Challenges Associated with Integrating Data from Multiple Scales to Assess Relationships

Linda J. Young[1], Carol A. Gotway[2], Kenneth K. Lopiano[1]

[1] University of Florida, Gainesville FL USA
[2] U.S. Centers for Disease Control, Atlanta GA USA

Accuracy 2010
July 22, 2010

Environmental Public Health Tracking in the United States

“CDC’s goal is to develop a tracking system that integrates data about environmental hazards and exposures with data about diseases that are possibly linked to the environment. This system will allow federal, state, and local agencies, and others to do the following:

monitor and distribute information about environmental hazards and disease trends

advance research on possible linkages between environmental hazards and disease

develop, implement, and evaluate regulatory and public health actions to prevent or control environment-related diseases.”

http://www.cdc.gov/nceh/tracking/background.htm

Purpose of This Study

To model the spatial and temporal association between myocardial infarctions (MIs) and the changing levels of ambient ozone in Florida

Initial focus: August 2005

Hospital Admission Data

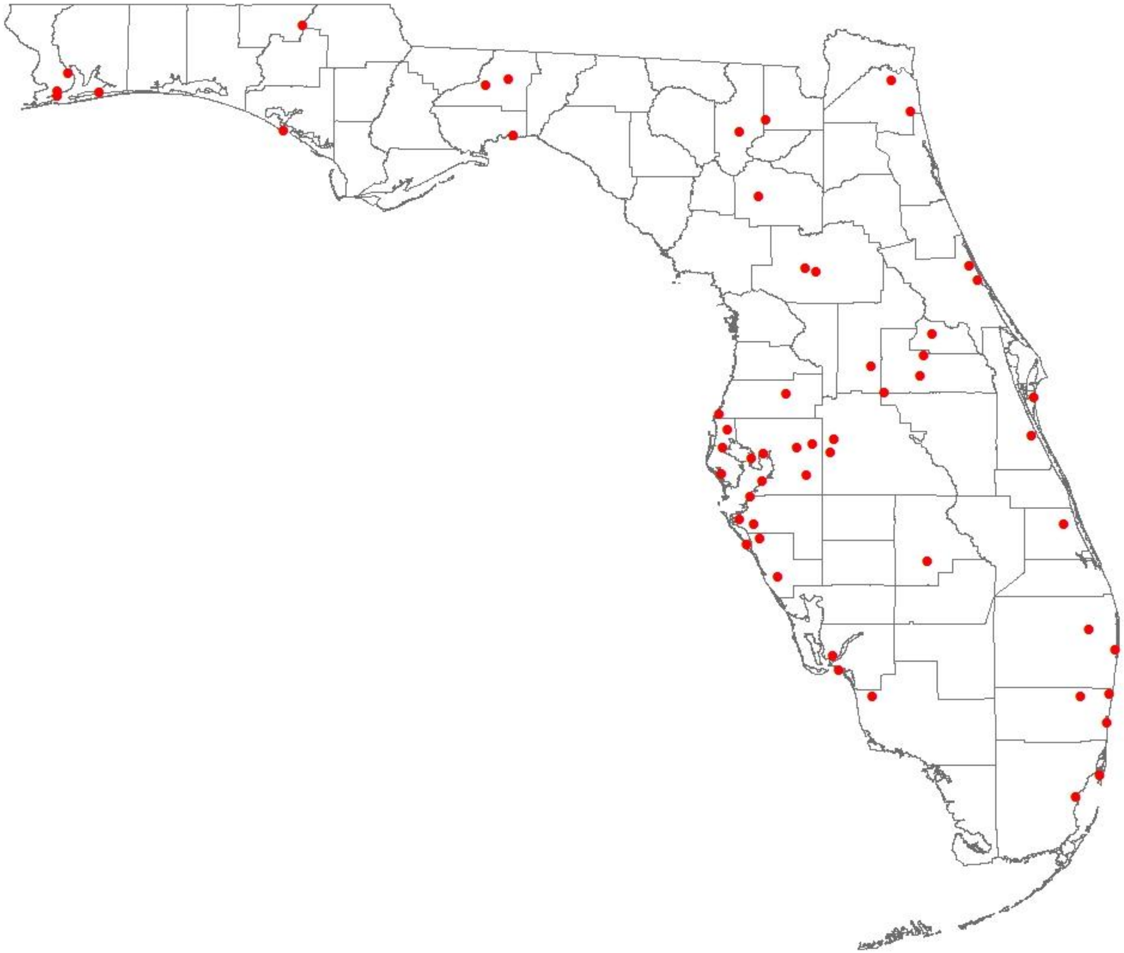
Data collected by AHCA

Data sharing agreement

Available 3 to 6 months after end of quarter

Information on patient's zip code, county, age, ethnicity, sex

Florida Ozone Monitors in August 2005



56 Monitors

Data collected by FDEP

Sometimes monitor malfunctions and data are missing for one or more days

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

Meteorological data

Population Socio-Demographic Data

Available from Census and BRFSS

Data available at various scales

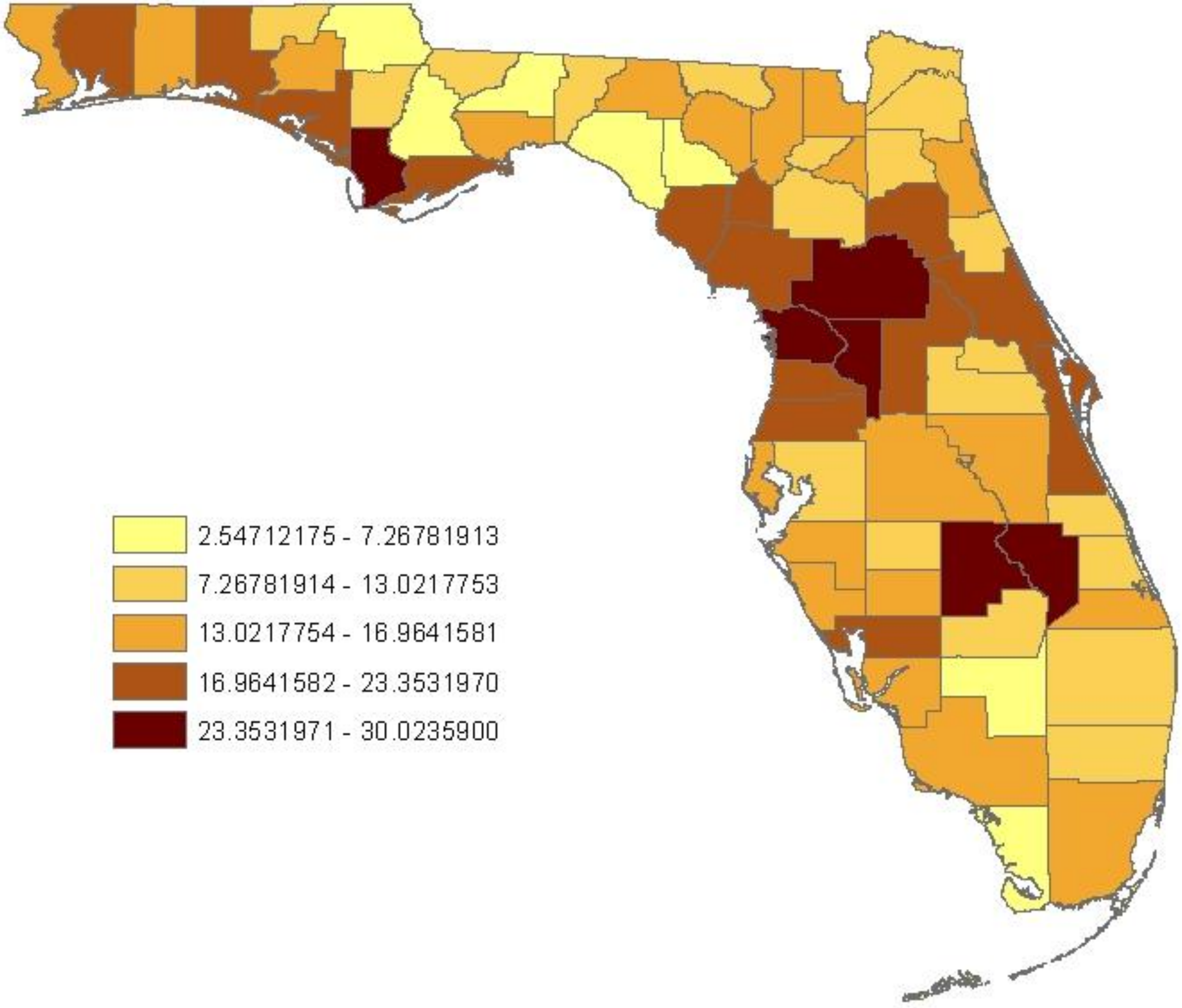
Scale of Analysis

Want the smallest possible geographical and temporal units while satisfying confidentiality requirements

Decided to analyze monthly county data

Need to link the monthly data at the county level

MI Cases Per 10,000 Population During August 2005



Indirect Standardization

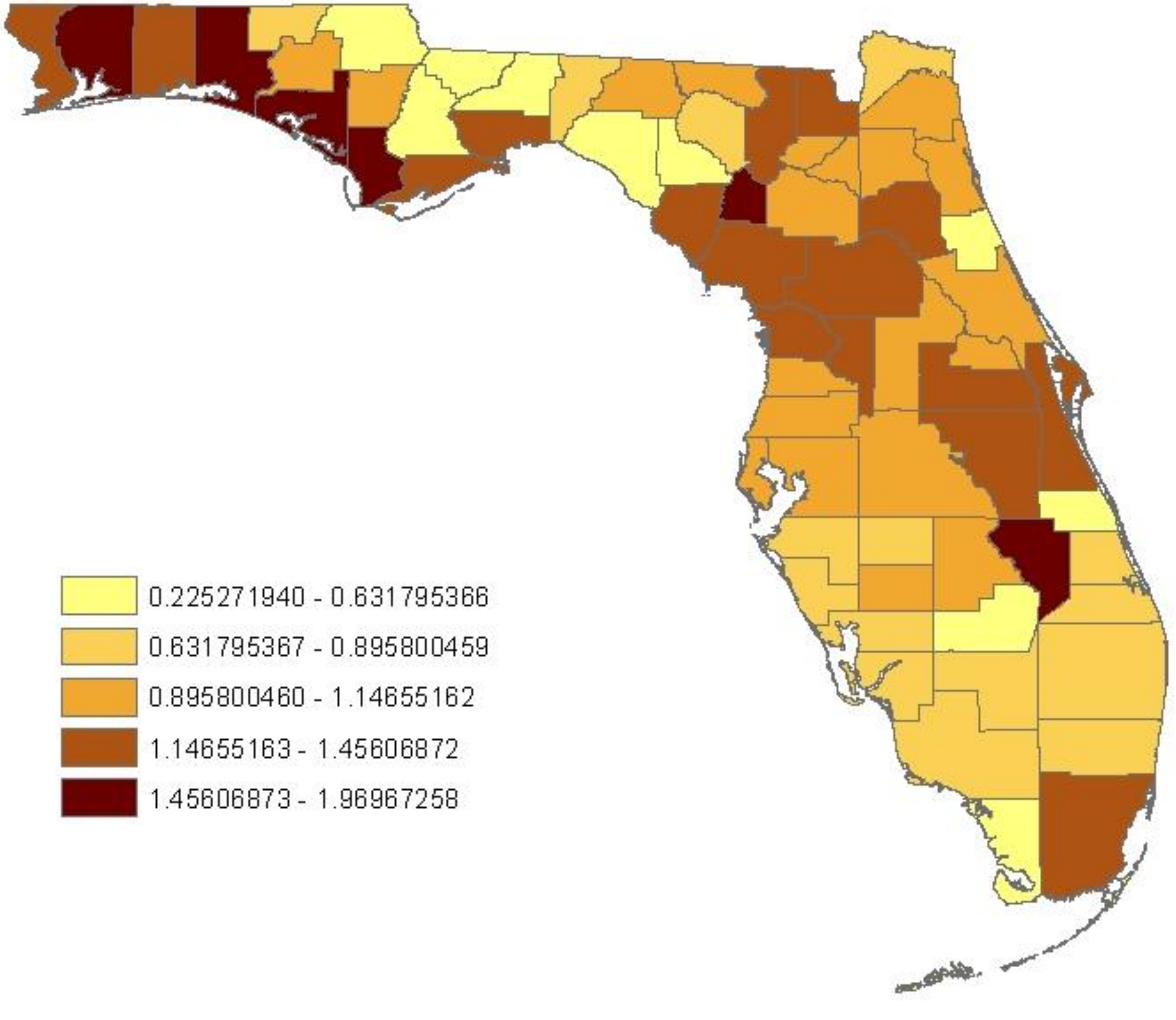
Obtain Standardized Event Ratio (SER)

Adjust for

- Age (aged ≤45, 45–55, 55–65, and >65 years)
- Sex (Female, Male)
- Ethnicity (Black, White, Other)

Uses Florida as the Standard Population

MI SER for August 2005



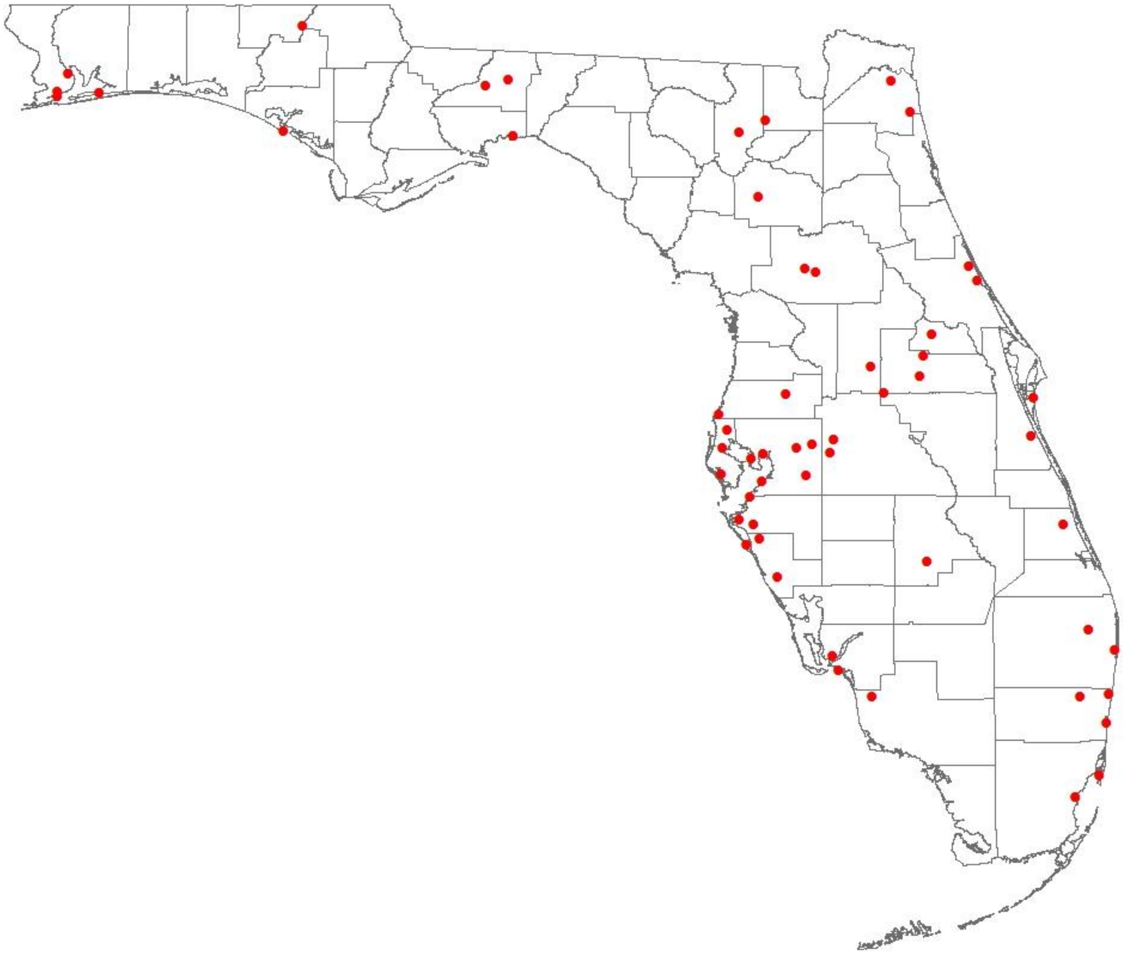
Ozone Exposure

EPA’s National Ambient Air Quality Standards are based on the maximum 8-hour average each day. The daily average ozone value is used here.

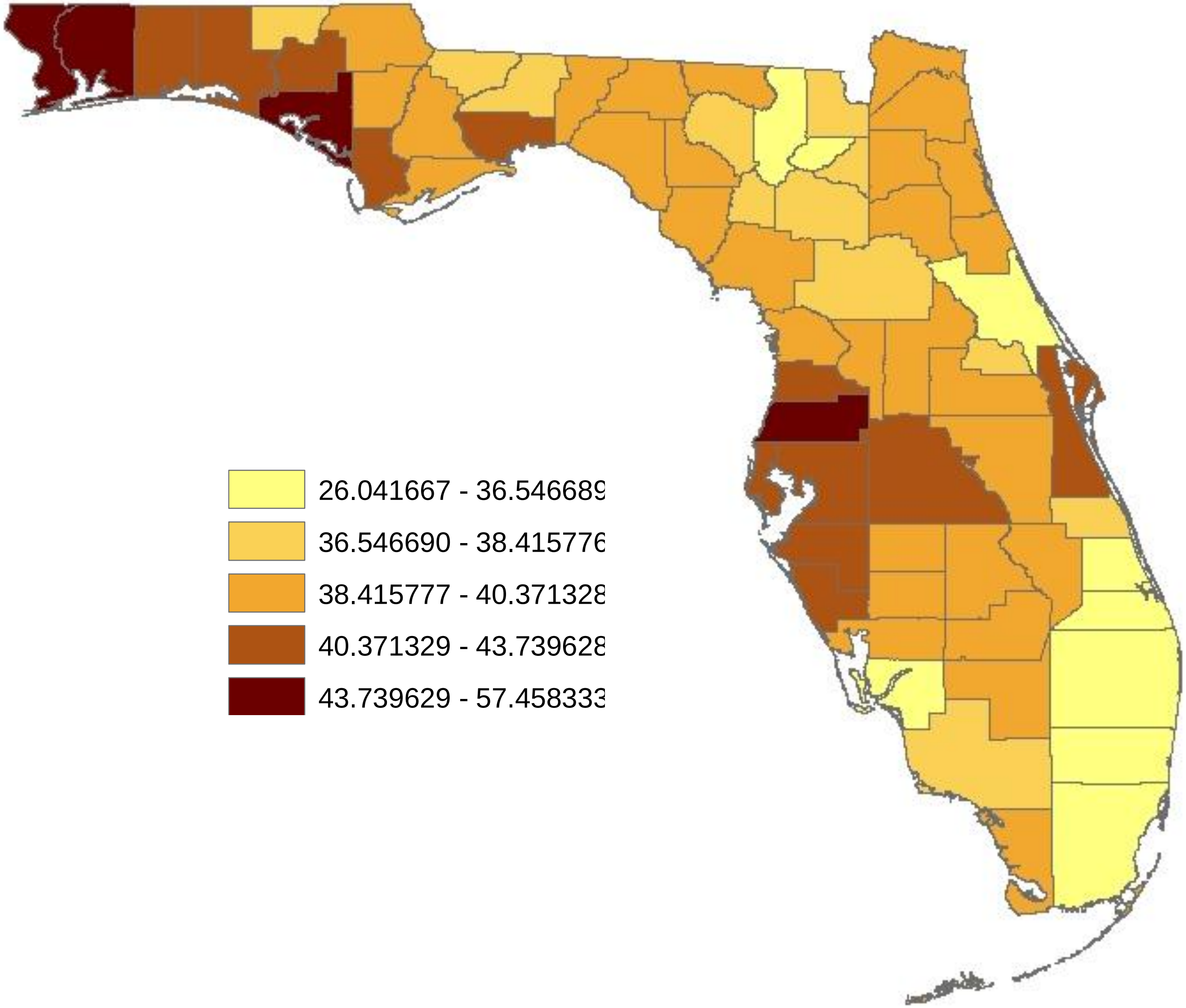
Because ozone levels decline at night, daytime peaks might not be evident in daily averages.

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

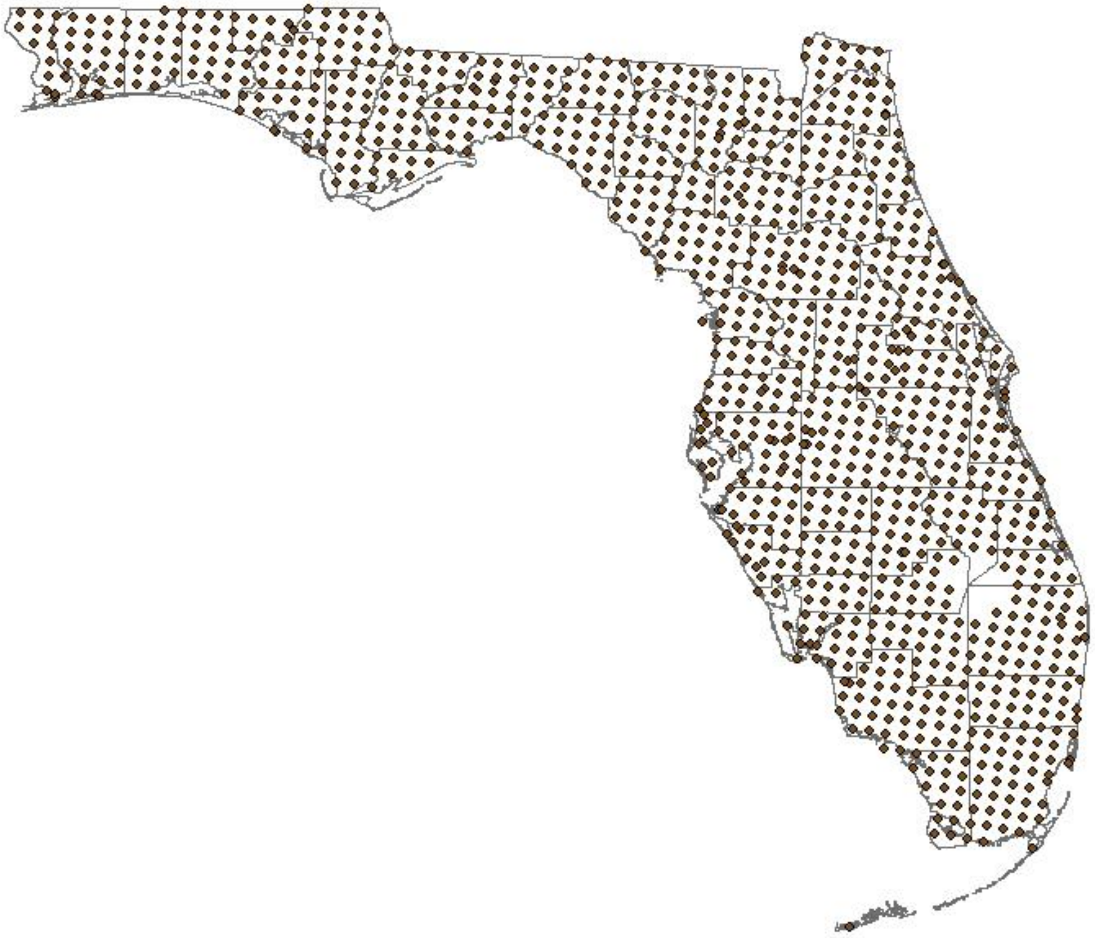
Florida Ozone Monitors in August 2005



Ozone Predicted at Centroids



Support-Adjusted Approach

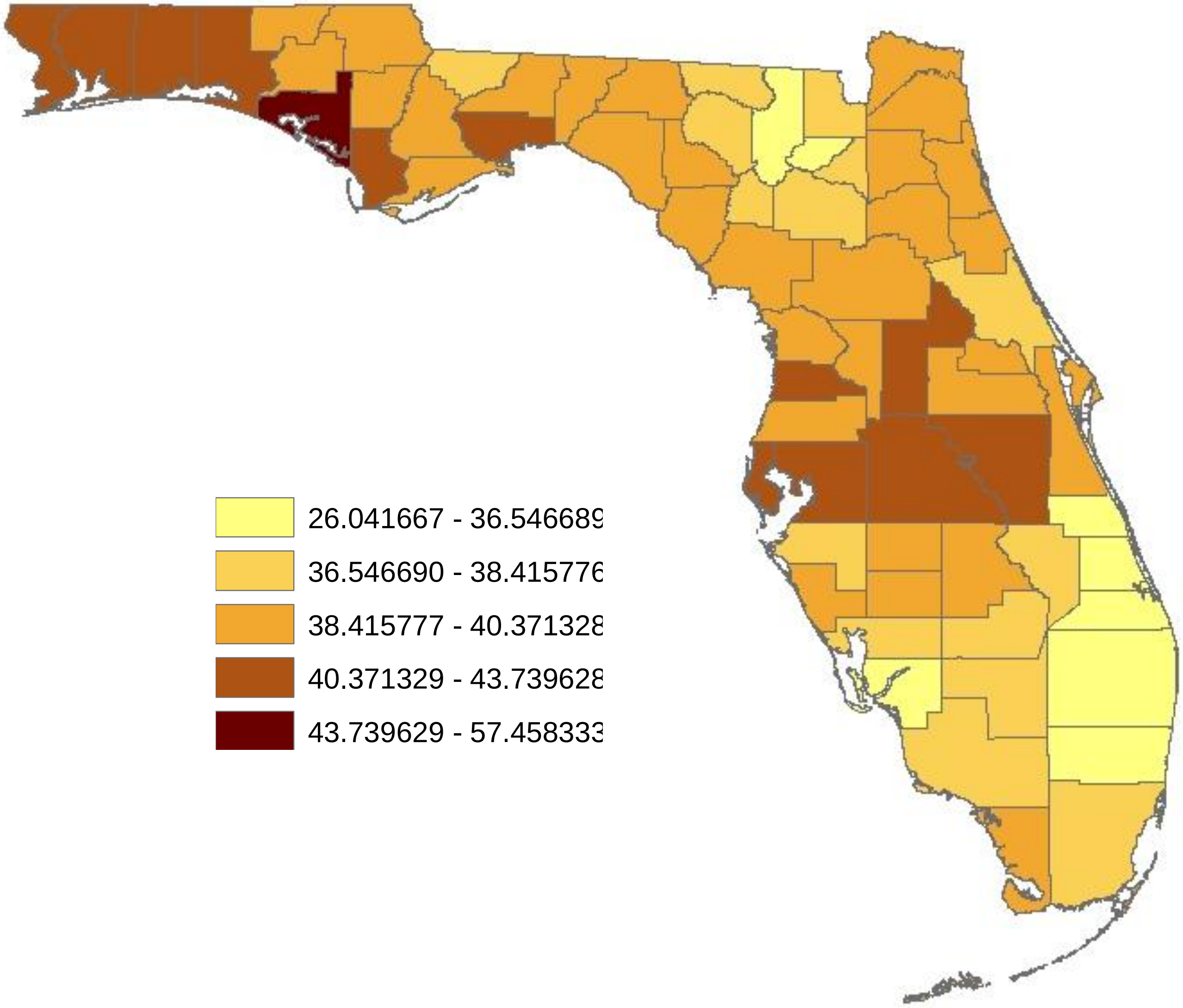


Use block kriging to predict county ozone levels

Process:

- Krige to predict at a grid of points
- Average over the points to obtain the county prediction
- Find the prediction error

Support-Adjusted Prediction of Ozone



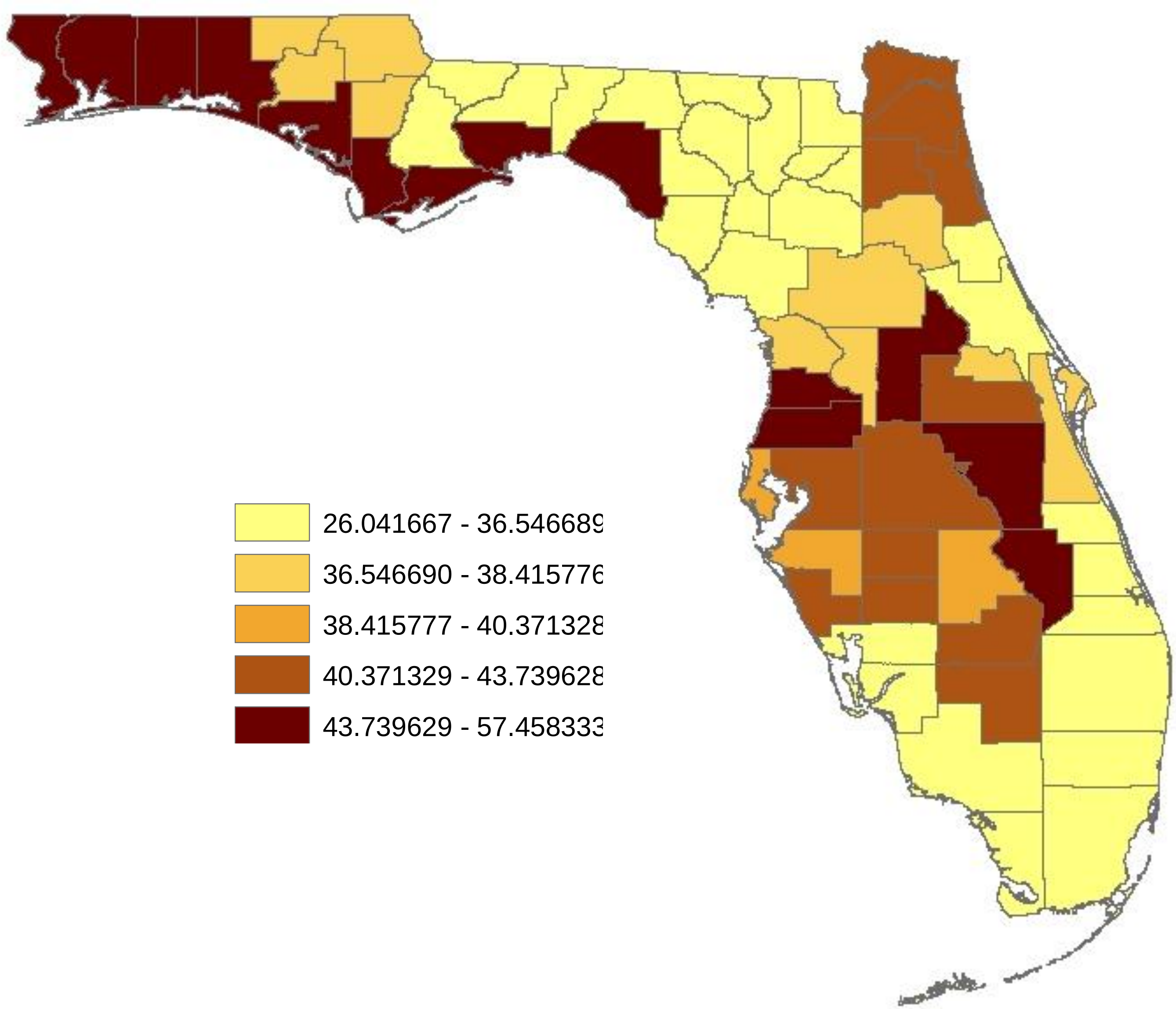
Modeled Prediction of Ozone

Hierarchical Bayesian fusion space-time statistical model used to combine information from the Air Quality System (AQS) monitoring data, and predictions from the Community Multi-scale Air Quality model (CMAQ). Predictions available on 12 and 36-km grid.

AQS data are obtained from air monitors, which tend to be located in more densely populated areas. These measurements are assumed to have some measurement error, but no bias.

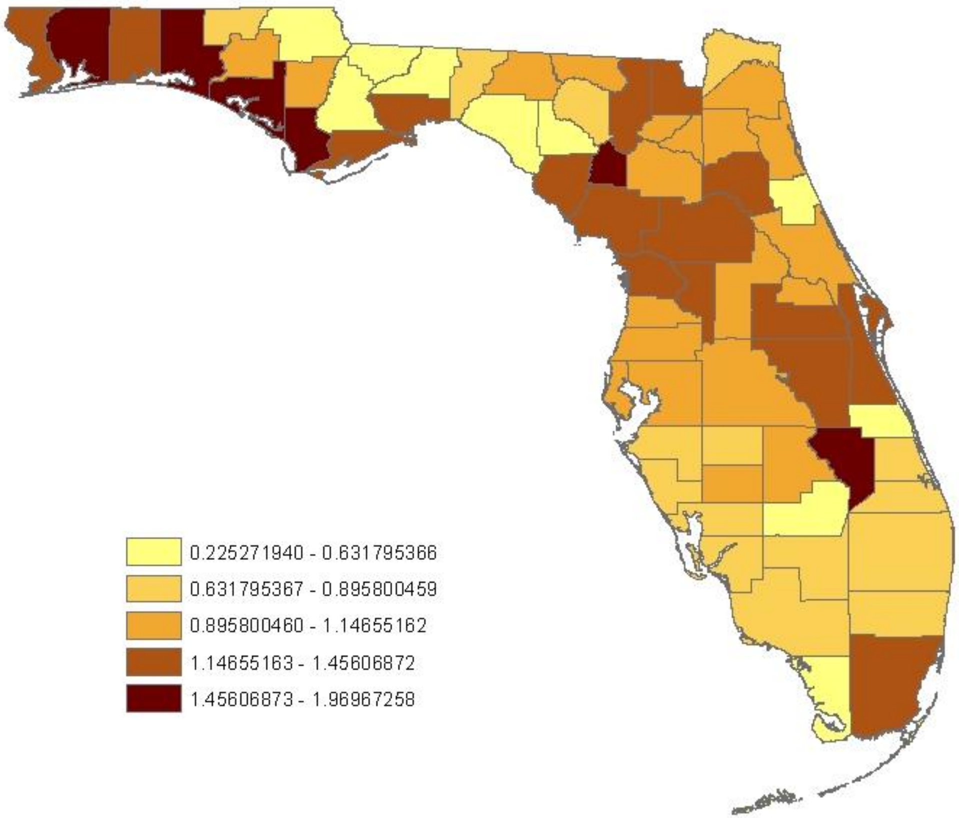
CMAQ model allows for covariates, such as population density and wind, so that the output approximates the variability of the true surface, but exhibits both measurement error and bias.

Modeled Prediction of Ozone

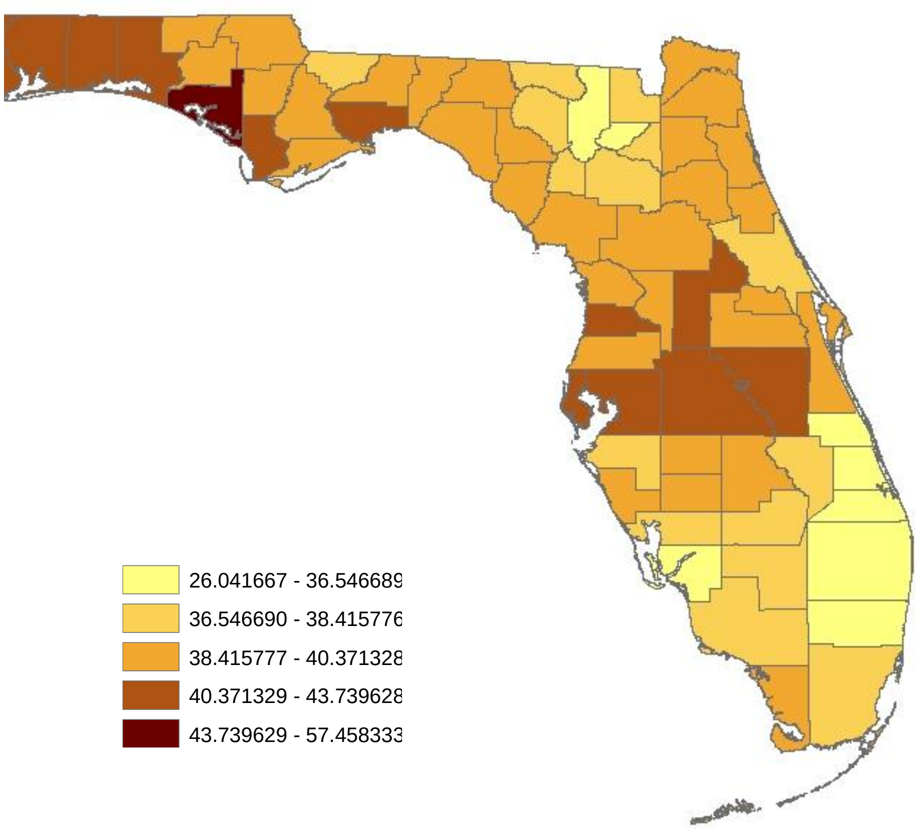


Association Between MI SER and Ozone?

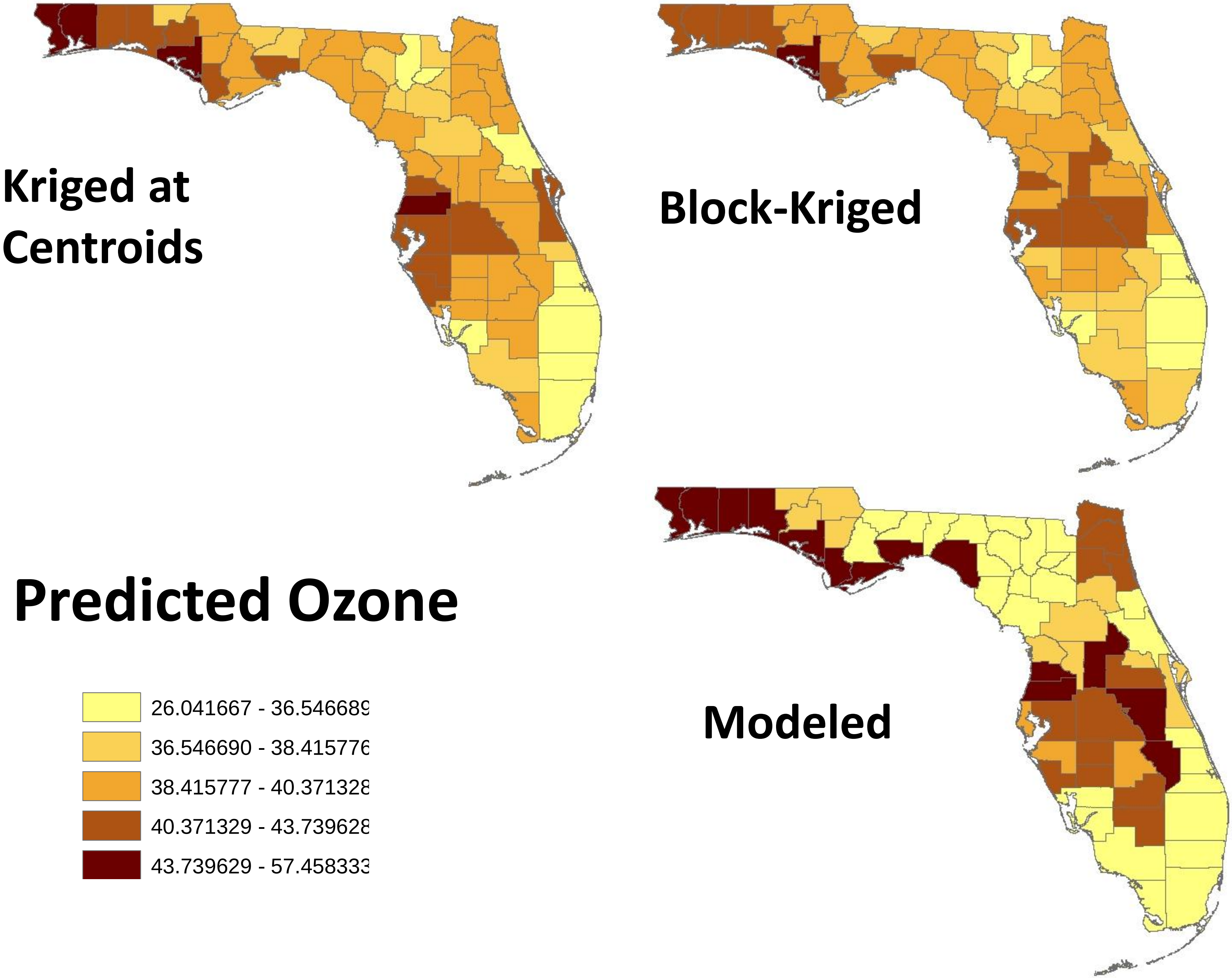
MI SER



Support-Adjusted Predicted Ozone



Predicted Ozone



Kriged at Centroids

Block-Kriged

Modeled

Relating MI to Ozone: Krige and Regress

$$\ln(SER_i) = \beta_0 + \beta_1 x_i + \mathbf{v}_i'\boldsymbol{\beta}_{\mathbf{v}} + e_i$$

where

SER_i = SER for county *I*

x_i is the maximum ozone level for county *i*

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

The relative MI SER is then e^{β_1}

Does the Uncertainty in Ozone Matter?

For kriging, predicted ozone $\hat{x}_i$ results in a smoother surface than the true ozone x_i . We can write

$$x_i = \hat{x}_i + u_i$$

where $u_i = x_i - \hat{x}_i$ is the error associated with predicting ozone. This error is **Berkson error** and affects the covariance structure of the model.

$$\ln(SER_i) = \beta_0 + \beta_1 \hat{x}_i + \mathbf{v}_i' \boldsymbol{\beta}_{\mathbf{v}} + \eta_i, \qquad i = 1, 2, \ldots, n$$

Krige and Regress with General Covariance Structure

If ambient ozone is unknown, the model becomes

$$\begin{aligned}\ln(SER_i) &= \beta_0 + \beta_1 x_i + \mathbf{v}_i'\boldsymbol{\beta}_{\mathbf{v}} + e_i \\ &= \beta_0 + \beta_1(\hat{x}_i + u_i) + \mathbf{v}_i'\boldsymbol{\beta}_{\mathbf{v}} + e_i \\ &= \beta_0 + \beta_1\hat{x}_i + \mathbf{v}_i'\boldsymbol{\beta}_{\mathbf{v}} + (\beta_1 u_i + e_i) \\ &= \beta_0 + \beta_1\hat{x}_i + \mathbf{v}_i'\boldsymbol{\beta}_{\mathbf{v}} + \eta_i\end{aligned}$$

where $\boldsymbol{\eta} = \beta_1\mathbf{u} + \mathbf{e}$ and $\mathrm{var}(\boldsymbol{\eta}) = \beta_1^2\boldsymbol{\Sigma}_{\mathbf{u}} + \boldsymbol{\Sigma}_{\mathbf{e}}$

Will using a general covariance structure lead to appropriate standard errors?

Partial Parametric Bootstrap

In addition to the Berkson error arising from kriging, classical measurement error arises from estimation of the kriging parameters (Madsen, et al. 2008). Assuming the classical measurement error is negligible, a partial parametric bootstrap can be used to obtain an improved estimate of the standard error of $\hat{\beta}$ (Szpiro, et al. 2009)

Approach:

Estimate β_1 as before

Simulate bootstrap samples using estimated exposure model parameters

Calculate the empirical standard deviation of the bootstrap $\hat{\beta}_1$ to obtain standard error of $\hat{\beta}_1$

What Changes when Ozone is Modeled?

Suppose the modeled estimate $\hat{x}_i$ is unbiased and has random variation about the true value x_i ; that is,

$$\hat{x}_i = x_i + e_i'$$

where e_i' is the error associated with predicting ozone. This error is **classical measurement**.

When fitting the model,

$$\ln(SER_i) = \beta_0 + \beta_1 \hat{x}_i + \mathbf{v}_i'\boldsymbol{\beta}_{\mathbf{v}} + e_i, \qquad i = 1, 2, \ldots, n$$

the estimate of β_1 and it standard error are both biased.

Relating MI to Ozone: Florida Data

Estimated trend surface using an exponential covariance structure with a range of 1 and a variance of 51.

Predicted ozone

- Kriged at centroids
- Block kriged
- Modeled and averaged over grid in county

Estimating Association between MI and Ozone: Florida Data

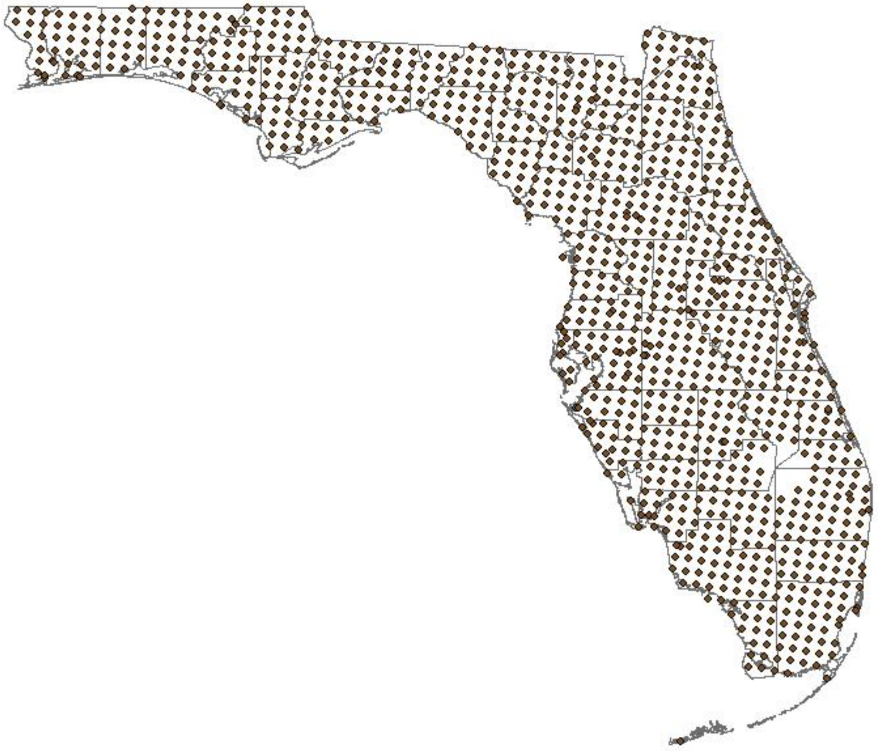
Estimated Association between MI and Ozone;

- CR: Kriged at centroids and regressed, assuming independent error structure
- CRGC: Kriged at centroids and regressed using a general exponential covariance structure
- KR: Block-kriged and regressed, assuming independent error structure
- KRGC: Block-kriged and regressed using a general exponential covariance structure
- PPB: Block-kriged and regressed with partial parameter bootstrap to compute standard error
- MR: Modeled values averaged over county and regressed, assuming independent error structure
- MRC: Modeled values averaged over county and regressed using a general exponential covariance structure

Estimating Association between MI and Ozone: Florida Data

Method	$\hat{\beta}_1$	$s_{\hat{\beta}_1}$	$e^{\hat{\beta}_1}$	$s_{e^{\hat{\beta}_i}}$
CR	0.015	0.0062	1.015	0.0063
CRGC	0.012	0.0069	1.012	0.0070
KR	0.025	0.015	1.025	0.015
KRGC	0.038	0.017	1.039	0.018
PPB	0.025	0.015	1.025	0.015
MR	0.0063	0.0039	1.0063	0.0039
MRGC	0.00087	0.0049	1.00087	0.0049

Simulating Health and Ozone



Generate realizations of ozone for the grid, centroid and monitor values using estimated trend surface as truth and adding error generated from an exponential covariance structure with a range of 1 and a variance of 51.

Given the simulated ozone values, health is simulated as

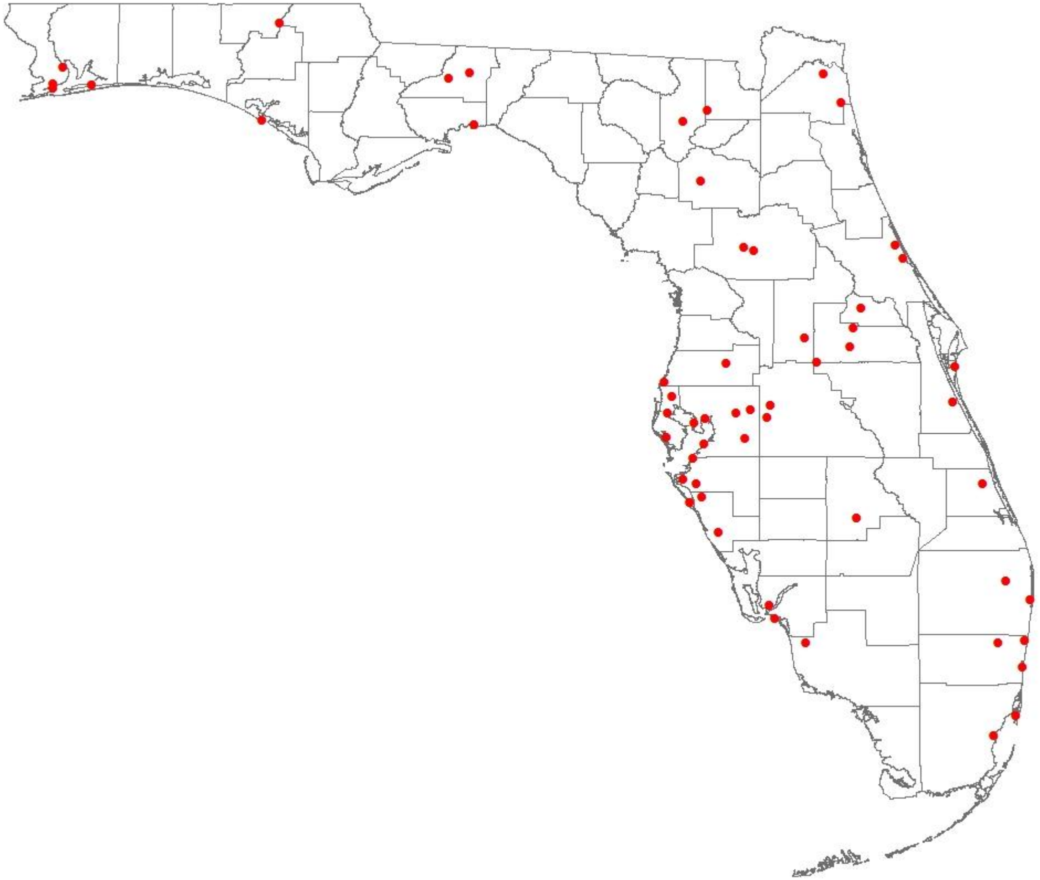
$$y_i = \beta_0 + \beta_1 x + e_i$$

where $\mathbf{e} \sim N(0, \sigma^2 \mathbf{I})$; $\beta_0 = -0.8$;

$\beta_1 = 0.2; \sigma^2 = 2.3$

Health is block-kriged (averaged over points within county).

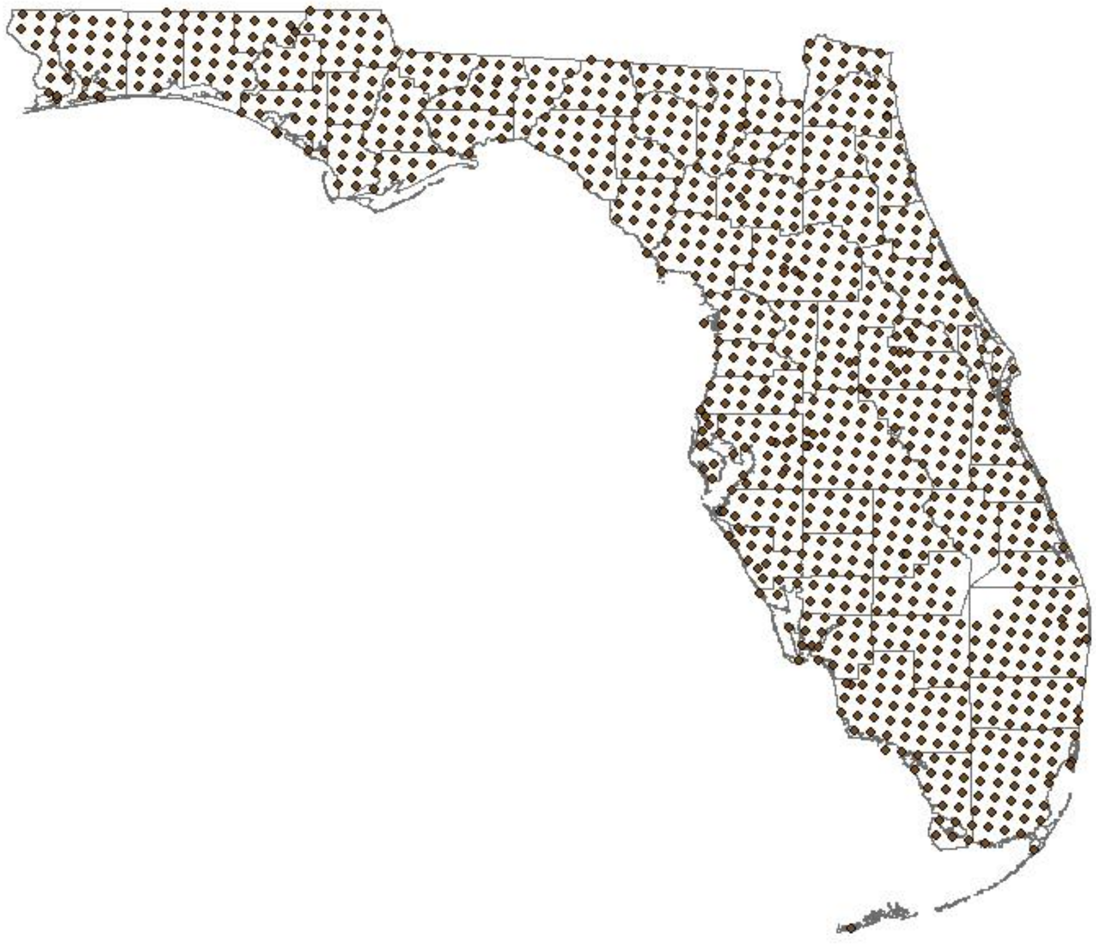
Simulation of Ozone: Kriging



For each realization of ozone generated, all but the simulated values at the monitors are deleted.

Predict ozone (1) at centroids or (2) using block-kriging.

Simulation of Ozone: Modeling



For each realization of ozone generated, keep only simulated ozone at grid points.

To simulate an unbiased model with some random error, add independent N(0, 7.5^2) errors to each point and average points within counties.

Simulation Results: Estimating Association Between MI SER and Ozone

Method	$\overline{\hat{\beta}_1}$	$\overline{s}^2_{\hat{\beta}_1}$	$s^2_{\hat{\beta}_1}$ (truth)	Coverage Probability
CR	0.18	0.00100	0.00060	0.76
CRGC	0.18	0.00097	0.00070	0.77
KR	0.20	0.0012	0.00068	0.84
KRGC	0.20	0.0012	0.00079	0.87
PPB	0.20	0.0012	0.0012	0.94
MR	0.18	0.00037	0.00044	0.78
MRGC	0.18	0.00039	0.00047	0.77

Conclusions

- When regressing health outcomes on predicted environmental exposure, the method used to predict ozone matters.
- If environmental exposure is predicted using block-kriging, the estimate of the association between health and environmental exposure obtained through regression is unbiased.
- The estimates are biased if centroids or modeled values (even those for which support is considered) are used to predict environmental exposure.

Conclusions

- For all methods, the standard errors obtained from regressing health outcomes on predicted environmental exposure are under-estimated.
- The Partial Parametric Bootstrap is a method for correcting the standard errors. Sometimes it seems to work well but, as was the case here, it often tends to over-estimate the standard errors.
- To date, no method proposed provides unbiased estimates of standard errors.

Conclusions

- Exposure of persons to ozone is the association of interest. Two problems:
  - ✓ Ambient ozone levels serve to approximate ozone exposure.
  - ✓ Data have been linked by month on the county level, but we want to draw inferences regarding a person's risk for MI.
- Goal of EPHT is on-going monitoring. Existing space-time models are not readily extendable to this setting.
- Bayesian models tend to be problem-specific and can not readily be adapted for different variables, locations, time, etc.

Conclusions

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