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 \( \mu_i^* = E(Y_i^*|X_i^*, Z_i^*) \). Then the following model is

\[
g(\mu_i^*) = \beta_0 + \beta_1 X_i^* + \beta_z Z_i^*, \quad i = 1, 2, \ldots, n_y
\]

where \( g(x) \) is a link function, \( U_i \sim N(0, \sigma_u^2), i = 1, 2, \ldots, n_w \) and \( Z_i^* \) is a \( q \times 1 \) vector of covariates measured without error.

- To estimate exposure, we get predictions \( \mathbf{S}^* = (S_1^*, S_2^*, \ldots, S_{n_y}^*)^T \) 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 $X(\cdot)$. Then,

$$
\left( \begin{array}{c} X \\ X^* \end{array} \right) \sim N \left[ \left( \begin{array}{c} \mu_1 \\ \mu_2 \end{array} \right), \left( \begin{array}{cc} R_{11} & R_{12} \\ R_{21} & R_{22} \end{array} \right) \right]
$$

The interim posterior before any health analysis for the distribution of $X^*$ given $W$ is

$$
X^*|W \sim N(\mu_2 + R_{21}(R_{11} + \sigma_u^2 I)^{-1}(W - \mu_1), \\
R_{22} - R_{21}(R_{11} + \sigma_u^2 I)^{-1}R_{12}).
$$

- The OLS estimator based on a regression model using $E(X_i^*|W)$
is unbiased. Thus we can write

$$X^*_i = E(X^*_i|W) + V^*_i,$$

where $V^*$ has mean zero and variance-covariance matrix $\Sigma^*$ equal to the posterior variance above.

- Other smoothing techniques create the similar structure, $X^* = S^* + V^*$, where $S^* \perp V^*$ and the residual term $V^*$ has a non diagonal covariance structure.

- The uncertainty in $X^*$ by the covariance matrix $\Sigma^* = Var(X^*|W)$ is spatially correlated and heteroscedastic.

- Focus on the linear regression model

$$Y^*_i = \beta_0 + \beta_1 X^*_i + \beta_z Z^*_i + \epsilon_i, i = 1, 2, \cdots, n_y$$

where $\epsilon_i \sim N(0, \sigma^2_\epsilon)$ and independent of the measurement errors $U_i$. 
Plug-in approach

- Direct use of smoothed predicted values of exposure as a covariate.

- First, fit the exposure model for $[X^*|W]$ and use $S^*$ instead of $X^*$ in health model:

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

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

- The OLS estimator $\hat{\beta}_1$ 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 $X^*_t = S^* + V^*_t$, $t = 1, 2, \cdots, M$ from the estimated distribution of $[X^*|W]$ and each $M$ samples is used as a predictor to fit the health model.

- $\hat{\beta}_1(t)$: average of overall estimate and
  
  $Var(\hat{\beta}_1) = Var(E(\hat{\beta}_1(t))) + E(Var(\hat{\beta}_1(t)))$.

- The approach goes back to the classical measurement error structure, which produces biased estimates.

- The size of the bias depends on the size of $\Sigma^*$. 
Out-of-sample regression calibration estimator

- Use the held-out data to fit a calibration of $X^{**}$, where $(\cdot)^{**}$ is the values at locations where exposure is observed but held out of the main model fitting.

- $S^{**}$ is the smoothed estimates of those locations based on the remaining exposure data and $Z^{**}$ is the matrix of covariates measured without error for those locations. Then the model frame is

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

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, $\hat{X}_i^* = \hat{\gamma}_0 + \hat{\gamma}_1 S_i^* + \hat{\gamma}_z Z_i^*$. 


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 \([X^*, \beta|Y^*, W, Z^*]\).

- 2-stage Bayesian approach:
  The first-stage for the exposure model is \([X^*|W] \propto [W|X^*][X^*]\) and the second-stage for the health model is \([X^*, \beta|Y^*, W, Z^*] \propto [Y^*|X^*, W, Z^*, \beta][X^*|W][\beta]\).
Applications

Simulations

• $N = 500$ simulated data sets with $n_w = 82$ monitoring stations

• Generate $W$ without instrument error, $W = X = g + \delta$ with $g \sim N(\mu 1, R(\rho, v))$ and $\delta \sim N(0, \sigma_\delta^2 I_{n_w})$.

• 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

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

$^\dagger$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

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

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

Table 4. **Results for greater Boston birth weight data**

<table>
<thead>
<tr>
<th>Method</th>
<th>Estimate (in g)</th>
<th>SE</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plug-in</td>
<td>−7.27</td>
<td>3.78</td>
<td>(−14.68, 0.14)</td>
</tr>
<tr>
<td>Exposure simulation</td>
<td>−0.48</td>
<td>3.40</td>
<td>(−7.13, 6.18)</td>
</tr>
<tr>
<td>RC-OOS</td>
<td>−9.46</td>
<td>4.38</td>
<td>(−18.05, −0.88)</td>
</tr>
</tbody>
</table>
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

