

# Web-Based Collaborative Tools for Geospatial Data Exploration

Dai, Xiping; Guo, Diansheng; MacEachren, Alan; Zhou, Biliang; Chen, Jin; and Macgill, James  
GeoVISTA Center & Department of Geography  
302 Walker Building, Pennsylvania State University  
University Park, PA 16802-5011

Federal government agencies have collected large and highly complex federal statistical summaries and geo-reference data. Methods and tools for identifying patterns and uncovering multivariate relationships in the complex datasets have drawn attention in multiple fields. The approach we present here extends from and integrates perspectives from multiple fields, including exploratory data analysis (EDA), geovisualization, information visualization, and data mining. Important components of our research are supported by a NSF Digital Government grant (#99883451). Other support comes from a National Cancer Institute (NCI) contract. Our DG research is now being extended to develop comprehensive multivariate analysis methods and tools, through a grant from NCI (CA95949).

Over the past year, (in addition to enhancement of existing tools), new multivariate analysis and visualization tools and functionalities added include: (a) feature selection and multivariate clustering tools for identifying interesting subspaces from high dimensional datasets, (b) a custom parallel coordinate plot (PCP) application, which supports the analysis requirements of NCI, (c) extensions of ColorBrewer for bivariate mapping, and (d) integration of GeoVISTA *Studio* and GeoTools, an open source Java library for developing OpenGIS solutions to geospatial data access, analysis, and presentation tasks.

The feature selection and multivariate clustering tools support work that develops a human-centered, component-based, exploratory spatial data analysis environment for discovering patterns in large and high-dimensional data, e.g., various census data, public health data, etc. The implemented system includes a suite of computational and visual components, each of which focuses on a specific task or step in the overall data exploration process and together they can communicate with each other and collaboratively address complex problems. Specifically, this part of our suite of methods and tools includes: (1) an interactive feature selection method for identifying interesting subsets of variables, (2) an interactive, hierarchical clustering method for searching multivariate clusters, and (3) a set of coordinated visual and computational components centered around the above two methods to facilitate an visually-enabled, efficient, effective, human-led exploration of multivariate spatial patterns. Figure 1 shows a snapshot of the

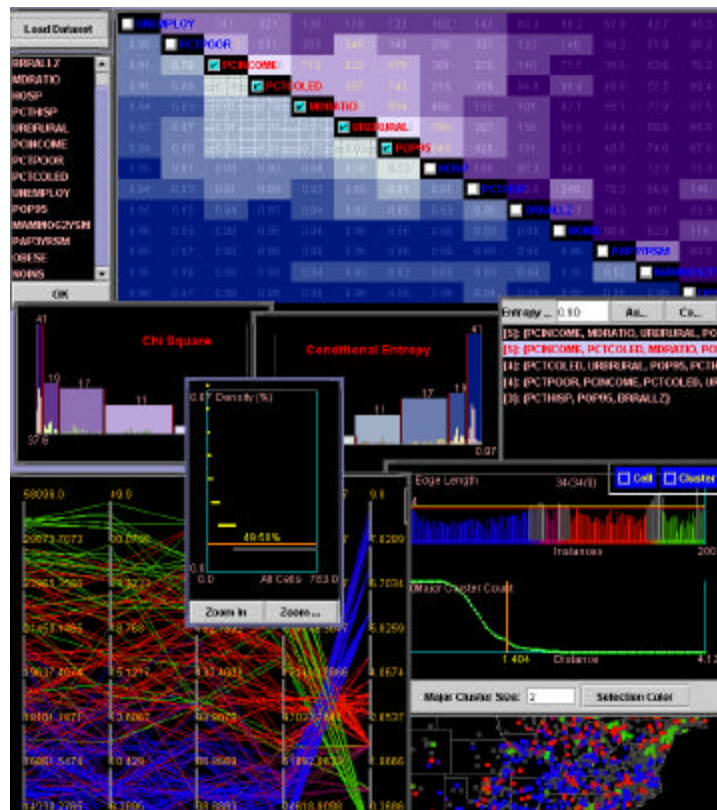


Figure 1: Integrated feature selection and multivariate clustering tools (Guo, 2003)

integrated system. A normal cycle within the iterative exploration process can be: loading data, cleaning the data, interactively selecting interesting subsets of variables for further analysis, identifying hierarchical clusters of the data (using selected variables), interactively exploring and interpreting those clusters, visualizing the clusters in a map and examining the spatial distribution of discovered multivariate spatial relationships.

New parallel coordinate plot (PCP) and time series analysis components were implemented under the contract with NCI and packaged as a stand-alone application that includes a scatterplot and bivariate map. This application is designed for NCI data analysis needs, which include temporal data analysis, interactive data range setup, box plot analysis on each variable, and others. The application's PCP supports display of multiple variables at the same time by mapping an n-dimensional dataset to a two-dimensional space where variables are listed as parallel axes, and each observation is visualized as a polyline, connecting the points on axes, which are the observation's values on those axes.

Color schemes play a critical role in displaying the patterns within multivariate data in geovisualization environments. An extension to univariate color schemes, ColorBrewer, has been developed to represent bivariate data. A set of scheme choices for each two-variable combination is implemented, such as sequential-qualitative or diverging-diverging. The schemes are constructed by sampling the surface of geometric objects within CIE L\* a\* b\* color space, a perceptually scaled 3D color space. Two tools, the Color scheme design board and the Color scheme coordinator, are implemented in GeoVISTA *Studio*. The Color scheme design board is aimed at advanced users who would like to explore the parameters of bivariate color schemes and/or to design custom schemes. The ColorBrewer stores the information about recommended color schemes and can communicate with client components in GeoVISTA *Studio* to apply the color schemes.

Two key architectural changes have been made to the GeoVISTA *Studio* environment, which is the backbone platform that the research uses to integrate various components. The first is the development of 'invisible' adapters, which automatically translate data structures between different components and thus greatly ease the process of integrating components into an analysis environment. The second adds the ability to attach notes (or metadata) to *Studio* designs and thus help non-developer users (e.g., epidemiologists) understand the use and purpose of the design. The research also supported experimentation and re-engineering of interfaces that conforms to standard GIS and database data formats, including Open GIS Consortium standards for describing data and describing the visual appearance of maps and displays.

The applications described above follow a component based software design standard; they are implemented as independent but coordinated Java beans. These components are "coordinator-aware" and make use a coordinator bean to communicate with each other, as well as other existing tools, such as maps, matrices, spread sheets, etc. The concept of dynamic coordination, supported by these independent components, is implemented as data sharing, selections, classifications and focusing, and enables a dynamic and comprehensive multivariate analysis of geospatial data.

The source code of these tools, along with GeoVISTA *Studio* is available through SorceForge. Stand-alone applications have been distributed to collaborating agencies and some applications are available as Java Applets, at [www.geovista.psu.edu](http://www.geovista.psu.edu). All of the multivariate analysis tools and applets developed by our group will be demonstrated at the System Demonstration session during the dg.o2004 conference.

Guo, D. (2003). Coordinating Computational and Visual Approaches for Interactive Feature Selection and Multivariate Clustering. *Information Visualization*, 2(4): 232-246.