Analyzing Health and Environmental Data: Statistical Methods and Inferential Considerations

Carol A. Gotway Crawford
Division of Environmental Hazards and Health Effects
National Center for Environmental Health
Centers for Disease Control and Prevention
Atlanta, GA

EPA and CDC Symposium on Air Pollution Exposure and Health
September 19, 2006

Acknowledgements

- **Many ideas are based on joint work with:**

Linda J. Young
Associate Chair of Biostatistics
Department of Statistics
University of Florida
Gainesville, FL

Combining incompatible spatial data. *Journal of the American Statistical Association,* 2002.

Linking spatial data from different sources: The effects of change of support. *Stochastic Environmental Research and Risk Assessment*, 2006.

Very Low Birth Weight Study

Assess the association between:

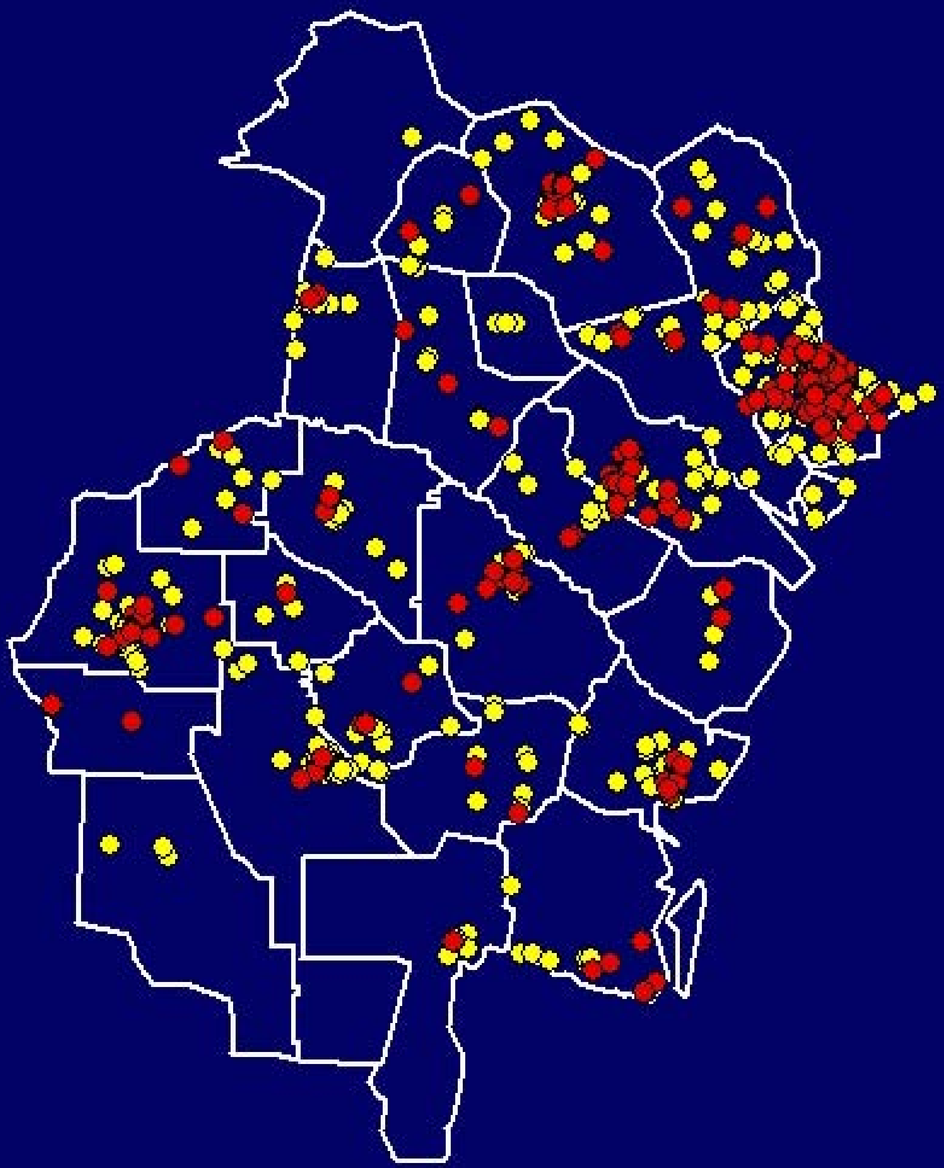
maternal exposure to air pollution
(here Total Suspended Particulates (TSP))

and

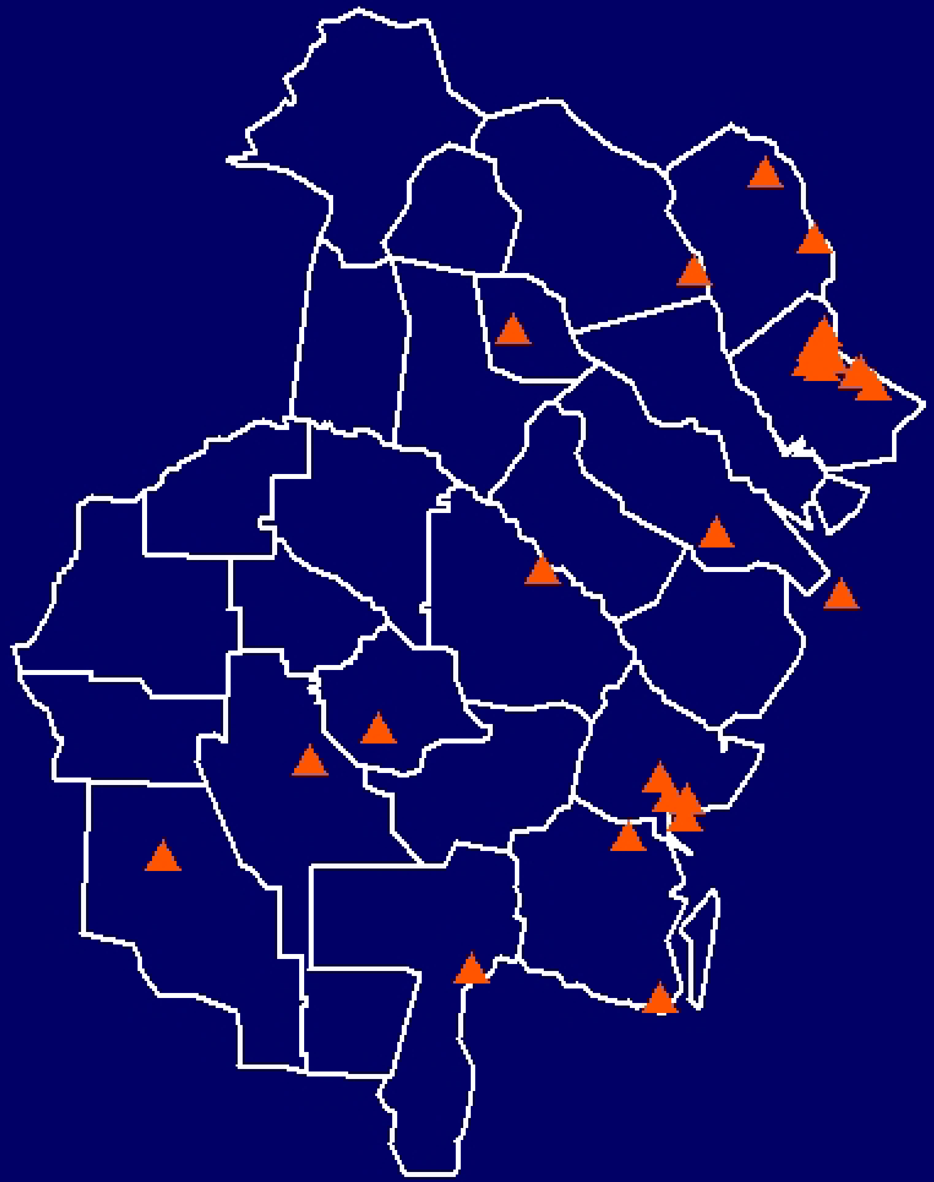
the risk of a very low birth weight baby
(weighs less than 1500 grams at birth)

Rogers, JF et al. (2000). *American Journal of Epidemiology*

VLBW Data Locations
Cases and **Controls**

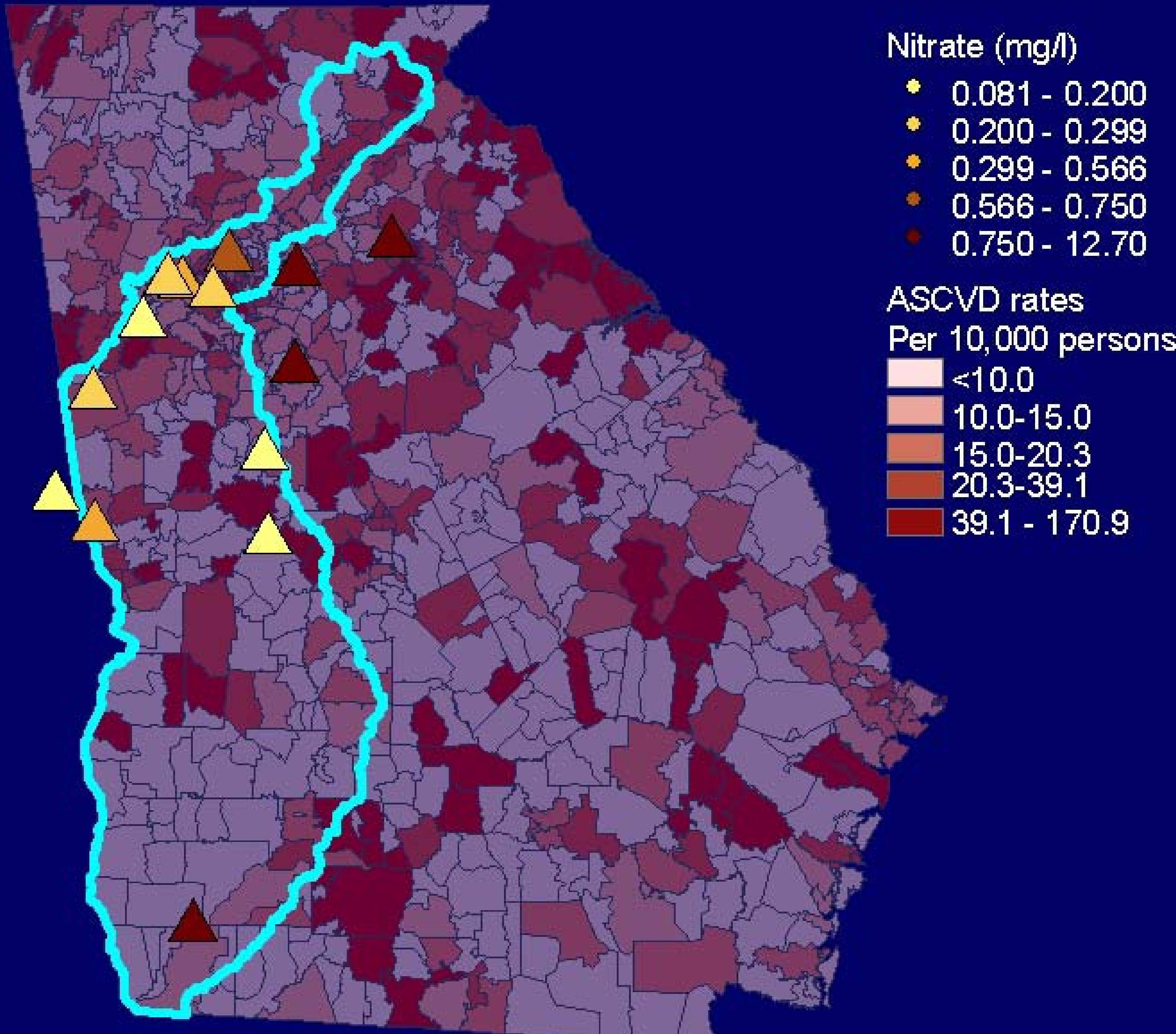


Emissions Data Locations

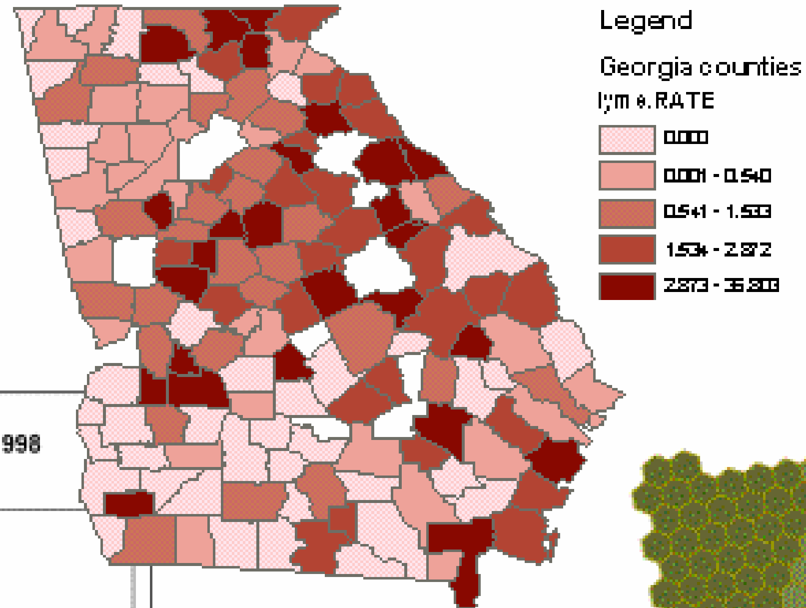


Linking Hospital Discharge and Water Quality Data

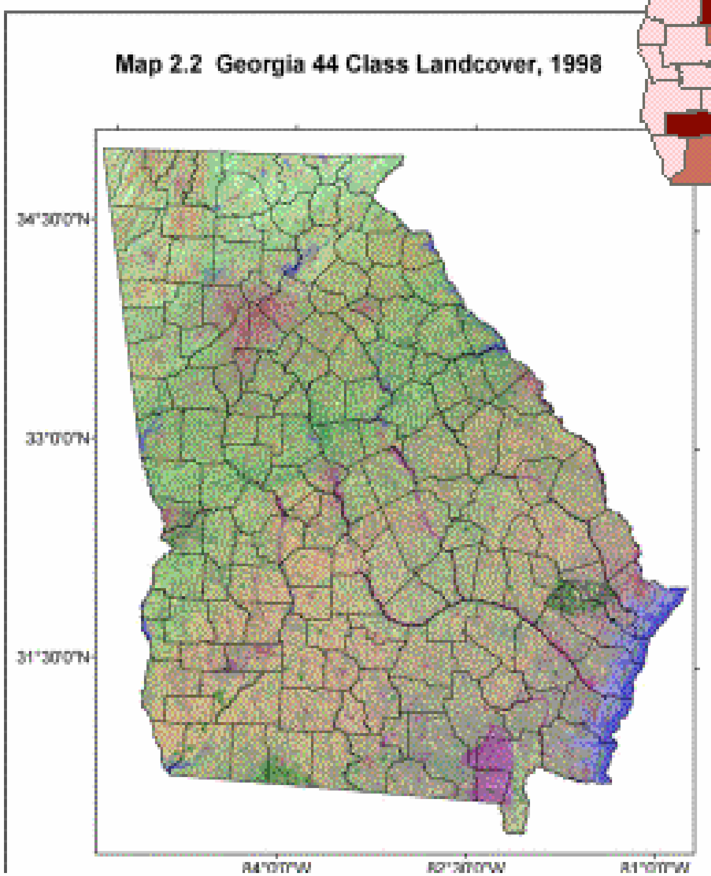
- Hospital discharge data: 1999-2003, GA, ZIP codes
- Environmental data from NAWQA
- Is there an association between **atherosclerotic cardiovascular disease** and **surface water nitrate concentration** in the Apalachicola-Chattahoochee-Flint river basin?
- How to monitor such potential associations?

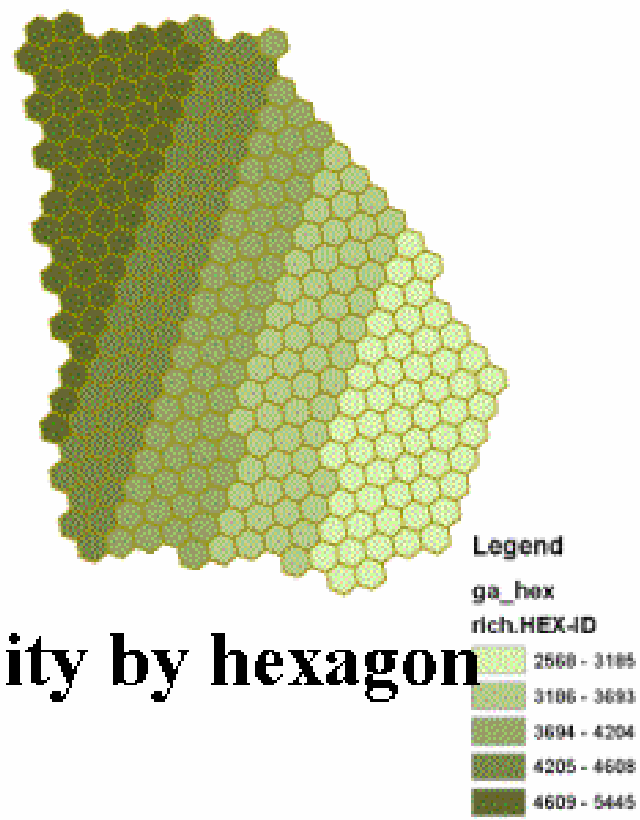

Nitrate (mg/l)
0.081 - 0.200
0.200 - 0.299
0.299 - 0.566
0.566 - 0.750
0.750 - 12.70
ASCVD rates
Per 10,000 persons
<10.0
10.0-15.0
15.0-20.3
20.3-39.1
39.1 - 170.9

Lyme disease rates by county



Landcover 30m





Species Diversity by hexagon

Key Features

- Data collected for different purposes
- Rarely recorded at the same time and place
- Many different spatial units involved:

Health Data:

- Census Units:
  - Tracts, counties, states
- ZIP code units
- Geocoded addresses

Environmental Data:

- Monitors
- Satellites
- Sampling Units

The Common Goal

- Use all the data to make inference about an outcome associated with one particular set of spatial units (e.g., health of individuals)
- Must involve upscaling (aggregation), downscaling (disaggregation), or side scaling (overlapping units or points)
- Statistically, this means making **predictions** of data associated with one set of spatial units from data associated with other sets of spatial units.

Main Statistical Issues

- Choice of geographic/spatial units
  - Modifiable Areal Unit Problem (MAUP)
  - Ecological Fallacy
- Spatial support
- Uncertainty

The Modifiable Areal Unit Problem

- Results from any statistical analysis depend on how the data are aggregated geographically.

- Openshaw and Taylor (1979):
  - Different geographical aggregations of the same data can produce ``**a million or so**'' correlation coefficients.
  - Could produce correlations ranging from -0.97 to +0.99!

- **The Ecological Fallacy:**
  - Analyses based on grouped data often lead to conclusions different from those based on individual data.

Spatial Support

- The size, shape and orientation of the spatial units.
- Measurements associated with areal units are inherently aggregates (totals, averages).
- The statistical and spatial properties of averages are different from those of the individual measurements (**Change of Support Problem**).
- Predicting at a point in the center of an areal unit is not the same as predicting an average value over that unit.

Implications for Data Linkage and Analysis

- Method of spatial prediction needed for linkage depends on the support of the data involved:
  - **Case 1:** VLBW: All data have point support.
  - **Case 2:** ACF Basin: Health data have supports defined by ZIP code units, environmental data have point support.
  - **Case 3:** Lyme Disease: Health data have county-level supports, land use data have areal support (30m pixels) and species diversity data have hexagonal support.

Spatial Prediction of Point-Support Data

Can obtain a predicted value at any point in space.

- **Deterministic:**
  - Given inputs get output
  - Does not use data to estimate parameters or determine form
  - No measure of uncertainty
- **Probabilistic:**
  - Based on a statistical model
  - Estimate unknown parameters and/or model using the data
  - Provide prediction standard errors.
- **Combination**

Prediction Methods

- **Deterministic interpolation methods**
  - Closest point/triangulation/Voronoi polygons
  - Inverse distance
- **Deterministic process models**
  - Gaussian plume models
  - Numerical methods
- **Statistical Methods**
  - Trend surface analysis
  - Regression
  - Kriging
  - Splines

Atmospheric Transport Model

$$\chi = \frac{Q\sqrt{2/\pi}\,f}{(2\pi x/16)\sigma_z u}\exp\left(-\frac{h_e^{\,2}}{2\sigma_z^{\,2}}\right)$$

u = wind speed (m s^{-1})

f = fraction of the time the wind blows into the receptor sector

x = distance (m) from the source to the receptor

σ_z = vertical dispersion coefficients (m) as a function of downwind distance from the source

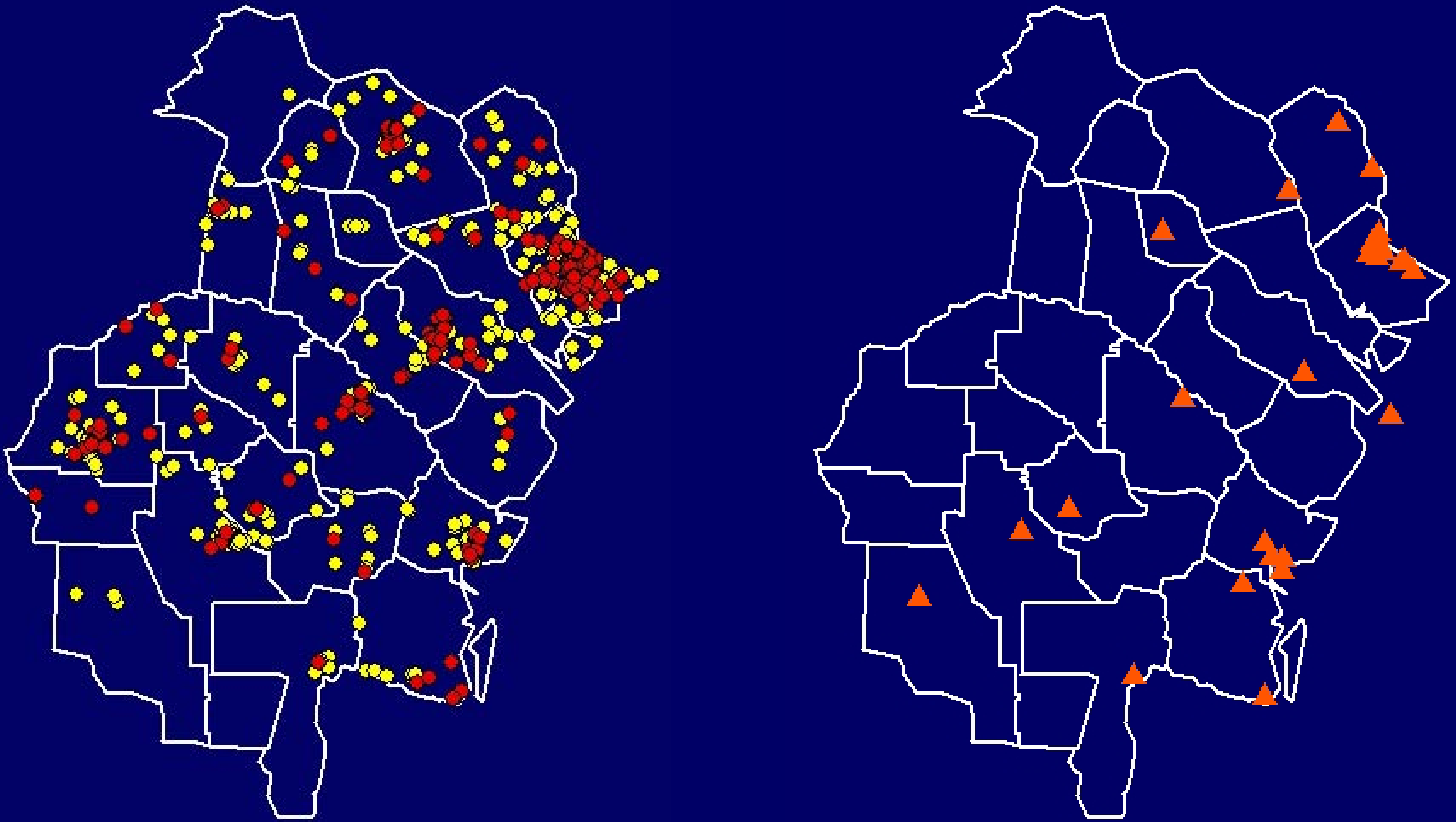
h_e = effective stack height of the source(m)

Q = release rate of TSP from the source (μg s^{-1})

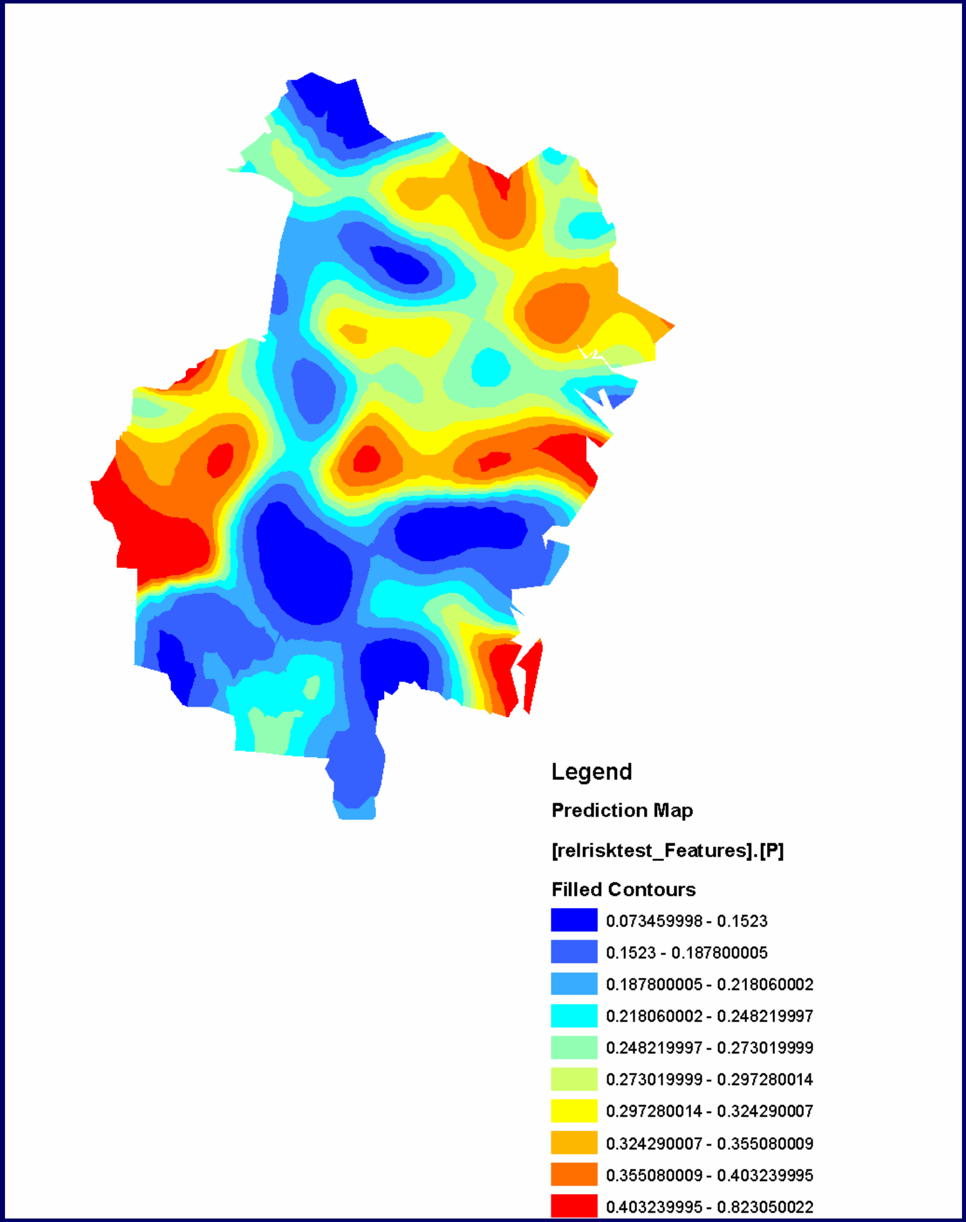
χ = average ground level concentration (μg m^{-3})

Case 1: Predict ground-level concentrations at each case and control location

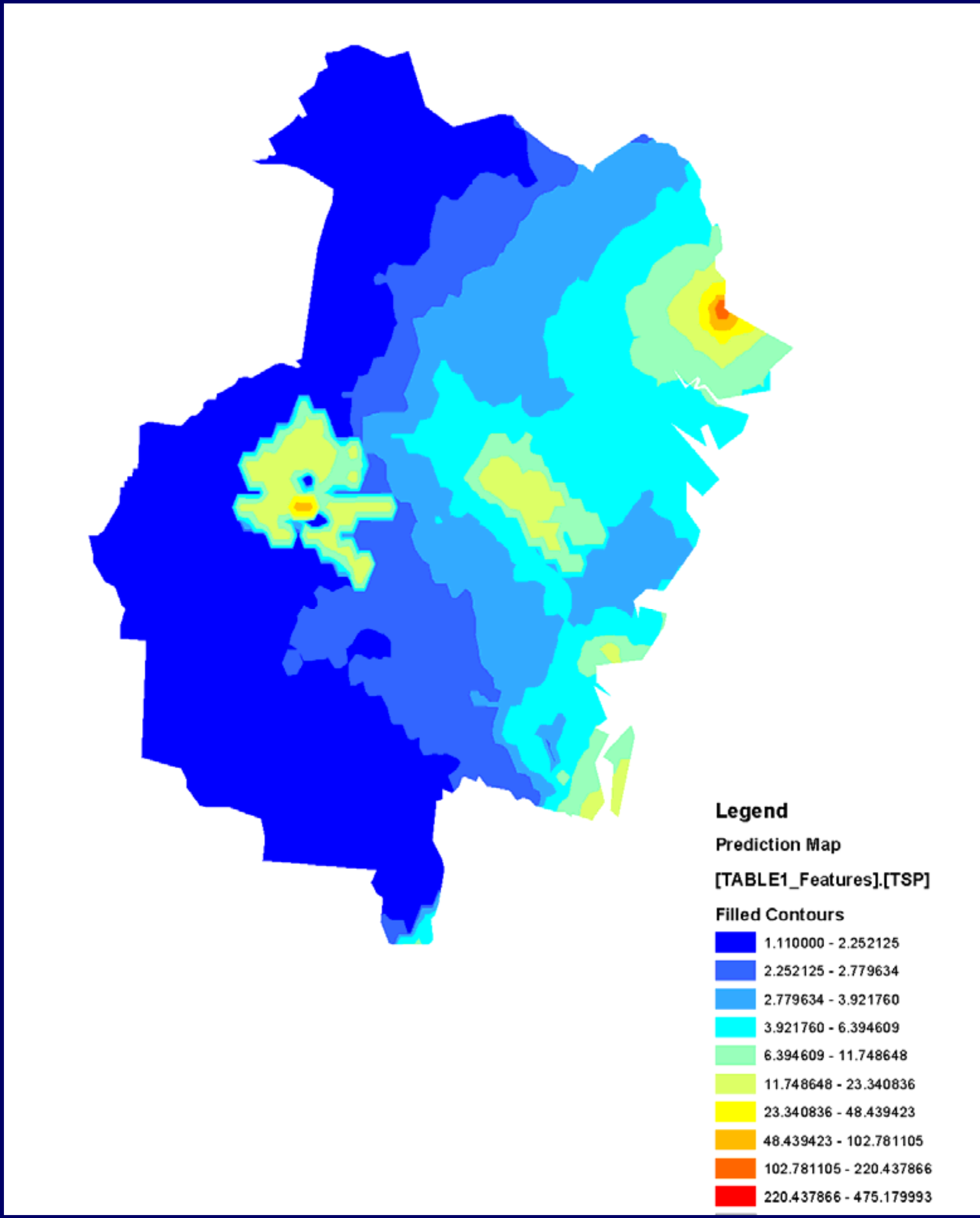
Can also predict the risk of having a VLBW birth at each industrial facility. Makes post-linkage analysis more difficult (?).



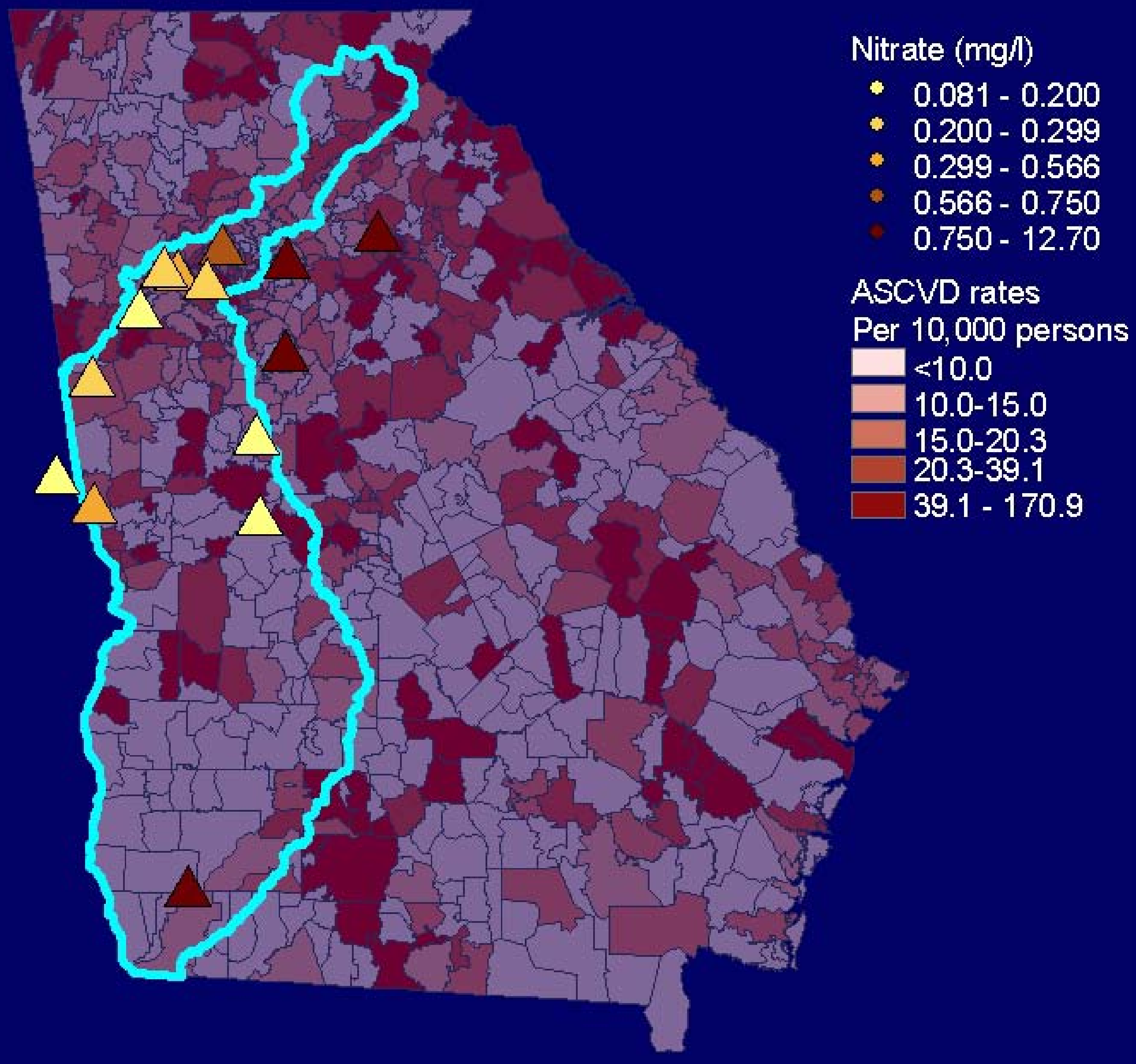
Relative Risk Map



Ground-level Exposure Estimates from Transport Model



Case 2: Nitrate data have point support, but health data are aggregated over ZIP code units (rates)



Options

- Predict a nitrate value at the “center” of each ZIP code unit (“traditional” approach)
- Support-adjusted prediction:
  - Predict many values in the ZIP code and then average the results
  - Need point-in-polygon codes or “zonal” analysis capabilities
  - Formal statistical technique is called “**block** kriging” and is used to get correct standard errors
- Can also “downscale” health data and predict the relative risk of ASCVD at any point.

Does Support Really Matter?

Poisson regression used after linkage:

Method	Rate Ratio	95% CI
Point Prediction at Centroids	0.85	(0.77, 0.94)
Support-adjusted Prediction	0.72	(0.58, 0.89)

Accounting for support effects may be even more important when we account for prediction uncertainty.

Case 3 and Beyond

- Many more sophisticated solutions
- Most are complex spatial and space-time models
- Many are situation specific
- Overview in Gotway and Young (2002)
- Area of current statistical research

Traditional Post-Linkage Analysis

- **Poisson/Logistic regression analysis**
  - Include variables for seasonal trends; can adjust for temporal autocorrelation
- **Non-/semi-parametric regression**
  - Allows more flexible models for seasonal variation in both health and environmental data
  - GAMs, Regression splines (B-splines, P-splines), GLASS (generalized linear additive smooth structures, Eilers and Marx, 2002)
- **Case-Crossover analysis**
  - Exposures of cases just prior to an event are compared to the exposures for the same cases from some separate time period.
  - Avoids geographic aggregation, but may introduce locational error.

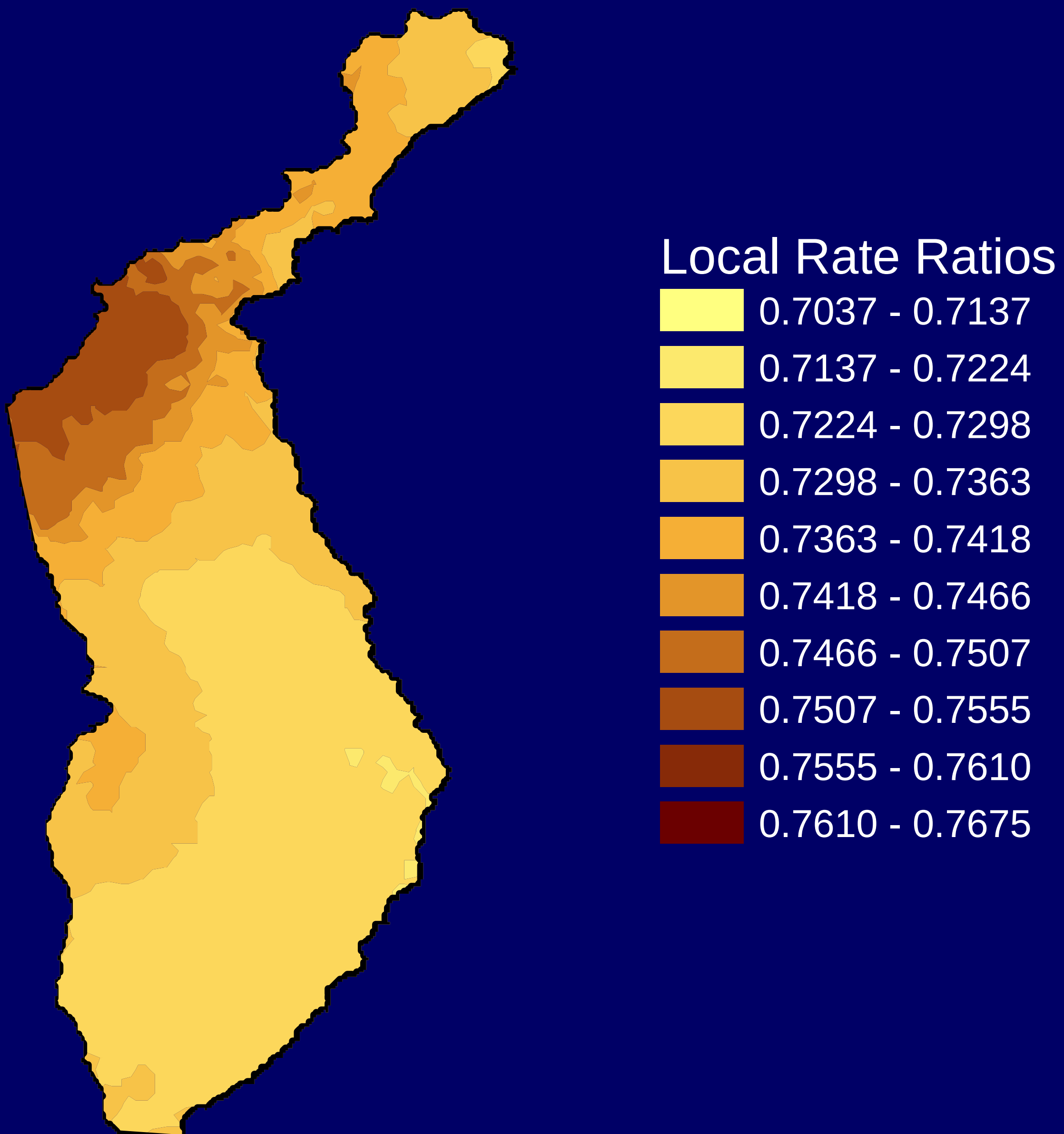
- Focus mostly on temporal variation in the data. Little focus on geographic variation.

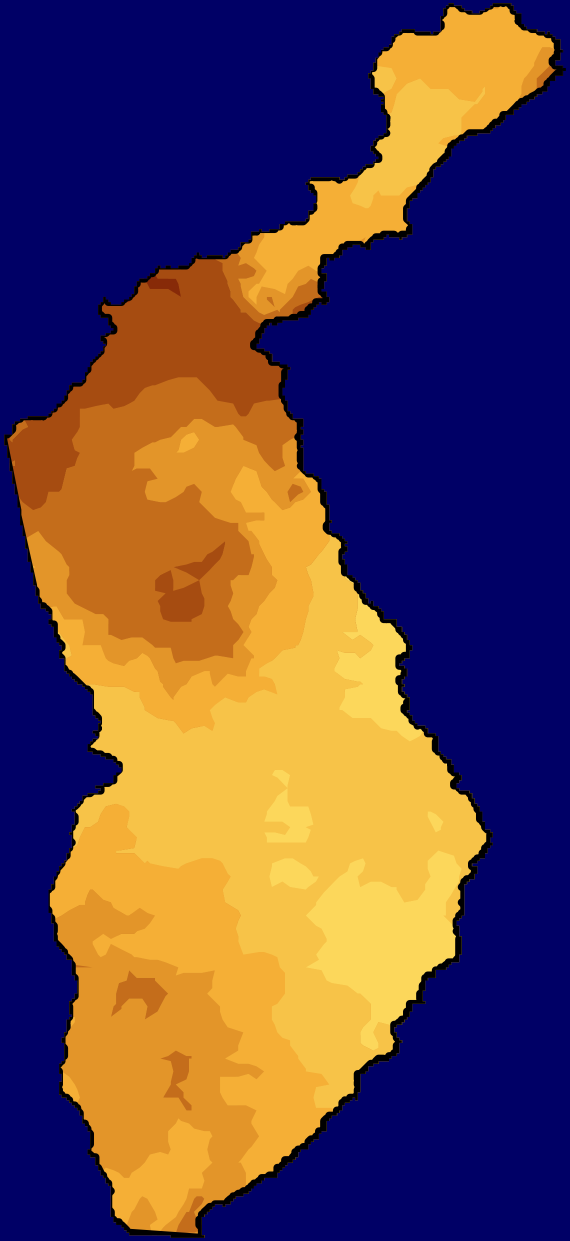
Beyond the Traditional

- **Quality control models**
  - Routinely used in health surveillance
    - EARS, Lori Hutwagner, NCID
  - Environmental monitoring, multivariate, regression, spatial, and space-time adaptations

- **Spatially-varying coefficient models**
  - Geographically weighted regression
    (Fotheringham et al 2002, Nakaya et al. 2005)
  - Random coefficient models
    (Gelfand et al. 2003; Congdon, 2003)

Map the association between the environmental hazard and the health outcome of concern
(Map the correlation, odds ratio, or risk)





Monitor through time

THE Most Important Issue: Uncertainty

- Post-linkage analysis must account for the uncertainties that arise from prediction during linkage, as well as any uncertainties in the initial data.
- Otherwise, confidence intervals are too narrow, p-values are too small, and conclusions are probably wrong.
- Probabilistic prediction methods provide a measure of prediction uncertainty (standard errors), but these cannot be easily used in subsequent analyses.

Methods for Quantifying Uncertainty

- Model the variability in the input, simulate from this model, analyze the resulting variation in output.
  - Monte Carlo/Geostatistical simulation
  - EM algorithm with maximum likelihood
  - Bayesian hierarchical models
- Must account for differing supports.
- Accounting for temporal variation may require complex space-time models.
- Computationally challenging.

Case 1: VLBW Study Regression Results (adjusted for maternal covariates)

- **Original logistic regression:**

Exposure Categories	Odds Ratio	95% C.I.
High	2.88	(1.16, 7.13)
Medium	1.27	(0.68, 2.37)
Low	0.99	(0.99, 1.92)

- **Accounting for uncertainty in modeled TSPs:**

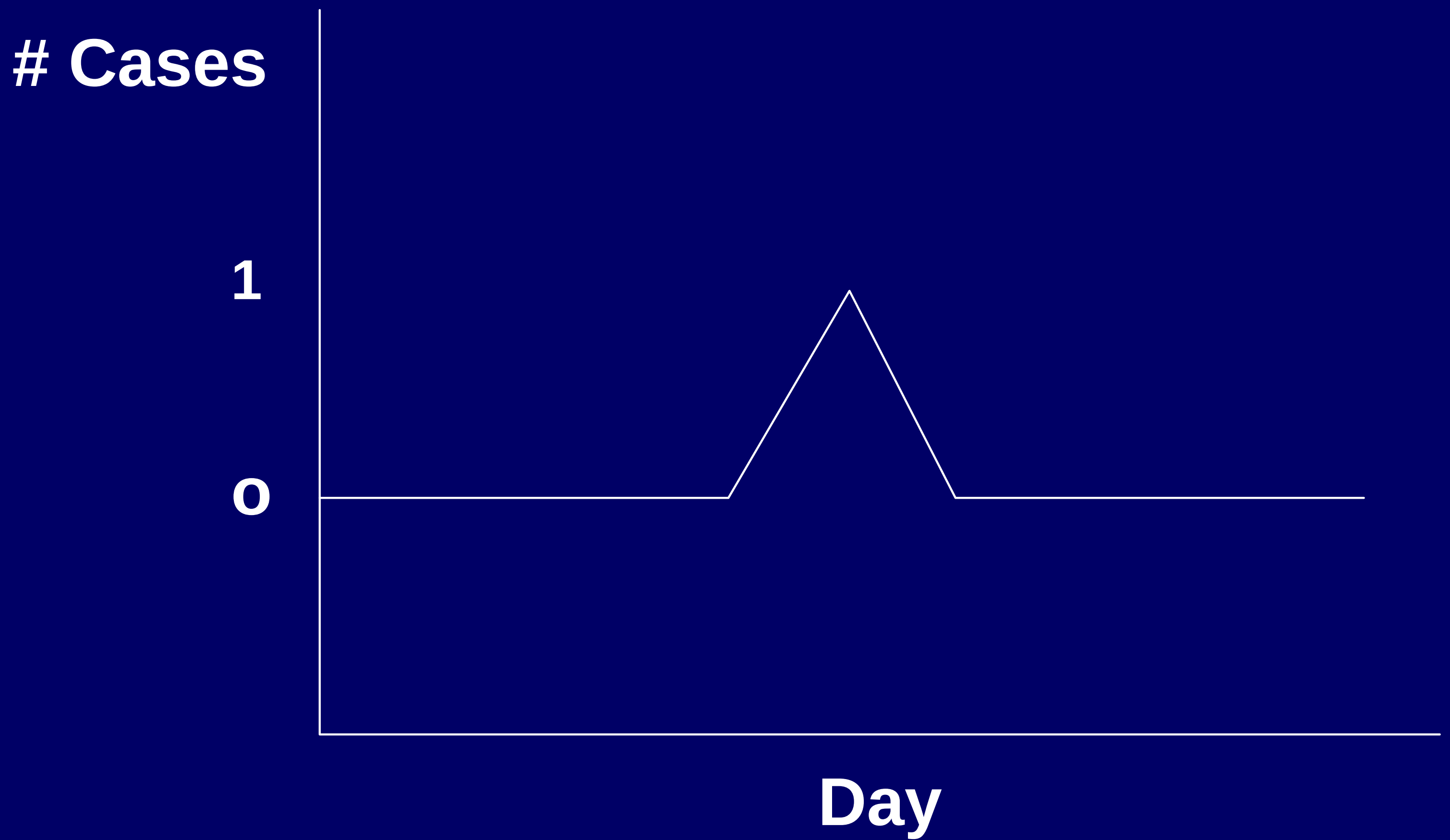
Exposure Categories	Odds Ratio	95% C.I.
High	4.18	(1.82, 9.08)
Medium	2.13	(1.09, 4.13)
Low	1.45	(0.72, 2.87)

Beyond The Three-Step Approach

- Link, Analyze, Uncertainty Assessment
- All-in one: Bayesian Hierarchical Models with Kalman Filtering. Theoretically, should be able to simultaneously:
  - take the data and models used in PHASE
  - combine them with health data
  - get a map of associations with estimated uncertainties
  - provide an algorithm for routine updating as new data become available
  - produce forecasts in space and time

Pre-Linkage Issues

- Data management
  - PHASE dataset has 6 million (**?!!!**) records
- Ensuring confidentiality of health data
  - Personal liability
  - Lack of secure infrastructure
- Controls (?)
- Other data (temperature, humidity)?
- Determination of meaningful temporal and spatial scales:
  - Fine stratification
  - Unstable rates (small number problem)
  - Lots of zeros



Is one case important if it is asthma in Atlanta in July?
If it is a rare birth defect?

If 95% of the values are 0, are the results from Poisson/Logistic regression valid? Meaningful?

Final Thoughts

- Just because we can does not mean we should:
  - Data not collected for this purpose; Spurious associations; Beyond usual observational study.
- Do we need to formally link the data?
  - What will we learn that we couldn't learn from monitoring each separately?
  - What will we learn that we don't already know?
- **Scale issues are extremely important:**
  - What temporal changes are important (Hourly? Daily? Yearly)?
  - Over what geographic extent are models valid?
  - How much geographic variability can we expect?

References

Congdon P. 2003. Modeling spatially varying impacts of socioeconomic predictors on mortality outcomes. *Journal of Geographic Systems* 5:161-184.

Eilers, P.H.C. and Marx, B. D. 2002. Generalized Linear Additive Smoothers. *Journal of Computational and Graphical Statistics,* 11: 758-783.

Fortheringham, A.S., C. Brandson, and M. Charlton 2002. *Geographically Weighted Regression: The Analysis of Spatially-Varying Relationships*. West Sussex: John Wiley & Sons.

Gelfand, AE, Kim H-J, Sirmans CF, and Banerjee, S. 2003. Spatial Modeling with Spatially Varying Coefficient Processes. *Journal of the American Statistical Association,* 98:387-396.

Gotway, C.A**.**, and Young, L.J. 2002. Combining incompatible spatial data. *Journal of the American Statistical Association,* 97: 632-648.

Gotway, C.A. and Young, L.J. 2006. A geostatistical approach to linking geographically-aggregated data from different sources. *Journal of Computational and Graphical Statistics,* In press.

Hutwagner LC**,** Thompson WG, Seeman GW, and Treadwell, T. 2003. The bioterrorism preparedness and response Early Aberration Reporting System (EARS). *Journal of Urban Health: Bulletin of the New York Academy of Medicine,* 80(2), supplement 1: pp. i89-i96.

Nakaya T, Fotheringham AS, Brunsdon C, and Charlton M. 2005. Geographically weighted Poisson regression for disease association mapping. *Statistics in Medicine* 24**:** 2695-2717.

Openshaw, S. and Taylor, P. 1979. A million or so correlation coefficients. Pages 127-144 In *Statistical Methods in the Spatial Sciences*, N. Wrigley, editor. Pion, London.

Rogers, J. F., Thompson, S. J., Addy, C. L., McKeown, R. E., Cowen, D. J. and DeCoulf'e P. 2000. The association of very low birthweight with exposures to environmental sulfur dioxide and total suspended particulates. *American Journal of Epidemiology ,*151:602-613.

Young, L.J. and Gotway, C.A. 2006. Linking spatial data from different sources: The effects of change of support. Invited for *Stochastic Environmental Research and Risk Assessment*, In press.