Measurement Error in Spatial Modeling of Environmental Exposures

Chris Paciorek, Alexandros Gryparis, and Brent Coull

August 9, 2005

Department of Biostatistics

Harvard School of Public Health

www.biostat.harvard.edu/~paciorek

LYX - FoilTEX - pdfLATEX

Outline

- Spatial exposure estimation and environmental epidemiology
- Spatial modelling of exposure
- Prediction-induced measurement error
- Methods for accounting for measurement error
- Simulation results

Spatial exposure estimation in environmental health

1. retrieval of spatio-temporal data from monitoring networks or site visits

2. space-time modelling, plus use of GIS-derived covariates

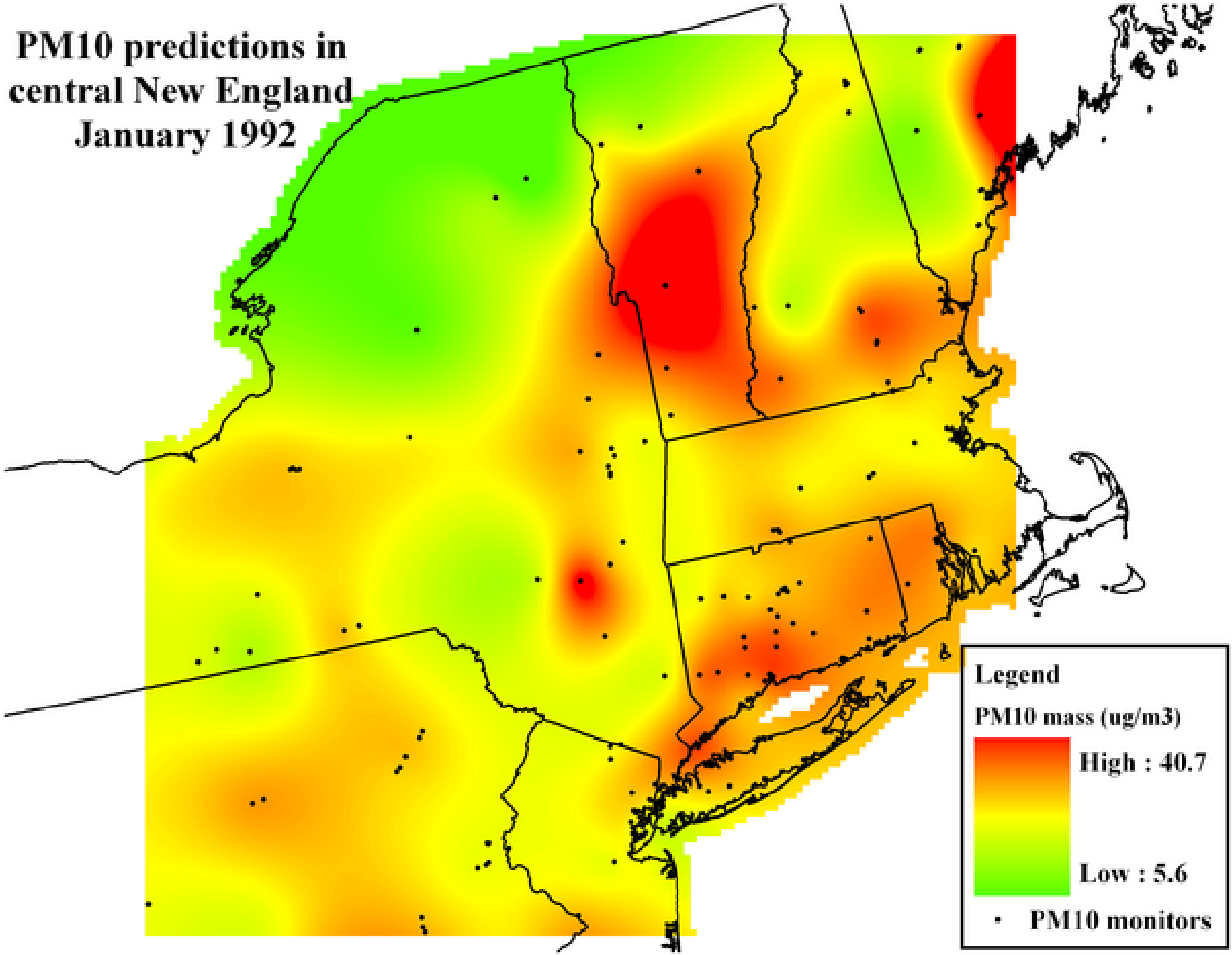
3. prediction at locations of individuals in health study
   (e.g., large cohort study)

4. epidemiological investigation with exposure predictions
   as a covariate

Cardiopulmonary disease in the Nurses' Health Study

- Hypothesis: coronary and respiratory disease are associated with chronic exposure to particulate matter ($\mathrm{PM}_{2.5}$ and PM_{10})
- Prospective cohort study of 122,000 female nurses
- PM data taken from EPA and government monitoring networks: 1985-2002
- Predictive space-time model with GIS-derived covariates
- Predictions made for each nurse's geocoded residence for each month, 1988-2002
- Proportional hazards survival modelling of health outcomes based on predicted exposure and personal covariates

Nurses’ Health Study prediction



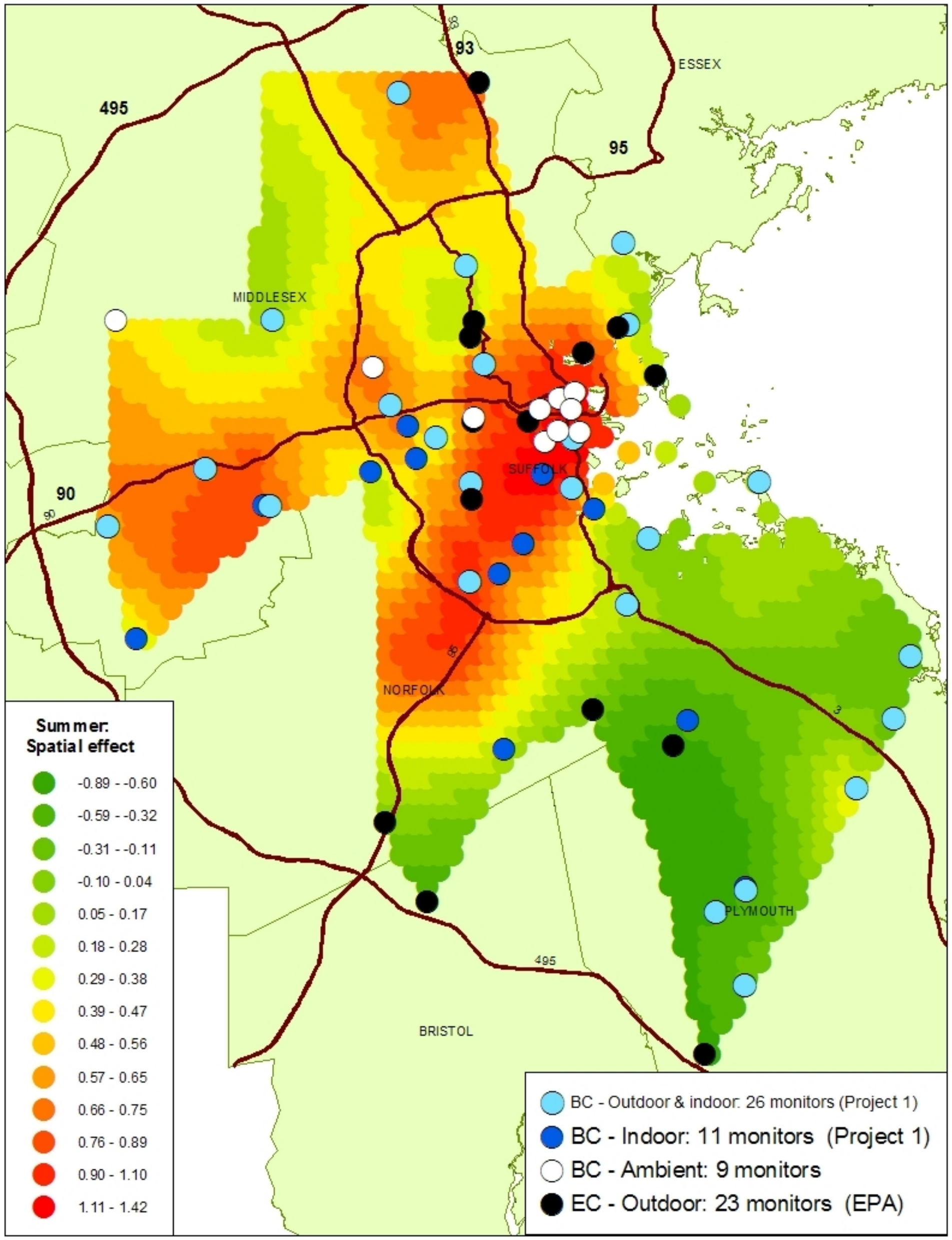
Latent variable modelling of traffic exposure in Boston

- Spatial latent variable model relating several pollutants to latent measure of traffic particles

- Predictive Bayesian space-time model with GIS-derived covariates

- Goal is to relate traffic exposure score to health outcomes in several local cohort studies

  - Birthweights in Boston
  - Diabetes cohort: heart rate variability (HRV) and inflammation markers (CRP/IL6)
  - Normative aging study: HRV and inflammation markers (CRP/IL6)
  - ...


93
ESSEX
495
95
MIDDLESEX
90
SUFFOLK
NORFOLK
3
PLYMOUTH
495
BRISTOL
Summer:
Spatial effect
-0.89 - -0.60
-0.59 - -0.32
-0.31 - -0.11
-0.10 - 0.04
0.05 - 0.17
0.18 - 0.28
0.29 - 0.38
0.39 - 0.47
0.48 - 0.56
0.57 - 0.65
0.66 - 0.75
0.76 - 0.89
0.90 - 1.10
1.11 - 1.42
BC - Outdoor & indoor: 26 monitors (Project 1)
BC - Indoor: 11 monitors (Project 1)
BC - Ambient: 9 monitors
EC - Outdoor: 23 monitors (EPA)

Predictive Modelling

- Nurses' Health Study model:

$$Y_{i,t} \quad \sim \quad N(g_t(s_i) + \sum_{p=1}^{P} f(z_{i,p}), \sigma^2)$$

  - $g_t(\cdot)$ represented as a thin-plate regression (knot-based) spline
  - individual spatial surfaces for each month (large scale heterogeneity)
  - Smooth terms of GIS covariates such as distance to roads, land use (small scale heterogeneity)
  - fit via gam() and backfitting in R

- Boston model:

$$Y_{i,t} \sim N(g(s_i) + h(t) + \sum_{p=1}^{P} f(z_{i,p}), \sigma^2)$$

  - $g(\cdot)$ represented as a thin-plate smoothing spline
  - single spatial surface with smooth terms of time and GIS covariates
  - fit via MCMC

Classical and Berkson measurement error

- Classical measurement error:
  - covariate, X, is measured with error as W

$$\begin{aligned} H_i &= \beta_0 + \beta_1 X_i + \beta_2 Z_i + \epsilon_i \\ W_i &= X_i + U_i \\ X &\perp U \\ Var(W) &= Var(X) + Var(U) \end{aligned}$$

- Berkson measurement error
  - covariate, X, is instead centered around a proxy

$$\begin{aligned} H_i &= \beta_0 + \beta_1 X_i + \beta_2 Z_i + \epsilon_i \\ X_i &= S_i + V_i \\ S &\perp V \\ Var(X) &= Var(S) + Var(V) \end{aligned}$$

Regression Calibration

- In classical measurement error, replace X with $E(X|W,Z)$
  - simple setting: $E(X|W) = \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}W$
  - linear regression: regression calibration is unbiased for β_1
  - logistic regression:
    - approximate bias, $\frac{\beta_1}{(1+\beta_1^2 0.59^2 \sigma_{X|W}^2)^{1/2}}$, is small if β_1 is small
  - survival analysis: bias is small if effect is small (e.g., relative risk < 2)

- Berkson error: $S = E(X)$
  - unbiased in linear regression
  - bias should be small for logistic and Cox regression if effect sizes are small
  - regression calibration produces a Berkson structure ($X = E(X|W) + V$)

Spatial smoothing as regression calibration

- The principle
  - Kriging/Gaussian process modelling/Bayesian smoothing act as regression calibration
    - $S = E(X|Y); \;\; X = S + V$
  - Mixed model prediction acts as regression calibration
    - the BLUP is the expected value of the spatial random effects, $S = E(X|Y)$
  - Other smoothers are likely to give similar predictions, so should mimic regression calibration

- The practice (in the Nurses' Health Study)

$$Var(X) = 0.18 \qquad Var(S) = 0.15$$

$$Cor(X,U) = -0.44 \qquad Cor(S,V) = -0.18$$

$X = S + V$ is a better model than $S = X + U$

Adjusting for measurement error

1. Use the smoothed estimates directly

2. Joint Bayesian modelling of health outcomes and exposure data, accounting for heteroscedasticity and correlation of smoothed covariate estimates

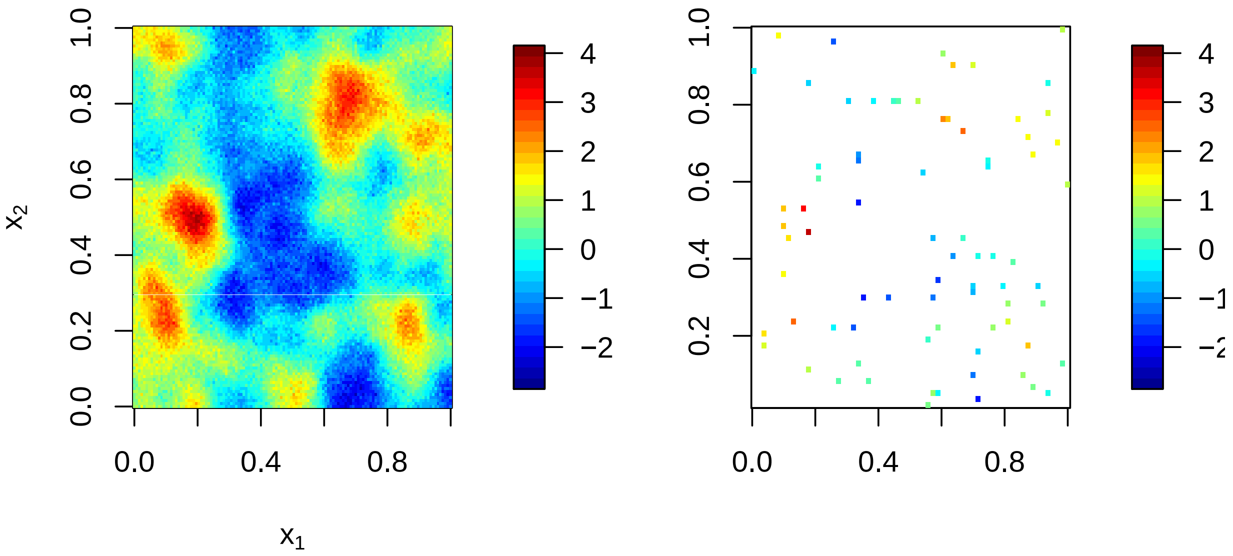
3. Sample from the exposure distribution and fit multiple health models to account for uncertainty
   – a bad idea as the sampling moves the situation from Berkson error back to classical error and induces bias

4. Cross-validation to assess under- or over-smoothing and adjust the naive estimate:
   possible model: $X = \gamma_0 + \gamma_1 S + V$
   $\hat{\beta}_{1,adj} = \hat{\beta}_1/\hat{\gamma}_1$ where $\hat{\gamma}_1$ is estimated as the slope from regressing held-out observations on smoothed predictions

   – for NHS, $\hat{\gamma}_1 =$0.88 (need to adjust for smoothing bias)

Simulation results

	Easy			Harder		
	bias	MSE	coverage	bias	MSE	coverage
true exposure				-0.002	0.003	95.2%
classical smoother	-0.013	0.011	93%	0.070	0.031	74.4%
classical with sampling	0.124	0.028	86%	0.591	0.375	1.2%
classical with γ correction				-0.018	0.028	95.2%
Bayesian	-0.031	0.016	99%	-0.114	0.083	89.0%



Conclusions

- Predictive space-time modelling of exposure induces measurement error
- Error is of the Berkson type, which in principle induces limited bias
- For continuous outcomes, some adjustments can improve estimation, particularly if smoothing problem is hard (e.g., sparse data)
- Further work is needed in the case of survival outcomes