Design and Usability of an Enhanced Geographic Information System for Exploration of Multivariate Health Statistics*

Robert M. Edsall
Arizona State University

The multidimensional nature of many types of data in modern geography calls for creative and innovative approaches to their analysis. Statisticians have recently developed methods for exploring and visualizing large, multivariate datasets, but cartographers and geographers in general have only recently begun to integrate these methods for use with spatial and spatiotemporal datasets that are multivariate in character. This article will present an example of such an integration—an environment for visualization of health statistics—as a case study to demonstrate the philosophical and practical advantages of geovisualization systems for the exploration of complex spatiotemporal information. Emphasis is placed on the encouragement of creative thinking about geographic phenomena through the use of such data-rich graphical tools. **Key Words: cartography, epidemiology, geovisualization, information visualization.**

Introduction

The analysis of datasets that are highly multivariate in character is common in modern geographical inquiry. A climatologist attempting to understand a hurricane, for example, is interested in its maximum sustained wind speed, its proximity to populated areas, the sea surface temperature underneath it, the synoptic characteristics of the air masses around it, and so on. A planner concerned with perception of place might survey residents who live around a park about the park's accessibility, its safety, its facilities, its proximity to their homes, and other variables. A transportation specialist might find the understanding of multidimensional characteristics of roads—such as their width, the volume of traffic traveling on them, their maximum speed limit, their accessibility, and other features—instrumental in the understanding of the transportation network as a whole.

The analysis of multidimensional geographic data is facilitated through the construction of and interaction with visual representations. Griffith and Amrhein (1997) hold that visualizing data is a basic rule of thumb for developing "statistical wisdom" about multivariate data. Many statisticians hold that visualization itself, even without more traditional confirmatory statistics, is a feasible method for gaining valuable insight into datasets, particularly datasets about which little is known or datasets that are massive and multidimensional (Tukey 1977; Hurley and Buja 1990; Wegman 2000). The use of maps and other graphic devices as means of exploring spatial data—that is, as part of the analysis process itself—is a cornerstone of the emerging research theme of geographic visualization, or *geovisualization* (DiBiase 1990; MacEachren et al. 1992; MacEachren and Kraak 1997). There is great potential in the synthesis of geographic visualization with statistical exploratory data analysis, particularly targeted toward multivariate spatial and spatiotemporal data, a link that has been established in practice in recent years by a number of projects and authors (Dykes 1997; Andrienko and Andrienko 1999; MacEachren et al. 1999).

Epidemiologists are among the many types of researchers for whom the understanding of the datasets they frequently use is made challenging by the multidimensional nature of those data. Linking cancer mortality rates, for example, to demographic, socioeconomic, and other possible risk factors is a matter of the

* I am pleased to acknowledge support for portions of this work provided by a contract from the NIH-National Cancer Institute and a grant from the U.S. National Science Foundation (#EIA9983451). Special thanks to both Dr. Alan M. MacEachren and Dr. Linda J. Pickle for their generous assistance through all phases of this research. The ArcView 3.2 application (including all scripts and interface tools) described herein is available for no charge. Please contact the author at robedsall@asu.edu, or by standard mail.

The Professional Geographer, 55(2) 2003, pages 146–160 © Copyright 2003 by Association of American Geographers.
Initial submission, May 2002; revised submission, November 2002; final acceptance, November 2002.
Published by Blackwell Publishing, 350 Main Street, Malden, MA 02148, and 9600 Garsington Road, Oxford OX4 2DQ, UK.

examination and exploration of tens or hundreds of variables, often over many spatial observational units and time steps. Though statisticians are well versed in methods for analyzing such datasets, visualization offers an attractive alternative (or supplement) to, for example, hundreds of logistic regression permutations. This article will report on the design and implementation of a visualization environment that links a geographic information system (GIS) to custom-designed interactive statistical representations, with the specific conceptual goal of exploring health-statistics data of many dimensions. A two-part usability assessment, also described here, lends support to the assertion that multiple linked views—in particular the linked parallel coordinate plot, scatterplot, and choropleth map—facilitate the construction of knowledge about multivariate health statistics data.

Multivariate Visualization and Geographic Visualization

The exploration and understanding of spatial, temporal, and attribute features trends in health-statistics data is facilitated through the use of interactive visual representations (Plaisant 1993; MacEachren et al. 1998). Such displays and tools encourage and support a creative and intuitive search for patterns and structures in the data that are difficult to detect through nonvisual means. Interactive computer displays provide the means of representing complex data in a variety of representational styles and symbolization choices. With each style and choice, different insights might be gained about a particular characteristic of the dataset (e.g., a temporal trend). In tandem, the interactive displays and tools in a visualization environment can combine with the user to construct knowledge about the complex data in the representations (Rogers 1999).

Dynamic statistical graphics in support of exploratory data analysis have roots in the work of Becker, Cleveland, and Wilks (1988), which has stimulated many other statisticians to develop novel methods for analyzing abstract multidimensional data. Buja, Cook, and Swayne (1996) summarize some of these multivariate representation techniques via a taxonomy (or, as they call it, a "zoology"). The general methods developed for basic multidimensional representation that they cite include: *scatterplots*, in which observations are represented by locations of points; *traces*, in which observations are represented as lines or functions, as in parallel coordinate plots (Inselberg 1985) and Andrews curves; and *glyphs*, in which characteristics of complex symbols such as Chernoff faces and stick figures (Erbacher et al. 1995) are functions of the observed values.

Providing geographers and other researchers who study spatial and spatiotemporal data with these sorts of tools for exploratory analysis of their data has been a theoretical and practical goal of geographic-information scientists for at least the last decade (MacDougall 1992; Monmonier 1992; Dykes 1997). Not unlike exploratory data analysis (EDA), the research initiative known as geovisualization grew out of issues concerning the representation of and interaction with large amounts of complex data (MacEachren et al. 1992). It also grew out of a rejection of a goal of previous cartographic research: to find single optimal ways of representing geographic information (MacEachren and Ganter 1990). Many geovisualization environments have been developed that utilize the capabilities of computer graphics for the display of geographical datasets consisting of many variables (using multiple forms of representation). In their review of multivariate display in for geographical information, DiBiase, Reeves, and colleagues (1994) propose that effective geovisualization systems consist of three complementary characteristics: (1) an interface tailored to a specific set of users (novices or domain experts, for example), (2) interactivity that affords users the ability to experiment with different symbolization methods, and (3) a design that fosters knowledge discovery over confirmation.

With respect to spatially referenced health-statistics data, the guidelines and tools described above have been implemented in dynamically linked maps and graphs by Plaisant (1994) and in an enhanced geographic information system and dynamic map by MacEachren and colleagues (1998). More recently, Wang and colleagues (2002) created a Web-based application designed for health statistics that features large geographical datasets represented simply and interactively through maps linked to statistical graphics. The work presented in this article is an extension of the

implementation, known as HealthVis, reported in MacEachren and colleagues (1998). The updated system, referred to hereafter as "HealthVisPCP," extends the capabilities of HealthVis by adding a method for multivariate spatiotemporal visualization using a dynamically linked parallel coordinate plot.

The Linked Parallel Coordinate Plot

A typical observation of demographic, socioeconomic, and health data from a census tract or other spatial enumeration unit for a given time period consists of a long list of variables. For a given spatial unit, these observations might include a vector of values for, perhaps, median income, rate of lung-cancer mortality in white females, number of hospitals, number of oncologists, average level of education, and so on. Together, this vector makes up a unique multivariate "signature," with a high value in one variable, a low value in another, and so on. For our purposes in developing a system for the visualization of these statistics, Inselberg's (1985) parallel coordinate plot (PCP) representation seemed particularly well suited.

The PCP employs a novel methodology to visualize beyond three dimensions by representing each observation, not as a *point* (as in a scatterplot), but as a series of unbroken *line segments* connecting parallel axes, each of which represents a different variable. The line segments are constructed so that they intersect the axes at a point representative of the relative observed value of that variable. As an example, consider the vector of epidemiological data described above. The observation could be represented by a series of points, one on each axis, positioned near the top of the axis if that variable's observed value were relatively high or near the bottom if the value were relatively low. These points could then be connected by line segments, resulting in a distinct (perhaps unique) signature of observed values for that observation (Figure 1). This would form one of many multivariate traces, one for each observation (e.g. each county, state, census tract, weather station, etc.).

The PCP can be used to represent many variables observed at a particular time, or a series of time steps for a given variable (e.g., six parallel axes each represent one year of a measurement of a cancer mortality rate at a given location). Thus, temporal trends and multivariate signatures can be easily discerned.

Using the PCP, with each axis depicting a variable, interactions among variables can be quickly identified. Observations with similar data values across all variables will share similar signatures; thus, clusters of like observations can be discerned. Two variables directly related to one another would appear on the PCP as two axes connected by a series of parallel (or at least noncrossing) line segments. Median income and average education, for example, tend to be directly related to one another; a parallel coordinate representation of these two variables would clearly show this pattern of uncrossing line segments. Conversely, an inverse relationship between two variables would be displayed as a series of line segments that cross

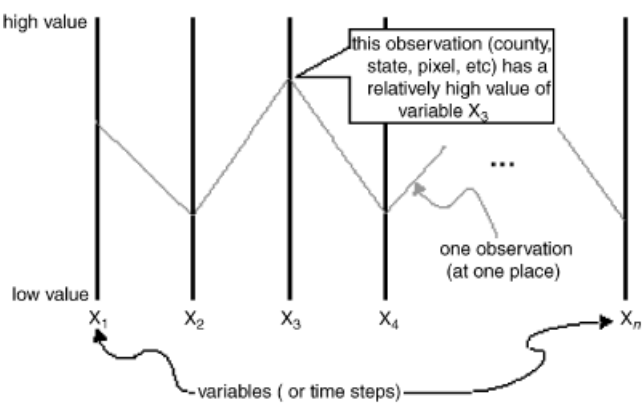

Figure 1 *Parallel coordinate geometry.*

each other between the axes (Wegman 1990). Temporal persistence can also be noted when adjacent axes are consecutive time steps of the same variable.

Though the PCP is a clever data representation technique, it has several disadvantages when compared to the scatterplot and other statistical graphics that may limit its effectiveness for visualization of data relationships. In particular, the use of lines as opposed to points to represent individual observations requires the use of a great deal more "display real estate" (or ink on a piece of paper), so that the result may quickly (with increasing number of observations) become a confusing tangle of colors and pixels that has little explanatory power. However, the addition of dynamic elements—such as brushing, focusing, coloring, zooming, and other graphical manipulation techniques—may be effective for overcoming these shortcomings of the parallel coordinate representation.

HealthVisPCP

Spatializing the PCP

At the time of the development of the dynamic PCP, all previous implementations of the PCP in the literature had been two-dimensional.[1] For geovisualization, extensions need to be made to the PCP in order to highlight the interrelation of the two dimensions of space; implementing latitude and longitude—two columns of data in a database—as typical "variables" in a PCP would place them on two separate axes and thus de-emphasize the spatial relationships crucial to geographic analysis. Conceptually, what is needed is the addition of a third display dimension in order to embed a two-dimensional geographic representation as a reference axis. This would amount to a "parallel plane" plot (Figure 2), as each axis could be extruded into two dimensions, each of which could represent a variable of the user's choosing (the obvious default would be latitude and longitude). In the development of the application described here, this parallel-plane plot was prototyped, but it was scrapped in favor of a computationally less intensive link between the PCP and a two-dimensional map in a separate window. This particular application of the dynamic PCP linked to a map (and a scatterplot) is known as HealthVisPCP.

HealthVisPCP

The system was developed using ArcView® GIS, version 3.2, with customized interface tools created using ArcView's scripting language "Avenue" (ESRI 1999). MacEachren and colleagues (1998) developed the original HealthVis system to (visually) analyze spatial, temporal, and attribute features of health statistics for the National Center for Health

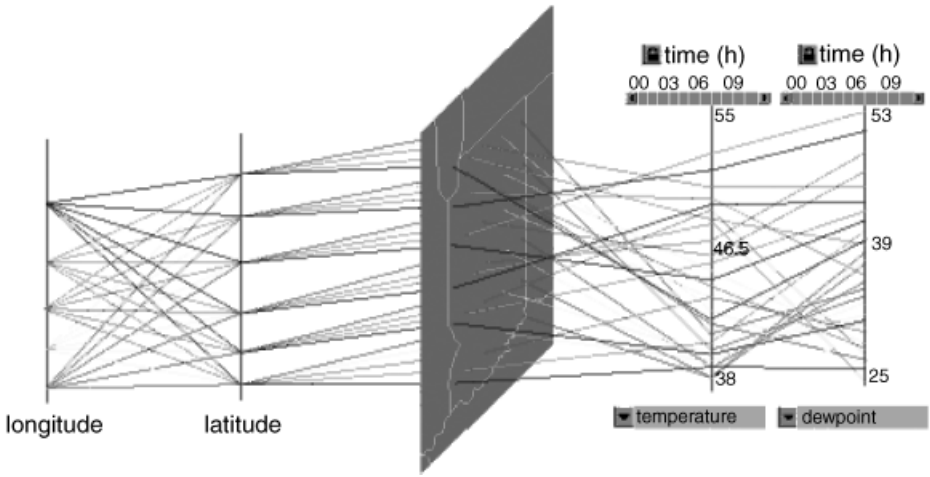

Figure 2 *The parallel plane concept.*

Statistics. Extensions to HealthVis, described here, were added with a subsequent contract from the National Cancer Institute (NCI). In HealthVis, a great deal of functionality was added to the default ArcView documents in order to accomplish a series of goals specific to epidemiological visualization. These goals were: (1) enabling the visualization of highs and lows, (2) detecting regions and clusters, (3) establishing relationships between mortality and risk factors (socioeconomic or demographic statistics), and (4) facilitating the exploration of association between two variables. These goals were operationalized with interactive extensions to the representations above, such as a focus-by-percentile tool, a dynamic classification tool, a brushing tool, and a specialized bivariate map.

HealthVisPCP expands the HealthVis system both conceptually and operationally. A new conceptual goal not listed among the four above is the facilitation of multidimensional exploration among several variables, accomplished by developing a version of the dynamic PCP in Avenue. The application (Figure 3) consists of a set of four linked representations:

1. a choropleth map of the contiguous forty-eight United States,[2] within which are approximately 800 "health service areas" (HSAs), multicounty units used by federal agencies for analyzing health statistics data;
2. a scatterplot, within which may be plotted one or two variables (usually, in the NCI application, mortality rates and risk factors), each point on the scatterplot representing a single HSA;
3. a map legend; and
4. a dynamic PCP (each line of which represents a single HSA).

The choropleth map is colored according to a statistic (such as "prostate cancer mortality, white male") and a year (such as "1982–84"; in

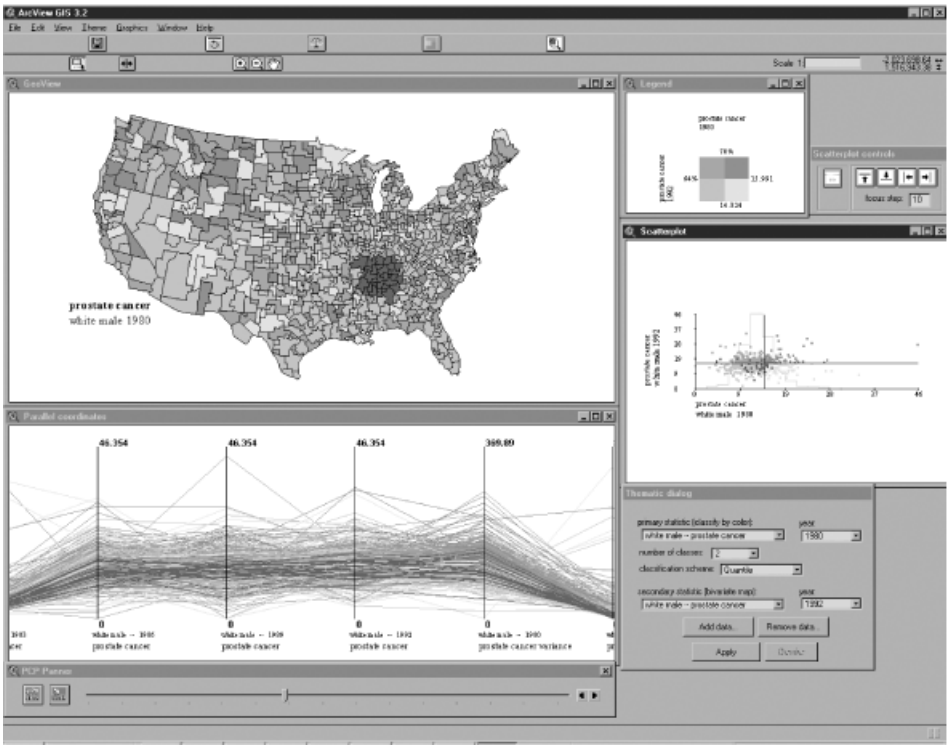

Figure 3 *The HealthVisPCP environment, showing choropleth map (upper left), legend (upper right), scatterplot (center right), and parallel coordinate plot (lower left).*

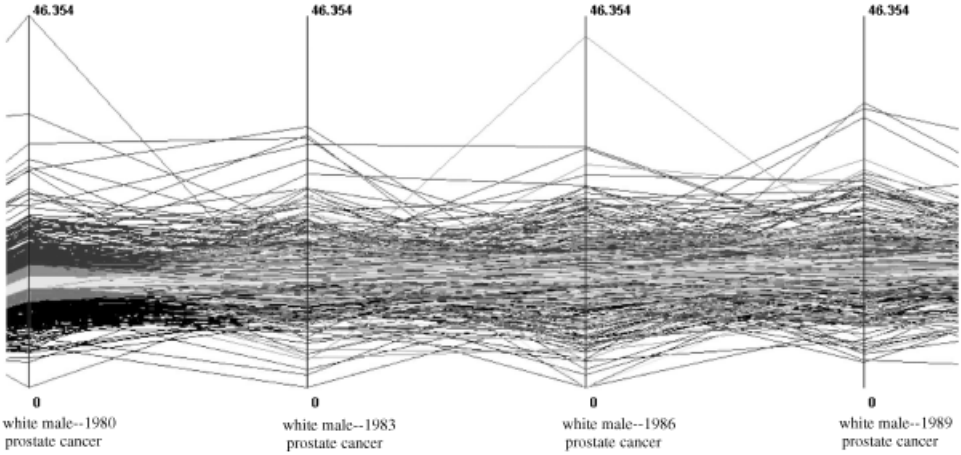

Figure 4 *Detail of PCP representation in HealthVisPCP.*

the datasets obtained, the individual rates for each HSA are three-year averages). The statistic and the year can be selected by the user in a dialog box known as the thematic mapper. In this dialog box, a user can also specify the number of classes (two, five, or seven). In addition, he or she is given the option of creating a bivariate map. Any change in classification of the choropleth map also changes the characteristics (e.g., color) of the corresponding objects in the PCP and the scatterplot.

The representations were linked such that when an object or set of objects (HSAs in the choropleth map, points on the scatterplot, or lines on the PCP), is selected by the user by clicking on the object or dragging a box in the display, the corresponding object(s) in the other two representations is (are) also highlighted. Dynamically linked views such as these give the graphics added power in the understanding of both the representations and the data represented within them.

The PCP (Figure 4) was added to the original HealthVis system as a new "View" document in ArcView, in a similar fashion to that used for the scatterplot. As a View, the tools for manipulating the three representations (PCP, scatterplot, and map—all View documents) behaved similarly across the representation types. For example, the zoom feature worked in all three representations: on the scatterplot and PCP, a zoom-in not only allowed more detail to be seen, but also allowed accurate value-reading, because axis labels were also updated to the new zoom extent. An additional useful feature of the original HealthVis system is the dynamic classifier, which allows users to drag lines on the plot that represent class breaks. The same concept was implemented on the dynamic PCP as a means to allow users to reorder the axes in cases where a user wishes to compare variables represented on axes that, in the default ordering, are not adjacent.

Building upon experiences in designing and using a separate PCP implementation (developed for another project in our lab, described in Edsall 1999), other useful enhancements were identified and incorporated in the implementation of the PCP for HealthVisPCP. Since at any given time the PCP is able to show a limited number (three or four) variables at its default zoom extent, a mechanism for panning the display left and right to reveal other axes was necessary. The PCP panner is a slider bar that pans the PCP horizontally, emphasizing the linear nature of the plot and representing how much of the PCP is "off" the screen in either direction at a given time. Also, a button was added that zooms the PCP to a default extent that maximizes the amount of information contained in the PCP display. This function is similar to a button (contained in the original HealthVis system) that simply

restores the default display extents of all of the windows, refreshing the entire display.

HealthVisPCP was developed not only to assist health researchers gain a new perspective on their data, but also to serve as a vehicle for the testing of the usability of PCPs and their relative utility for display and analysis of multidimensional information. The assertion that the PCP is an effective tool for prompting multivariate thinking seems reasonable, but has never been tested, although most authors who have written about or implemented PCPs assert their effectiveness for that (or a similar) purpose (Inselberg 1985; Wegman 1990; Edsall 1999). In the following section, I report on a usability assessment that tests this assertion. The assessment addresses narrow pieces of a larger research endeavor to understand how representations of multivariate spatiotemporal data should be designed and implemented.

Investigating Linked Statistical Graphics: Usability Assessment of HealthVisPCP

The usability assessment was designed to evaluate the dynamic PCP in the context of HealthVisPCP, a visualization environment designed for the exploration of multivariate health statistics. The specific goal of the assessment was to compare the relative effectiveness of each representation form, the scatterplot and the PCP, for a series of narrowly defined tasks and for unrestricted exploration of multivariate spatiotemporal data.

Challenges for Assessment of Exploratory Visualization Systems

A system's "usability" is defined by the International Organization for Standardization as "the extent to which the system can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (Karat 1997, 691). Even a definition as vague as this seems to miss key issues when applied to design for exploratory analysis. Since data-exploration tasks are themselves broad and hard to define, the evaluation of a system's usability in achieving those tasks is a challenge. With interactive systems for exploratory data analysis, software engineers are unlikely to come up with questions—much less solutions—for the "specified users" (domain experts) to investigate in an evaluation of the system.

Knapp (1995) describes a task-analysis model applied to the design of environments for "visualization" of geographic data. She proposes an iterative approach to constructing a system through discussions with potential users of the system, who gave the designers goals (such as "I want to be able to describe the current climate situation in the context of the expected climate") that were then mapped to physical and mental actions, data, and visual operators. This task-oriented approach to interactive system design, however, neglects the fact that there are tasks and goals of a visualization system designed for data exploration that are not definable, no matter what the expertise of the potential user, because the datasets and potential insights from them are unexamined and unknown. However, there may be ways of defining generic tasks that are typical for exploratory visualization, such as the recognition of clusters and trends or the comparison of features. The ability of the system to facilitate these generic tasks might serve as a surrogate for the effectiveness of the system for data-exploration tasks in general. Examples of these surrogate tasks were developed for the first phase of the assessment described in the following sections.

In a visualization environment, designer attention, if not emphasis, needs to be placed on testing of the use of the system by domain experts to explore, hypothesize, observe, and gain insight into their complex data. User exploration of information cannot be dependent on predefined tasks or goals other than the very general goal of gaining insight and understanding. Insight can be evaluated potentially by some measure of effective hypothesis generation or observation. In the end, deeper user understanding of the represented phenomena, not simply the understanding of the representations themselves or the ability to solve a narrow problem using those representations, should be the ultimate goal of any system designed for exploratory visualization. The evaluation of the HealthVisPCP system for multivariate health statistics considers these issues in phase two of the assessment described below.

Assessment

In June 2000, thirty-one subjects at Penn State and six subjects at the National Cancer Institute participated in the assessment sessions. The Penn State subjects were undergraduate and graduate students in the geography department, selected because they were thought to represent a typical level of expertise of intended users: researchers—in this case, epidemiologists—with significant domain expertise. Specifically, subjects had experience both in examining datasets for spatial features and in representing those datasets on maps. The individuals who made up the sample were also all relatively familiar with other graphical representations—such as scatterplots and histograms—and GIS, but were relatively unfamiliar with the parallel coordinate representation. These characteristics were also thought to be representative of the target user group. The sessions lasted between ninety minutes and two hours. After an initial training tutorial, the assessment consisted of two phases, during which data was recorded.

Assessment: Phase One. The first phase in testing the usability of the PCP for exploratory visualization compared the use of the PCP and the scatterplot in isolation from each other. The strategy for comparing these representations consisted of comparing the performance of users in accomplishing a set of tasks. These tasks were systematically developed to represent tasks considered typical for use of the system to explore health statistics.

The development of the tasks used in the assessment was based on a typology that focused on three characteristics: the dimensionality of the task (the number of variables involved); the spatial extent, or scale, of the task (whether the task was local to one HSA or regional, incorporating several HSAs); and task complexity (the difficulty of the task). This task typology was modeled after a similar structure in an experiment developed by researchers in both cartography and epidemiology, a structure thought to represent the map-reading and use tasks typical of epidemiological analysis (MacEachren, Brewer, and Pickle 1998).

Given the geometry and purpose of the PCP relative to the scatterplot and other statistical graphics of lower dimensionality, one reasonable assertion is that the PCP would outperform the scatterplot in questions regarding multivariate tasks. The reverse may be true for those queries requiring bivariate interpretations. It was also considered likely that other task type distinctions and combinations might reveal that one representation was favored over the other, depending on the question type. Because of this possibility, it is reasonable to expect that users need to be given—and prefer working with—multiple representations forms in order to explore complex geospatial data fully and comprehensively.

In phase one, the participants were presented with sixteen tasks in the form of multiple-choice questions. These questions dealt with data preloaded into HealthVisPCP concerning prostate cancer and lung cancer for white males in the 1980s and early 1990s (part of the set of health data supplied by the National Cancer Institute as sample data). Each task was carefully created to represent one of the tasks in the typology above. There were two questions of each of the eight types tested, for a total of sixteen questions. In this phase, participants were shown *only one* of the two statistical graphic representations at a time, the scatterplot *or* the PCP. After the first set of eight questions, the representation that they had been shown disappeared and was replaced by the other, which was then available to use in answering the remaining eight questions. Four subject groups were used to counterbalance representation and question-set order. Both the participants' answers to the questions and the interactions the participants made with the environment were recorded for analysis.

Assessment: Phase Two. In phase two, participants were presented with a more open-ended task. They were given a new set of data—heart disease and lung cancer mortality rates in the 1980s and early 1990s for white females in the U.S.—and asked to play the role of an epidemiologist, looking for patterns and structure in these health statistics. The subjects were able to use any of the HealthVisPCP representations (scatterplot, PCP, map, thematic mapper, etc.) to look for interesting spatial, temporal, spatiotemporal, or attribute trends in the data. Their interactions were logged as in phase one. Participants were asked to provide

written commentaries of their observations in a text-entry dialog box.

This second phase was designed to assess the effectiveness of the system in performing exploratory tasks of observation and hypothesis generation and to investigate differences in the strategies employed by the participants. In effect, the interaction log in this phase became an independent variable in the analysis. By comparing the sophistication of observations made to the sophistication of the interaction strategies employed, it is possible to consider whether specific interaction strategies or representation use led to specific types of commentary and insights, thereby altering the way the researcher thinks about a given problem.

Usability Results

Accuracy. The interaction logs recorded in phase one were carefully designed to facilitate querying using relational-database software. In order to analyze the accuracy of responses, a series of cross-tabulations were designed, yielding multivariate contingency tables. The dependent variable in the tables is the number (and percent) of correct responses, given a representation form (PCP or scatterplot) and a question type. Since the focus of this particular assessment is the effect of the representation type on question-answering accuracy given specific types of questions, multiway tables were also created to list the counts of correct and incorrect answers by representation type and then by question type.

Some of these tables showed potentially significant differences in performance between the representation types, given a state of variable dimensionality in the question. Table 1, for example, shows effects such as "Given a multivariable question, users performed better using the PCP." A logistic regression model, appropriate for binary categorical response variables (in this case, "correct" or "incorrect" answers) was fit to the observed values. Despite the apparent interactions in the contingency tables such as the one mentioned above, the model specified in the analysis (via a backwards-elimination process) did not include interactions that would have resulted if the representation type, given a specific question type, had a significant influence on the odds of a correct or incorrect answer. Although the raw data seem to indicate such an interaction, the statistical evidence for such a relationship is not present.

These results (and others; see Edsall 2001 for a complete discussion) demonstrate that there is no objectively measurable difference in response accuracy between the two representation types. The parameter estimates associated with the logistic regression indicate some differences in accuracy when questions are grouped in other ways. For example, the odds of correctly answering a question regarding a single HSA are better than those of correctly answering a question regarding multiple HSAs. The intercept term is also very significant ($p < 0.05$), indicating that the odds of a correct answer are much greater (regardless of representation type or any of the other explanatory variables) than those of an incorrect answer. This latter result provides relatively strong support for the overall contention that geovisualization environments that incorporate multiple linked representations forms are usable and useful.

Interaction Log Visualizations. Analysis of the interaction logs involved the investigation of the actual interaction that led to the responses analyzed above. In a geovisualization environment, a variety of interaction strategies can often be used in an investigation of a task or the solution to a problem. Is it possible to observe a "sophisticated" interaction strategy that would lead to a correct answer? Did individuals who performed particularly well

Table 1 *Multiway Contingency Table Showing Potential Differences in User Performance (Percent of Answers Answered Correctly) Given Different Task Types and Representation Types*

	Parrallel Coordinate Plot				Scatterplot			
Type	No. Correct	No. Incorrect	Total No. Questions	Percent Correct	No. Correct	No. Incorrect	Total No. Questions	Percent Correct
Bivariate	62	24	86	72.09	68	19	87	78.16
Multivariate	63	19	82	76.83	59	28	87	67.82
Univariate	43	15	58	74.14	47	12	59	79.66

share ways of attacking a problem and working toward a solution? Likewise, did individuals who performed relatively poorly with the environment do so in part because their *interaction* with the environment was somehow inferior?

To address these possibilities, interactions a user made with the environment in the experiment described above were logged: each mouse click, each pan or zoom—each physical manipulation of the display or the data represented. From those records, important differences can be observed and patterns can be extracted distinguishing the interaction record of a high-scoring participant from that of a low-scoring participant.

Interaction logs may be analyzed in a variety of ways. For the purposes of this assessment, interaction strategies were of primary interest, and since interaction strategies are likely to be influenced by the type of question asked, separating interactions by individual (to examine individual subjects' strategies) and by question type (to determine if there is a preferred interaction method given a certain question type) was logical. These interaction strategies are analogous to signatures, for individuals *or* for question types: of interest is not just which interaction methods were used, but when and in what sequence (see Table 2 for interaction-method coding in the interaction-log graphs).

Table 2 *Interaction-Log Coding*

Window	Interaction
Question	**1. Answer:** question is answered
	2. Open: new question presented to users
System (interactions affecting multiple windows using buttons or menu items)	**4. ResetWins**: **R** tool was used to restore the representations to default zoom extent
Thematic Mapper	**6. Apply**: adjustments were made to the thematic mapper (variables displayed on choropleth map, number of classes), and apply button was depressed
Choropleth map/GeoView (a.k.a. ''Health Service Areas'')	**8. Brushing**: selection of HSA(s) on the choropleth map using the brushing tool
	9. ZoomIn: ArcView® zoom in tool was used to magnify the choropleth map
	10. PanHand: ArcView® ''pan hand'' tool was used to pan across the choropleth map
	11. ZoomOut: ArcView® zoom out tool was used to reduce the choropleth map
Parallel coordinate plot	**13. Brushing**: selection of HSA(s) on the PCP using the brushing tool
	14. ZoomIn: zoom tool was used to magnify the PCP
	15. ZoomOut: zoom tool was used to reduce the PCP
	16. PanSlider: PCP panner slider bar tool was used to pan on the PCP
	17. ZoomToExtent: PCP restore zoom tool was used to return to default extent of PCP view
	18. Reorder: PCP axes were reordered using dynamic reclassify tool
	19. PanHand: ArcView® ''pan hand'' tool was used to pan across the PCP
	20. RestoreAxes: PCP axes restored to original default order
Scatter plot	**13. Brushing**: selection of HSA(s) on the SP using the brushing tool
	14. ZoomIn: zoom tool was used to magnify the SP
	15. ZoomOut: zoom tool was used to reduce the SP
	16. Reclassify: dynamic reclassifier used to reposition class breaks on the SP*
	17. Focus (Left, Right, Up, Down): focus buttons were used to adjust the class breaks in the SP (and choropleth map and PCP indirectly)
	18. PanHand: ''pan hand'' tool was used to pan across the SP
Legend	**Brushing**: selection of legend elements: considered inadvertent
	ZoomIn: zoom tool was used to magnify the legend
	ZoomOut: zoom tool was used to reduce the legend
	PanHand: ''pan hand'' tool was used to pan across the legend

Note: The numbers associated with each interaction correspond to numbers on the y-axis of the interaction logs (Figure 5). These unordered (qualitative) interaction methods (right column) are grouped together according to the window within which the interaction occurs (left column); because the scatter plot and parallel coordinate plot were never visible at the same time, these numbers can be repeated in the interaction log visualizations. Interactions with the legend, though recorded, were likely accidental and were not included in the numerical coding.

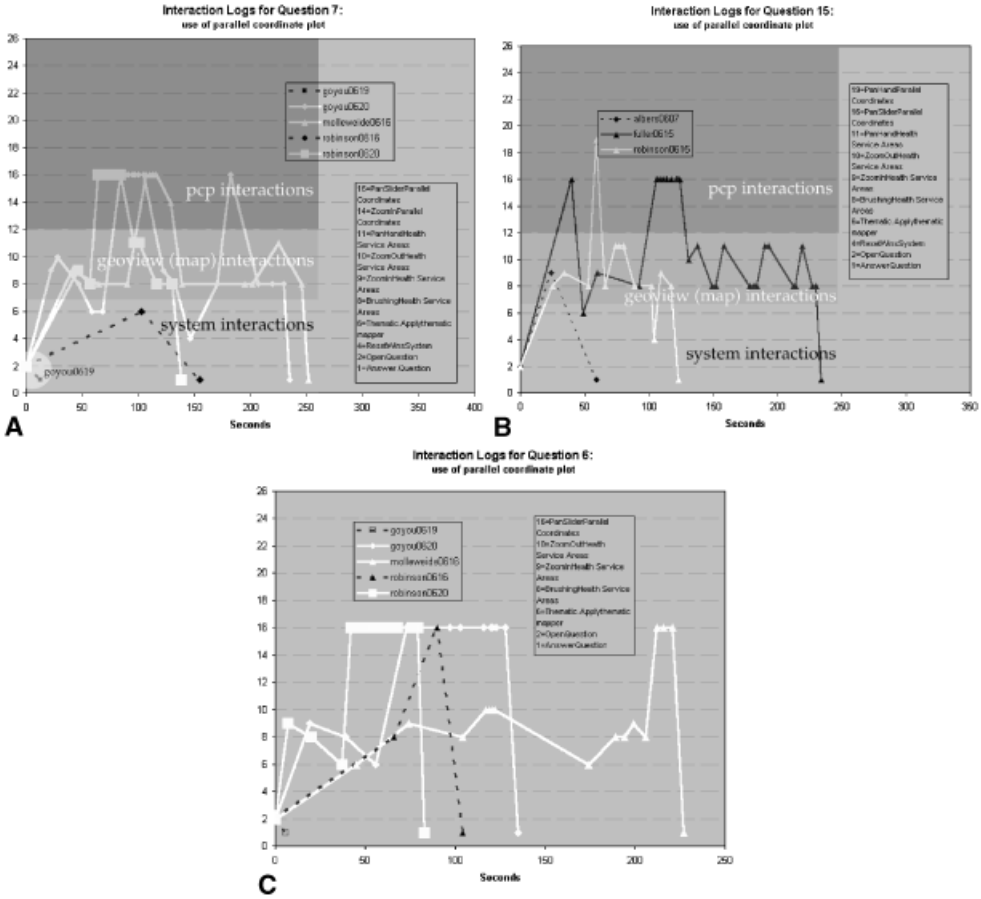

Figure 5 *Sample interaction log graphs. Dashed traces indicate an incorrect answer, solid traces indicate correct answers, black traces indicate low scorers, white traces indicate high scorers; for example, the interaction strategy of a low-scoring subject who answered the question correctly is represented by a solid black line (see text for details).*

To discover whether there was a discernable difference between interaction strategies leading to either successful or unsuccessful task accomplishment, eight of the thirty-one interaction logs were selected for visual analysis. The eight consisted of two groups: the four who scored the highest on the sixteen-question phase-two questionnaire (four subjects scored 15 out of 16: no one answered every question correctly) and the four who scored the lowest (between 7 and 9 questions correctly answered).

The graphs that are produced are timelines (time on the x-axis) of interactions, with each interaction marked on the timeline and mapped onto the y-axis in an order chosen by the researcher (Figure 5 shows sample interaction-log visualizations). For this analysis, interactions were grouped on the y-axis according to the window that was manipulated (Table 2). The example above (with system, map, and PCP interactions grouped together) is the ordering I chose to emphasize, since I was primarily interested in the representations being manipulated.

Much can be discerned by comparing the black (low scorers) and white (high scorers) traces. During the entire assessment, the four low scorers tended not to use features of the scatterplot or the PCP when it was made available to them, whereas the four more

successful subjects tended to use the statistical representations and several of their features with each question. In Figure 5A, it can be seen that goyou0619,[3] the subject with the fewest correct responses, employed very unsophisticated interaction strategies. In fact, the strategy employed by that subject was simply to open the question and answer it, with little or no interaction with the environment. Robinson0616 was occasionally more ambitious, adjusting the variables mapped, brushing the map, or using the PCP panner sporadically. Both subjects tended to finish answering questions much more rapidly than their successful counterparts. In Figure 5B, however, Fuller0615 stands out as a highly curious (though low-scoring) subject, using a wide variety of interactions, often spending a great deal longer than his/her lower-scoring counterparts deciding on an answer.

A careful look at several of the questions indicates that it may be possible to extract a prototypical "successful" interaction strategy: an amalgam of the interaction signature of the three successful participants using the same representation (goyou0620, molleweide0616, and robinson0620). This is most evident in the log visualization for question 6 (Figure 5C). Each of the three participants (all of whom answered this question correctly) panned the PCP extensively just before answering, and before that, they each manipulated the choropleth map in various ways. Though each subject spent a different amount of time with the question, the interaction strategy—the tools used and the sequence in which they were used—was similar. It can also be seen that those who got the questions wrong employed remarkably different (and simpler) strategies for that question.

The preceding provide examples of general patterns. Successful subjects generally exhibited more interaction with the system before deciding on each response than did their less successful counterparts. Visualizing these interaction logs reveals other evidence of specific sequences and strategies that are more likely to lead to a correct answer. In addition, those subjects who interacted with the most representations tended to answer more accurately.

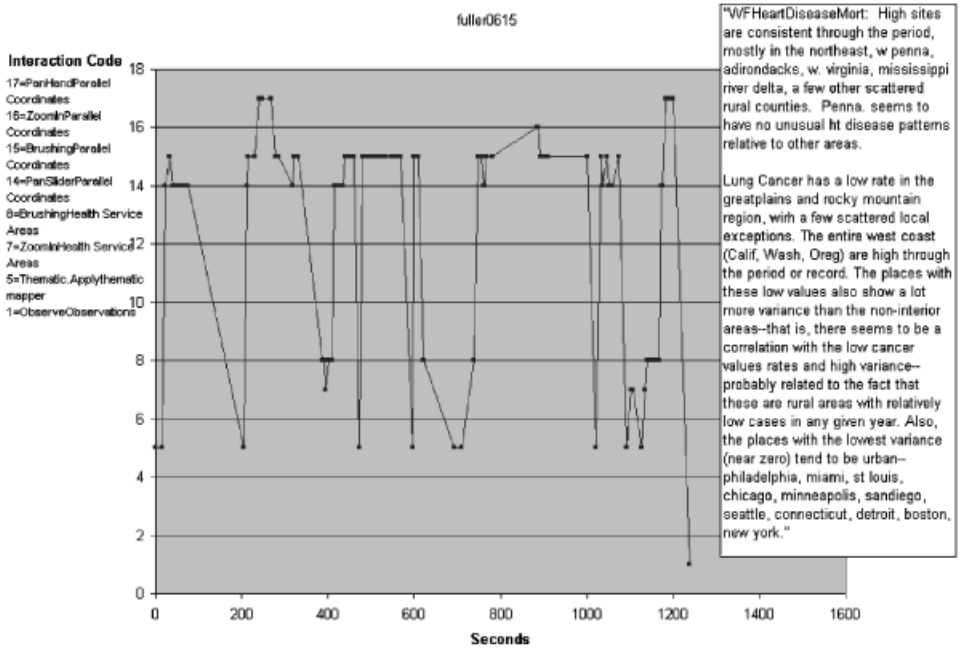

Figure 6 *An example exploration interaction log, with corresponding user observations and commentary.*

Exploration Logs. The interaction logs described above also serve as useful tools to understand how subjects used the tools and the environment for more free-form exploration of data about which they knew little or nothing (and for which they were not prompted by a request to perform specific tasks). What useful information can be gleaned from these interaction logs and their corresponding commentary statements?

The commentaries of observations, like that found in the text box of Figure 6, revealed some interesting similarities and differences. For example, all seven observation sets are similar, in that they include some recognition of spatial patterns. Such recognition is illustrated by phrases such as "Concentration of heart disease notable in the east, especially the Appalachian Region" and "Lung cancer seems to be a problem especially in Florida and Northern California." These observations are possible only through the use of the choropleth map, either alone or in tandem with other representations. This use is reflected in the frequent use of the thematic mapper tool, which changes all representations to varying degrees but has an obvious and direct effect on the map. The consistency of the recognition of spatial patterns—be they correct or incorrect—is probably a reflection of the perceived focus of the study (which took place in a geography lab, was conducted by a geographer, and consisted of an extension of a geographic information system). It is thus difficult to extrapolate that this system is necessarily best for observing spatial patterns (as opposed to other kinds of patterns) simply because a majority of observations dealt with space.

Temporal trends were observed *only* by those subjects who interacted with the PCP. Observations such as "General decrease in heart disease mortality (HDM)/General increase in lung cancer mortality (LCM)" and "There is a decline in heart disease mortality over time" represent the recognition of temporal features. As described previously, the most direct way in which temporal trends are noticed in the environment is through the PCP. *Spatio*temporal trends are probably best observed through an animation of the choropleth map through time, a feature that was not available to the users. However, some users were able to notice spatiotemporal features; for example, fuller0615 observed that "High sites [of white female heart disease] are consistent through the period, mostly in the northeast, W Penna., Adirondacks, W. Virginia, Mississippi River delta, a few other scattered rural counties." Notice that this complex observation consists of the identification and classification of specific locations and regions according to their temporal characteristics. The interactions of this user are among the most complex of the subjects. S/he panned and zoomed the PCP, zoomed the scatterplot, brushed the map, and frequently adjusted the thematic mapper. Through this wide variety of interactions, s/he made complex observations.

Only two subjects stated what was, unambiguously, a "hypothesis." Those two considered relationships and factors beyond the data represented for possible explanations about the data: molleweide0614 noted that "There is a greater incidence in those areas with an older population" and fuller015 said that patterns s/he saw were "possibly related to the fact that these are relatively rural populations." Although it is impossible to draw conclusions from a sample of this size, if the concept of a hypothesis is expanded to include any observation that includes the word "seem" (e.g., "Heart disease seems to be a particular problem in Florida"), since use of that word might indicate active and curious searching and investigating, more than half of the subjects were using the system to generate statements that indicate an exploratory curiosity. This larger percentage is encouraging, since the greatest test for a visualization environment's success is its ability to prompt such noted but unconfirmed patterns and to facilitate curious and creative thought.

Hypothesis generation is an extension of the observation of trends, patterns, and outliers and is a high-level conceptual goal of visualization systems. It is reasonable to expect that those who developed hypotheses about the data represented are those who interacted with the system in a better-planned way. However, such a qualitative assessment depends to a great extent not only on the interpretation of the researcher, but also on the level of knowledge about the domain (in this case, epidemiology) of the participant. Certainly, individuals who are trained to look for spatial trends and to speculate about possible correlations are more likely to use the system for that purpose than individuals who are novices in

the domain. This expert-novice distinction will deserve close attention in future, similar assessment exercises.

Through the assessments described in the section above, it can be said that the PCP and the scatterplot play important roles in successful exploration: the most complex and comprehensive commentaries—those that discuss spatial, temporal, spatiotemporal, and attribute trends, patterns, and hypotheses—were those that made extensive use of the interactive capabilities of the statistical representations.

Conclusions

This article presents a system for exploring health statistics. The multidimensional nature of health statistics and their analysis calls for creative and useful techniques for their visual representation. A highly interactive system such as the one presented here allows multiple perspectives on complex information, which can lead to deeper and more comprehensive knowledge construction about the data represented. This assertion is examined in a usability assessment focused on the efficacy of the PCP and the scatterplot when linked to a geographic representation such as a choropleth map. The findings of the assessment include a lack of support for the superiority of the PCP or scatterplot to accomplish narrowly defined tasks of any specific type, at least according to the task typology developed for the evaluation. Nevertheless, interaction logs and written commentary of participants in the study indicate that a visualization environment is most effective for data exploration when a variety of tools are presented to and used by the researcher. The linked PCP is one of a suite of tools and representations that, in tandem, serve to create an important visual link between the human analyst and the represented multidimensional data. ■

Notes

[1] Extensions to the Starlight project developed recently at Pacific Northwest National Laboratory include three-dimensional PCPs linked with maps.

[2] Alaska and Hawaii were added in subsequent versions, but the application used for the tests described in the fifth section of this article did not include those states.

[3] Subject names were randomly assigned and are in no way related to the subjects' names or identities.

Literature Cited

Andrienko, G. L., and N. V. Andrienko. 1999. Interactive maps for visual data exploration. *International Journal of Geographic Information Science* 13 (4): 355–74.

Becker, R. A., W. S. Cleveland, and A. R. Wilks. 1988. Dynamic graphics for data analysis. In *Dynamic graphics for statistics*, ed. W. S. Cleveland and M. E. McGill, 1–49. Belmont, CA: Wadsworth and Brooks.

Buja, A., D. Cook, and D. Swayne. 1996. Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics* 5 (1): 78–99.

DiBiase, D. 1990. Visualization in the earth sciences. *Bulletin of the College of Earth and Mineral Sciences* 59 (2): 13–18.

DiBiase, D., C. Reeves, A. M. MacEachren, J. Krygier, J. Sloan, and M. Detweiller. 1994. Multivariate display of geographic data: Applications in earth systems science. In *Visualization in modern cartography*, ed. A. M. MacEachren and D. R. F. Taylor, 287–312. Oxford, U.K.: Pergamon.

Dykes, J. A. 1997. Exploring spatial data representation with dynamic graphics. *Computers and Geosciences* 23 (4): 345–70.

Edsall, R. M. 1999. Development of interactive tools for the exploration of large geographic databases. In *Proceedings of the 19th International Cartographic Conference*, ed. C. P. Keller. Canadian Institute of Geomatics, Ottawa, 14–20 August. CD-ROM.

———. 2001. Interactive with space and time: Designing dynamic geovisualization environments. Ph.D. diss., Department of Geography, The Pennsylvania State University.

Environmental Systems Research Institute (ESRI). 1999. ArcView GIS 3.2. Redlands, CA.

Erbacher, R. F., G. Grinstein, J. P. Lee, L. Masterman, R. Pickett, and S. Smith. 1995. Exploratory visualization research at the University of Massachusetts at Lowell. *Computers and Graphics* 19 (1): 131–39.

Griffith, D. A., and C. G. Amrhein. 1997. *Multivariate statistical analysis for geographers*. Upper Saddle River, NJ: Prentice-Hall.

Hurley, C., and A. Buja. 1990. Analyzing high-dimensional data with motion graphics. *SIAM Journal of Statistical Computing* 11 (6): 1193–211.

Inselberg, A. 1985. The plane with parallel coordinates. *The Visual Computer* 1: 69–97.

Karat, J. 1997. User-centered software evaluation methodologies. In *Handbook of human-computer interaction*, ed. M. Helander, T. K. Landauer, and P. Prabhu, 689–704. Amsterdam: Elsevier.

Knapp, L. 1995. A task analysis approach to the visualization of geographic data. In *Cognitive aspects of human-computer interaction for geographic information systems*, ed. T. Nyerges, D. Mark, R. Laurini, and M. Egenhofer, 355–72. Dordrecht, The Netherlands: Kluwer.

MacDougall, E. B. 1992. Exploratory analysis, dynamic statistical visualization, and geographic information systems. *Cartography and Geographic Information Systems* 19 (4): 237–46.

MacEachren, A. M., F. P. Boscoe, D. Haug, and L. W. Pickle. 1998. Geographic visualization: Designing manipulable maps for exploring temporally varying georeferenced statistics. In *Proceedings of the Information Visualization '98*, ed. J. Hollan, 87–94. Raleigh-Durham, NC: IEEE Computer Society.

MacEachren, A. M., C. A. Brewer, and L. W. Pickle. 1998. Visualizing georeferenced data: Representing reliability of health statistics. *Environment and Planning A* 30 (9): 1547–61.

MacEachren, A. M., B. Buttenfield, J. Campbell, D. DiBiase, and M. Monmonier. 1992. Visualization. In *Geography's inner worlds: Pervasive themes in contemporary American geography*, ed. R. Abler, M. Marcus, and J. Olson, 99–137. New Brunswick, NJ: Rutgers University Press.

MacEachren, A. M., and J. Ganter. 1990. A pattern identification approach to geographic visualization. *Cartographica* 27 (2): 64–81.

MacEachren, A. M., and M.-J. Kraak. 1997. Exploratory cartographic visualization: Advancing the agenda. *Computers and Geosciences* 23 (4): 335–43.

MacEachren, A. M., M. Wachowicz, R. Edsall, D. Haug, and R. Masters. 1999. Constructing knowledge from multivariate spatiotemporal data: Integrating geographic visualization (GVis) with knowledge discovery in databases (KDD). *International Journal of Geographic Information Science* 13 (4): 311–34.

Monmonier, M. 1992. Summary graphics for integrated visualization in dynamic cartography. *Cartography and Geographic Information Systems* 19 (1): 23–36.

Plaisant, C. 1993. Facilitating data exploration: Dynamic queries on a health statistics map. In *Proceedings of the Annual Meeting of the American Statistical Association, Government Statistics Section*, 18–23. San Francisco.

Rogers, Y. 1999. What is different about interactive graphical representations? *Learning and Instruction* 9: 419–25.

Tukey, J. W. 1977. *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Wang, X., J. X. Chen, D. B. Carr, B. S. Bell, and L. W. Pickle. 2002. Geographic statistics visualization: Web-based micromap plots. *Computing in Science and Engineering* 4 (3): 90–94.

Wegman, E. J. 1990. Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association* 85 (411): 664–75.

———. 2000. Visions: New techniques and technologies in statistics. *Computational Statistics* 15: 133–44.

ROBERT M. EDSALL is an assistant professor in the Department of Geography at Arizona State University, Tempe, AZ 85287-0104. E-mail: robedsall@asu.edu. His current research focuses on the design and usability of geographic visualization tools for a wide variety of geographic research and classroom applications.