

## Collaborative Research: Quality graphics for federal statistical summaries – results

Alan M MacEachren<sup>1&2</sup>, Daniel Carr<sup>3</sup>, David Scott<sup>4</sup>

<sup>1</sup>GeoVISTA Center, Penn State

<sup>2</sup>Dept. of Geography, Penn State

<sup>3</sup>Department of Applied and Engineering Statistics, George Mason University

<sup>4</sup>Department of Statistics, Rice University

correspondence to: [maceachren@psu.edu](mailto:maceachren@psu.edu); [dcarr@galaxy.gmu.edu](mailto:dcarr@galaxy.gmu.edu); [scottdw@stat.rice.edu](mailto:scottdw@stat.rice.edu)  
[www.geovista.psu.edu/grants/dg-qg](http://www.geovista.psu.edu/grants/dg-qg)

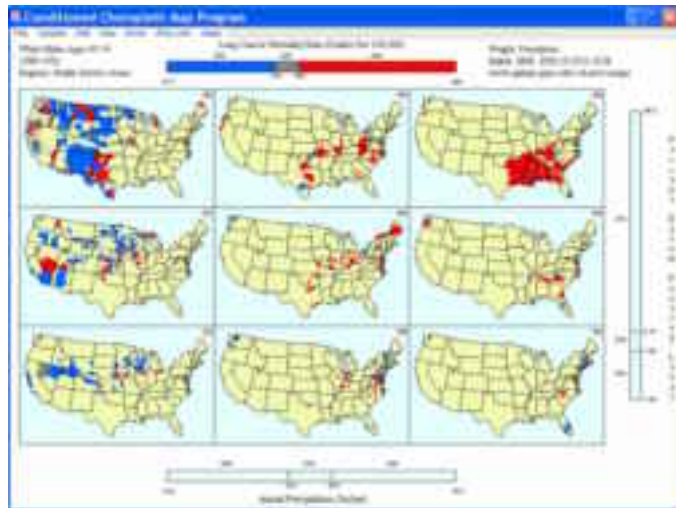
The **specific** objective of our *dgQG* research is to develop and assess quality graphics for **federal statistical summaries**. The goal has been to develop methods for generating quality graphics that facilitate **exploration** by agency users evaluating data quality and looking for emergent trends, **decision making** by public policy makers, and **communication** of statistical summaries to the public. Among the project accomplishments, efforts can be categorized into three general foci: developing ESDA methods, supporting public communication, and facilitating internal data quality review by agency staff. Examples of ESDA advances are highlighted below.

### Developing ESDA Methods

One major focus of work on this project has been development of a range of exploratory spatial data analysis (ESDA) methods and tools that integrate visual and statistical approaches to exploring data to identify multivariate patterns and relationships. Three related developments are highlighted: conditioned choropleth maps, ESDA applications created through GeoVISTA Studio, and spatial smoothing and outlier detection.

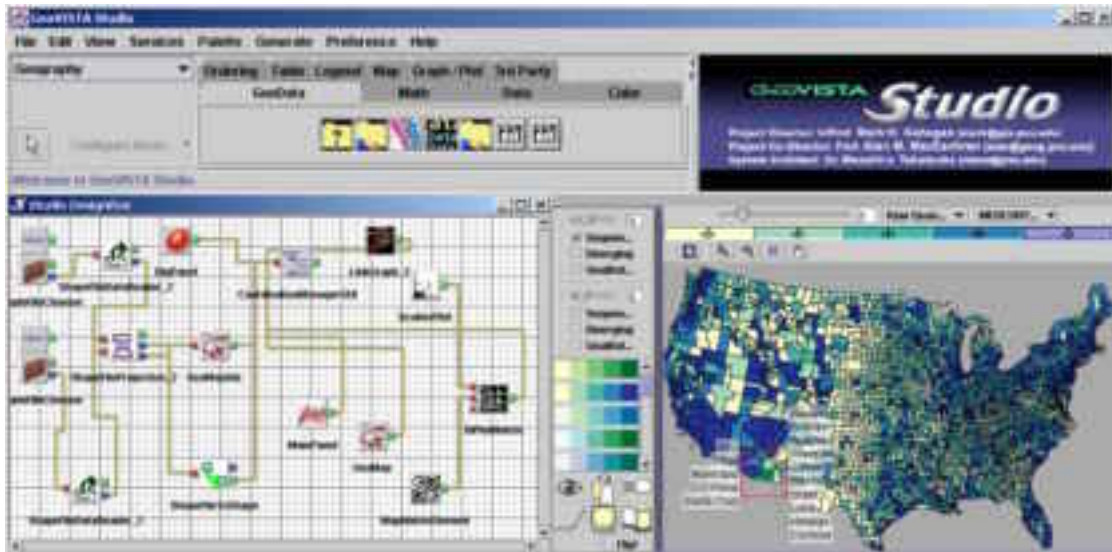
#### *Conditional Choropleth Maps*

(CCmaps) provide a two-way layout of maps designed to facilitate comparisons. Our recent, highly interactive implementations of CCmaps help users to compare the distributions of conditioned subsets of geospatial data (e.g., data for U.S. counties). Patterns evident across subsets indicate the association of conditioning variables with the dependent variable. One goal of this tool is to prompt hypothesis about scientific relationships behind the apparent associations.



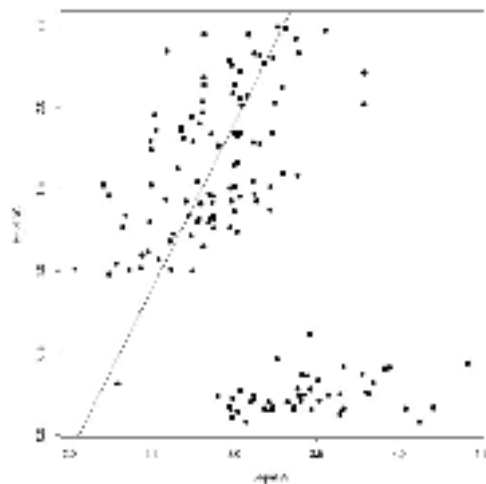
*GeoVISTA Studio* has been leveraged to support rapid development of new visual and computational analytical methods through its facilities for integrating independently developed components. Recent work has focused on a Multiform bi-variate Matrix tool and dynamically connected LinkGraph. The former generalizes the well-know scatterplot matrix to support any bivariate representation forms (we have implemented bi-variate choropleth maps and space-filling visualizations). The latter uses a minimum spanning tree to define relative position of geographic places (e.g., counties) in attribute space (e.g., to find all counties in the U.S. that are similar in demographic-health space to Centre county in PA). *Studio* is open source software distributed through

SorceForge: <http://geovistastudio.sourceforge.net/>. The view below shows the design window (left) where applications are built and the view on the right depicts three integrated components, a choropleth map, a Java implementation of the ColorBrewer color scheme selection tool (see: [www.colorbrewer.org](http://www.colorbrewer.org)), and an excentric labelling tool added from work by Fekete and Plaisand (see: <http://www.cs.umd.edu/hcil/excentric/>).



### *Spatial Smoothing and Outlier Detection.*

In numerical analysis, specialized algorithms can find the first few eigenvectors of a large covariance matrix. Suppose the data are not from a single cluster. Can you find the largest eigenvalue/vector of *one* of the clusters? Our most recent advance is the L2E algorithm, which tries to find a line that goes through a subset of the data with minimal variance. This so-called "skewer" is essentially the largest eigenvector of the local points. For example, The L2E algorithm line shown is estimated to fit 66.7% of the data. These well-known data on iris flowers have multiple clusters. Without modeling any other clusters, the algorithm finds the largest eigenvector of one of the clusters. This is a completely new capability that means far less complex mixtures models need be fit than by ordinary clustering algorithms.



For more information, see:

- Carr, D. B., MacPherson, D., White, D., & MacEachren, A. M. in press, Conditioned choropleth maps and hypothesis generation. *Annals of the Association of American Geographers*.
- MacEachren, A., Dai, X., Hardisty, F., Guo, D., & Lengerich, G. 2003, Exploring High-D Spaces with Multiform Matrices and Small Multiples. *Proceedings of the International Symposium on Information Visualization*, Seattle, WA, Oct. 19-21, 2003, pp. 31-38.
- Scott, D.W. & Christian, J.B. 2003. Finding outliers in models of spatial data. *Proceedings of the Third National Conference on Digital Government*, E. Hovy, Ed, Digital Government Research Center, Boston.