

# Spatial Methods for Election Prediction: An Interdisciplinary Study

A.C. Thomas\*

*Harvard Department of Statistics*

(Dated: January 14, 2005)

As evening approached on November 2nd this past year, Democratic households across the country were giddy. John Kerry's election was all but certain after exit polls from a well-regarded pollster, Zogby International, predicted Kerry victories in Ohio, Florida and Pennsylvania, three highly sought electoral prizes. Winning just two of these states would cement a victory for either candidate.

And win two of these is exactly what President Bush managed to do, after taking Ohio and Florida, and with them the possibility of a change of management.

Were the polls inaccurate? Definitely not; the final tallies all fell within the 3 percent margin of error. But there must be other ways to approach the problem that give us new ways of thinking about the problem. And statewide exit polls don't make full use of one vital set of information: geography.

Geostatistics was pioneered by geologists searching largely for mineral concentrations and other valuable deposits.

Geostatistical methods of interpolation are most often used to extract a continuous spectrum of information based on a set of points, but this does not limit our ability to predict values at any set of locations as well. County borders are a natural set of partitions within which results are collected; by approximating each county with a single point, carrying voting data for the region, we can theoretically use this as a base for interpolation.

Besides, politicians consider voter bases to be little more than commodities, so what's stopping statisticians from thinking the same way?

---

\*Electronic address: [athomas@stat.harvard.edu](mailto:athomas@stat.harvard.edu); URL: <http://www.fas.harvard.edu/~acthomas/>

## I. EMPIRICAL SEMIVARIOGRAMS AND KRIGING

In this section, I'll give a short briefing thing on bin-fitted semivariograms, and how we use this information to make predictions through universal kriging, with credit where credit is due.

## II. STARTING WITH CROSS-VALIDATION

As an important first step to developing this method, we need to see how well these methods predict themselves. One simple method we can use is leave-one-out cross-validation, predicting a value at each data point, using all other points, and finding the difference from the actual value. This most closely resembles the Conditional Autoregressive Model in lattice statistics. (It can be explored more closely.)

Next would be to try a bigger number to leave out, say 10, 20, or 30 counties at a time, and check the accuracy of this prediction.

## III. PLACES AND TIMES

Started with Iowa, 1996, 2000, 2004. Showed that fitted semivariogram values were all comparable. This suggests that we can use snapshots from previous years in order to make predictions.

Next, we expanded to surrounding states. Can we use Minnesota, Wisconsin, Illinois (etc.) counties to improve our results? (Depends on what data can be obtained.) Also, how do our predictions fare in each of these other areas?

Clearly, the best case for this sort of analysis is a "rising tide" model, where an additive rise in a percentage of votes in one county indicates the same rise in every other county. (Of course, this is dreaming, since it would represent a very high correlation coefficient.)

## IV. MAJOR OBSERVATIONS

Most of the error will come from poorly predicting results in big counties. In Iowa, the biggest difference between counties is roughly 75-fold in population.

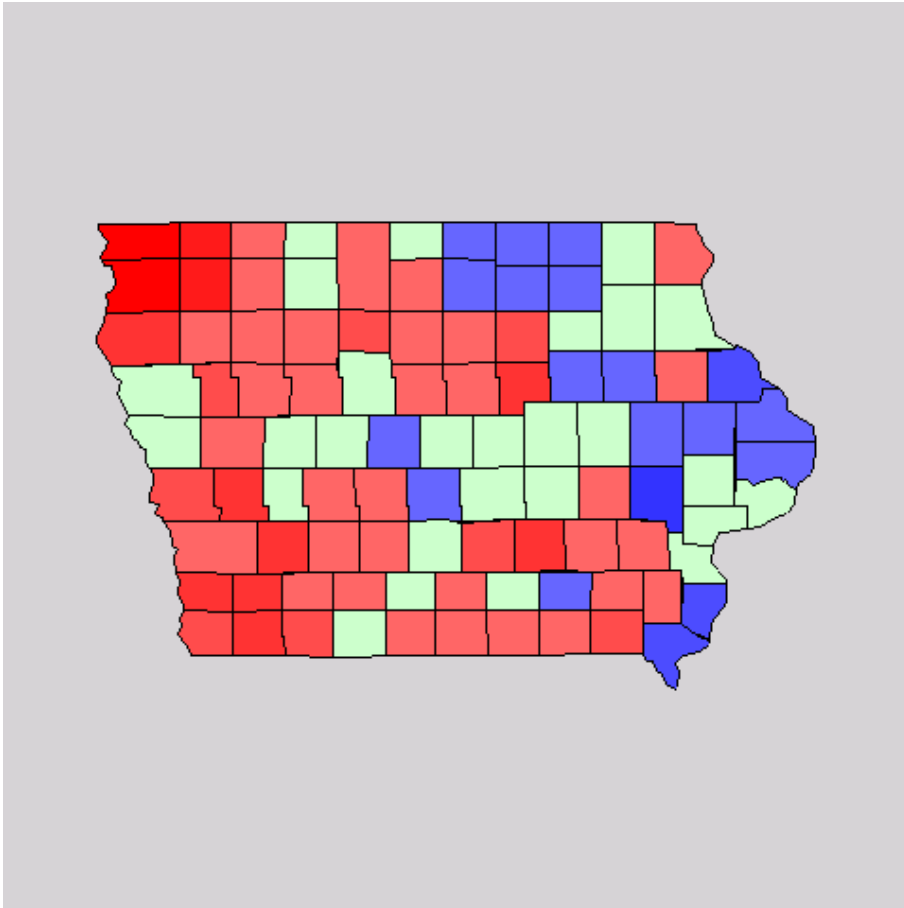


FIG. 1: The results of the 2000 Presidential Election by county in Iowa.

There is also a clearly dominating east-west trend in Iowa, though it's more pronounced in the predicted data. This suggests that this anisotropy is smoothing out the data to too much of a degree.

## V. CONCLUSIONS

Here's where I state whether these methods are better predictors than using the same ones as others. I did these analyses first in GeoR, then in gstat (as I found cokriging in the latter.) I would need to make a point of comparison, but I'm not sure how to "cokrige" lattice data. I certainly don't know how to do it in R, though I did manage to do an easy CAR model LOOCV in R as well.

The main point that I wish to bring across is the marriage of different methods used to solve the problem, as to my knowledge this particular technique hasn't been used (though I

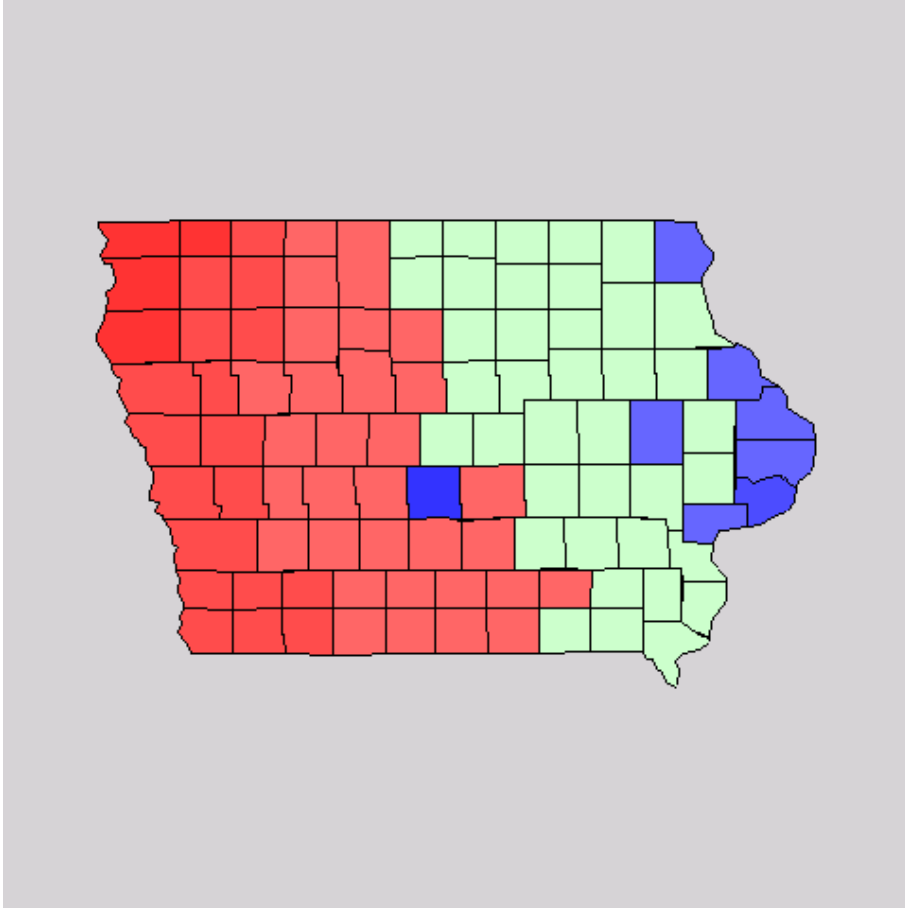


FIG. 2: The results predicted through leave-one-out cross-validation for Iowa in 2000.

could easily be mistaken.)