh. 2009. Practical large-scale spatio-temporal modeling of particulate matter concentrations. Annals of Applied Statistics 3:370-397.

Other References (cont'd)

- Peng, Dominici, and Louis. 2006. Model choice in time series studies of air pollution and mortality. Journal of the Royal Statistical Society Series A 169: 179-203.
- Puett, Schwartz, Hart, Yanosky, Speizer, Suh, Paciorek, Neas, and Laden. 2008. Chronic particulate exposure, mortality and cardiovascular outcomes in the Nurses' Health Study. AJE 168:1161-1168.
- Puett, Hart, Yanosky, Paciorek, Schwartz, Suh, Speizer, and Laden. 2009. Chronic fine and coarse particulate exposure, mortality and coronary heart disease in the Nurses' Health Study. EHP 117:1697-1701.
- Szpiro, Sheppard, and Lumley. 2009. Efficient measurement error correction with spatially misaligned data. Univ. Washington Biostatistics Tech Report 350.
- Yanosky, Paciorek, and Suh. 2009. Predicting chronic fine and coarse particulate exposures using spatio-temporal models for the northeastern and midwestern US. EHP, 117:522-529.
- Zeger, Dominici, McDermott, and Samet. 2007. Mortality in the Medicare population and chronic exposure to fine particulate air pollution. Johns Hopkins Biostatistics Tech Report 133.