

Supporting visual integration and analysis of geospatially-referenced data through web-deployable, cross-platform tools

Alan M. MacEachren, Frank Hardisty, Mark Gahegan,
Michael Wheeler, Xiping Dai, Diansheng Guo, Masa Takatsuka

GeoVISTA Center, Department of Geography, 302 Walker, Penn State University
alan@geog.psu.edu; www.geovista.psu.edu/geovistastudio

1 INTRODUCTION

Federal government agencies generate massive volumes of statistical data. A substantial proportion of these data include geospatial referencing (in the form of geographic coordinates, zip codes, addresses or other location specifications). This “georeferencing” can be the key to integrating data across agencies, because it often provides the only common link through which diverse data sets can be joined. Joining statistical summaries produced by different agencies can result in critical insights about relationships among seemingly separate events and processes, relationships that might otherwise be ignored because the relevant statistics are generated by different federal agencies. As one example, consider demographic change (with statistics produced by the Bureau of the Census), evolution of changes in public health (with statistics collected and synthesized by the Centers for Disease Control), and shifting locations for human-induced environmental hazards (tracked by the Environmental Protection Agency). Integrating these data and supporting flexible visually analysis can prompt innovative hypotheses (and support subsequent rigorous analysis) about problems such as the evolution of disease-risk factor relationships.

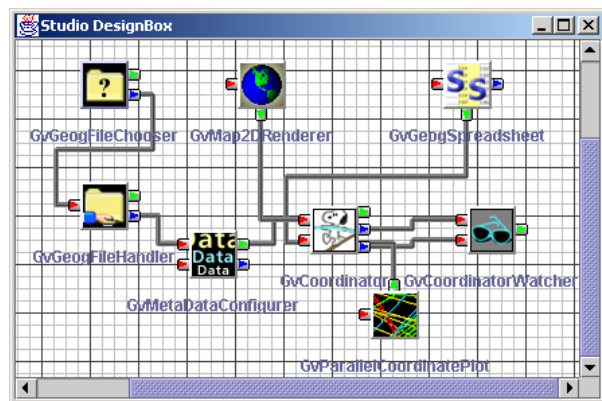
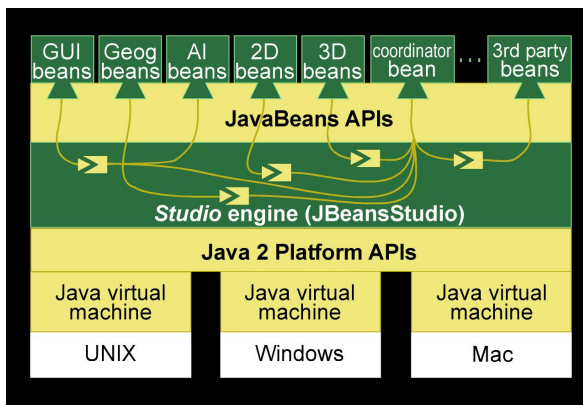
As part of one National Science Foundation (NSF) supported Digital Government (DG) Collaborative Research project (*Quality Graphics for Federal Statistical Summaries*, #9983451; PIs: Dan Carr, George Mason, Alan MacEachren, Penn State, David Scott, Rice) MacEachren, Hardisty, Dai, and Guo are working with divisions in eight federal agencies that generate statistical summaries to develop visual-analytical methods that can support agency missions by enabling integrated analysis of geospatial data from within and across agencies. In a second NSF supported DG project (in which Gahegan is a collaborator, *Knowledge management over time-varying datasets*, PI, Peggy Agouris, Maine, #EIA-9983445), work is underway with four agency partners to develop approaches and tools that facilitate the mining of information from geospatial datasets across space and time and to improve (by better integration) knowledge management over these datasets. In both projects, we are leveraging a separately funded web-deployable, cross-platform, Java-based software development effort, GeoVISTA *Studio*. In this paper, we outline capabilities of *Studio*, discuss how it is being extended to support our joint Digital Government geospatial data integration and analysis goals, and provide examples of the kinds of multivariate geospatial data integration and analysis being undertaken.

2 GeoVISTA *Studio* – AN OVERVIEW

GeoVISTA *Studio* (subsequently referred to as *Studio*) is a component-based environment designed to support the fusing of diverse visual and analytical capabilities into custom analysis tools that enable a multi-perspective (‘mixed initiative’ ((Amant and Cohen, 1998)) approach to knowledge construction and dissemination (Takatsuka and Gahegan, in press). *Studio* provides a visual programming environment that allows an analyst to package assembled functionality into a working program (in the form of a cross-

platform, JavaBeans component, an applet, or an application). The result can be easily disseminated or deployed on the Internet. Like past visual programming environments, designed to support rapid development of scientific visualization applications, *Studio* allows users to quickly combine components into flexible applications. But unlike many such environments to date, the available components address a range of activities that span statistical analysis, visualization and machine learning, i.e. a broad range of approaches to exploring and analyzing a dataset.

Another major difference from most past visual programming environments for building visual analysis tools is that the modules a user links together, in the program's *design box*, are true "components" – units that can be deployed independently and combined with third party applications. Thus, components developed by our research team to work within *Studio* can also function independently. These components can be used in other JavaBean applications or by third parties; and components developed by others can be used within *Studio* (as long as they meet the JavaBeans Application Programming Interface (API) standards). This flexibility is enabled through the "builder" (a component-oriented application construction



system), which connects components together at runtime, without the need for recompilation, linking or any other form of 'preparation'. By using Java's 'introspection' function, the builder obtains a syntactic description of all the services (methods) that a bean provides, and can expose these methods for linking to other beans. Thus it is not necessary to have source code available before a bean can be assimilated into the Studio environment, nor any prior knowledge of its methods. The figures above depict the system architecture (left) and the user's view of a set of components and their connections shown in a typical design box (right).

A third major difference between past visual programming environments and *Studio* is the potential for multi-way connections among components. Although the *Studio* builder supports a standard pipeline architecture, other forms of data communication and sharing are also possible. This difference is best illustrated by the "coordinator" bean developed to support the kinds of multivariate analysis critical for the two Digital Government projects cited above (second bean from the right in architecture view, above left, and center component in the design box, above right). The coordinator acts as a multiplexor that arbitrates the passing of events between beans that require a high degree of specialized, coordinated behavior without the need for highly coupled software. Without the need for a highly elaborate set of bean interconnections, it arbitrates the sharing of data, metadata, and several display and event properties among components, including selections, categorizations, and visual mappings (e.g., assignment of color, line weight, or other features to categories). Developers can choose to what degree beans coordinate their behavior and their data with other beans, easing the problem of providing many linked views onto the same dataset, via different visual and analytical methods. Section 4 provides some examples of 'coordination in action'. For

Studio applications built using the coordinator, beans can be designed specifically to take advantage of the ability to share parameters among components. Alternatively, for beans not originally designed to support coordinated behaviors, a wrapper must be added to make the bean “coordinator aware”. In the application example described below, a commercial spreadsheet bean was downloaded from the web, a wrapper was added, and the spreadsheet is now able to share features such as changes in color scheme (assigned to value ranges within an attribute) and visual selection (and highlighting) of data subsets. As described above, the spreadsheet’s capabilities were effectively customized for our own uses without requiring source code to be available.

3 EXPLORING MULTIVARIATE AND TIME SERIES GEOSPATIAL STATISTICS

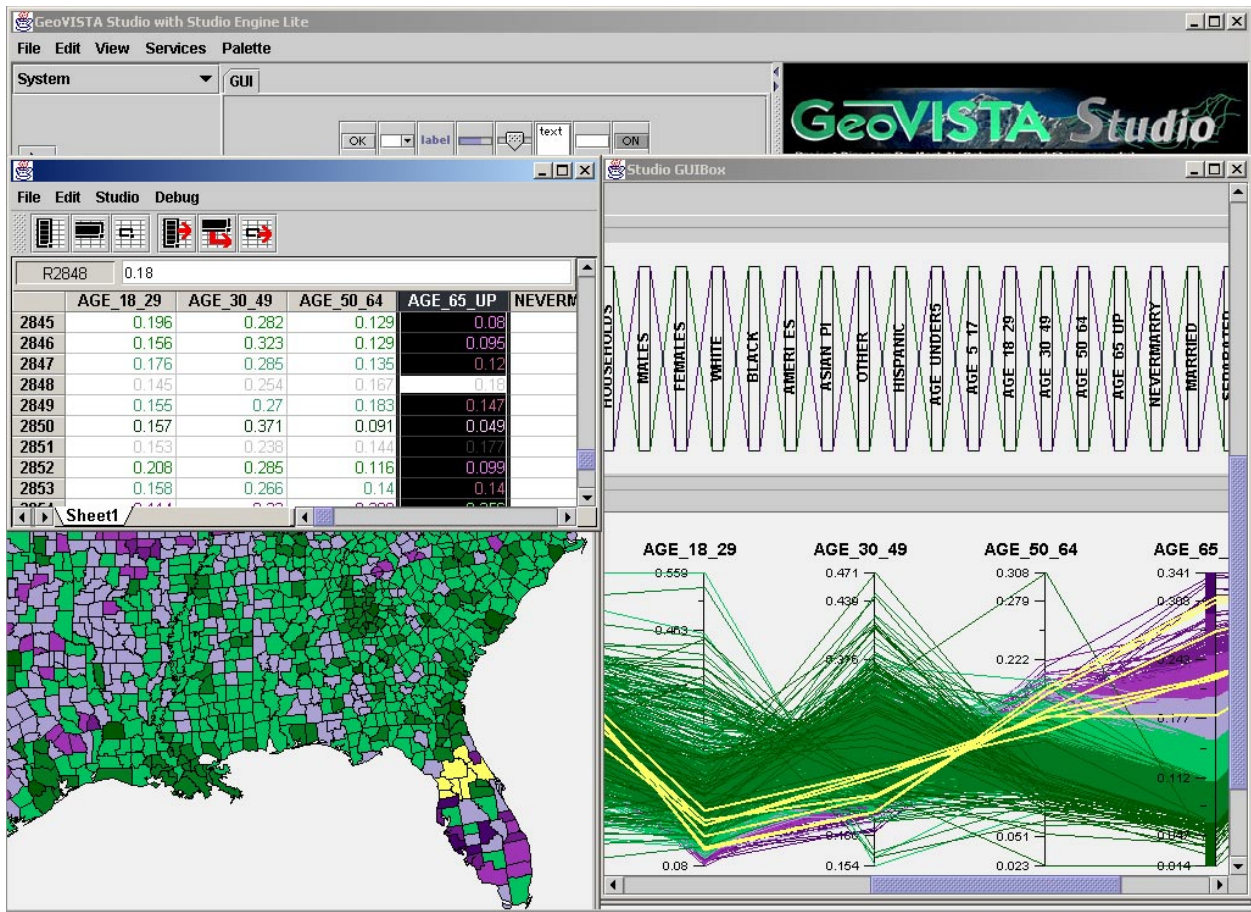
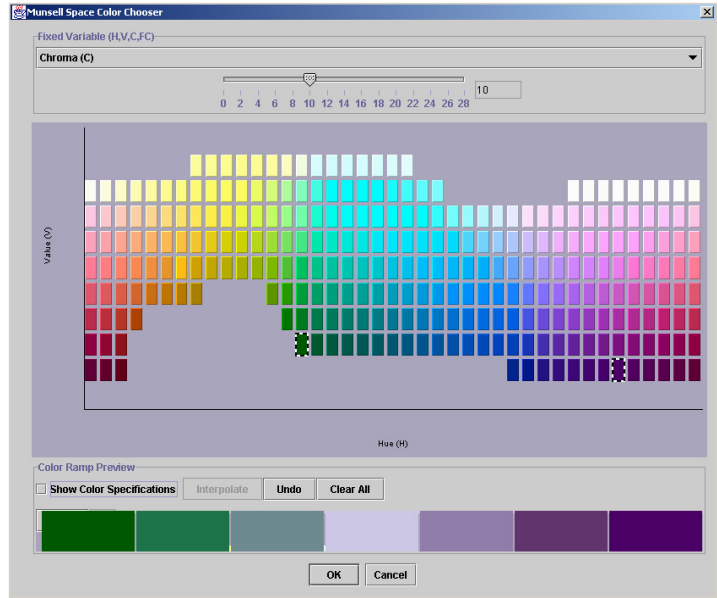
In the context of the two DG projects cited above, our research teams have developed or extended a suite of visual analysis components, for use within *Studio*, that support dynamic visual analysis of highly multivariate statistical data produced by our partner agencies. These tools build, in part, on our past work to develop, and explore the cognitive-usability implications of, a range of methods for multivariate visual (and computational) analysis of geospatial data (e.g., (Gahegan, 1998; MacEachren et al., 1998; MacEachren et al., 1999) While particular emphasis in our work is on data that include geospatial referencing, many of the visual and computational analysis tools we are developing are suited to analysis of any multivariate data.

To support visual data exploration and analysis, we are implementing (as JavaBeans) a suite of highly interactive representation forms that support different perspectives on data, along with several components that support manipulation of parameters to control the *data-to-display* mappings for each representation form independently, or globally for all representation forms. As noted above, our recent extensions to the *Studio* environment focus on use of the coordinator bean to support dynamic, multi-parameter links among different representation forms. Here, we sketch the functionality of several of these components and the dynamic linking supported between them, specifically the following beans: *2D renderer* (map display tool), *spreadsheet*, *parallel coordinate plot (PCP)*, *visual data classification*, and *color scheme*. Other components implemented thus far or under construction include: a scatterplot matrix, a map matrix, a 3D renderer, image analysis tools, self-organizing map neural network, K-means clustering, and regression analysis.

The user controls the data-to-display mapping for display components through two other beans. First, the *visual classifier* allows the user to set the number of categories into which data are grouped and the method applied for determining where the category breaks are in the range of values (e.g., equally spaced values along the range, quantiles: breaks that group data so that the same number of counties is in each category, Jenks optimal: minimum variance categories in which all data values in a category are as similar to one another as possible, (Slocum, 1998)). Then, the *color scheme* bean is used to select color assignments for each category. In addition to allowing users to build color schemes using an RGB color picker, we have implemented a color picker that allows users to build schemes using the Munsell perceptually balanced color space (below-top), and to interpolate color assignments within this space between two (or more) chosen colors. In the figure, the endpoints of a diverging color scheme were selected in the “chroma” view (in which colors of the same chroma/ saturation that vary in value/lightness and hue are displayed) and a midpoint was selected in the “value” view (in which colors of different hue and chroma are depicted – not shown). The four intermediate colors are interpolated between these three anchor points in 3D perceptual color space. The color scheme selected can be applied to variables displayed in map, PCP, spreadsheet and other views (below-bottom). The figure shows the color scheme applied to a map of

the proportion of the U.S. 2000 population by county that is 65 years of age or older (in this case, the map is a choropleth map that depicts aggregate interval or ratio level statistics for counties using color fills assigned to each county polygon).

Dynamic links between the different perspectives on multivariate datasets provided through different interactive representation forms in *Studio* (e.g., maps, scatterplot matrices, spreadsheet, PCPs) support exploration of many kinds of relationship. As one example, the screen capture above illustrates dynamic sharing of: (a) color schemes (here, assigned to one axis in the PCP and propagated to both the map and spreadsheet views); (b) axis selected (in the PCP and spreadsheet), and (c) user selection of an



interesting spatial cluster by “brushing” on the map (highlighted in yellow and propagated to both other views). Here, the high proportion of elderly residents noticed by the user on the map (for the section of South Florida, highlighted) can be seen to have a quite different age distribution than is typical for other states.

Our implementation of a PCP, as described above, supports interactive categorizing and coloring of the strings (where each string represents a data instance). In addition, several other interactive capabilities have been added that are designed to make the PCP more useable for exploratory analysis and to support larger data volumes than used typically with this graphical analysis tool in the past (e.g., the PCP above is displaying data for the 3000+ counties in the U.S.). Among the extensions implemented, users can: isolate the strings for one category of a variable (e.g., those counties with the lowest proportion of 18-29 year olds) by clicking on that category along the axis depicting the variable (which causes all other strings to temporarily be turned off), change the category breaks manually by grabbing the break point between categories on an axis and dragging it to the desired value, and reorder axes by picking up their icons in the inset above the PCP and dragging them to a new location. Users can also animate the rendering of the strings within the PCP in a couple of ways, to help overcome perceptual bias due to string drawing order and overwriting. In the first of these, strings are drawn onto a blank canvas in a specific order, defined by the user. It is then possible to get a sense of the number of strings that have been obscured, and their values. In the second, strings are highlighted, one at a time, in rapid succession, and again in a pre-specified order. This is very useful for exploring the range of data values that might make up a particular category of interest.

In addition to the multivariate visual analysis methods detailed above, *Studio* facilitates the multivariate analysis of geospatial attributes by supporting the linkage of analysis components and visualization components. Thus, the output of statistical tools like K-means clustering or regression analyses can be visualized as the parameters are set, making previously “black-box” techniques more transparent. The manner in which data are being visualized and analysed can be reflected across components, shortening the iterative visualize – analyse – visualize knowledge construction loop. This aspect of the environment is discussed in (Gahegan et al., 2000).

4 APPLICATION EXAMPLE

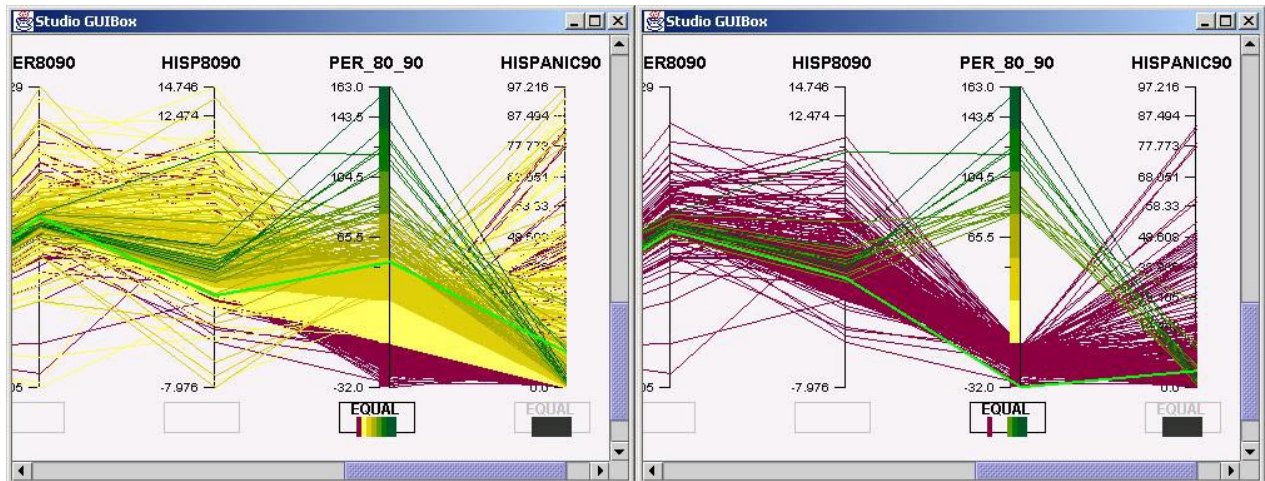
Here, we describe a case study application of the multivariate analysis environment described above, to illustrate possibilities raised by coordination of numerical analysis and visual analysis across components. This study focuses on changes in the spatial structure of the U.S. population, and particularly the relationship between the Hispanic population and population growth, as reflected in the 1980, 1990 and 2000 Censuses.

Many governmental and business organizations depend on accurate analysis of demographic trends for their planning efforts. The primary source for population information in the United States is the decennial census, so the release of data from the 2000 Census has occasioned a great deal of interest. The application of visual-analytical methods to these data sets can help investigators uncover unexpected relationships. For example, from reading the major news media, one might expect that among those counties with a quickly growing population, those with a high proportion of Hispanics would show the greatest growth in population. However, by using *Studio*, we uncovered exactly the opposite relationship between the 1980 and 1990 Censuses, which we hypothesize will extend to the 2000 Census.

One of the “big stories” of the 2000 Census so far is the growing importance of the Hispanic population in the demographic makeup of the United States. From accounts in important news outlets, one would expect that those counties which are growing quickly would have an especially high proportion of Hispanics, or a quickly growing proportion of Hispanics (Janofsky, 2001). County level demographic data was not yet available for the 2000 Census when we completed this analysis, so in this case study we examine the relationship between proportion of Hispanics and population growth at the county level between 1980 and 1990. The analysis will be extended as 2000 data becomes available.

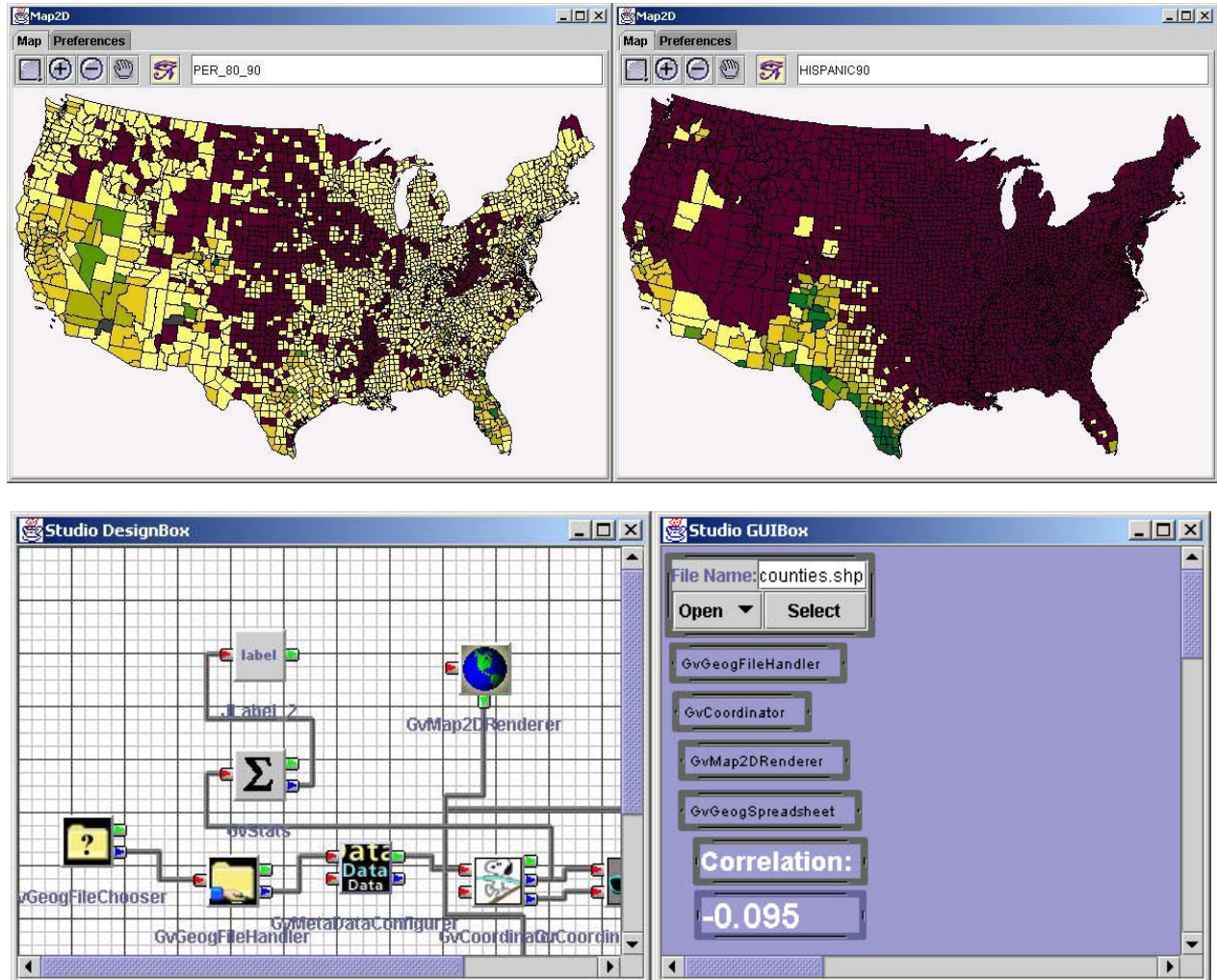
A strength of GeoVISTA *Studio* as an analysis tool is that we can examine, visually and numerically, the geographic and statistical relationships between phenomena of interest. Using *Studio* to examine the proportion of Hispanics in quickly growing U.S. counties and changes in population counts in those counties enables us to discover and confirm surprising relationships. One of these is that, between the 1980 and 1990 censuses, in the subset of counties that exhibited extremes in population growth (specifically, the top 100 and bottom 100) the relationship between the proportion of Hispanic and the rate of population growth is *negative*. The same surprising relationship holds between the growth in the proportion of Hispanics and the population growth overall. Both the discovery and the confirmation of these relationships are enabled by the use of *Studio*.

An appealing property of a PCP is that correlation between variables on adjacent axes is depicted visually (by the extent to which lines cross) (Inselberg, 1985). However, if there are thousands of lines connecting two axes, the relationship may be obscured. As discussed above, the ability to classify data depicted in the PCP and to selectively highlight (or repress) subsets of the data make our implementation of the PCP useable for larger data sets. Below, the patterns of lines for the counties with particularly high and low percentages of growth are obscured in the first PCP (left), but clearer in the second (right – in which three intermediate classes are turned off).



In the PCP at right, the crossing of the lines indicate a negative relationship between percentage population change for counties between 1980 and 1990 (PER_80_90) and percentage of population identifying as Hispanic (HISPANIC90). High change (the classes in green) corresponds with low percent Hispanic and the reverse. Similarly, there appears to be a negative relationship between percent change overall (PER_80_90) and change in the percentage of population identifying as Hispanic (HISP8090).

The surprising nature of this relationship suggests that the next step in analysis should be to check the accuracy of the data. We do this first visually by mapping the data on separate maps. Below, a map of the percentage change in population between 1980 and 1990, and a map of proportion of Hispanics in 1990, both conform to well-known distributions. To statistically confirm that the relationship we suspect actually exists, we can add functionality to a *Studio* design at run-time. In the second figure below, a statistical bean and a label bean have been added to the design. With these additions, the user can set the display properties for the label, and the resulting statistics are displayed. In this case, the correlation between the proportion of Hispanics in 1990 in the 100 counties with the highest population growth and the population growth in those counties is shown to be -0.095 which is a weak, but negative relationship.



As we would expect, the relationship between population growth and percentage of Hispanics for the U.S. as a whole, expressed as a correlation coefficient, is positive, 0.118 (although not strong). Similarly, the correlation coefficient between population growth and growth in the Hispanic proportion of the total is also positive at 0.187. However, if we examine the 100 counties with the highest and lowest growth, the picture changes dramatically to being a weak or negative relationship. For the 100 counties with the highest growth, the correlation coefficient of overall population growth with proportion Hispanic in 1990 is (–

0.095), and for the 100 counties with the lowest population growth it is (-0.345). The relationship between the change in percentage of Hispanics in the county and population growth for the 100 counties with the fastest growing proportion of Hispanics is positive but very weak at 0.036, while for the 100 counties with the largest shrinkage in proportion of Hispanics the relationship between population growth and change in the proportion of Hispanics is negative at -0.323. The reversal of the national trend in the extreme cases could be due to a number of factors. One such factor could be that Hispanic immigrants have been migrating to cities in the "rustbelt" with declining populations. Another factor is that there is a strong relationship between wealth, as measured by median housing values and population growth in counties (0.45). People from Hispanic backgrounds may be less likely to self-identify as Hispanic as they integrate into wealthier counties with different dominant cultural values. These demographic trends will be further explored in reference to other socio-economic data sets to explain these relationships.

GeoVISTA *Studio* is being disseminated through use of Java Web Start. This mechanism will make it possible for users to download the software once (through a standard web browser) and automatically retrieve software updates whenever the program is launched (on a machine with an active web connection). See www.geovista.psu.edu for more information.

References

- Amant, R.S. and Cohen, P.R., 1998. Intelligent support for exploratory data analysis. *Journal of Computational and Graphical Statistics*, 7(4): 545-558.
- Gahegan, M., 1998. Scatterplots and scenes: visualization techniques for exploratory spatial analysis. *Computers, Environment and Urban Systems*, 21(1): 43-56.
- Gahegan, M., Takatsuka, M., Wheeler, M. and Hardisty, F., 2000. GeoVISTA Studio: A Geocomputational Workbench, *Proceedings: GeoComputation 2000*, University of Greenwich, Medway Campus, UK, Aug. 2000.
- Inselberg, A., 1985. The plane with parallel coordinates. *The Visual Computer*, 1: 69-97.
- Janofsky, M., 2001. Arizona Owes Growth Spurt Largely to an Influx of Hispanics, *New York Times*, Denver, CO, pp. 13.
- MacEachren, A.M., Boscoe, F.P., Haug, D. and Pickle, L.W., 1998. Geographic Visualization: Designing Manipulable Maps for Exploring Temporally Varying Georeferenced Statistics, *Proceedings, Information Visualization '98*. IEEE Computer Society, Reliegh-Durham, NC, Oct. 19-20, 1998, pp. 87-94.
- MacEachren, A.M., Wachowicz, M., Haug, D., Edsall, R. and Masters, R., 1999. Constructing Knowledge from Multivariate Spatiotemporal Data: Integrating Geographic Visualization with Knowledge Discovery in Database Methods. *International Journal of Geographic Information Science*, 13(4): 311-334.
- Slocum, T., 1998. *Thematic Cartography and Visualization*. Prentice Hall, Upper Saddle River, NJ.
- Takatsuka, M. and Gahegan, M., in press. GeoVISTA Studio: A codeless visual programming environment for geoscientific data analysis and visualization., preprint can be downloaded from <http://www.geog.psu.edu/~mark/>.