Relating Health Outcomes to Environmental Factors: Accounting for Prediction Error in Environmental Exposure

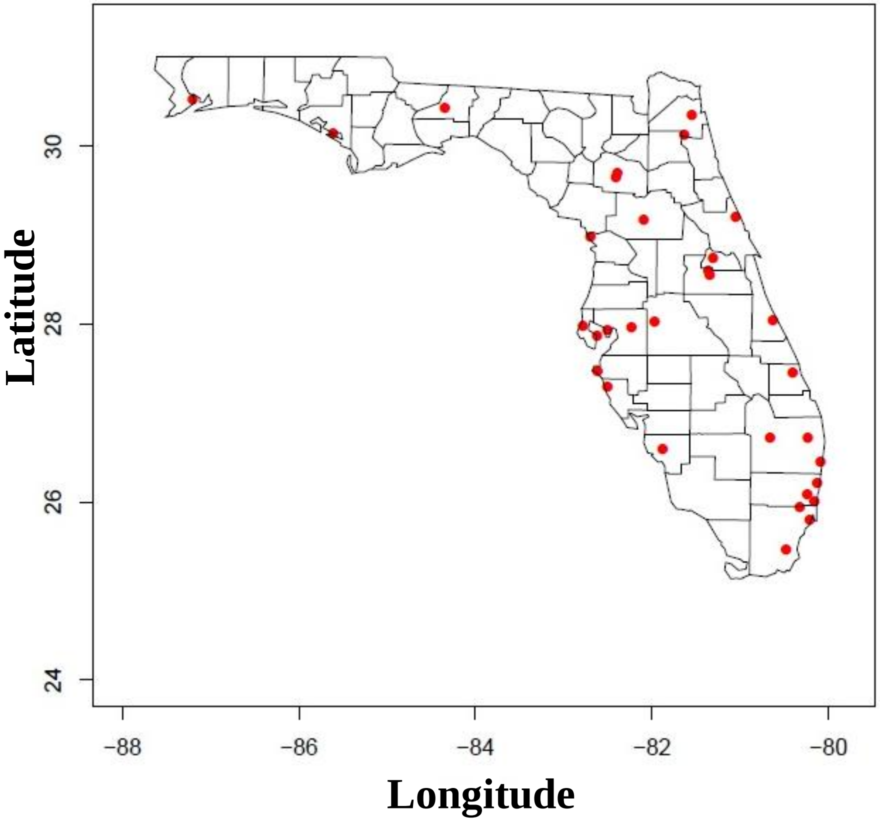
Linda J. Young
University of Florida
Gainesville FL USA

EMR 2011
May, 2011

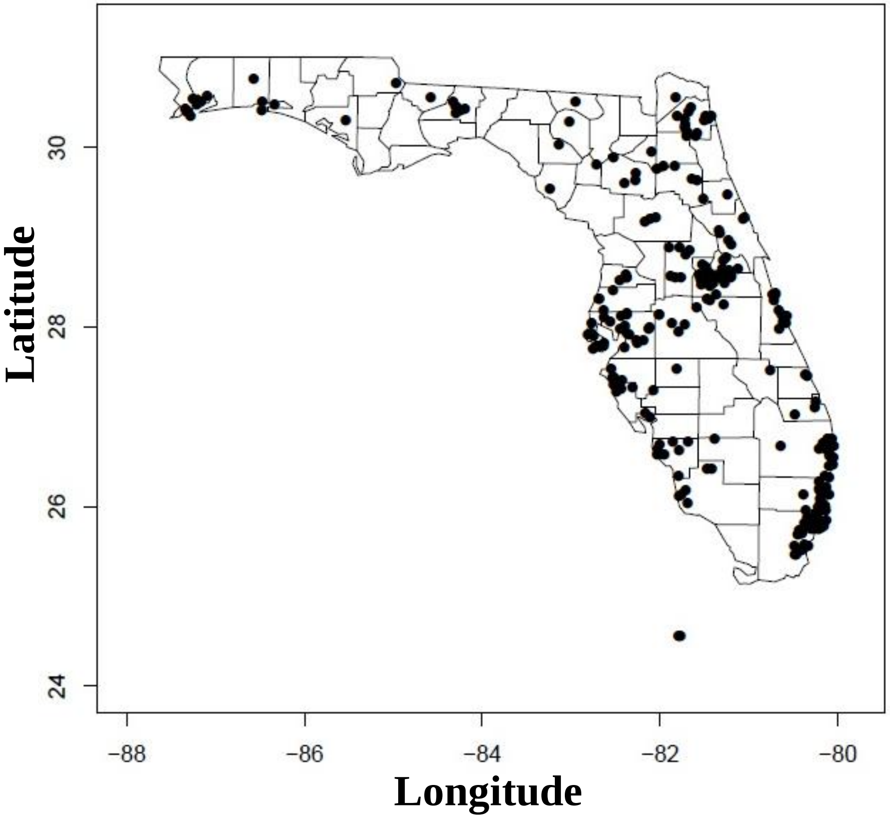
Joint Work with Kenneth K. Lopiano and Carol A. Gotway

Purpose of Study: What is the association between health outcomes and environmental exposure?

Exposure observed at 32 monitor locations



Health outcomes observed on different geographical units



Regression Relating Health Outcome to Environmental Exposure

Consider the simple linear regression

$$\mathbf{y}(\mathbf{s}_u) = \beta_0 + \beta_1 \mathbf{x}(\mathbf{s}_u) + \mathbf{e}(\mathbf{s}_u)$$

where $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_e)$ and

$$\mathbf{x}(\mathbf{s}) \sim \mathrm{N}(C(\mathbf{s})\xi, \boldsymbol{\Sigma}_{\mathbf{x}}) \;\; \forall \mathbf{s} \in \mathrm{D} \subset \Re^2$$

Note:

$\mathbf{y}(\mathbf{s}_u)$ is observed health outcome

$\mathbf{x}(\mathbf{s}_u)$ is unobserved exposure

$\mathbf{x}(\mathbf{s}_o)$ is observed exposure

Classical Measurement Error

Suppose that a model is used to predict exposure at the points of observed health outcomes. Further assume that the model provides unbiased predictions with normally distributed measurement error; that is,

$$\tilde{\mathbf{x}}(\mathbf{s}_u) = \mathbf{x}(\mathbf{s}_u) + \mathbf{w}$$

where $\mathbf{w} \sim \mathrm{N}(\mathbf{0}, \sigma_w^2\mathbf{I})$. This is **classical measurement error**. The predicted exposure is more variable than the true exposure.

Ignoring Prediction Error: Model and Regress

Ordinary Least Squares:

$$\hat{\boldsymbol{\beta}}_{\mathrm{M}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

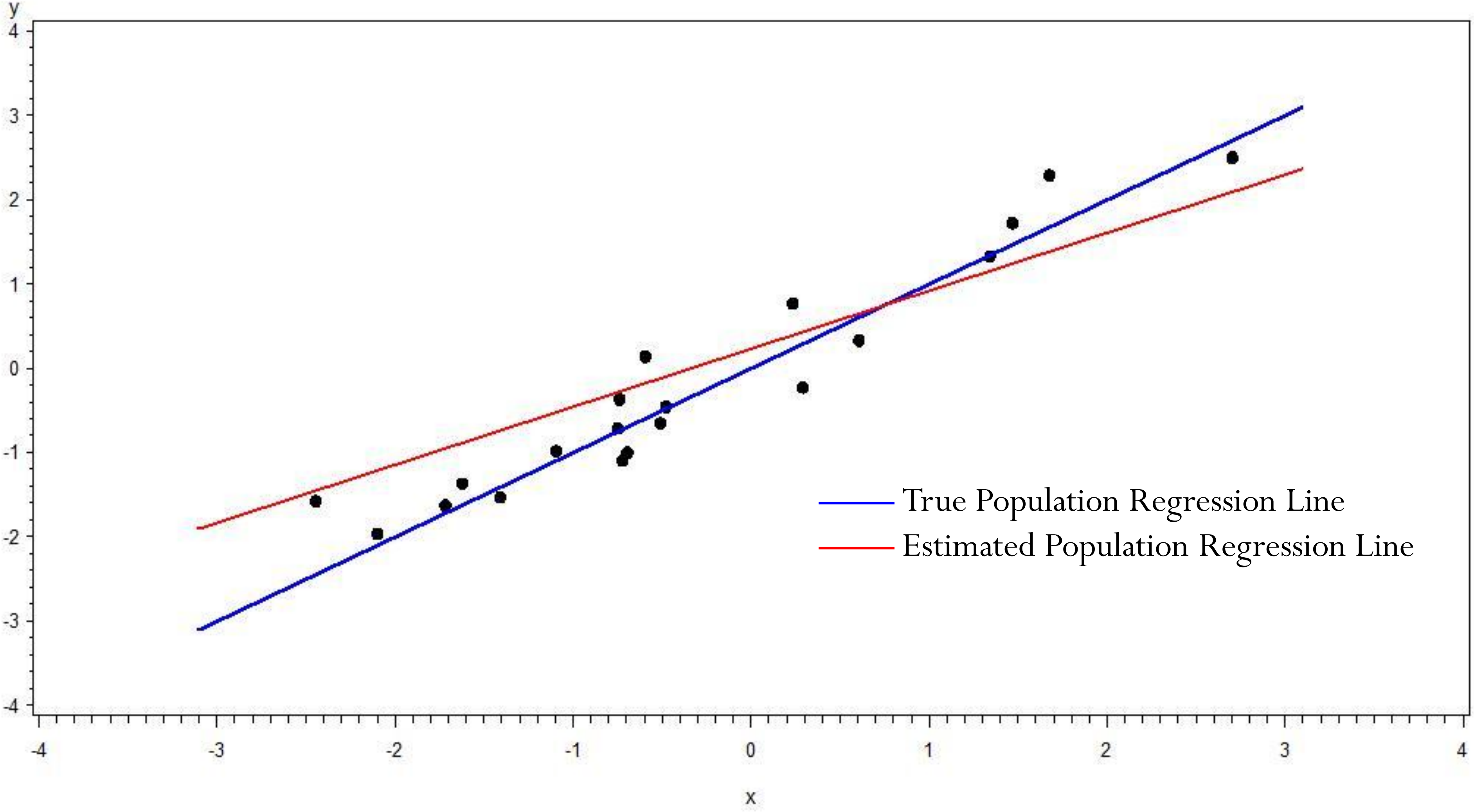
where

$$\mathbf{X} = (\mathbf{1}_{n\times 1} \quad \tilde{\mathbf{x}}(\mathbf{s}))$$

$\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ is

- Biased estimator of $\boldsymbol{\beta}$
- Uncertainty associated with predicting exposure results in standard errors being under-estimated

Classical Measurement Error



Berkson Error

Suppose kriging is used to predict exposure at the points of observed health outcomes.

$$\mathbf{x}(\mathbf{s}_u)\,|\,\mathbf{x}(\mathbf{s}_0) \sim \mathrm{N}(\boldsymbol{\mu}_k(\mathbf{s}_u), \boldsymbol{\Sigma}_k(\mathbf{s}_u))$$

$$\boldsymbol{\mu}_k(\mathbf{s}_u) = \mathbf{C}(\mathbf{s}_u)\boldsymbol{\xi} - \boldsymbol{\Sigma}_{uo}\boldsymbol{\Sigma}_{oo}^{-1}(\mathbf{x}(\mathbf{s}_o) - \mathbf{C}(\mathbf{s}_o)\boldsymbol{\xi})$$

$$\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_{uu} - \boldsymbol{\Sigma}_{u0}\boldsymbol{\Sigma}_{00}^{-1}\boldsymbol{\Sigma}_{ou}$$

where $\boldsymbol{\xi}$ is a known mean parameters of $\mathbf{x}$

$\boldsymbol{\Sigma}_{uu}$ is the known var-cov matrix among unobserved locations

$\boldsymbol{\Sigma}_{oo}$ is the known var-cov matrix among observed locations

$\boldsymbol{\Sigma}_{u0}$ is the known var-cov matrix among observed and unobserved locations

That is, $\mathbf{x}(\mathbf{s}_u)\,|\,\mathbf{x}(\mathbf{s}_0) = \boldsymbol{\mu}_k(\mathbf{s}_u) + \mathbf{v}, \quad \mathbf{v} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_k)$

Berkson Error

Using the predicted exposure, $\hat{\mathbf{x}}(\mathbf{s}_u) = \boldsymbol{\mu}_k(\mathbf{s}_u)$, results in a smoother surface than the true exposure $\mathbf{x}(\mathbf{s}_u)$; that is,

$$\mathbf{x}(\mathbf{s}_u) = \boldsymbol{\mu}_k(\mathbf{s}_u) + \mathbf{v} = \hat{\mathbf{x}}(\mathbf{s}_u) + \mathbf{v}$$

Thus,

$$\begin{aligned} \mathbf{y}(\mathbf{s}_u) \mid \mathbf{x}(\mathbf{s}_o) &= \beta_0 \mathbf{1}_{n\times 1} + \beta_1(\boldsymbol{\mu}_k(\mathbf{s}_u) + \mathbf{v}) + \mathbf{e} \\ &= \beta_0 \mathbf{1}_{n\times 1} + \beta_1 \boldsymbol{\mu}_k(\mathbf{s}_u) + (\beta_1 \mathbf{v} + \mathbf{e}) \\ &= \beta_0 \mathbf{1}_{n\times 1} + \beta_1 \boldsymbol{\mu}_k(\mathbf{s}_u) + \boldsymbol{\eta} \end{aligned}$$

where $\boldsymbol{\eta} = \beta_1 \mathbf{v} + \mathbf{e}$. The error $\mathbf{v}$ associated with the prediction of exposure is known as Berkson error.

Ignoring Prediction Error: Krige and Regress

Ordinary Least Squares:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

where

$$\mathbf{X} = (\mathbf{1}_{n\times 1} \quad \boldsymbol{\mu}_k(\mathbf{s}))$$

$\hat{\boldsymbol{\beta}}$ is

- Unbiased estimator of $\boldsymbol{\beta}$
- Uncertainty associated with predicting exposure results in standard errors being under-estimated

How does one account for the additional uncertainty induced by using kriging predictions in linear regression models?

Prior Work

- Hierarchical Bayesian Models

  Mugglin and Carlin, 1998; Mugglin, et al. 1999; Gelfand, et al. 2001; Zhu, et al. 2003

- Adjusted Krige and Regress Method

  Madsen, et al. 2009

- Bootstrap Methods

  Szpiro, et al. 2010

Comparison of Existing Methods

Gryparis, et al. 2009 and Lopiano, et al. 2010

Comparison of Existing Methods

Gryparis, et al. 2009 and Lopiano, et al. 2010

Conclusion: Existing frequentist approaches do not provide unbiased estimates of uncertainty

Goal: Develop an easy to implement frequentist approach

Developing an Estimator

Recall:
$$\mathbf{y}(\mathbf{s}_u) \mid \mathbf{x}(\mathbf{s}_o) = \beta_0 \mathbf{1}_{n\times 1} + \beta_1 \boldsymbol{\mu}_k(\mathbf{s}_u) + \boldsymbol{\eta}$$

$$\boldsymbol{\eta} = \beta_1 \mathbf{v} + \mathbf{e}$$

Note:
$$\mathrm{E}[\mathbf{y}(\mathbf{s}_u) \mid \mathbf{x}(\mathbf{s}_o)] = \beta_0 \mathbf{1}_{n\times 1} + \beta_1 \boldsymbol{\mu}_k(\mathbf{s}_u)$$

$$\mathrm{V}[\mathbf{y}(\mathbf{s}_u) \mid \mathbf{x}(\mathbf{s}_o)] = \beta_1^2 \boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_e$$

Thus,

$$\mathbf{y}(\mathbf{s}_u) \mid \mathbf{x}(\mathbf{s}_o) \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta} = \beta_0 \mathbf{1}_{n\times 1} + \beta_1 \boldsymbol{\mu}_k(\mathbf{s}_u), \Sigma_{\mathbf{y}} = \beta_1^2 \boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_{\mathbf{e}})$$

Developing an Estimator

$$\mathbf{y}(\mathbf{s}_u) \mid \mathbf{x}(\mathbf{s}_o) \sim \mathrm{N}(\mathbf{X\beta} = \beta_0 \mathbf{1}_{n\times 1} + \beta_1 \boldsymbol{\mu}_k(\mathbf{s}_u), \boldsymbol{\Sigma}_{\mathbf{y}} = \beta_1^2 \boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_{\mathbf{e}})$$

$\boldsymbol{\mu}_k(\mathbf{s}_u)$ and $\boldsymbol{\Sigma}_k$ are known

The generalized least squares estimator of $\boldsymbol{\beta}$ is the best linear unbiased estimator:

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{y}$$

The variance of this estimator is

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_{GLS}) = (\mathbf{X}'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{X})^{-1}$$

An Estimated Generalized Least Squares (EGLS) Estimator

Assuming $\boldsymbol{\Sigma}_{\mathbf{e}} = \sigma^2 \mathbf{I}$, the challenge is that σ^2 is unknown so that $\boldsymbol{\Sigma}_{\mathbf{y}} = \beta_1^2 \boldsymbol{\Sigma}_k + \boldsymbol{\Sigma}_{\mathbf{e}}$ is unknown (recall $\boldsymbol{\Sigma}_k$ is assumed known)

Let $\mathbf{P} = \mathbf{X}(\mathbf{s}_u)(\mathbf{X}'(\mathbf{s}_u)\mathbf{X}(\mathbf{s}_u))^{-1}\mathbf{X}'(\mathbf{s}_u)$. Then

$$\hat{Q} = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{P})\mathbf{y} - \text{trace}((\mathbf{I}-\mathbf{P})(\hat{\beta}^2_{1,OLS}\boldsymbol{\Sigma}_{\mathbf{y}})}{n-p}$$

is an unbiased estimator of σ^2.

Thus, $\hat{\Sigma}_{\mathbf{y}} = \hat{\beta}^2_{1,OLS}\boldsymbol{\Sigma}_k + \hat{Q}\mathbf{I}$ is an unbiased estimator of $\Sigma_{\mathbf{y}}$

An Estimated Generalized Least Squares (EGLS) Estimator

The EGLS estimator is

$$\hat{\boldsymbol{\beta}}_{EGLS} = (\mathbf{X}'(\mathbf{s}_u)\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}^{-1}\mathbf{X}(\mathbf{s}_u))^{-1}\mathbf{X}'(\mathbf{s}_u)\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}^{-1}\mathbf{y}(\mathbf{s}_u)$$

with variance

$$\text{var}(\hat{\boldsymbol{\beta}}_{EGLS}) = (\mathbf{X}'(\mathbf{s}_u)\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}^{-1}\mathbf{X}(\mathbf{s}_u))^{-1}$$

where

$$\hat{\Sigma}_{\mathbf{y}} = \hat{\beta}_{1,OLS}^2\boldsymbol{\Sigma}_k + \hat{Q}\mathbf{I}$$

Iteration Process for Final Estimate

1. Update

$$\hat{Q} = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{P})\mathbf{y} - \mathrm{trace}((\mathbf{I}-\mathbf{P})(\hat{\beta}^2_{1,EGLS}\boldsymbol{\Sigma}_{\mathbf{y}})}{n-p}$$

   where

$$\mathbf{P} = \mathbf{X}(\mathbf{s}_u)(\mathbf{X}'(\mathbf{s}_u)\hat{\boldsymbol{\Sigma}}_y^{-1}\mathbf{X}(\mathbf{s}_u))^{-1}\mathbf{X}'(\mathbf{s}_u)\hat{\boldsymbol{\Sigma}}_y^{-1}$$

   and

$$\hat{\Sigma}_{\mathbf{y}} = \hat{\beta}^2_{1,EGLS}\boldsymbol{\Sigma}_k + \hat{Q}\mathbf{I},$$

2. Update

$$\hat{\boldsymbol{\beta}}_{EGLS} = (\mathbf{X}'(\mathbf{s}_u)\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}^{-1}\mathbf{X}(\mathbf{s}_u))^{-1}\mathbf{X}'(\mathbf{s}_u)\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}^{-1}\mathbf{y}(\mathbf{s}_u)$$

   and

$$\mathrm{var}(\hat{\boldsymbol{\beta}}_{EGLS}) = (\mathbf{X}'(\mathbf{s}_u)\hat{\boldsymbol{\Sigma}}_{\mathbf{y}}^{-1}\mathbf{X}(\mathbf{s}_u))^{-1}$$

3. Iterate (1) and (2) until changes in estimates are sufficiently small.

Parameters Associated With X

Suppose $\boldsymbol{\xi}$ and $\boldsymbol{\tau}_{\mathrm{X}}$, the mean and covariance parameters, respectively, associated with **X**, are unknown.

Then,

$$\tilde{\boldsymbol{\xi}} = (\mathbf{C}'(\mathbf{s}_0)\hat{\boldsymbol{\Sigma}}_{00}^{-1}\mathbf{C}(\mathbf{s}_0))^{-1}\mathbf{C}'(\mathbf{s}_0)\hat{\boldsymbol{\Sigma}}_{00}^{-1}\mathbf{X}(\mathbf{s}_o)$$

and the estimated variance of $\mathbf{x}(\mathbf{s}_u) \mid \mathbf{x}(\mathbf{s}_0)$ (the kriging variance) is

$$\hat{\boldsymbol{\Sigma}}_{uu} - \hat{\boldsymbol{\Sigma}}_{uo}\hat{\boldsymbol{\Sigma}}_{oo}^{-1}\hat{\boldsymbol{\Sigma}}_{ou} +$$

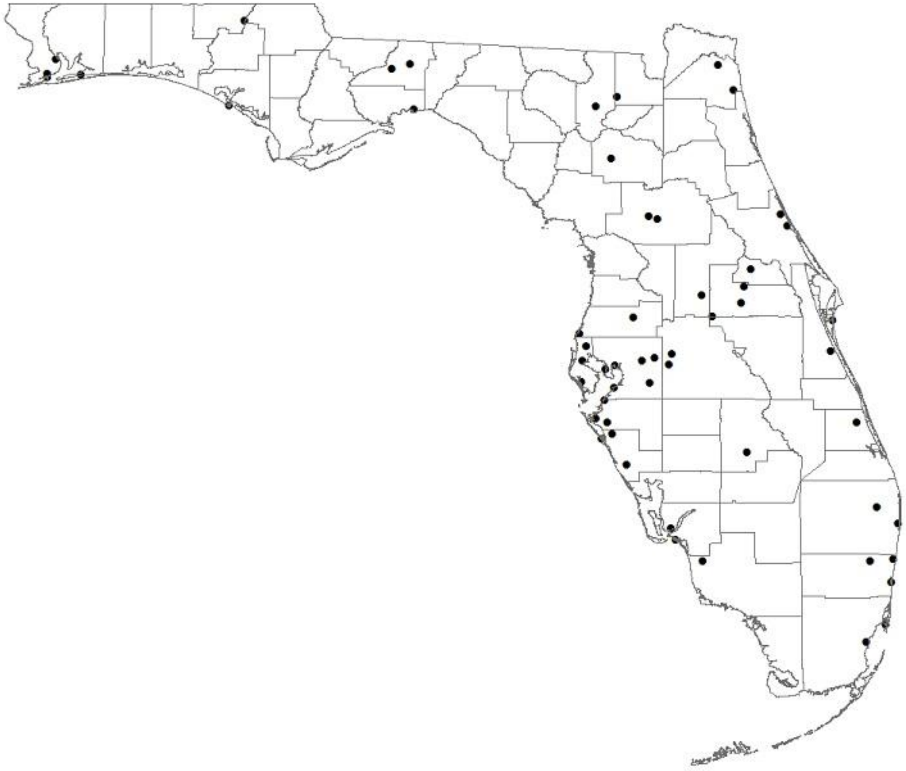
$$(\mathrm{C}(\mathrm{s}_{\mathrm{u}}) - \hat{\boldsymbol{\Sigma}}_{\mathrm{uo}}\hat{\boldsymbol{\Sigma}}_{\mathrm{oo}}^{-1}\,\mathbf{C}(\mathbf{s}_{\mathrm{o}}))(\mathbf{C}(\mathbf{s}_o)\hat{\boldsymbol{\Sigma}}_{00}^{-1}\mathbf{C}(\mathbf{s}_o))^{-1}(\mathbf{C}(\mathbf{s}_{\mathrm{u}}) - \hat{\boldsymbol{\Sigma}}_{\mathrm{uo}}\hat{\boldsymbol{\Sigma}}_{\mathrm{oo}}^{-1}\,\mathbf{C}(\mathbf{s}_{\mathrm{o}}))'$$

.

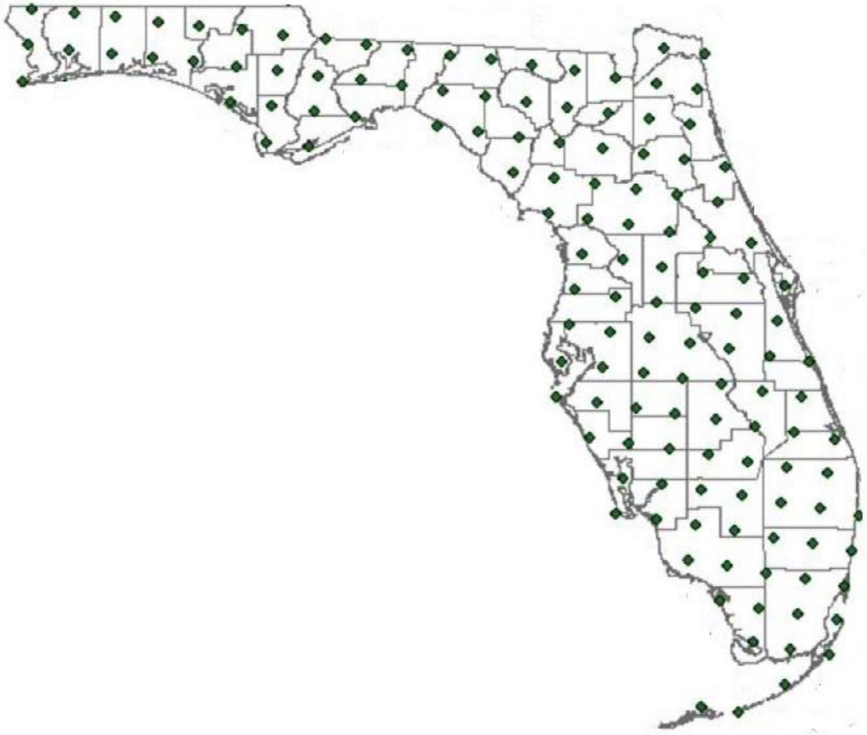
Note: No correction is being made for using the plug-in estimators of $\boldsymbol{\tau}_{\mathrm{X}}$

Simulation Study

Exposure measured at 56 monitor locations



Health outcomes observed on a grid



Simulation Study

10,000 Datasets Generated for Each Scenario

Exposure Model

- Mean
  - 40
  - -73 x 0.23 Northing – 0.00011 Northing^2
- Variance-Covariance
  - Exponential covariance structure with a scale of 70 and a range of 200

Study Results: Mean Only

True Model: $\mathbf{y}(\mathbf{s}_u) = -0.8 + 0.2\mathbf{x}(\mathbf{s}_u) + \mathbf{e}(\mathbf{s}_u)$

where $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, 2.3^2\mathbf{I})$

σ^2	Scale	Range	Mean	Iterations	$\overline{\hat{\beta}_1}$	$\sigma^2_{\hat{\beta}_1}$	$\bar{s}^2_{\hat{\beta}_1}$	Coverage
-	X	X	X	0	0.199	0.00212	0.00210	94.78
-	X	X	X	20	0.200	0.00215	0.00210	94.67
-	X	X	-	0	0.199	0.00213	0.00210	94.66
-	X	X	-	20	0.200	0.00216	0.00210	94.51
-	-	-	-	0	0.204	0.00242	0.00234	94.68
-	-	-	-	20	0.205	0.00246	0.00244	94.57

$\overline{\hat{\beta}_1}$ observed mean of $\hat{\beta}_1$

$\sigma^2_{\hat{\beta}_1}$ observed variance of $\hat{\beta}_1$

$\bar{s}^2_{\hat{\beta}_1}$ observed mean of estimated variance of $\hat{\beta}_1$

Coverage: observed coverage for nominal 95% confidence intervals

Study Results: Trend Surface

True Model: $\mathbf{y}(\mathbf{s}_u) = -0.8 + 0.2\mathbf{x}(\mathbf{s}_u) + \mathbf{e}(\mathbf{s}_u)$

where $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, 2.3^2\mathbf{I})$

σ^2	Scale	Range	Mean	Itera-tions	$\bar{\hat{\beta}}_1$	$\sigma^2_{\hat{\beta}_1}$	$\bar{s}^2_{\hat{\beta}_1}$	Cover-age
-	X	X	X	0	0.199	0.00163	0.00161	94.49
-	X	X	X	20	0.200	0.00166	0.00162	94.47
-	X	X	-	0	0.193	0.00178	0.00168	92.65
-	X	X	-	20	0.195	0.00180	0.00160	92.97
-	-	-	-	0	0.195	0.00192	0.00175	92.62
-	-	-	-	20	0.197	0.00196	0.00177	92.92

$\bar{\hat{\beta}}_1$ observed mean of $\hat{\beta}_1$

$\sigma^2_{\hat{\beta}_1}$ observed variance of $\hat{\beta}_1$

$\bar{s}^2_{\hat{\beta}_1}$ observed mean of estimated variance of $\hat{\beta}_1$

Coverage: observed coverage for nominal 95% confidence intervals

Study Results: Trend Surface, Modeled as Constant Mean

True Model: $\mathbf{y}(\mathbf{s}_u) = -0.8 + 0.2\mathbf{x}(\mathbf{s}_u) + \mathbf{e}(\mathbf{s}_u)$
where $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, 2.3^2\mathbf{I})$

σ^2	Scale	Range	Mean	Itera-tions	$\overline{\hat{\beta}_1}$	$\sigma^2_{\hat{\beta}_1}$	$\bar{s}^2_{\hat{\beta}_1}$	Cover-age
-	X	X	-	0	0.205	0.00182	0.00193	95.75
-	X	X	-	20	0.206	0.00185	0.00191	95.58
-	-	-	-	0	0.207	0.00203	0.00209	95.53
-	-	-	-	20	0.208	0.00230	0.00208	95.38

$\overline{\hat{\beta}_1}$ observed mean of $\hat{\beta}_1$

$\sigma^2_{\hat{\beta}_1}$ observed variance of $\hat{\beta}_1$

$\bar{s}^2_{\hat{\beta}_1}$ observed mean of estimated variance of $\hat{\beta}_1$

Coverage: observed coverage for nominal 95% confidence intervals

Berkson Error and Classical Measurement Error

When kriging is used to predict environmental exposure at points at which health outcomes are observed, Berkson error occurs.

When the kriging parameters are not known, but must be estimated, classical measurement error is introduced

Estimated generalized least squares adjusts for Berkson error, but what about classical measurement error?

Accounting for Classical Measurement Error

Szipiro, et al. (2010) suggested three approaches for correcting for classical measurement error when predicting exposure and then using ordinary least squares to relate health outcomes to environmental exposure:

- Parametric bootstrap
- Parameter bootstrap
- Partial parametric bootstrap

Can these be extended to generalized least squares?

Parametric Bootstrap

The steps for the estimation process are as follows.

1. Estimate the exposure model parameters, $\boldsymbol{\xi}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$
2. Use $\hat{\mu}_{\mathbf{X}}(\mathbf{s}_u)$ to estimate model parameters $\hat{\boldsymbol{\beta}}_{EGLS}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$
3. Repeat the steps below for each j = 1, 2, …, M

   (a) Simulate a new set of observations $\mathbf{Y}_j(\mathbf{s}_u)$ and $\mathbf{X}_j(\mathbf{s}_o)$ based on the models for exposure and health outcomes using $\hat{\boldsymbol{\xi}}$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}}$, $\hat{\boldsymbol{\beta}}_{EGLS}$, and $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$.

   (b) Estimate new **X** parameters $\hat{\boldsymbol{\xi}}_j$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{X},j}$ using $X_j(\mathbf{s}_u)$.

   (c) Derive $\hat{\mu}_{\mathbf{X},j}(\mathbf{s}_u)$ using $\hat{\boldsymbol{\xi}}_j$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{X},j}$, and $X_j(\mathbf{s}_u)$

   (d) Calculate $\hat{\boldsymbol{\beta}}_{EGLS,1.j}$

4. Calculate the parametric bootstrap standard error

$$\sigma_{\hat{\beta}_{EGLS,1}} = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}\left(\hat{\beta}_{EGLS,1,j} - \frac{1}{M}\sum_{j=1}^{M}\hat{\beta}_{1.j}\right)^2}$$

Parametric Bootstrap

The method is computationally intensive

Estimating ξ and $\mathbf{\Sigma}_{\mathbf{X}}$ requires nonlinear optimization, and this is required for each of the M steps in the parametric bootstrap.

However, the sampling distribution $\hat{p}(\cdot,\cdot)$ of $\hat{\xi}$ and $\hat{\mathbf{\Sigma}}_{\mathbf{e}}$ can be estimated without much additional cost.

Parameter Bootstrap

The steps for the estimation process are as follows.

1. Estimate the exposure model parameters, $\boldsymbol{\xi}$ and $\boldsymbol{\Sigma}_{\mathbf{X}}$, and their sampling distribution $\hat{p}(\cdot,\cdot)$

2. Use $\hat{\mu}_{\mathbf{X}}(\mathbf{s}_u)$ to estimate model parameters $\hat{\boldsymbol{\beta}}_{EGLS}$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$

3. Repeat the steps below for each j = 1, 2, …, M

   (a) Simulate a new set of observations $\mathbf{Y}_j(\mathbf{s}_u)$ and $\mathbf{X}_j(\mathbf{s}_o)$ based on the models for exposure and health outcomes using $\hat{\boldsymbol{\xi}}$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{X}}$, $\hat{\boldsymbol{\beta}}_{EGLS}$, and $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$.

   (b) Sample the parameters $\hat{\boldsymbol{\xi}}_j$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{X},j}$ from the probability distribution defined by $\hat{p}(\cdot,\cdot)$

   (c) Derive $\hat{\mu}_{\mathbf{X},j}(\mathbf{s}_u)$ using $\hat{\boldsymbol{\xi}}_j$, $\hat{\boldsymbol{\Sigma}}_{\mathbf{X},j}$, and $X_j(\mathbf{s}_u)$

   (d) Calculate $\hat{\boldsymbol{\beta}}_{EGLS,1.j}$

4. Calculate the parametric bootstrap standard error

$$\sigma_{\hat{\beta}_{EGLS,1}} = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}\left(\hat{\beta}_{EGLS,1,j} - \frac{1}{M}\sum_{j=1}^{M}\hat{\beta}_{1.j}\right)^2}$$

Partial Parameter Bootstrap (PPB)

To reduce computations, the PPB was suggested

Instead of estimating ξ and $\mathbf{\Sigma}_{\mathbf{X}}$ for each bootstrap sample, the estimates from the data are used.

This ignores the classical measurement error, but accounts for Berkson error

The EGLS estimator fully accounts for Berkson error so the PPB is not considered further

Study Results: Trend Surface

True Model: $\mathbf{y}(\mathbf{s}_u) = -0.8 + 0.2\mathbf{x}(\mathbf{s}_u) + \mathbf{e}(\mathbf{s}_u)$ where $\mathbf{e} \sim \mathrm{N}(\mathbf{0}, 2.3^2\mathbf{I})$

Method	$\overline{\hat{\beta}_1}$	$\sigma^2_{\hat{\beta}_1}$	$\overline{s}^2_{\hat{\beta}_1}$	Coverage
OLS	0.208	0.00373	0.00265	90.70
OLS & Parameter Bootstrap	0.208	0.00189	0.00370	96.40
EGLS with Iteration	0.197	0.00196	0.00177	92.92
EGLS & Parametric Bootstrap	0.197	0.00220	0.00206	93.33
EGLS & Parameter Bootstrap	0.197	0.00202	0.00380	96.72

$\overline{\hat{\beta}_1}$ observed mean of $\hat{\beta}_1$

$\sigma^2_{\hat{\beta}_1}$ observed variance of $\hat{\beta}_1$

$\overline{s}^2_{\hat{\beta}_1}$ observed mean of estimated variance of $\hat{\beta}_1$

Coverage: observed coverage for nominal 95% confidence intervals

Conclusions from Simulation

- Both OLS and EGLS estimators exhibit some bias when the exposure parameters must be estimated, though the bias with EGLS is less than that with OLS
- For both OLS and EGLS, the variance of $\hat{\beta}_1$ is under-estimated, but the bias is more pronounced for OLS than for EGLS
- The parameter bootstrap results in an upwardly biased estimate of the variance of $\hat{\beta}_1$
- The parametric bootstrap reduces the bias in the estimated variance of $\hat{\beta}_1$, but it is still present
- The coverage probabilities suggest using the parametric bootstrap

Is There an Association Between $PM_{2.5}$ and Birth Weight?

To model the spatial and temporal association between birth weights and the changing levels of $PM_{2.5}$ in Florida

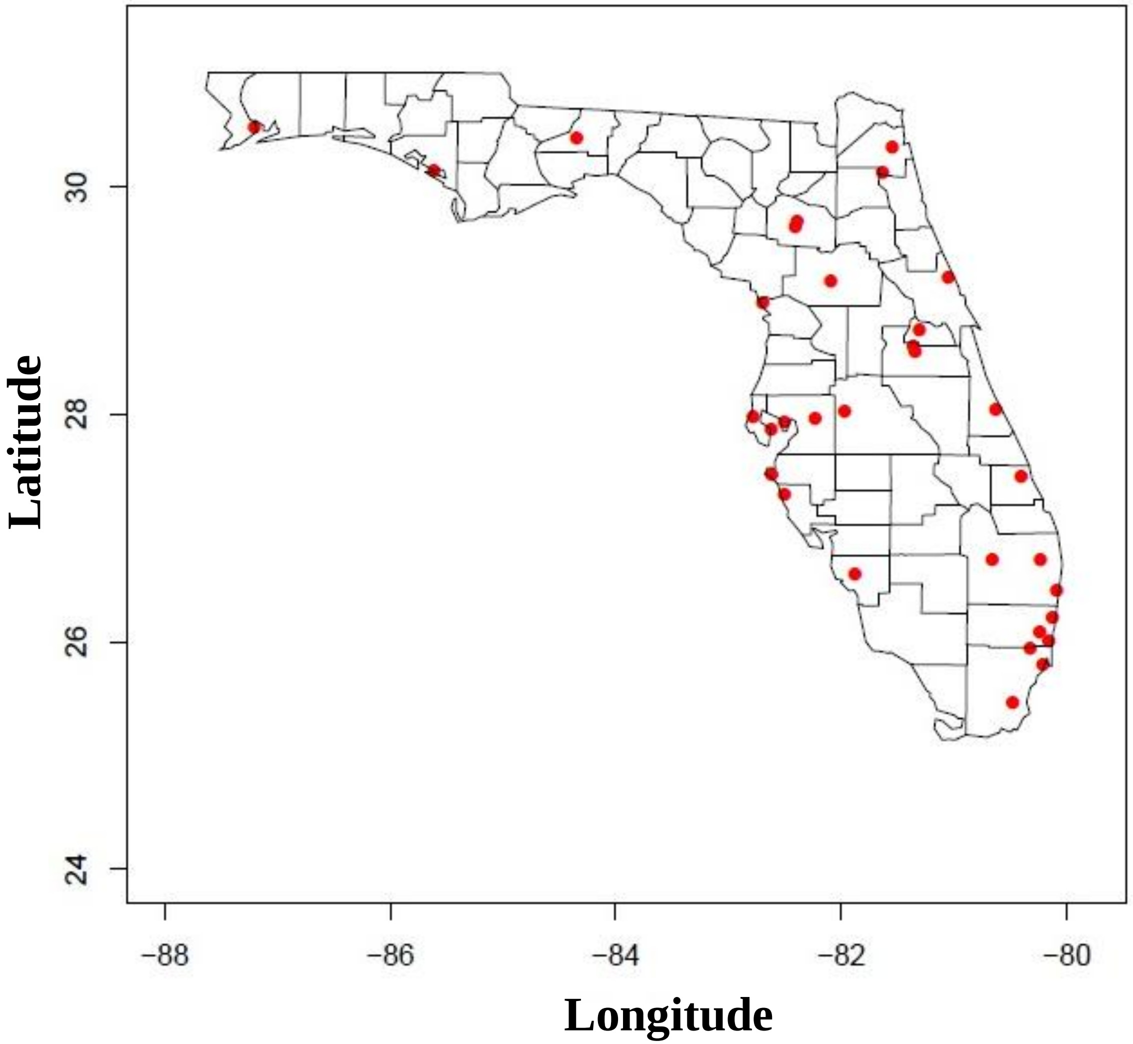
Initial focus: $PM_{2.5}$ Data from August 2005
Birth weights during first 12 hours
of September 1, 2005

$PM_{2.5}$ Exposure

EPA’s National $PM_{2.5}$ Air Quality Standards are based on the 24-hour average and annual average. The daily average $PM_{2.5}$ value is used here.

To avoid peak $PM_{2.5}$ levels being reduced by averaging over days of the month, the maximum of the daily average $PM_{2.5}$ values during a month was used as the monthly data value for a particular monitor.

Florida $PM_{2.5}$ Monitors in August 2005

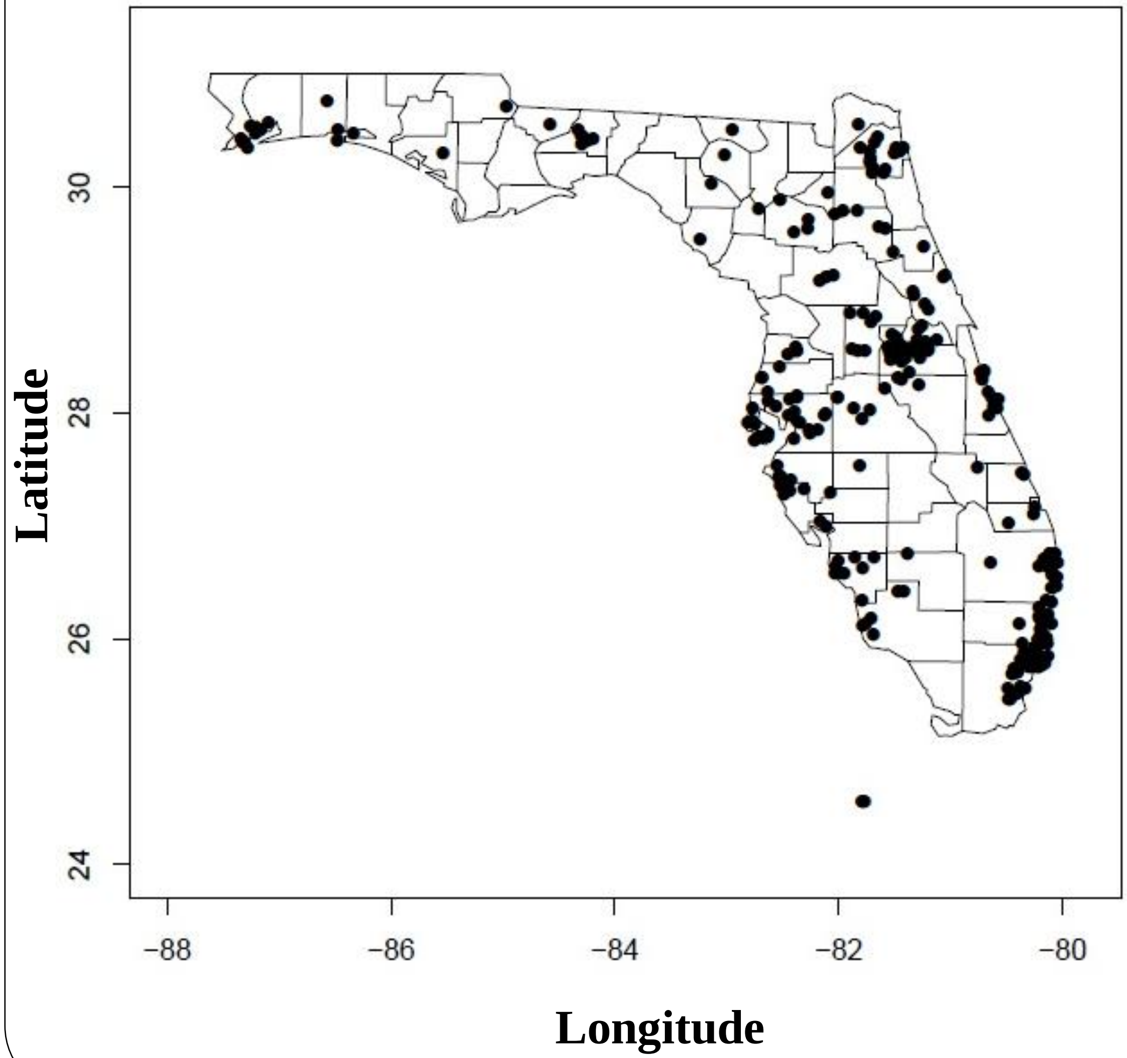


32 Monitors

Data collected by FDEP

About a 3-month lag between data collection and completion of quality assurance

Live Birth Weight Data



265 Live births during first 12 hours of September 1, 2005

Data collected by the Florida Department of Health's Office of Vital Statistics

Data sharing agreement

Geocoded addresses included in the file

Information on mother's age, ethnicity, etc.

Relating Birth Weight to PM$_{2.5}$

$$y_i = \beta_0 + \beta_1 \hat{x}_i + \mathbf{v}_i' \boldsymbol{\beta}_{\mathbf{v}} + e_i$$

where

y$_i$ = birth weight of *i*th baby

$\hat{x}_i$ is the predicted maximum PM$_{2.5}$ level for baby *i* assuming a constant mean and exponential covariance structure

$\mathbf{v}_i$' = (v_{i1}, …, v_{ik}) are covariates for baby *i*

$\beta_0, \beta_1, \boldsymbol{\beta}_{\mathbf{v}}$ are the unknown parameters

e_i is the error associated with county *i*

Suppose that the errors are assumed to be iid N(0, σ^2).

Covariates

- Gestational Age (in weeks)
- Age of Mother (in years)
- Weight gain of Mother (in pounds)
- History Factor - Diabetes/Prepregnancy - Yes/No
- History Factor - Diabetes/Gestational - Yes/No
- Tobacco Use - Yes/No
- History Factor - Hypertension/Prepregnancy - Yes/No
- History Factor - Hypertension/Gestational - Yes/No
- History Factor - Previous Preterm Birth - Yes/No
- A calculated code identifying the mother into one of 6 Hispanic origins
- A calculated code identifying the mothers race into one of 14 races
- History Factor - Mother Received Food Stamps - Yes/No
- The principal source of payment - 4 Categories
- The mothers education – 8 Categories

Methods for Relating Birth Weight to $PM_{2.5}$

- Krige and Regress
  - $PM_{2.5}$ is predicted for each residence associated with the birth of a baby
  - OLS is used to estimate the association between birth weight and $PM_{2.5}$ ignoring prediction error

- Krige and EGLS
  - $PM_{2.5}$ is predicted for each residence associated with the birth of a baby
  - EGLS is used to estimate the association between birth weight and $PM_{2.5}$ accounting prediction error

Results

MSE from OLS: 200,770

$\hat{\sigma}^2$ from EGLS: 199,402

Method	$\hat{\beta}_1$	$s^2_{\hat{\beta}_1}$
OLS	-12.67	66.92
OLS with Parameter Bootstrap	-12.67	215.90
EGLS with No Iteration	-12.75	69.06
EGLS with Iterations	-12.75	69.08
EGLS with Parametric Bootstrap	-12.75	87.54
EGLS with Parameter Bootstrap	-12.75	327.67

Conclusions

- If environmental exposure is predicted and the parameters of **X** are known, the estimate of the association between health and environmental exposure obtained through regression is unbiased, but the standard error tends to be under estimated.
- The EGLS estimator is unbiased and provides appropriate standard errors if the parameters of **X** are known.
- If the parameters of **X** are estimated, classical measurement error is introduced, and the estimate of the association between health and environmental exposure obtained through regression may be biased, and the standard errors may be biased.
- As the number of observed health outcomes increases relative to the number of observed exposures, measurement error becomes more dominant than Berkson error.

Conclusions

- Bootstrap approaches can be used to adjust the standard errors for classical measurement error, but
  - The estimated parameter may be biased
  - The standard errors remain biased
  - The coverage probability is close to the nominal level
- Ideally, the exposure and health outcomes should be modeled together instead of the stepwise approach considered here.

Conclusions

- Exposure of persons to $PM_{2.5}$ is the association of interest. Two problems:
  - ✓ Ambient $PM_{2.5}$ levels serve to approximate $PM_{2.5}$ exposure.
  - ✓ Exposure is predicted at the residence.
- Goal is on-going monitoring. Existing space-time models are not readily extendable to this setting.
- Bayesian models tend to be problem-specific and cannot readily be adapted for different variables, locations, time, etc.

Conclusions

- The process of relating health outcomes to environmental factors, from data collection through interpretation, is challenging.
- Standardized analytical approaches should be adopted if the process is to become routine.