

Measurement Error caused by Spatial Misalignment in Environmental Epidemiology

By Gryparis, A., Paciorek, C. J., Zeka, A., Schwartz, J. and Coull, B. A.

Soyoung Jeon

December 8, 2009

Introduction

- Spatial Misalignment: The locations of exposure data and health assessments do not match.
- Measurement Error: The health effects analysis often use the predictions from an exposure model, which contains some measurement error as predicted value, unequal with the true exposures.
- Existing Approaches:
 - i) Plug-in approach
 - ii) Exposure simulation approach
 - iii) Out-of-sample regression calibration estimator (RC-OOS)
 - iv) Bayesian approaches

Modeling Framework

- Let \mathbf{X} be the vector of the true exposures, \mathbf{W} be the vector of (not misaligned) measurements, $\mathbf{U} = \mathbf{W} - \mathbf{X}$ the vector of measurement errors, \mathbf{S} be the vector of smoothed estimates of \mathbf{X} , $\mathbf{V} = \mathbf{X} - \mathbf{S}$ the vector of the error after smoothing and \mathbf{Y} the health response.
- Let $(\cdot)^*$ indicates the values at locations without exposure observations and $\mu_i^* = E(Y_i^* | X_i^*, \mathbf{Z}_i^*)$. Then the following model is

$$g(\mu_i^*) = \beta_0 + \beta_1 X_i^* + \beta_z \mathbf{Z}_i^*, i = 1, 2, \dots, n_y$$

where $g(x)$ is a link function, $U_i \sim N(0, \sigma_u^2)$, $i = 1, 2, \dots, n_w$ and \mathbf{Z}_i^* is a $q \times 1$ vector of covariates measured without error.

- To estimate exposure, we get predictions $\mathbf{S}^* = (S_1^*, S_2^*, \dots, S_{n_y}^*)^T$ using the smoothing procedure.

Modeling Framework (cont'd)

- Exposure estimates are generated often using one of approaches to spatial smoothing.
- Consider a Bayesian framework with GP prior on $\mathbf{X}(\cdot)$. Then,

$$\begin{pmatrix} \mathbf{X} \\ \mathbf{X}^* \end{pmatrix} \sim N \left[\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix} \right]$$

The interim posterior before any health analysis for the distribution of \mathbf{X}^* given \mathbf{W} is

$$\mathbf{X}^* | \mathbf{W} \sim N(\boldsymbol{\mu}_2 + \mathbf{R}_{21}(\mathbf{R}_{11} + \sigma_u^2 \mathbf{I})^{-1}(\mathbf{W} - \boldsymbol{\mu}_1), \mathbf{R}_{22} - \mathbf{R}_{21}(\mathbf{R}_{11} + \sigma_u^2 \mathbf{I})^{-1} \mathbf{R}_{12}).$$

- The OLS estimator based on a regression model using $E(X_i^* | \mathbf{W})$

is unbiased. Thus we can write

$$X_i^* = E(X_i^*|\mathbf{W}) + V_i^*,$$

where \mathbf{V}^* has mean zero and variance-covariance matrix Σ^* equal to the posterior variance above.

- Other smoothing techniques create the similar structure, $\mathbf{X}^* = \mathbf{S}^* + \mathbf{V}^*$, where $\mathbf{S}^* \perp \mathbf{V}^*$ and the residual term \mathbf{V}^* has a non diagonal covariance structure.
- The uncertainty in \mathbf{X}^* by the covariance matrix $\Sigma^* = Var(\mathbf{X}^*|\mathbf{W})$ is spatially correlated and heteroscedastic.
- Focus on the linear regression model

$$Y_i^* = \beta_0 + \beta_1 X_i^* + \beta_z \mathbf{Z}_i^* + \epsilon_i, i = 1, 2, \dots, n_y$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ and independent of the measurement errors U_i

Plug-in approach

- Direct use of smoothed predicted values of exposure as a co-variate.
- First, fit the exposure model for $[\mathbf{X}^*|\mathbf{W}]$ and use \mathbf{S}^* instead of \mathbf{X}^* in health model:

$$\mathbf{Y}^* = \beta_0 + \beta_1 \mathbf{X}^* + \epsilon = \beta_0 + \beta_1 (\mathbf{S}^* + \mathbf{V}^*) + \epsilon = \beta_0 + \beta_1 \mathbf{S}^* + \eta$$

where $\eta = \beta_1 \mathbf{V}^* + \epsilon$, nondiagonal error structure.

- The OLS estimator $\widehat{\beta}_1$ is unbiased, but the variance estimator is not correct due to the correlated, heteroscedastic error structure (Carroll and *others*, 1995).
- In practice, bias or unbiased of OLS estimator occur in different situations such as the degree of smoothness, sparse monitoring data and correlation of confounders and exposure.

Exposure simulation approach

- Use the simulated exposures as an attempt to correct the variance of the plug-in estimator.
- Generate M samples $\mathbf{X}_{(t)}^* = \mathbf{S}^* + \mathbf{V}_{(t)}^*$, $t = 1, 2, \dots, M$ from the estimated distribution of $[\mathbf{X}^* | \mathbf{W}]$ and each M samples is used as a predictor to fit the health model.
- $\widehat{\beta}_{1(t)}$: average of overall estimate and
$$Var(\widehat{\beta}_1) = Var(E(\widehat{\beta}_{1(t)})) + E(Var(\widehat{\beta}_{1(t)})).$$
- The approach goes back to the classical measurement error structure, which produces biased estimates.
- The size of the bias depends on the size of Σ^* .

Out-of-sample regression calibration estimator

- Use the held-out data to fit a calibration of \mathbf{X}^{**} , where $(\cdot)^{**}$ is the values at locations where exposure is observed but held out of the main model fitting.
- \mathbf{S}^{**} is the smoothed estimates of those locations based on the remaining exposure data and \mathbf{Z}^{**} is the matrix of covariates measured without error for those locations. Then the model frame is

$$X_i^{**} = \gamma_0 + \gamma_1 S_i^{**} + \gamma_z^T \mathbf{Z}_i^{**} + \epsilon_{x,i},$$

where $E(\epsilon_{x,i}) = 0$ and $Var(\epsilon_{x,i}) = \sigma_x^2$.

- We obtain $\hat{\gamma} = (\hat{\gamma}_0, \hat{\gamma}_1, \hat{\gamma}_z^T)^T$ and calculate new estimated exposures, $\hat{X}_i^* = \hat{\gamma}_0 + \hat{\gamma}_1 S_i^* + \hat{\gamma}_z^T \mathbf{Z}_i^*$.

Bayesian approaches

- Way for uncertainty associated with using the predicted exposure values in the health model and getting a correct variance estimate

- Fully Bayesian approach:

Use the samples from the distribution $[\mathbf{X}^*, \boldsymbol{\beta} | \mathbf{Y}^*, \mathbf{W}, \mathbf{Z}^*]$.

- 2-stage Bayesian approach:

The first-stage for the exposure model is $[\mathbf{X}^* | \mathbf{W}] \propto [\mathbf{W} | \mathbf{X}^*][\mathbf{X}^*]$

and the second-stage for the health model is

$[\mathbf{X}^*, \boldsymbol{\beta} | \mathbf{Y}^*, \mathbf{W}, \mathbf{Z}^*] \propto [\mathbf{Y}^* | \mathbf{X}^*, \mathbf{W}, \mathbf{Z}^*, \boldsymbol{\beta}][\mathbf{X}^* | \mathbf{W}][\boldsymbol{\beta}]$.

Applications

Simulations

- $N = 500$ simulated data sets with $n_w = 82$ monitoring stations
- Generate \mathbf{W} without instrument error, $\mathbf{W} = \mathbf{X} = \mathbf{g} + \boldsymbol{\delta}$ with $\mathbf{g} \sim N(\mu\mathbf{1}, \mathbf{R}(\rho, v))$ and $\boldsymbol{\delta} \sim N(0, \sigma_\delta^2 \mathbf{I}_{n_w})$.
- Scenario A: very smooth surface
 - Scenario B: moderately smooth surface
 - Scenario C: more heterogeneous and the roughest surface
 - Scenario D: the same as Scenario C,
but exposure is not causally related to health ($\beta_1 = 0$)

Table 1. Results of simulation study for $\hat{\beta}_1$: bias, average model-based SE, Monte Carlo standard deviation, MSE, and coverage of 95% CIs or credible intervals, over 500 simulations, for Scenarios A–D

| Scenario | Method | Bias | $E(\text{SE}(\beta_1))$ | $\text{SD}(\hat{\beta}_1)$ | MSE | Coverage (%) |
|----------|---------------------|--------|-------------------------|----------------------------|-------|--------------|
| A | True exposure | −0.000 | 0.093 | 0.096 | 0.009 | 94.8 |
| | Plug-in | 0.004 | 0.105 | 0.122 | 0.015 | 91.6 |
| | Exposure simulation | −0.068 | 0.118 | 0.119 | 0.019 | 91.2 |
| | RC-OOS | 0.006 | 0.122 | 0.122 | 0.015 | 96.4 |
| | Fully Bayesian | 0.002 | 0.109 | 0.122 | 0.015 | 92.8 |
| | 2-stage Bayes | 0.000 | 0.108 | 0.123 | 0.015 | 93.2 |
| B | True exposure | 0.002 | 0.059 | 0.059 | 0.003 | 95.2 |
| | Plug-in | −0.085 | 0.091 | 0.149 | 0.029 | 69.8 |
| | Exposure simulation | −0.254 | 0.116 | 0.126 | 0.080 | 42.2 |
| | RC-OOS | 0.036 | 0.197 | 0.251 | 0.064 | 95.6 |
| | Fully Bayesian | 0.011 | 0.107 | 0.151 | 0.023 | 86.4 |
| | 2-stage Bayes | 0.004 | 0.105 | 0.150 | 0.023 | 83.8 |
| C | True exposure | 0.004 | 0.058 | 0.058 | 0.003 | 95.2 |
| | Plug-in | −0.140 | 0.130 | 0.211 | 0.064 | 63.4 |
| | Exposure simulation | −0.591 | 0.141 | 0.146 | 0.371 | 0.4 |
| | RC-OOS [†] | 0.039 | 0.340 | 0.367 | 0.136 | 92.6 |
| | Fully Bayesian | 0.029 | 0.155 | 0.177 | 0.032 | 93.0 |
| | 2-stage Bayes | 0.039 | 0.1646 | 0.239 | 0.059 | 90.8 |
| D | True exposure | 0.003 | 0.059 | 0.062 | 0.004 | 93.4 |
| | Plug-in | 0.001 | 0.090 | 0.095 | 0.009 | 94.2 |
| | Exposure simulation | 0.000 | 0.068 | 0.054 | 0.003 | 98.8 |
| | RC-OOS | 0.001 | 0.111 | 0.115 | 0.013 | 95.6 |
| | Fully Bayesian | 0.000 | 0.159 | 0.140 | 0.019 | 94.0 |
| | 2-stage Bayes | 0.000 | 0.148 | 0.135 | 0.018 | 94.4 |

[†]One simulation with anomalous estimate omitted.

Table 2. Results of logistic regression simulation study for $\hat{\beta}_1$: bias, average model-based SE, Monte Carlo standard deviation, MSE, and coverage of 95% CIs or credible intervals, over 500 simulations, for Scenarios A and C

| Scenario | Method | Bias | $E(\text{SE}(\beta_1))$ | $\text{SD}(\hat{\beta}_1)$ | MSE | Coverage (%) |
|----------|---------------------|-------|-------------------------|----------------------------|--------|--------------|
| A | True exposure | -1.24 | 0.070 | 0.073 | 0.0054 | 95.0 |
| | Plug-in | -0.55 | 0.094 | 0.102 | 0.0103 | 95.6 |
| | Exposure simulation | -0.91 | 0.101 | 0.101 | 0.0102 | 95.6 |
| | RC-OOS | -0.35 | 0.098 | 0.107 | 0.0114 | 100.0 |
| C | True exposure | -1.23 | 0.030 | 0.029 | 0.0009 | 95.8 |
| | Plug-in | -6.72 | 0.036 | 0.048 | 0.0027 | 81.8 |
| | Exposure simulation | -13.2 | 0.042 | 0.043 | 0.0035 | 78.4 |
| | RC-OOS | -1.22 | 0.046 | 0.050 | 0.0025 | 100.0 |

Traffic Particles and Birth Weight in Boston Area

- The association between traffic-related particulate matter by motor vehicles and birth weight in the greater Boston area
- Exposure data: BC and elemental carbon (EC) particles, use the exposure model suggested by Gryparis and *others* (2007).
- Health outcome: Birth weights in greater Boston area over Jan.1, 1996-Dec.31, 2002

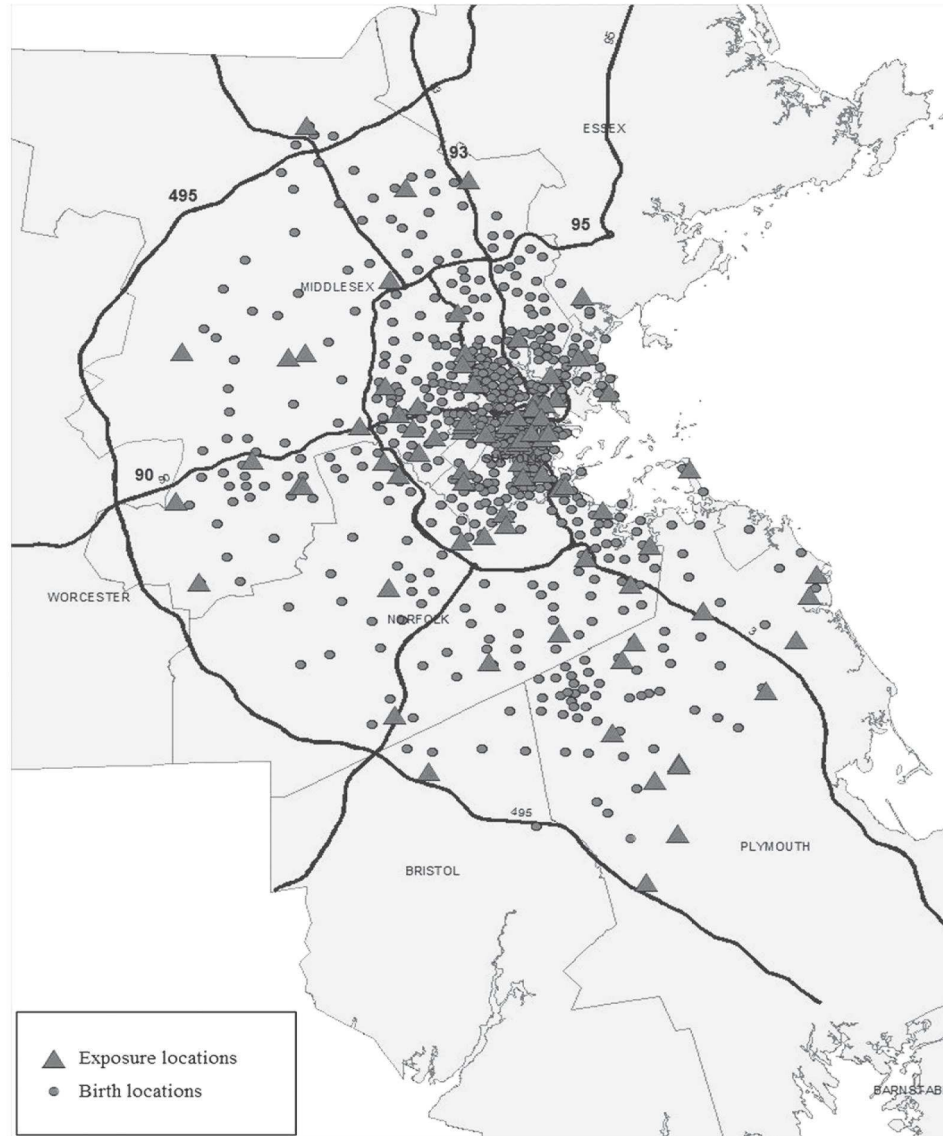


Fig. 2. Map of the locations of the residences of the birth weight study subjects and their positioning relative to the 82 exposure monitors.

Table 3. RC-OOS estimates for greater Boston birth weight data

| Method | Estimate | SE | 95% CI |
|----------------------------------|------------|-------|--------------------|
| Predicted BC | -9.46 | 4.38 | (-18.05, -0.88) |
| Mother's age | 6.36 | 0.20 | (5.97, 6.75) |
| Gestational age | 551.45 | 6.16 | (539.37, 563.52) |
| Gestational age squared | -5.72 | 0.08 | (-5.88, -5.55) |
| Number of cigarettes | -28.91 | 0.84 | (-30.56, -27.26) |
| Number of cigarettes squared | 0.69 | 0.04 | (0.61, 0.78) |
| Previous infant weighing >4000 | 480.10 | 11.56 | (457.43, 502.77) |
| Previous preterm | -242.10 | 12.82 | (-267.23, -216.97) |
| Maternal condition | -29.89 | 3.40 | (-36.56, -23.23) |
| CT median income (1000 K) | 0.15 | 0.04 | (0.07, 0.24) |
| Maternal education (<12 years) | 8.57 | 6.74 | (-4.63, 21.77) |
| Maternal education (12-16 years) | 1.00 (ref) | — | (—, —) |
| Maternal education (>16 years) | 16.63 | 2.52 | (11.70, 21.57) |
| Race (Caucasian) | 1.00 (ref) | — | (—, —) |
| Race (African American) | -131.01 | 3.64 | (-138.15, -123.87) |
| Race (Asian) | -192.72 | 3.99 | (-200.54, -184.90) |
| Race (other) | -93.15 | 3.85 | (-100.69, -85.61) |
| Sex (male) | 132.62 | 2.06 | (128.58, 136.66) |
| Sex (female) | 1.00 (ref) | — | (—, —) |
| 1996 | 19.37 | 3.96 | (11.61, 27.14) |
| 1997 | 16.52 | 4.36 | (7.97, 25.06) |
| 1998 | 23.73 | 3.85 | (16.18, 31.27) |
| 1999 | 17.02 | 3.78 | (9.61, 24.43) |
| 2000 | 10.49 | 3.77 | (3.09, 17.89) |
| 2001 | 3.36 | 3.75 | (-3.98, 10.70) |
| 2002 | 1.00 (ref) | — | (—, —) |
| Kotelchuck index (inadequate) | -70.39 | 4.31 | (-78.85, -61.94) |
| Kotelchuck index (intermediate) | -51.16 | 4.36 | (-59.71, -42.61) |
| Kotelchuck index (appropriate) | 1.00 (ref) | — | (—, —) |
| Kotelchuck index (appropriate +) | -16.17 | 2.43 | (-20.92, -11.41) |

Table 4. Results for greater Boston birth weight data

| Method | Estimate (in g) | SE | 95% CI |
|---------------------|--------------------|------|-----------------|
| Plug-in | -7.27 | 3.78 | (-14.68, 0.14) |
| Exposure simulation | -0.48 | 3.40 | (-7.13, 6.18) |
| RC-OOS | -9.46 | 4.38 | (-18.05, -0.88) |

Discussion

- Exposure simulation can show very poor performance under some realistic situations.
- The performance of the different approaches depends on the underlying exposure surface such as the amount of spatial heterogeneity.
- To explain relative practical performance of the different methods, traditional concepts of measurement error can be helpful.

References

- Gryparis, A., Paciorek, C. J., Zeka, A., Schwartz, J. and Coull, B. A. (2009) Measurement Error caused by Spatial Misalignment in Environmental Epidemiology. *Biostatistics* **10**, 258-274.
- Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995) *Measurement Error in Nonlinear Models*. New York: Chapman & Hall.
- Gryparis, A., Coull, B. A., Schwartz, J. and Suh, H. H. (2007) Semiparametric latent variable regression models for spatio-temporal modeling of mobile source particles in the greater Boston area. *Journal of the Royal Statistical Society, Series C* **56**, 183-209.