

# 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

## 1 Introduction

In some environmental epidemiology studies, the locations of exposure data and health assessments do not coincide. To overcome the misalignment problem, the health effects analysis often use the predictions from an exposure model, which contains some measurement error as predicted value but is unequal with the true exposures. Gryparis et al. (2009) focus on the framework for spatial measurement error modeling caused by spatial misalignment, describe and compare several existing approaches for continuous and binary health outcomes; plug-in approach, exposure simulation approach, out-of-sample regression calibration estimator (RS-OOS) and Bayesian approaches. And the different methods is applied to the data on the association between traffic particles and birth weights in the greater Boston area.

## 2 Modeling Framework

The model considered in this paper is under the generalized linear model to illustrate measurement error issues associated with spatially misaligned exposure and health data. To explain the framework of modeling in detail, 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 procedure and  $\mathbf{Y}$  is the vector of the health response at locations without exposure data.

Let  $(\cdot)^*$  indicates the values at locations without exposure observations and assume that  $Y_i^*$  given  $X_i^*$  and  $\mathbf{Z}_i^*$  are independent random variables with a distribution in the exponential family. Let  $\mu_i^* = E(Y_i^* | X_i^*, \mathbf{Z}_i^*)$ . Then the following model is

$$g(\mu_i^*) = \beta_0 + \beta_1 X_i^* + \beta_2 \mathbf{Z}_i^*, i = 1, 2, \dots, n_y$$

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. The measurement errors  $U_i$  are independent of  $Y_i^*$  given  $X_i^*$  and  $\mathbf{Z}_i^*$ .  $X_i^*$  is the unobserved exposure and to estimate exposure, we get predictions  $\mathbf{S}^* = (S_1^*, S_2^*, \dots, S_{n_y}^*)^T$  using the smoothing procedure. The regression coefficient  $\beta_1$  is the interest of parameter relating exposure and health.

First, a very simple linear regression is considered as a model. The linear regression model is

$$Y_i^* = \beta_0 + \beta_1 X_i^* + \beta_2 \mathbf{Z}_i^* + \epsilon_i, i = 1, 2, \dots, n_y$$

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

The use of exposure predictions from spatially misaligned exposure may be applied for discrete health outcomes, e.g. a binary or a count variables, using a generalized linear model. Suppose the  $V_i^*$  are uncorrelated and homoscedastic. Then probit model for binary health outcomes is based on the mean of the estimated exposure given the observed data with measurement error,

$$p(Y_i^* = 1 | \mathbf{W}, \mathbf{Z}_i^*) = \Phi \left[ \frac{\beta_0 + \beta_1 E(X_i^* | \mathbf{W}) + \beta_z \mathbf{Z}_i^*}{(1 + \beta_1^2 \sigma_v^2)^{1/2}} \right]$$

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

## 2.1 Plug-in approach

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

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

where  $\boldsymbol{\eta} = \beta_1 \mathbf{V}^* + \boldsymbol{\epsilon}$ , 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 as mentioned in 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.

## 2.2 Exposure simulation approach

The exposure simulation approach use the simulated exposures as an attempt to correct the variance of the plug-in estimator.  $M$  samples  $\mathbf{X}_{(t)}^* = \mathbf{S}^* + \mathbf{V}_{(t)}^*$ ,  $t = 1, 2, \dots, M$  are generated from the estimated distribution of  $[\mathbf{X}^* | \mathbf{W}]$  and each  $M$  samples is used as a predictor to fit the health model. Then  $\widehat{\beta}_{1(t)}$  is averaged through an overall estimate and variance of  $\widehat{\beta}_{1(t)}$  is calculated using the formula  $Var(\widehat{\beta}_1) = Var(E(\widehat{\beta}_{1(t)})) + E(Var(\widehat{\beta}_{1(t)}))$ . This approach goes back to the classical measurement error structure, which produces biased estimates.

## 2.3 Out-of-sample regression calibration estimator

Out-of-sample regression calibration estimator (RC-OOS) 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  $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_0, \widehat{\gamma}_1, \widehat{\boldsymbol{\gamma}}_z^T)^T$  and calculate new estimated exposures,  $\widehat{X}_i^* = \widehat{\gamma}_0 + \widehat{\gamma}_1 S_i^* + \widehat{\boldsymbol{\gamma}}_z^T \mathbf{Z}_i^*$ .

## 2.4 Bayesian approaches

Bayesian approaches is a way for uncertainty associated with using the predicted exposure values in the health model and getting a correct variance estimate. First this paper proposed fully Bayesian approach which uses the samples from the distribution  $[\mathbf{X}^*, \beta | \mathbf{Y}^*, \mathbf{W}, \mathbf{Z}^*]$ . Another alternative method is 2-stage Bayesian approach. In this method, 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  $[\mathbf{X}^*, \beta | \mathbf{Y}^*, \mathbf{W}, \mathbf{Z}^*] \propto [\mathbf{Y}^* | \mathbf{X}^*, \mathbf{W}, \mathbf{Z}^*, \beta][\mathbf{X}^* | \mathbf{W}][\beta]$ .

## 3 Application

A simulation study with 500 simulated data sets is performed to compare the different methods under different 4 exposure scenarios. The results from the simulation study show that the performance of all methods depends on the different exposure surface. As the exposure surface gets more heterogeneous, the bias of the plug-in estimator increases and the exposure simulation approach shows very poor performance. The RC-OOS approach and Bayesian approaches performed well under all scenarios. See Table 1.

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 <sup>†</sup>	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

<sup>†</sup>One simulation with anomalous estimate omitted.

The data analysis was done to find the association between traffic-related particulate

matter by motor vehicles and birth weight in the greater Boston area. The results from the three different approaches are shown in Table 4. Outputs show that the relative performance of the various methods follows the patterns suggested by simulation studies. Exposure simulation attenuates grossly the estimated health effect comparing to the plug-in approach.

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)

## 4 Conclusion

The paper proposes several important insight. First, exposure simulation can show very poor performance under some realistic situations. And 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

- [1 ] 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.
- [2 ] Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995) *Measurement Error in Non-linear Models*. New York: Chapman & Hall.
- [3 ] 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.