Combining Incompatible Spatial Data

Carol A. Gotway Crawford

Office of Workforce and Career Development

Centers for Disease Control and Prevention

Invited for Quantitative Methods in Defense and National Security
George Mason University, February 2007

Acknowledgements

Linda J. Young
Department of Statistics
University of Florida

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.

A geostatistical approach to linking geographically-aggregated data from different sources *Journal of Computational and Graphical Statistics*, 2007.

Very Low Birth Weight Study

Assess the association between:

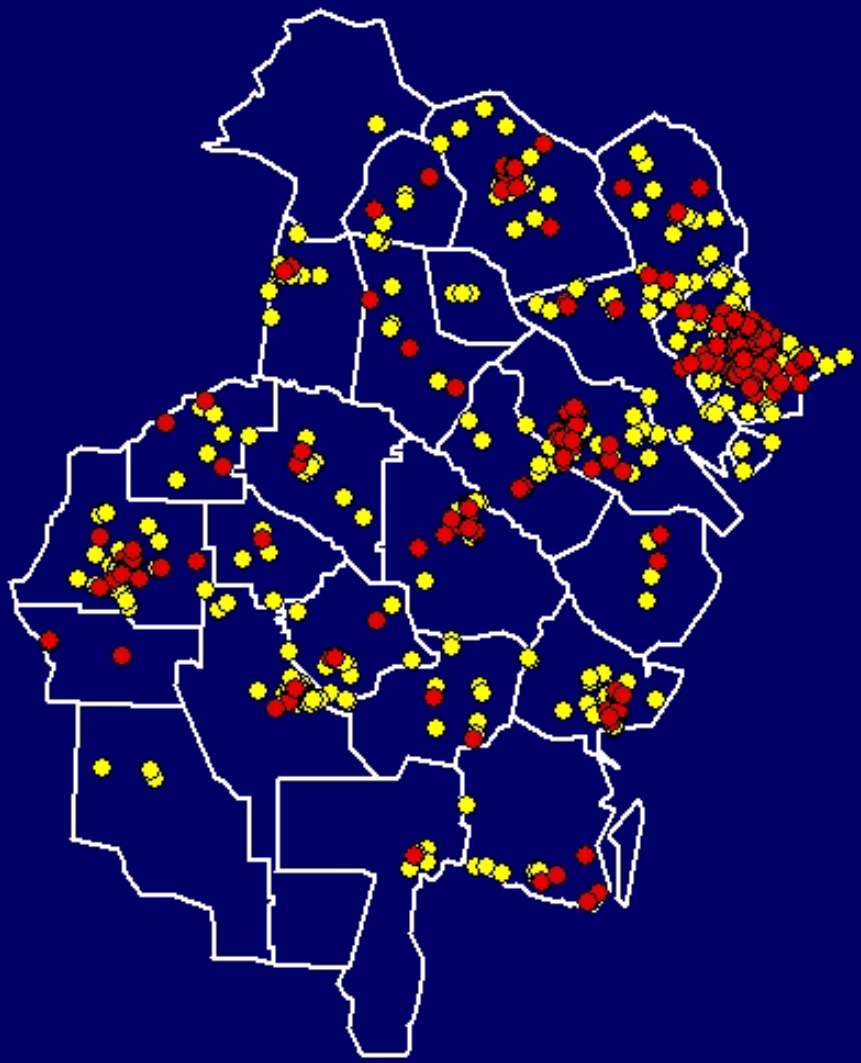
maternal exposure to air pollution
(here Total Suspended Particulates (TSP))

and

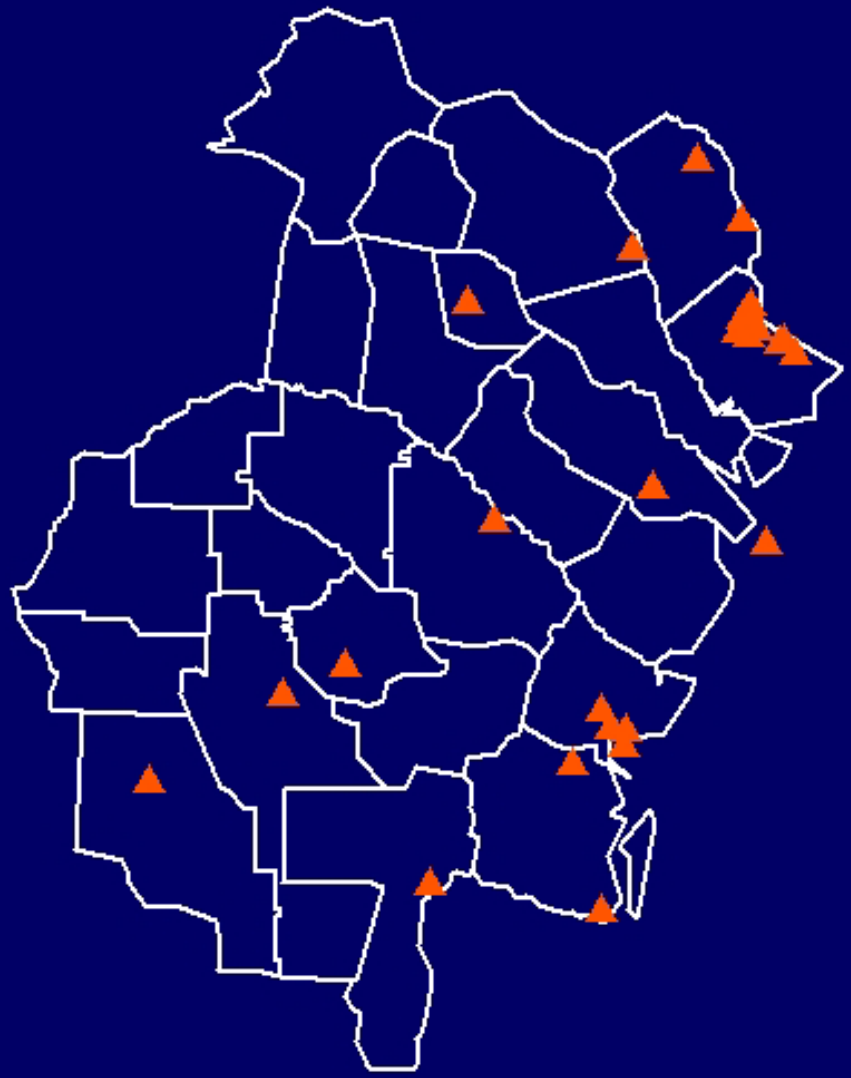
the risk of a very low birth weight baby
(weighs less than 1500 grams at birth)

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

VLBW Data Locations
Cases and **Controls**

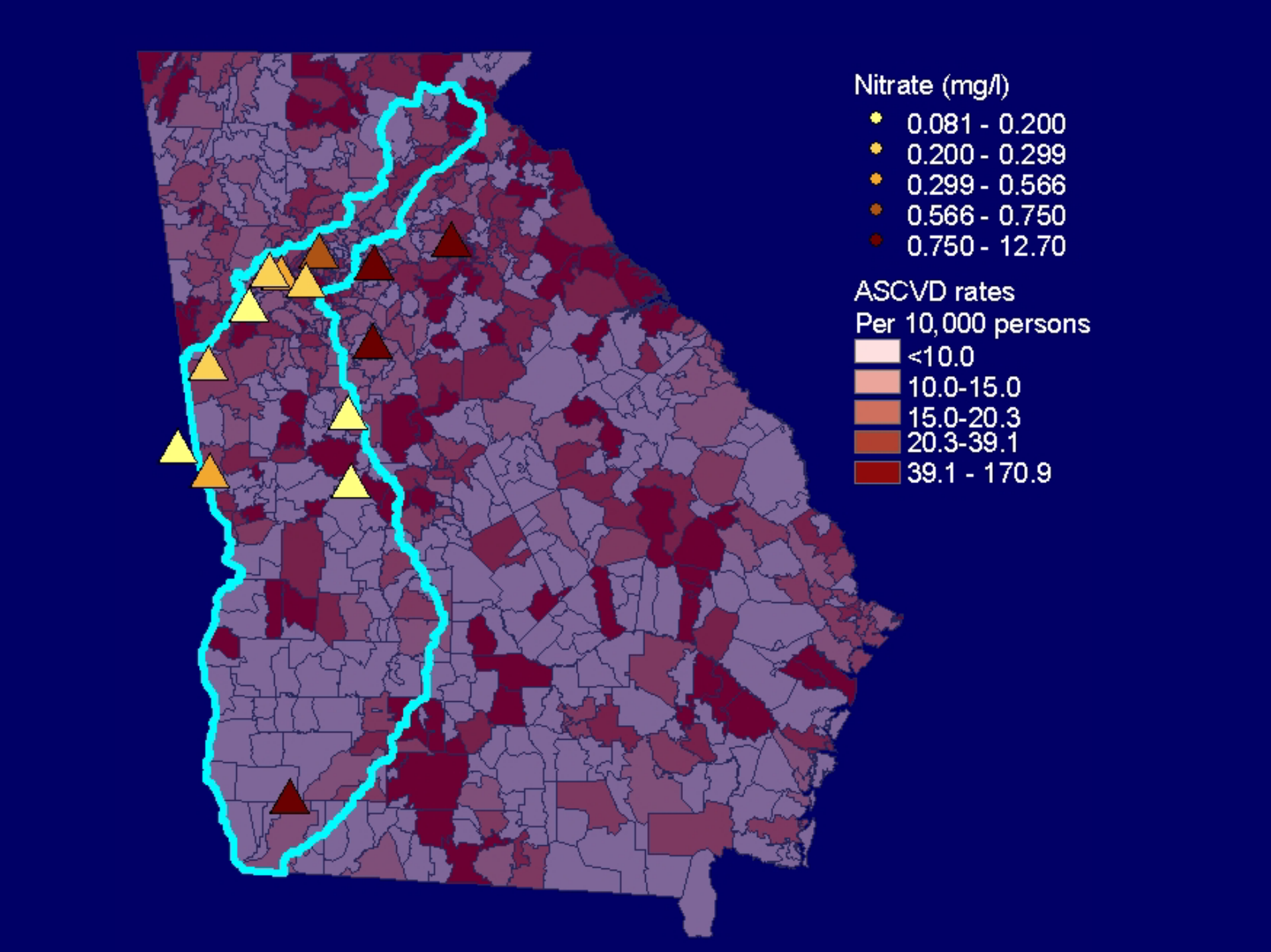


Emissions Data Locations

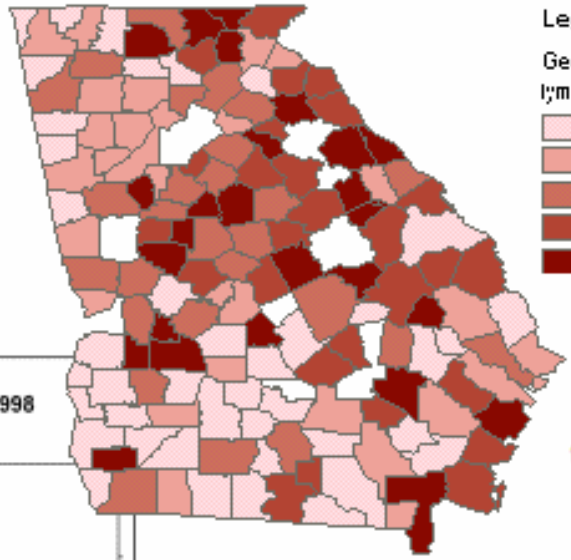


Linking Hospital Discharge and Water Quality Data

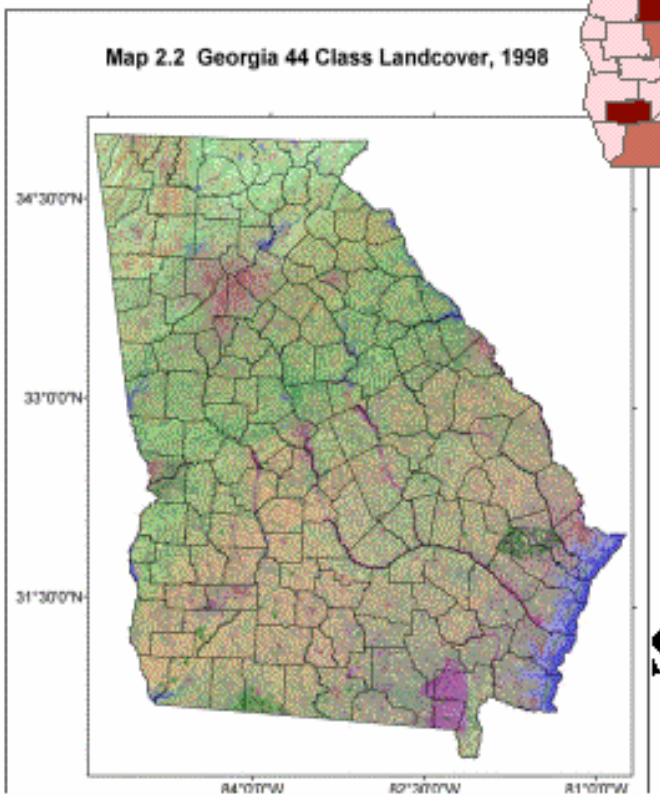
- Hospital discharge data: 1999-2003, GA, ZIP codes
- Environmental data from NAWQA
- Is there an association between **atherosclerotic cardiovascular disease** and **surface water nitrate concentration** in the Apalachicola-Chattahoochee-Flint river basin?

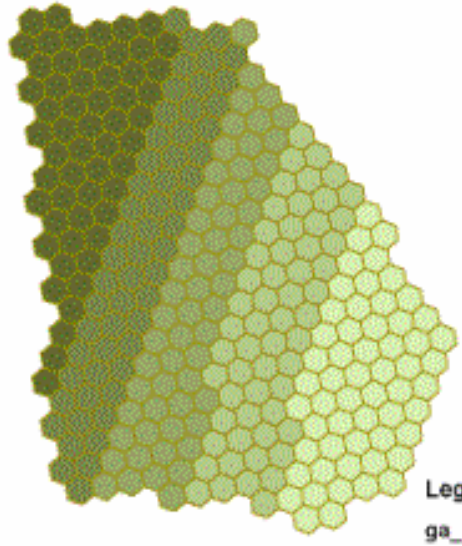

Nitrate (mg/l)
0.081 - 0.200
0.200 - 0.299
0.299 - 0.566
0.566 - 0.750
0.750 - 12.70
ASCVD rates
Per 10,000 persons
<10.0
10.0-15.0
15.0-20.3
20.3-39.1
39.1 - 170.9

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

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.

Two Aspects of the MAUP

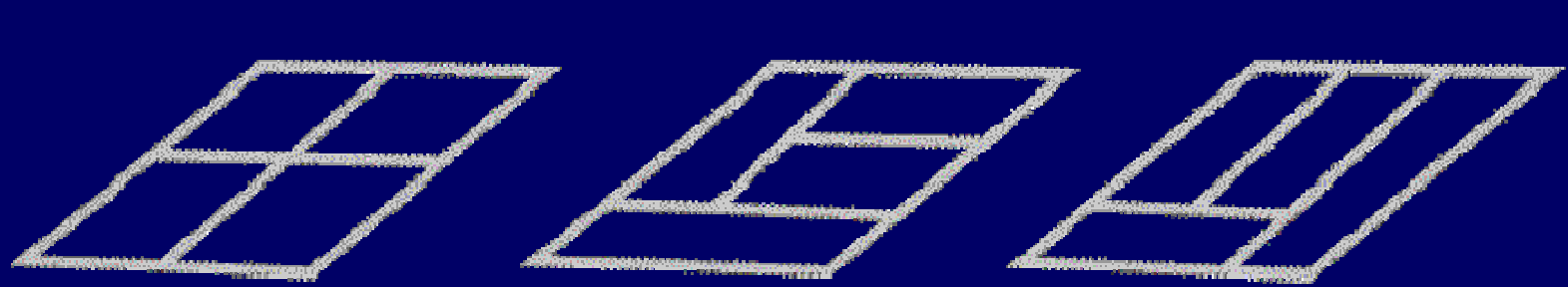
"Scale Effect" or the "Aggregation Effect":

Different results and inferences are obtained when the same set of data is grouped into increasingly larger areal units.

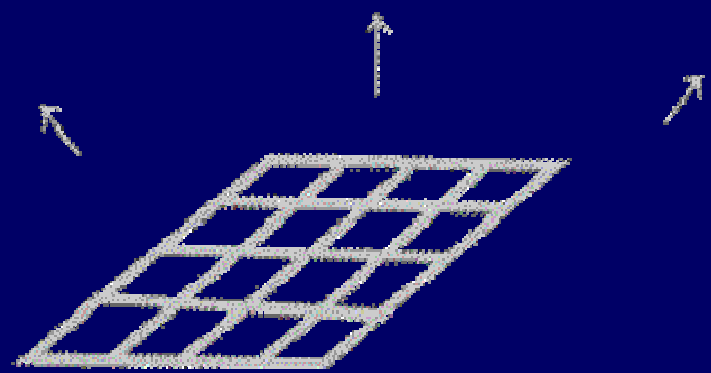
"Zoning Effect" or the "Grouping Effect":

The variability in results and inference due to alternative formations of the areal units.

Zoning Effect



Aggegation Effect

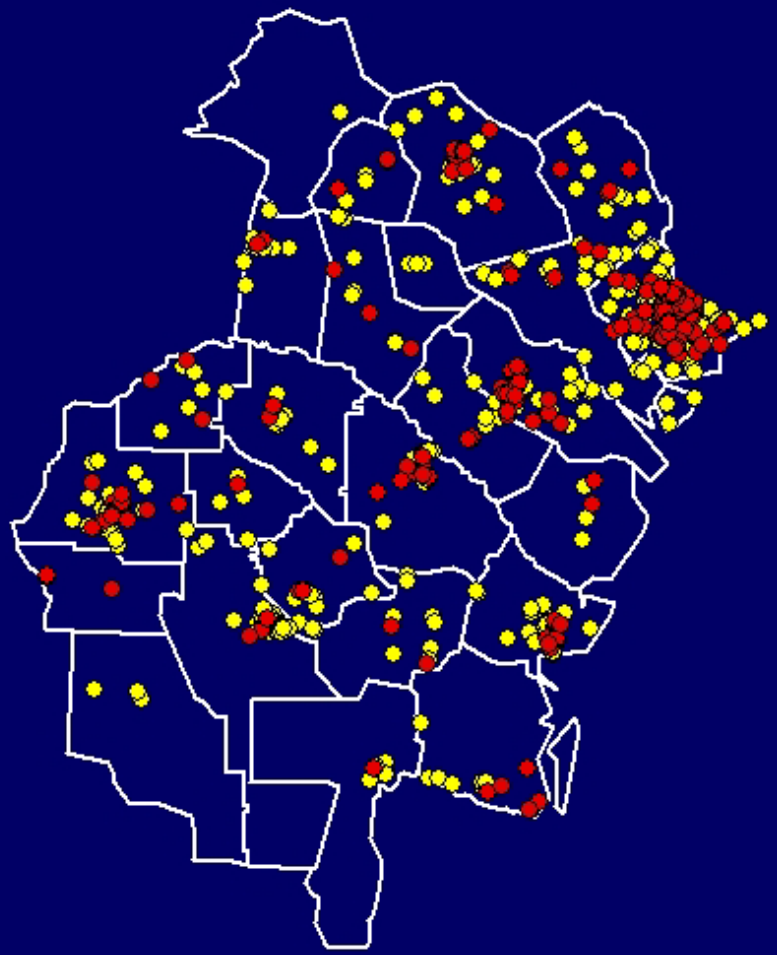


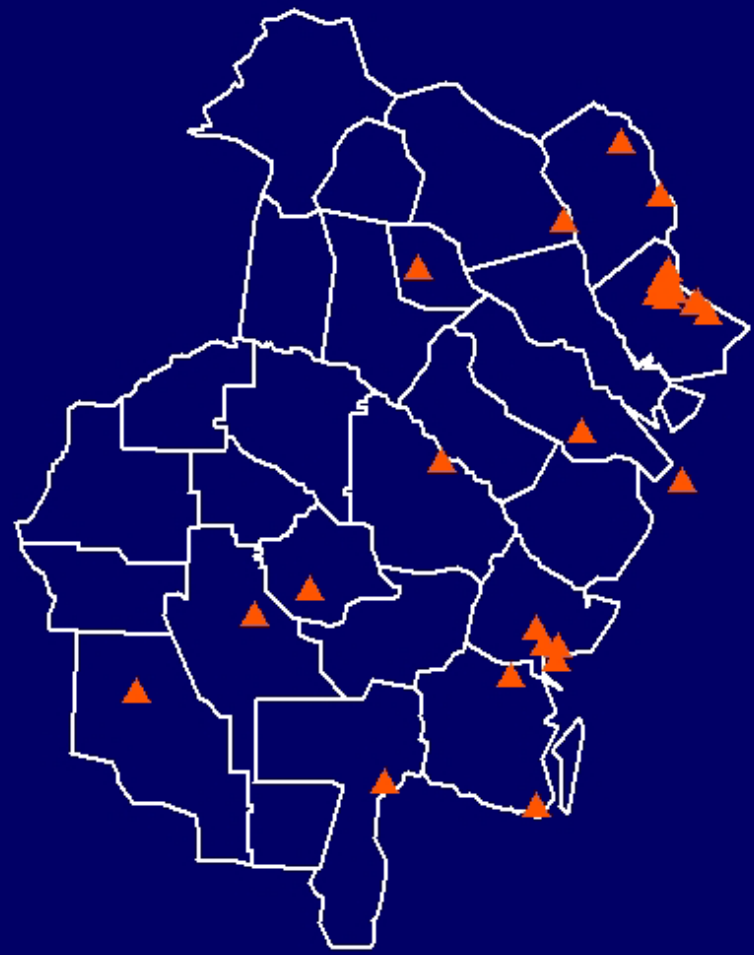
Spatial Support

- 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.

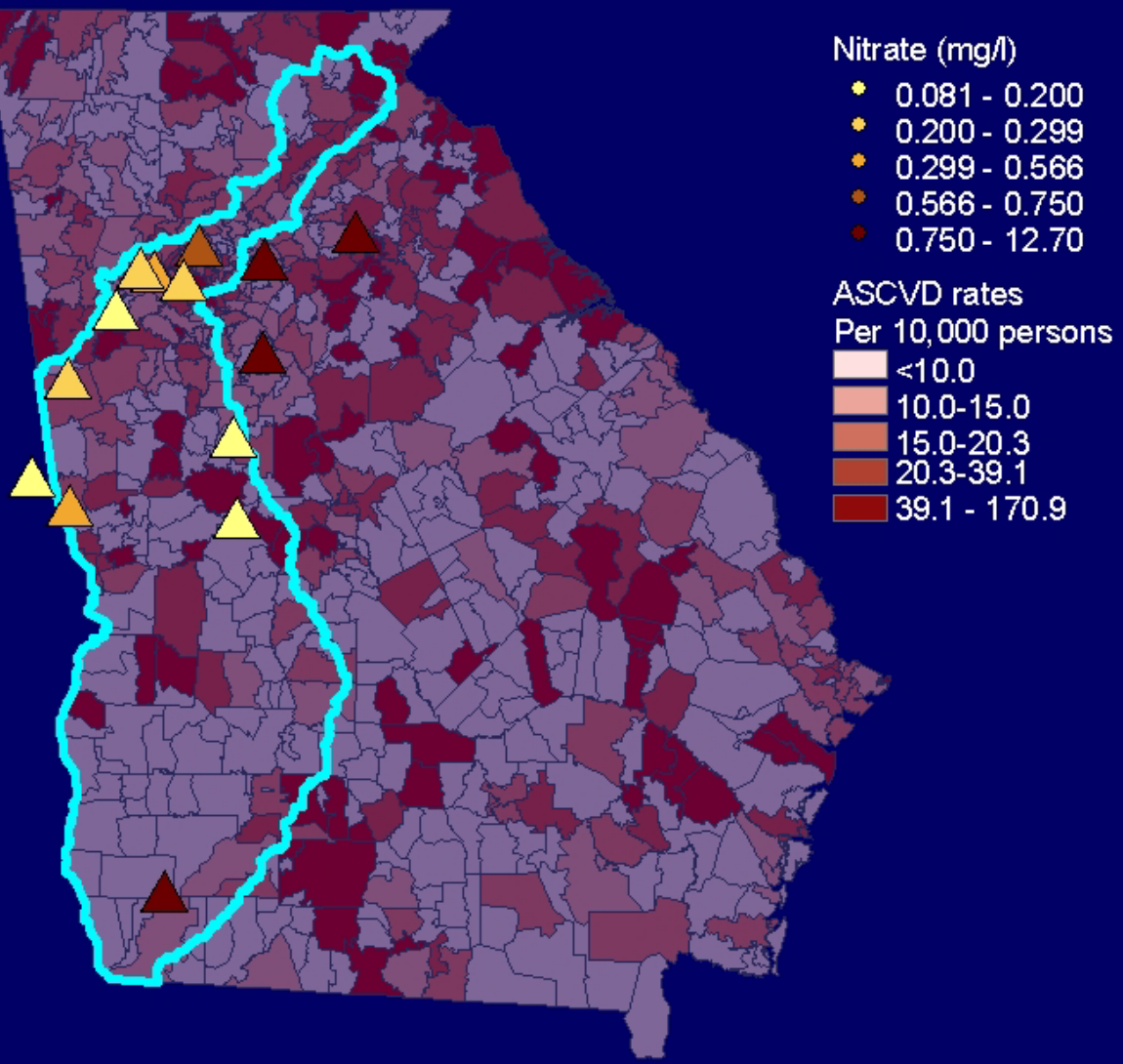
Case 1: Predict ground-level concentrations at each case and control location

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





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.

More General Approaches

- Regression Methods
- Multi-scale tree models
- Bayesian Hierarchical Models
- Geostatistical Methods

Gotway and Young (2002) *JASA*

Regression Methods

- Data on “source units” are modeled as functions of the areas of atomic units formed by intersection with “target units.” Flowerdew and Green (1992)

Advantages:

- can include covariates
- some give measure of uncertainty

Disadvantages:

- have to build atomic model
- may not be aggregation consistent
- need GIS for atomic areas
- often ignore spatial autocorrelation

Multi-Scale Tree Models

- Each level of the tree corresponds to a different spatial scale.
- Data are observed at some of the nodes of the tree and the goal is prediction at other nodes of the tree.
- Algorithms are based on the Kalman filter.
- Basseville et al. (1992), Chou et al. (1994), Huang et al. (2002)

Multi-Scale Tree Models

Advantages:

- can be computationally efficient
- work well with large data sets
- elegant statistical theory
- can get measure of uncertainty

Disadvantages:

- often ignore spatial support
- parameter estimation can be difficult
- not suitable for use with overlapping units?

Bayesian Hierarchical Models

- Specify a model for the data, given unknown variables, and then specify prior distributions for the unknown variables.
- Unknown variables may include data to be predicted.
- Kalman filtering and state space models can be special cases.

- Mugglin and Carlin (1998), Best et al. (2000), Wikle et al. (2001), Gelfand et al. (2001).

Bayesian Hierarchical Models

Advantages:

- can easily build complex models
- many multi-scale processes are inherently hierarchical
- can incorporate parameter uncertainty
- comprehensive description of uncertainty

Disadvantages:

- most use areal weighting only
- rely too heavily on Gaussian distributions and likelihoods
- prior choice and convergence issues
- Implementation difficult with large data sets

Geostatistical Approaches

- Predictions are optimal functions of the data.
- Optimal weights obtained by minimizing prediction mean-squared error (BLUPs).
- ``Block'' kriging, isofactorial models and ``co''-kriging.
- Can be implemented using Kalman filtering.

Journel and Huijbregts (1978), Matheron (1984), Daley (1992), Cressie (1993), Gotway and Young (2007)

Geostatistical Approaches

Advantages:

- geostatistics was invented to address the COSP
- active work in mining to increase profitability
- have optimal statistical properties
- can get a measure of uncertainty
- basic calculations can be done in GIS

Disadvantages:

- parameter estimation can be challenging
- requires simplified models for use with large data sets
- most based on linear models

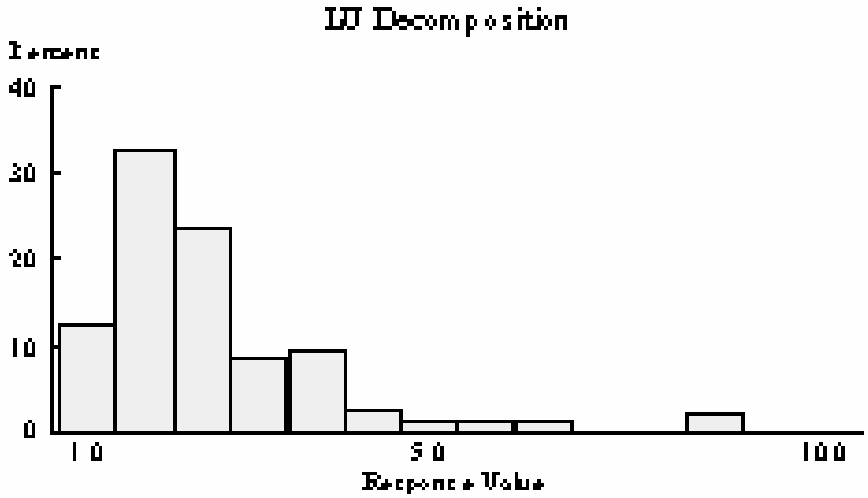
THE Most Important Issue: Uncertainty

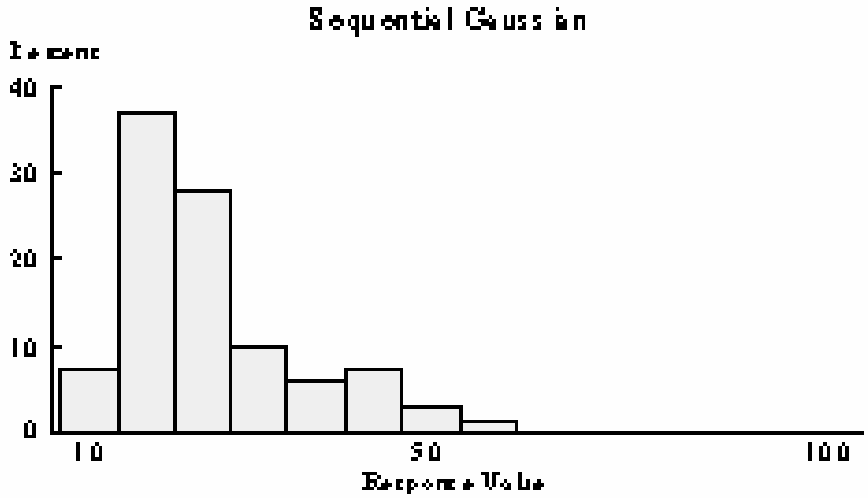
- Post-linkage analysis must account for the uncertainties that arise from prediction during linkage, as well as any uncertainties in the initial data.
- Otherwise, confidence intervals are too narrow, p-values are too small, and conclusions are probably wrong.
- Probabilistic prediction methods provide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

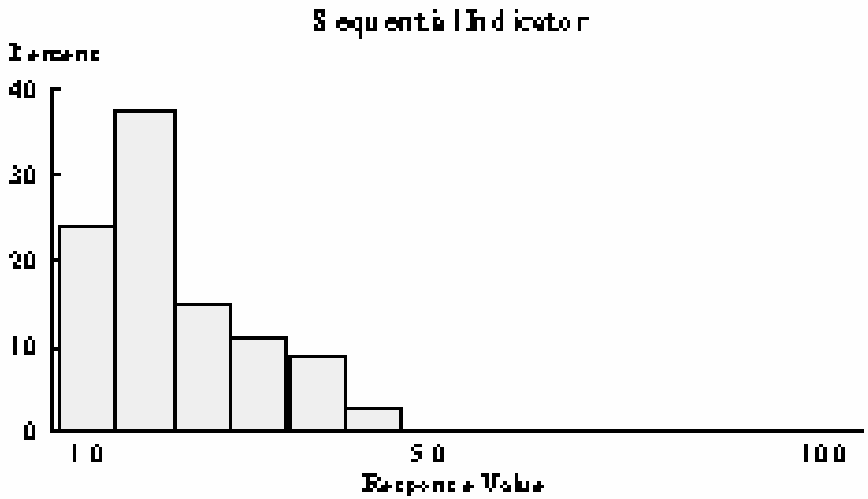
Methods for Quantifying Uncertainty

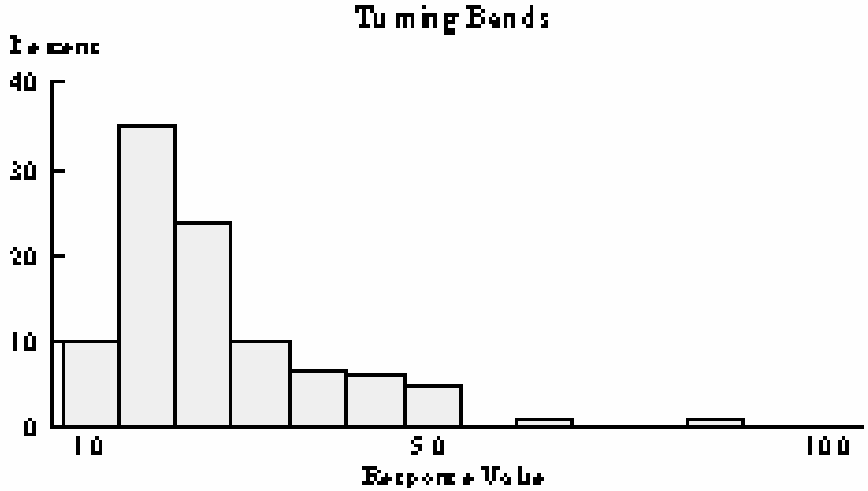
- Model the variability in the input, simulate from this model, analyze the resulting variation in output.
    - Monte Carlo/Geostatistical simulation
    - EM algorithm with maximum likelihood
    - Bayesian hierarchical models
- Computationally challenging particularly for COSPs.
- Properties of simulated data? Are uncertainty measures accurate?

Gotway and Rutherford (1994,1996).


LU Decomposition
Percent
40
30
20
10
0
10
50
100
Response Value


Sequential Gaussian
Percent
40
30
20
10
0
10
50
100
Response Value


Sequential Indicator
Percent
40
30
20
10
0
10
50
100
Response Value


Turning Bands
Percent
40
30
20
10
0
10
50
100
Response Value

Summary

- Spatial data are often more than just points in space. The support of the data is important.
- Corresponding to every spatial scale is a level of spatial aggregation that reflects a particular mixture of sub-units that comprise the larger units.
- Aggregation alters variability and impacts inference
- Quantifying uncertainty is very important. Assessing the accuracy of uncertainty measures is also important and has been largely overlooked.