A Geostatistical Approach to Linking Geographically-Aggregated Data From Different Sources

Carol A. Gotway Crawford
National Center for Environmental Health
Centers for Disease Control and Prevention, Atlanta, GA USA

and

Linda J. Young
Department of Statistics
University of Florida, Gainesville, FL USA

GeoSeer Workshop, May 10, 2004

Combining Incompatible Spatial Data

Digital spatial data are everywhere!

Data are often on different units and at different scales:

points, ZIP-code polygons, census tracts and blocks, counties, hydrogeologic regions, and user-defined.

Geographic information systems allow us to combine all this information rather easily:

overlay

buffering

geoprocessing: union and intersection

zonal averages

proportional allocation

Common Examples

1. Health outcomes recorded by zip-codes or counties;
   Sociodemographic data from Census tracts;
   Exposure estimates within a region of suspected source.

2. Cancer registries have geocoded data on cases;
   Control group obtained by population (Census blocks).

3. Election outcomes recorded by voting districts;
   Voter preferences linked to sociodemographic data (Census).

4. Boundaries change over time.

5. Species counts: spatial units are user-defined;
   Link to natural resource surveys, land use classification.

A Common Goal

Use all data to make valid inference for a particular set of units.

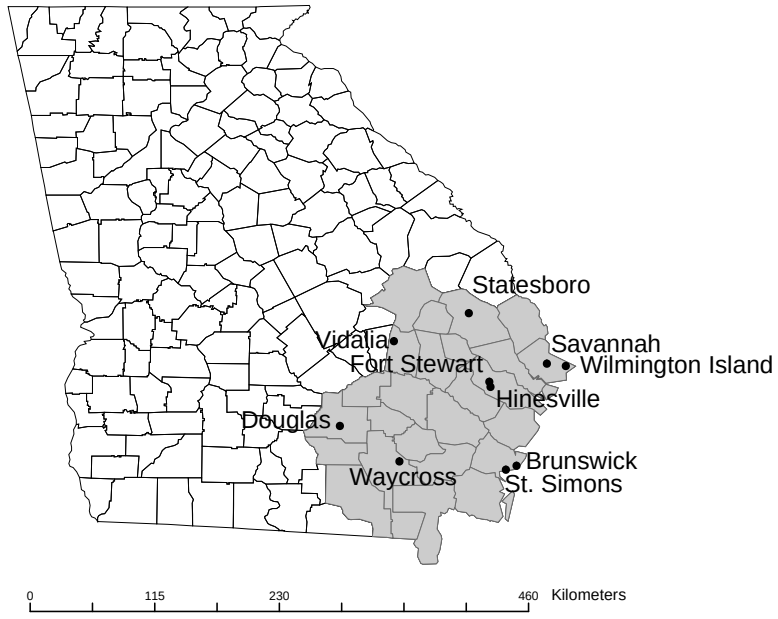
Often involves upscaling (aggregation), downscaling (disaggregation), or side-scaling (overlapping units).

Usually requires spatial prediction of data associated with one set of units based on data associated with another set of units.

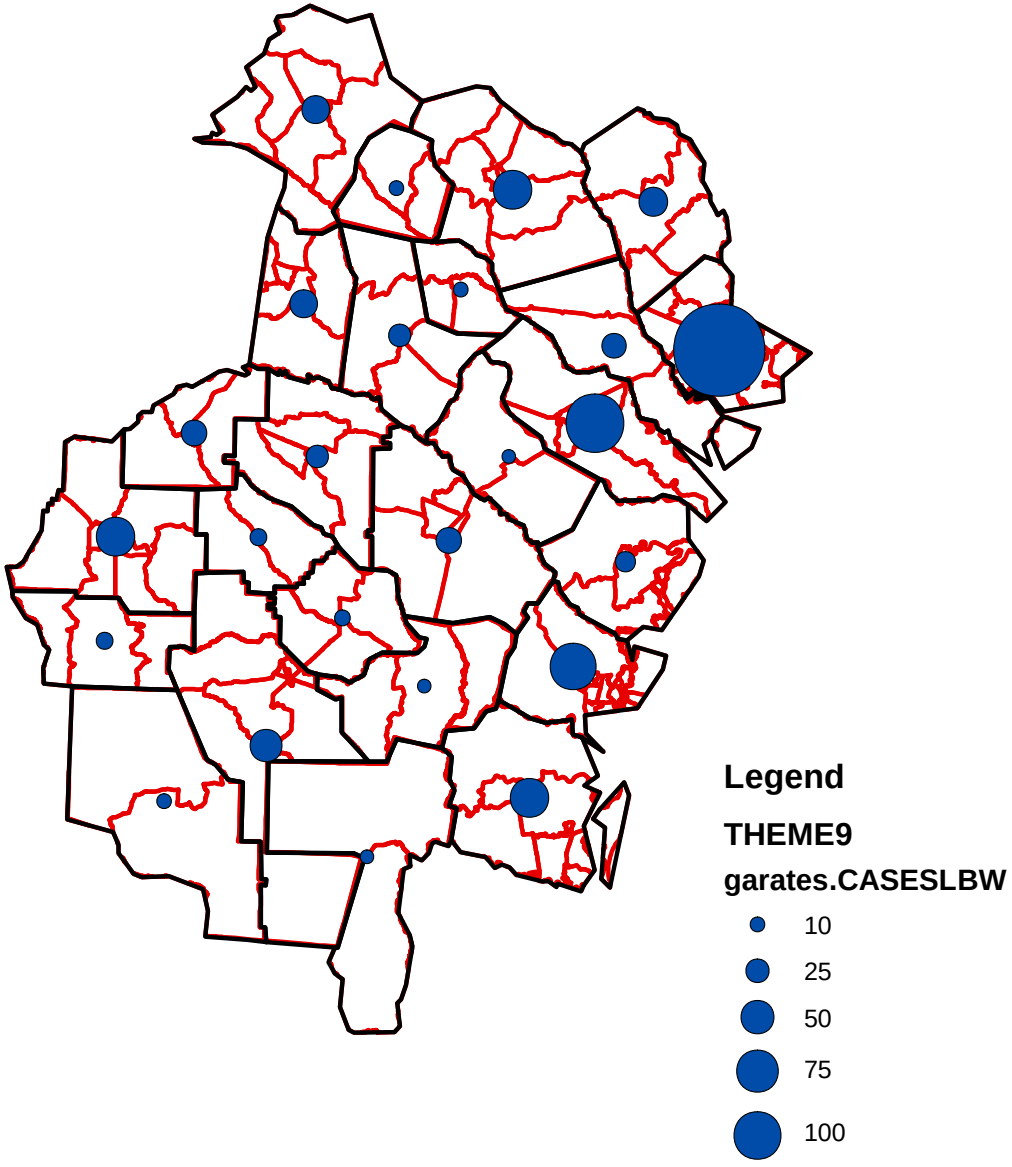
Georgia Health Care District 9

Georgia vital statistics reports include the number of low birth weight (LBW) babies per county

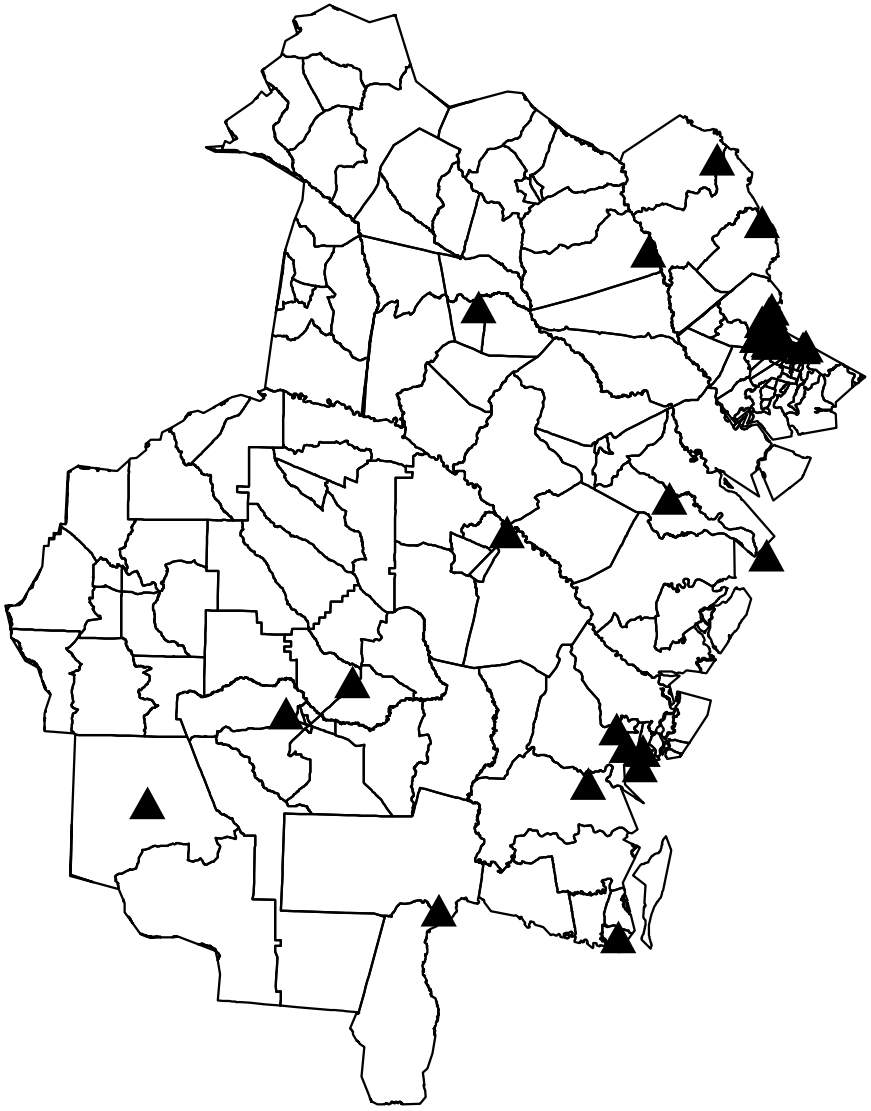
Most hypotheses about the factors contributing to LBW involve more local variables including sociodemographic factors measured for census tracts



LBW Counts in GHCD9



Industrial emissions facilities measure PM_{10} concentrations



Goal: Predict LBW counts at the Census tract level using LBW counts at the county level such that:

1. Predictions are nonnegative;
2. The total for tract predictions within a county is equal to the original county total;
3. Covariates can be used to improve predictions;
4. Standard errors of the predictions can be computed;
5. Spatial support is explicitly utilized;
6. No distributional assumptions;
7. Computations can be done within a GIS;
8. Approach generalizes to other problems (e.g., estimation of intensity surface that is smooth across boundaries).

A Geostatistical Framework

Let $\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \Re^2\}$, be a random field with point support.

$E(Z(\mathbf{s})) = \mu_s; \quad \text{cov}(Z(\mathbf{u}), Z(\mathbf{v})) = C(\mathbf{u}, \mathbf{v})$

We observe a realization of the aggregated process, $Z(B_1), Z(B_2), \ldots, Z(B_n)$, where

$$Z(B_i) = \int_{B_i} Z(\mathbf{s}) d\mathbf{s}$$

B_i are areal regions within D

GOAL: Predict data $Z(A_1), \ldots, Z(A_k)$.

Use best linear unbiased prediction:

$$\hat{Z}(A) = \sum_{i=1}^{n} w_i(A) Z(B_i)$$

weight $w_i(A)$ measures the influence of datum $Z(B_i)$ on the prediction of $Z(A)$.

Unbiasedness requires

$$\sum_{i=1}^{n} w_i(A)|B_i| = |A|.$$

Similarly, for any covariate x,

$$\sum_{i=1}^{n} w_i(A) x_{B_i} = x_{A_i}.$$

The optimal weights satisfy:

$$\mathbf{w}_c = (\Sigma_c)^{-1}\boldsymbol{\sigma}_c,$$

$$\Sigma_c = \begin{bmatrix} \Sigma_{BB} & X_B \\ \\ X_B' & 0 \end{bmatrix},$$

$$\mathbf{w}_c = (w_1(A), \ldots w_n(A)\ m)'$$

$$\boldsymbol{\sigma}_c = \begin{bmatrix} \sigma_{AB_i} \\ \\ X_A \end{bmatrix}.$$

The elements in Σ_{BB} are

$$C(B_i, B_j) = \text{cov}(Z(B_i), Z(B_j)) = \int_{B_j} \int_{B_i} C(\mathbf{u}, \mathbf{v}) d\mathbf{u} d\mathbf{v}$$

The elements in $\boldsymbol{\sigma}_{AB_i}$ are

$$C(A, B_i) = \text{cov}(Z(A), Z(B_i)) = \int_{A} \int_{B_i} C(\mathbf{u}, \mathbf{v}) d\mathbf{u} d\mathbf{v}$$

Each covariate can be related to point support via

$$x_B = \int_B x(\mathbf{s}) d\mathbf{s}$$

$$\mathbf{X}_B = (|\mathbf{B}|, \mathbf{x_{B_1}}, \ldots \mathbf{x_{B_j}})'$$

$$|\mathbf{B}|' = (|B_1|, |B_2|, \ldots, |B_n|)$$

$$X_A = (|A|, x_A)'.$$

The prediction mean-squared error, PMSE, is a measure of uncertainty associated with the prediction:

$$C_A(\mathbf{0}) - \mathbf{w}_c' \boldsymbol{\sigma}_c,$$

$$C_A(\mathbf{0}) = C(A, A) = \text{cov}(Z(A), Z(A)) = \int_A \int_A C(\mathbf{u}, \mathbf{v}) d\mathbf{u} d\mathbf{v}$$

- This is a variation on the universal block kriging predictor
- Consistency in aggregation is built in
- The approach is general; can be used for upscaling, downscaling, side-scaling and intensity estimation where A_i's are point locations $\mathbf{s}$.

Estimation of $C(\mathbf{u}, \mathbf{v})$

Assumed a parametric model $C(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta})$ to ensure positive definiteness.

Estimated $\boldsymbol{\theta}$ using *generalized estimating equations* (McShane et al. 1997).

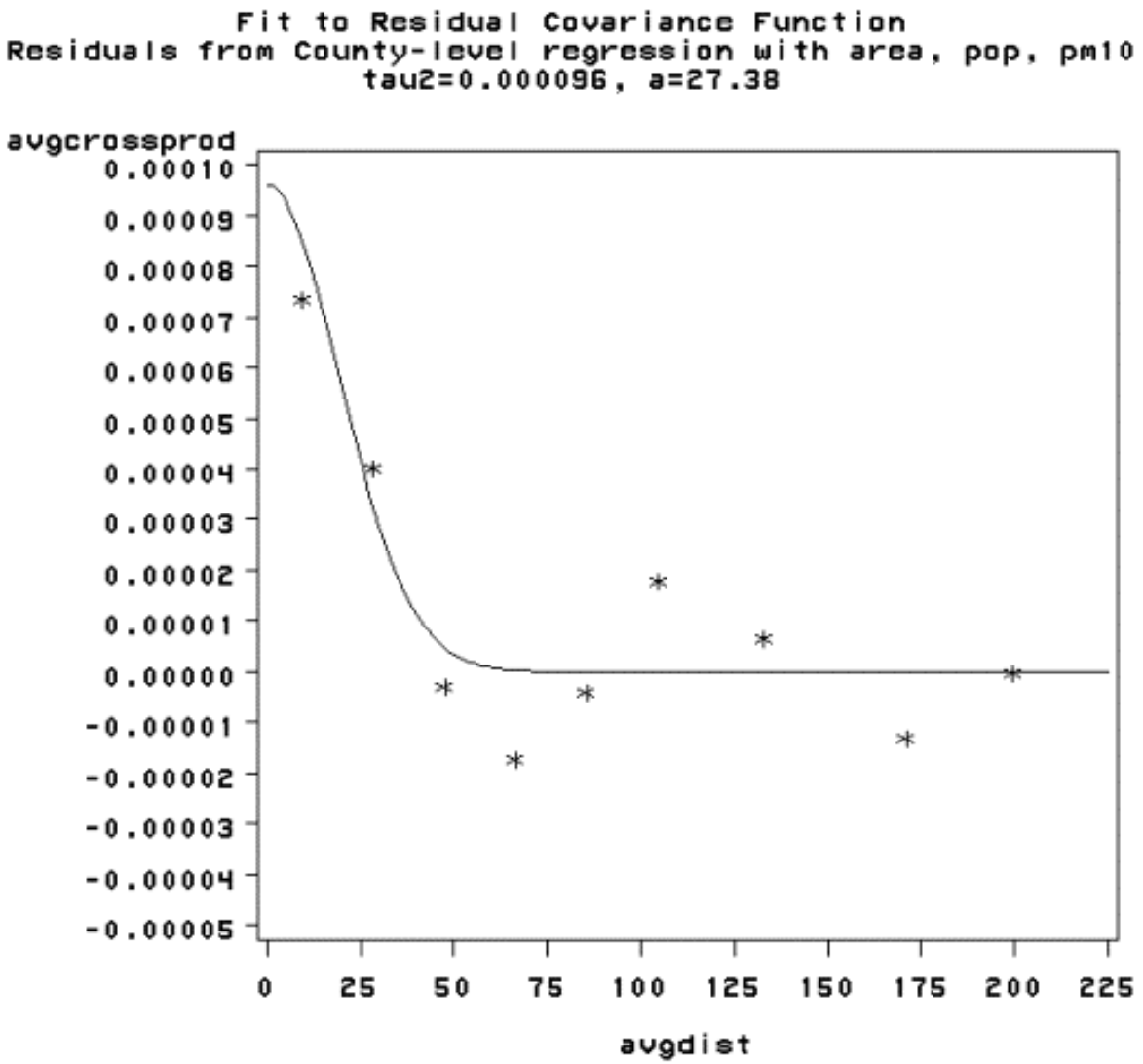
Algorithm is much like iteratively re-weighted generalized least squares, adjusted for support.

Implementation with GHCD9 Data

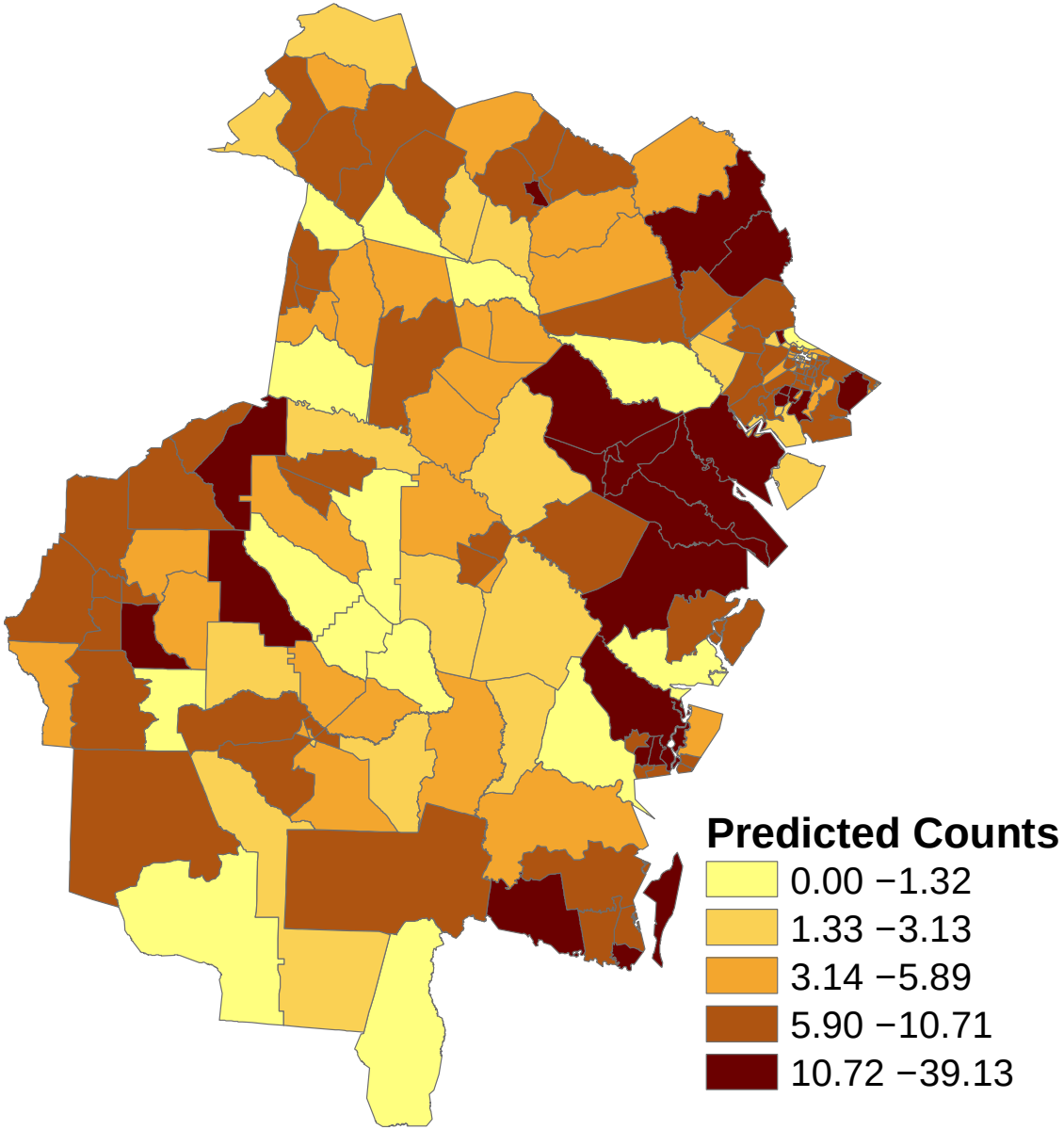
Covariates:

- x_1 : area, based on both county and tract support;
- x_2 : tract populations;
- x_3 : to account for dispersion and meteorological properties, an atmospheric transport model was used to obtain ground-level PM_{10} concentrations on a fine, regular grid; based on point support.

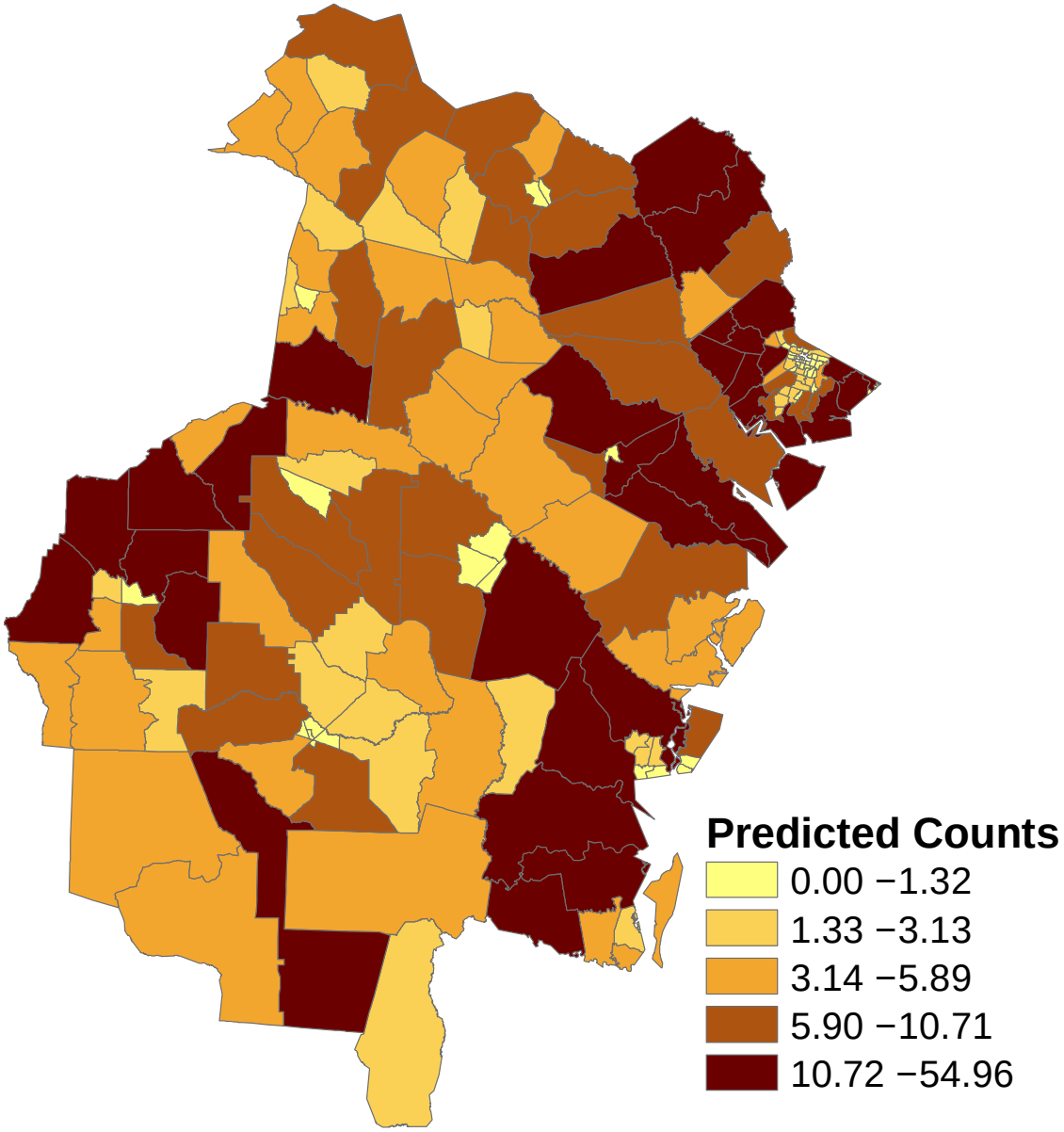
Compared results to Tobler's pycnophylactic interpolation and proportional allocation.


Fit to Residual Covariance Function
Residuals from County-level regression with area, pop, pm10
tau2=0.000096, a=27.38
avgcrossprod
0.00010
0.00009
0.00008
0.00007
0.00006
0.00005
0.00004
0.00003
0.00002
0.00001
0.00000
-0.00001
-0.00002
-0.00003
-0.00004
-0.00005
0
25
50
75
100
125
150
175
200
225
avgdist

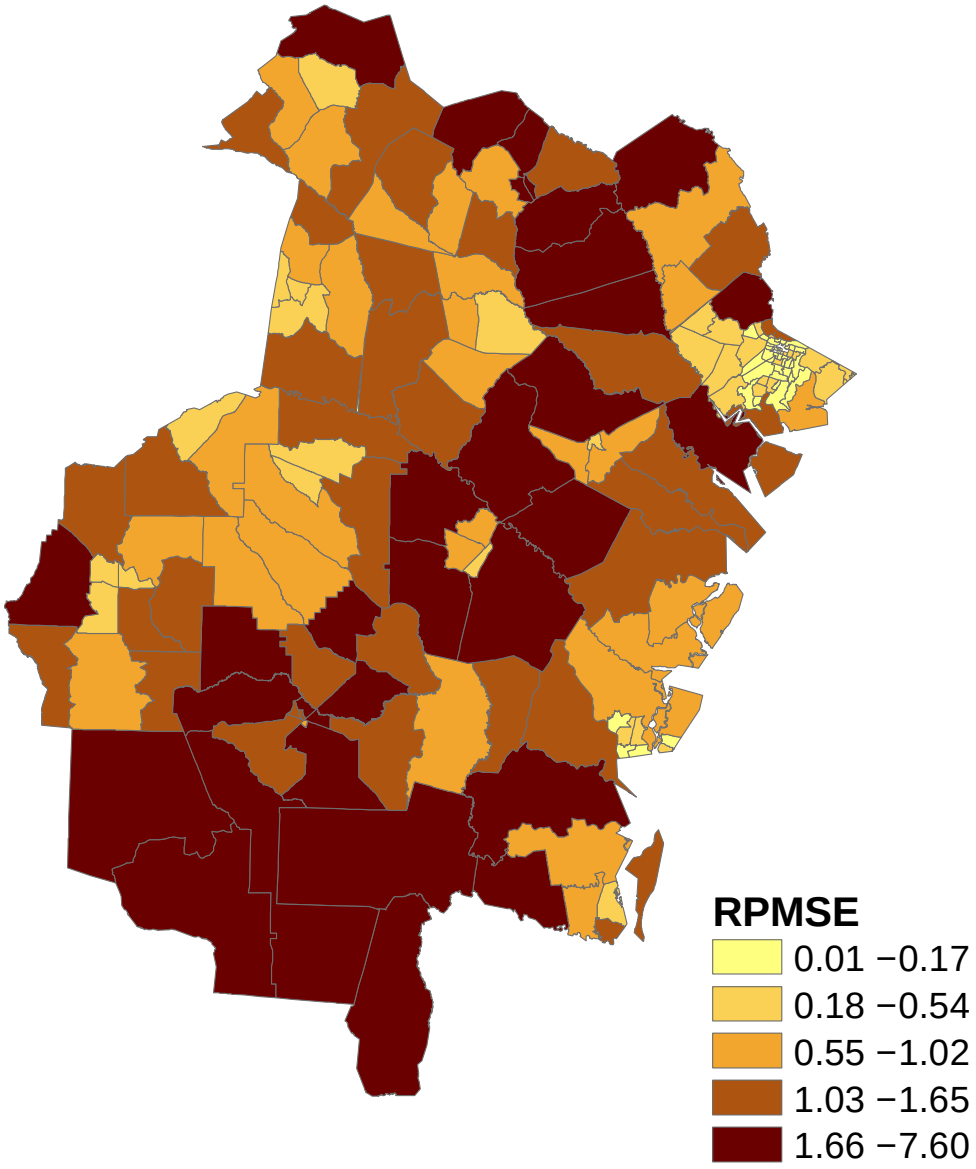
Predicted Tract LBW Counts from Geostatistical Method



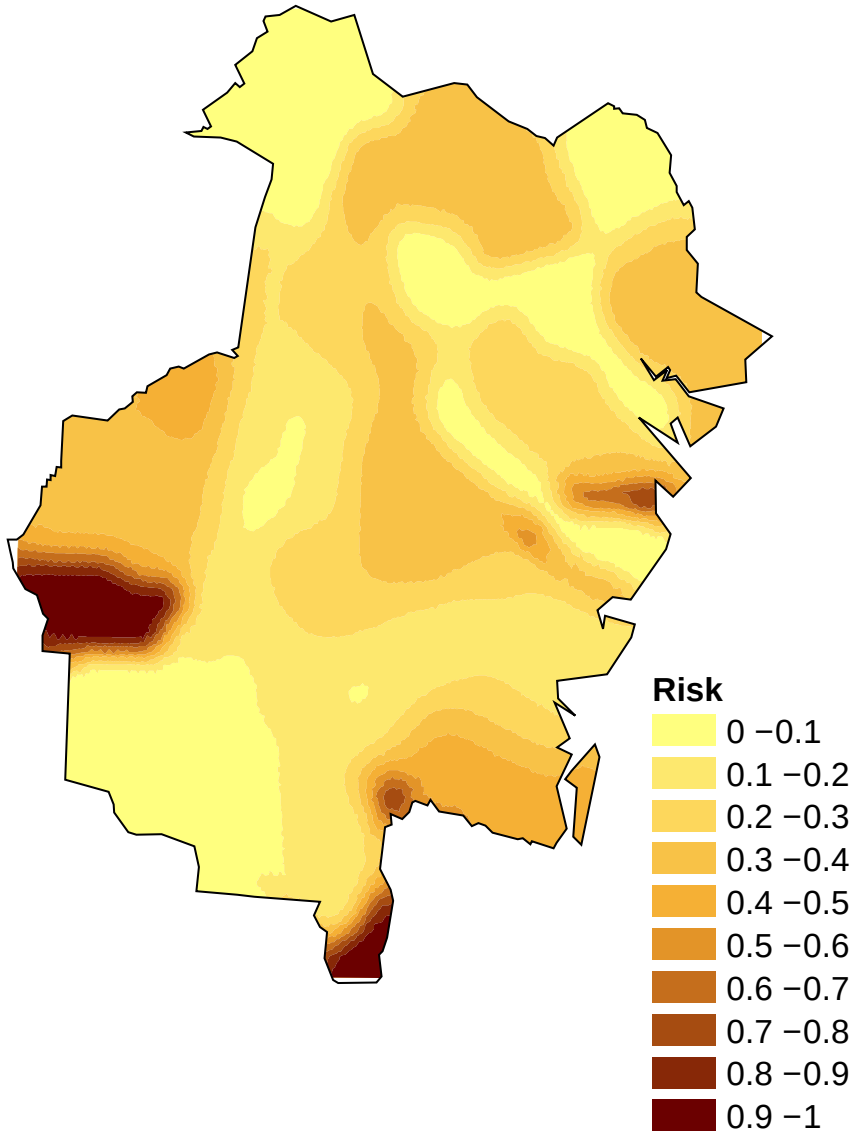
Estimated Tract LBW Counts Using Proportional Allocation



Root Mean-Squared Prediction Errors for Geostatistical Method



Relative Risk Based on Predicted Intensity



Geostatistical Approach: Summary and Conclusions

- General Framework:

  Upscaling: point $\uparrow$ area; or A $\uparrow$ B;

  Downscaling: area $\downarrow$ point or $B \downarrow A$;

  Sidescaling: $C \rightarrow$ D (e.g., ZIP codes to Census tracts).

- Allows data and predictions to be autocorrelated;

- Allows assessment of prediction uncertainty;

- Covariates can be used to enhance predictions;

- Makes more use of spatial information by explicitly incorporating spatial support in the analysis;

- Gives results similar to traditional methods when covariates are not included;
- Proportional allocation is a special case;
- Flowerdew and Green (1994)'s iterative, mass-balance adjusted, linear regression is a special case;
- No distributional assumptions;
- Feasible in GIS with current technology

Outstanding Issues and Future Research

- Smart calculation of complex integrals;
- Positivity. Was not as problematic as expected. An extra constraint may increase uncertainty and computational complexity;
- Estimation of point-point covariance function. Much is unknown;
- Uncertainty about modeled covariates. Can use geostatistical simulation;
- Automation and choice of defaults for "black box" GIS;
- **This is just a beginning. Other ideas from geostatistics may be applicable.**

References

Bracken, I. and D. Martin (1989). The generation of spatial population distributions from census centroid data. *Environment and Planning A 21*, 537–543.

Cressie, N. (1993). Aggregation in geostatistical problems. In A. Soares (Ed.), *Geostatistics Troia '92*, Dordrecht, pp. 25–35. Kluwer Academic Publishers.

Cressie, N. (1996). Change of support and the modifiable areal unit problem. *Geographical Systems, 3*, 159–180.

Flowerdew, R. and M. Green (1994). Areal interpolation and types of data. In S. Fotheringham and P. Rogerson (Eds.), *Spatial Analysis and GIS*, London, pp. 121–145. Taylor and Francis.

Gelfand, A. E., L. Zhu, and B. P. Carlin (2001). On the change of support problem for spatio-temporal data. *Biostatistics 2*, 31–45.

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

Huang, H.-C., N. Cressie, and J. Gabrosek (2002). Fast resolution-consistent spatial prediction of global processses from satellite data. *Journal of Computational and Graphical Statistics 11*, 1–26.

Kelsall, J. and J. Wakefield (2002). Modeling spatial variation in disease risk: A geostatistical approach. *Journal of the American Statistical Association 97*, 692–701.

McShane, L., P. Albert, and M. Palmatier (1997). A latent process regression model for spatially correlated count data. *Biometrics 53*, 698–706.

Mugglin, A. S. and B. P. Carlin (1998). Hierarchical modeling in geographic information systems: Population interpolation over incompatible zones. *Journal of Agricultural, Biological, and Environmental Statistics, 3*, 117–130.

Müller, H.-G., U. Stadtmüller, and F. Tabnak (1997). Spatial smoothing of geographically aggregated data, with application to the construction of incidence maps. *Journal of the American Statistical Association, 92*, 61–71.

Tobler, W. (1989). Frame independent spatial analysis. In M. Goodchild and S. Gopal (Eds.), *The Accuracy of Spatial Data Bases*, London, pp. 115–122. Taylor and Francis.

Wikle, C. K., R. F. Milliff, D. Nychka, and L. M. Berliner (2001). Spatio-temporal hierarchical Bayesian modeling: Tropical ocean surface winds. *Journal of American Statistical Association, 96*, 382–397.