Measurement error caused by spatial misalignment in environmental epidemiology: Supplementary Material

Alexandros Gryparis,[1,*] Christopher J. Paciorek,[1] Ariana Zeka,[2] Joel Schwartz[3] and Brent A. Coull[1]

[1] Department of Biostatistics, Harvard University, Boston, MA 02115, U.S.A.

[2] Institute for the Environment, Brunel University, West London, U.K.

[3] Department of Environmental Health, Harvard University, Boston, MA 02115, U.S.A.

August 13, 2008

A. Weighted least squares and generalized least squares

Here we describe two additional approaches to the problem of measurement error in exposure predictions induced by spatial misalignment.

A.1 *Weighted Least Squares (WLS)*

An approach that downweights anomalous exposure predictions having high uncertainty, which might be influential in the health model, is weighted least squares (WLS) with the weights based on the uncertainty estimates from the exposure model (e.g., Kunzli et al., 2005). Although this approach seems intuitive may be useful for downweighting estimates with larger error, these are not the correct weights in our modeling framework, as discussed in Section 3. As shown there, the correct covariance is $\beta_1^2\boldsymbol{\Sigma}^* + \sigma_\epsilon^2\boldsymbol{I}_{n_y}$. As an example of when this approach could perform poorly, when all values have large, but similar, uncertainty, this approach will give similar results to that from OLS without adjusting for the correlation. In practice though, there may exist some problematic spatial regions where the exposure has not been estimated well

[*] *email:* alexandros@post.harvard.edu

(e.g., at locations far from data, which give large prediction errors), in which case this method may improve upon the plug-in estimator.

A.2 *GLS based on the exposure covariance estimate (GLS)*

As mentioned in Section 3, if one is interested in directly using the predictions from an exposure model, one should use GLS with covariance $\beta_1^2\boldsymbol{\Sigma}^* + \sigma_\epsilon^2\boldsymbol{I}_{n_y}$. This is essentially the "Krige and Regress" estimator of Madsen et al. (2008), except for the implementation differences described below. Like the fully Bayesian approach, GLS accounts for the structure in the uncertainty in $\boldsymbol{X}^*$. To apply this approach, one needs a good estimate of $\boldsymbol{\Sigma}^*$, the prediction error variance at the health locations. This estimate can be obtained from the exposure model. Since we use $\boldsymbol{S}^* = \hat{E}(\boldsymbol{X}^*|\boldsymbol{W})$, we want an estimate of $\hat{\mathrm{Var}}(\boldsymbol{X}^*|\boldsymbol{W})$. This conditional variance is difficult to compute (Booth and Hobert, 1998), so we use the approximation considered by Ruppert et al. (2003, page 103):

$$\widehat{\boldsymbol{\Sigma}^*} = \boldsymbol{C}^*\widehat{Cov}\left(\left[\begin{array}{c}\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}}\\ \widehat{\boldsymbol{b}}_{\boldsymbol{w}}\end{array}\right]|\boldsymbol{b}_w\right)(\boldsymbol{C}^*)^T + \widehat{\sigma}_\delta^2\boldsymbol{I}_{n_y}.$$

Since β_1 appears in both the mean and the covariance in the health data, one cannot use a typical GLS approach. Madsen et al. (2008) chose to use an initial estimate of β_1 for the covariance matrix. We choose to maximize the likelihood of the health model using a direct optimization routine, such as *nlm* or *optim* in R. These functions require initial values for all unknown parameters, for which an obvious choice is the plug-in estimates. To estimate the standard error of the estimated coefficients, one can use the standard likelihood approach and invert the information matrix.

We emphasize that our GLS approach maximizes the likelihood of the health model treating an estimate of the covariance structure as fixed and known. Note that this approach is not a joint maximum likelihood approach to fitting the health and exposure models simultaneously (Madsen et al., 2008). Such an approach would be the frequentist analogue of the fully Bayesian approach. We applied that joint approach as well in our simulations and found results similar to those from the fully Bayesian approach (not shown). This finding is in contrast to the simulation results presented of Madsen et al. (2008), who found that the joint ML approach yields coverage of only 45% and suggested that the conditions required for the asymptotic variance estimator do not hold in their spatial setting. One possible reason for this difference is the fact that Madsen et al. (2008) considered relatively large residual correlation among second-stage outcomes, motivated by an ecological application in which the outcome represented an environmental variable (log chloride concentration in streams). In contrast, motivated by health outcomes that are likely to be much less spatially correlated, we considered independent residuals in the second-stage outcome and did not see any evidence that this assumption was violated in the Boston birthweight data.

A.3 *WLS and GLS simulation results*

In addition to the primary methods described in Section 6, we considered the WLS and GLS approaches in our simulations. For the WLS weights we used the inverse of the prediction variances from the penalized spline model used in the plug-in approach.

Table 1 duplicates Table 1 in the paper with added rows for the WLS and GLS methods. When the exposure is relatively smooth (Scenario A), all methods, including WLS and GLS, perform reasonably well. In the other scenarios, the WLS approach provides a slightly improved fit over the plug-in estimator, mostly by decreasing the bias, but overall it performs quite similarly to the plug-in approach. The GLS approach, which accounts for heteroscedasticity and correlation in $\boldsymbol{X}^*$, performs reasonably well. Under Scenario C, it decreases the bias of the plug-in estimator substantially and attains a coverage of 86%. However, we note that in Scenario C, this approach had numerical difficulties in the estimation procedure that resulted in very small (< 0.01) estimates for the variance of the health model for 3% of the datasets. With regard to type I error reflected in Scenario D, all the approaches perform well, with the exception of the WLS approach, for which the estimated type I error is 0.094, almost double the nominal type I error of 0.05. This occurs because the WLS approach uses incorrect weights when

$\beta_1 = 0$.

[Table 1 about here.]

A.4 *WLS in the application*

We also used WLS in the birthweight application, in which it gave a seemingly untrustworthy estimate of -55.25 for the health effect coefficient, well away from the estimates from the three approaches (Section 7), with very large standard error of 52.07 and 95% confidence interval of (-157.31, 46.81). This may have occurred because the weighting strategy systematically downweights suburban locations relative to urban locations (see Figure 2 in the paper), without taking into account the spatial structure of these weights. This may be a form of selection bias.

B. Simulation details

To generate the Gaussian processes in the simulations we used the Fourier basis approximation in the spectralGP package (Paciorek, 2007) in R using the Matérn correlation function with the parameterization,

$$\frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu}\tau}{\rho\pi} \right)^{\nu} K_{\nu} \left(\frac{2\sqrt{\nu}\tau}{\rho\pi} \right),$$

with distance τ, spatial range ρ (correlation decay) and differentiability parameter $\nu > 0$. Note that ν dictates the differentiability of the surface, with large values corresponding to smoother surfaces.

The settings that we used for the simulations were:

- Scenario A:

  $\boldsymbol{g} \sim N(\boldsymbol{0}, \boldsymbol{R}(1.6, 1))$, $\boldsymbol{\delta} \sim N(\boldsymbol{0}, \sigma_\delta^2 \boldsymbol{I}_{82})$, $\sigma_\delta^2 = 0.1^2$, $\sigma_\epsilon^2 = 0.8^2$

- Scenario B:

  $\boldsymbol{g} \sim N(\boldsymbol{0}, \boldsymbol{R}(0.3, 2))$, $\boldsymbol{\delta} \sim N(\boldsymbol{0}, \sigma_\delta^2 \boldsymbol{I}_{82})$, $\sigma_\delta^2 = 0.2^2$, $\sigma_\epsilon^2 = 0.8^2$

- Scenario C:

  $\boldsymbol{g} \sim N(\boldsymbol{0}, \boldsymbol{R}(0.3, 0.5))$, $\boldsymbol{\delta} \sim N(\boldsymbol{0}, \sigma_\delta^2 \boldsymbol{I}_{82})$, $\sigma_\delta^2 = 0.2^2$, $\sigma_\epsilon^2 = 0.8^2$

- Scenario D:

  $\boldsymbol{g} \sim N(\mathbf{0}, \boldsymbol{R}(0.3, 0.5))$, $\boldsymbol{\delta} \sim N(\mathbf{0}, \sigma_\delta^2 \boldsymbol{I}_{82})$, $\sigma_\delta^2 = 0.2^2$, $\sigma_\epsilon^2 = 0.8^2$ but we generated health data using $\boldsymbol{Y}^* \sim N(\mathbf{0}, \sigma_\epsilon^2 \boldsymbol{I}_{n_y})$

For Scenarios C and D, $\nu = 0.5$, giving the exponential correlation function, which corresponds to Gaussian processes with continuous, but not differentiable sample paths.

C. Mixed model spatial smoothing

In our simulations and application, we spatially smooth exposure using a mixed model representation of penalized regression splines (Ruppert et al., 2003). This approach is simple to implement, has low computational cost and is widely applicable. Consider the simple nonparametric regression model,

$$W_i = f(\boldsymbol{geog}_i) + U_i,\ 1 \le i \le n,\ \ U_i \sim N(0, \sigma_u^2),$$

where $\boldsymbol{geog}_i = (longitude, latitude)_i$. A mixed model representation of penalized regression splines for $f(\cdot)$ is:

$$f(\cdot) \equiv \boldsymbol{X} \equiv \boldsymbol{C}\boldsymbol{z}, \tag{1}$$

for a choice of basis functions and appropriate representation in terms of the parameters, $\boldsymbol{z}$ (Ruppert et al., 2003). The vector $\boldsymbol{z}$ consists of a fixed effects vector $\boldsymbol{\beta}$ of length p and random effects $\boldsymbol{b}$, with $b_i \sim N(0, \sigma_b^2)$, $i = 1, 2, ..., K$, where K is the number of knots and $\boldsymbol{C}$ is the corresponding design matrix. We use the thin plate spline generalized covariance to construct $\boldsymbol{C}$. Let $\tilde{\boldsymbol{z}} = (\frac{1}{\sigma_u^2}\boldsymbol{C}^T\boldsymbol{C} + \boldsymbol{B})^{-1}\frac{1}{\sigma_u^2}\boldsymbol{C}^T\boldsymbol{W}$ be the best linear unbiased predictor (BLUP) for $\boldsymbol{z}$, where

$$\boldsymbol{B} = \left[\begin{array}{cc} \mathbf{0}_{p\times p} & \mathbf{0}_{p\times K} \\ \mathbf{0}_{K\times p} & \frac{1}{\sigma_b^2}\mathbf{I}_K \end{array} \right].$$

Then the BLUP for $\boldsymbol{X}^*$, for known σ_u^2 and σ_b^2, is $\boldsymbol{S}^* \equiv \hat{E}(\boldsymbol{X}^*|\boldsymbol{W}) = \boldsymbol{C}^*\tilde{\boldsymbol{z}} = \boldsymbol{C}^*(\frac{1}{\sigma_u^2}\boldsymbol{C}^T\boldsymbol{C} + \boldsymbol{B})^{-1}\frac{1}{\sigma_u^2}\boldsymbol{C}^T\boldsymbol{W}$, which is a weighted average of the observed data $\boldsymbol{W}$. Note that $\boldsymbol{C}^*$ is the design matrix that corresponds to $\boldsymbol{X}^*$, for the same choice of basis functions and knots used in (1). The BLUP conditions on the available information, as in regression calibration, so that in this modeling framework the true covariate X_i^* is centered around around its BLUP, S_i^*. As in the Gaussian process framework, the smoothing reverses the conditioning, producing a Berkson structure, rather than the classical measurement error structure (Section 3). In implementation of the Bayesian approaches in the simulations, we used a Bayesian version of the mixed model representation with priors on the regression coefficients and variance components as described in Section 6.

References

Booth, J. G. and Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association* **93**, 262–272.

Kunzli, N., Jerrett, M., Mack, W. J., Beckerman, B., LaBree, L., Gilliland, F., Thomas, D., Peters, J. and Hodis, H. N. (2005). Ambient air pollution and atherosclerosis in Los Angeles. *Environmental Health Perspectives* **113**, 201–206.

Madsen, L., Ruppert, D. and Altman, N. S. (2008). Regression with spatially misaligned data. *Environmetrics* **19**, 453–467.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression.* Cambridge, UK: Cambridge University Press.

Scenario	Method	Bias	E(se(β_1))	sd($\widehat{\beta_1}$)	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
	WLS	0.005	0.110	0.124	0.015	91.6
	Exposure simulation	-0.068	0.118	0.119	0.019	91.2
	GLS	0.005	0.110	0.120	0.014	93.2
	RC-OOS	0.006	0.122	0.122	0.015	96.4
	Fully Bayesian	0.002	0.109	0.122	0.015	92.8
	Two-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
	WLS	-0.049	0.089	0.135	0.021	79.2
	Exposure simulation	-0.254	0.116	0.126	0.080	42.2
	GLS	-0.022	0.103	0.144	0.021	82.4
	RC-OOS	0.036	0.197	0.251	0.064	95.6
	Fully Bayesian	0.011	0.107	0.151	0.023	86.4
	Two-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
	WLS	-0.096	0.130	0.204	0.050	72.0
	Exposure simulation	-0.591	0.141	0.146	0.371	0.4
	GLS	-0.020	0.169	0.215	0.047	85.6
	RC-OOS*	0.039	0.340	0.367	0.136	92.6
	Fully Bayesian	0.029	0.155	0.177	0.032	93.0
	Two-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
	WLS	-0.002	0.072	0.084	0.007	90.6
	Exposure simulation	0.000	0.068	0.054	0.003	98.8
	GLS	0.001	0.066	0.066	0.004	96.4
	RC-OOS	0.001	0.111	0.115	0.013	95.6
	Fully Bayesian	0.000	0.159	0.140	0.019	94.0
	Two-stage Bayes	0.000	0.148	0.135	0.018	94.4

Table 1

Results of simulation study for $\widehat{\beta}_1$: Bias, average model-based standard error, Monte Carlo standard deviation, MSE, and coverage of 95% confidence or credible intervals, over 500 simulations, for Scenarios A-D. *One simulation with anomalous estimate omitted.