Case Study: Design and Assessment of an Enhanced Geographic Information System for Exploration of Multivariate Health Statistics

Robert M. Edsall
Department of Geography
Arizona State University
robedsall@asu.edu

Alan M. MacEachren
Department of Geography
The Pennsylvania State University
alan@geog.psu.edu

Linda Pickle
National Cancer Institute
picklel@mail.nih.gov

Abstract

An implementation of an interactive parallel coordinate plot linked with the ArcView® geographic information system (GIS) is presented. The integrated geographic visualization system was created for the exploratory analysis of mortality data from specific cancers as they relate, specifically spatially, to other mortality causes and to demographic and socioeconomic risk factors. The linked and interactive parallel coordinate plot was tested with and compared to a similarly interactive and linked scatterplot in usability assessments designed to assess each representation's relative effectiveness for exploration of these data sets. Evidence from these studies suggests that multivariate, spatial, and/or time series exploration is enhanced through the use of the parallel coordinate plot linked to maps.

1. Introduction

Epidemiologists are among the many types of researchers for whom the understanding of the data sets they frequently use is made challenging by the multidimensional nature of those data. This paper will report on an effort to design and implement 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.

2. HealthVisPCP

The visualization environment described in this paper is an extension of previous work to link GIS to statistical graphics for health statistics visualization. MacEachren, Boscoe, et al. [1] developed the original HealthVis system

to analyze (visually) spatial, temporal, and attribute features of health statistics for the National Center for Health Statistics (NCHS). The extensions to HealthVis described here were implemented as part of a contract from the National Cancer Intsiute (NCI).

Implementation of HealthVis, and of extensions specific to the NCI contract discussed in detail below, were accomplished using ArcView®'s scripting language "Avenue." The HealthVis system consisted of three ArcView® "View documents": (1) a choropleth map of the contiguous 48 United States, within which are approximately 800 "health service areas" (HSAs), multi-county units used by federal agencies for analyzing health statistics data; (2) a scatter plot, and (3) a map legend. By creating the application entirely within ArcView®, the developers were able to take advantage of the existing interaction capabilities of ArcView® View documents, like panning, zooming, and selecting.

In addition, a great deal of functionality was added to those documents in order to accomplish a series of goals specific to health statistics visualization. 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 discussed among those for the original implementation is the facilitation of multidimensional exploration among several variables, accomplished by developing a dynamic parallel coordinate plot (PCP) in Avenue.

The application (Figure 1) adds a dynamic PCP (each line on which represents a single HSA) to the three documents of HealthVis. The choropleth map is colored according to a statistic (such as "prostate cancer mortality, white male") and a year. The statistic and the year can be selected by the user in a redesigned dialog box known as the Thematic Mapper. Also in this dialog box, a user can specify the number of classes and create 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 scatter plot.
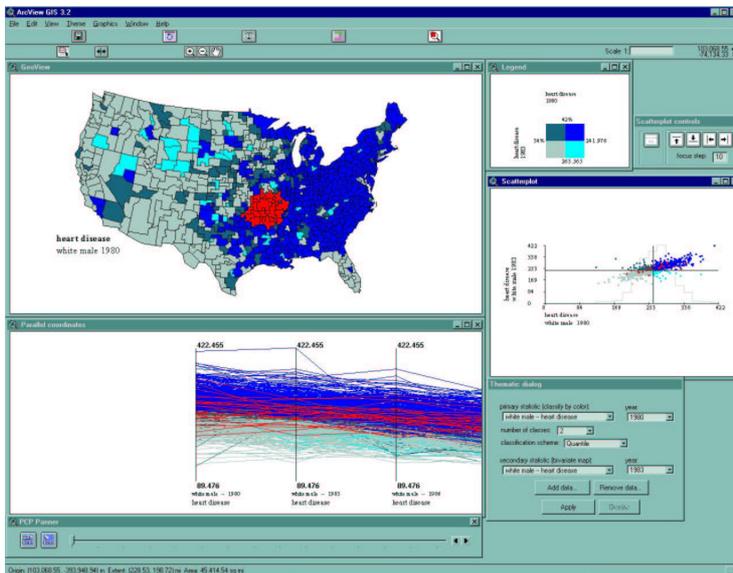
Figure 1. The HealthVisPCP environment.

Linked brushing was also implemented. When an object, or set of objects (HSAs in the choropleth map, points on the scatter plot, 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.

The PCP was added to the original HealthVis system as a new View document in ArcView®, in a similar fashion to that used for the scatter plot. As a View, the tools for manipulating the three representations (PCP, scatter plot, and map – all View documents) behaved similarly across the representation types. 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 [4]), other useful enhancements were identified and are 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 included. Also added was a button that zooms the PCP to a default extent that maximizes the amount of information contained in the PCP 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 [2, 3, 4].

Though the PCP is a clever data representation technique, it has several disadvantages over 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 (on a scatter plot) 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. A goal in the assessment (described in the next section) was to discover whether or not these disadvantages can be overcome by the PCP's use in an interactive environment that includes such refinements like brushing, focusing, coloring, zooming, and other graphical manipulation tools.

3. Investigating linked statistical graphics: usability assessment of HealthVisPCP

The specific goal of the assessment was to compare the relative effectiveness of each representation form (the scatter plot and the PCP) for a series of narrowly defined tasks (in phase one), and to test the effectiveness of the environment as a whole for unrestricted exploration of multivariate spatiotemporal data (in phase two).

In June of 2000, thirty-one subjects at Penn State and six subjects at the National Cancer Institute participated in the assessment sessions. After an initial training tutorial, the assessment consisted of two phases during which data was recorded.

3.1. Assessment: Phase One

The first phase in testing the usability environment compared the use of the PCP and the scatterplot in isolation from each other. The strategy for comparing these representations consisted of examining the performance of users in accomplishing a set of tasks. These tasks were systematically developed to represent (narrow) tasks considered typical for use of the system to explore health statistics.

In phase one, the participants were presented with sixteen tasks in the form of multiple-choice questions. These questions dealt with data pre-loaded into Health-

VisPCP concerning prostate and lung cancer over time (part of a sample set of health data supplied by the NCI). Each question was carefully created to represent one of the tasks in a *task typology*. This typology focused on three characteristics of the task: the dimensionality of the task (the number of variables involved), the spatial extent of the task (the number of spatial units involved), and task complexity (the difficulty of the task). 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 scatter plot in questions regarding multivariate tasks. The reverse may be true for those queries requiring bivariate interpretations. Because of this possibility (and others like it), 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.

There were two questions of each of the eight types tested (a subset of the possible types in the typology described above). In this phase, participants were shown only one of the two statistical graphic representations at a time, the scatter plot or the PCP (seeing each representation for eight questions, in random order). Both the participants' answers to the questions and the interactions the participants made with the environment to reach an answer were recorded for analysis.

3.2. Assessment: Phase Two

In phase two, participants were presented with an open-ended task. They were given a new set of data – heart disease and lung cancer mortality rates over time – 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).

3.3. Results: The Accuracy Analysis

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 "response" variable in the tables is the number (and percent) of correct responses, given a representation form (PCP or scatter plot) and given 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, multi-way tables were also created to list the counts of correct and incorrect answers according to representation type, then by question type.

Some of these multi-way tables seemed to show a potentially significant difference in performance between the representation types, given a state of variable dimensionality in the question. 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, the model specified in the analysis (via a backwards-elimination process in SAS) did not retain 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 interactions, the statistical evidence for such a relationship was not present. However, the intercept term in the model was very significant, 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 visualization environments that incorporate multiple linked representations forms are usable and useful.

3.4. Results: The Interaction Log Visualizations

Analysis of the interaction logs involved the investigation of the actual interaction that led to the responses analyzed above. In a visualization 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?

Interaction logs may be analyzed in a variety of ways. For the purposes of this assessment, interaction *strategies* were of primary interest. To discover whether there was a discernable difference between interaction strategies leading to either successful or unsuccessful task accomplishment, eight of the 34 interaction logs were selected for visual analysis. The eight consisted of two groups: the four who scored the highest on the 16-question phase-two questionnaire and the four who scored the lowest.

We produced graphs that serve as timelines of interactions, (Figure 2). For this analysis, interactions were grouped on the y-axis according to the window that was manipulated. The example above (with system, map, and PCP interactions grouped together) is the ordering we chose to emphasize, since we were primarily interested in the representations being manipulated.

Comparing the black (incorrect) and white (correct) traces, much can be discerned. Among other observations, successful subjects were found to exhibit more
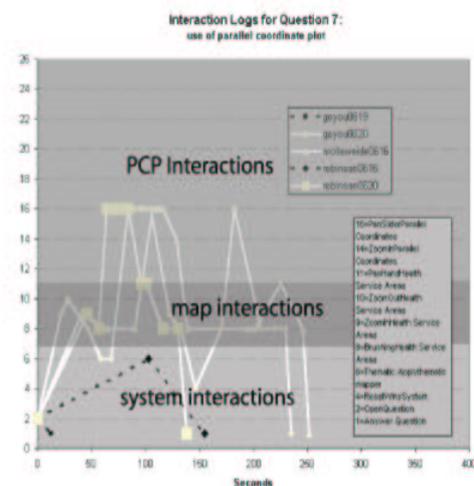
Figure 2. Example interaction log for 5 participants while using PCP. Traces of correct answers are solid, incorrect are dashed. Traces of adept users are white, less adept users are black.

interaction with the system before deciding on each response. Visualizing these interaction logs also reveals evidence of specific *sequences* and strategies that are more likely to lead to a correct answer.

3.5. Results: The 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?

Indeed, the sophistication and type of strategy employed in interacting with the system to explore data seem to have an effect on the type and quality of the observations in the commentary. For example, all of the recorded commentaries are similar in that each includes some recognition of spatial patterns. 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 Geo-View. Temporal trends, in contrast, were observed *only* by those subjects who interacted with the PCP. The most direct way temporal trends are noticed in the environment is through the PCP.

Through these assessments, it can be said that the parallel coordinate plot and the scatter plot play important roles in successful exploration: the most complex and comprehensive commentaries – those that discuss spatial, temporal, spatiotemporal, and attribute trends, patterns, and hypotheses – are those that made extensive use of the interactive capabilities of the statistical representations.

4. Conclusions

In this paper, a system for exploring health statistics is presented. The multidimensional nature of health statistics and their analysis calls for creative and useful techniques for their visual representation. A highly interactive system like 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 parallel coordinate plot and scatterplot when linked to a geographic representation (like a choropleth map). The findings of the assessment include the 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 which, in tandem, serve to create an important visual link between the human analyst and the represented multidimensional data.

[1] MacEachren, A.M., F. Boscoe, D. Haug, and L. Pickle. "Geographic Visualization: Designing Manipulable Maps for Exploring Temporally Varying Georeferenced Statistics." *Proceedings of Information Visualization '98*. IEEE Computer Society, Raleigh-Durham, NC, Oct. 19-20, 1998, pp. 87-94.
[2] Inselberg, A. "The Plane with Parallel Coordinates." *The Visual Computer* **1**: 69-97. 1985.
[3] Wegman, J. J. "Hyperdimensional Data Analysis Using Parallel Coordinates." *Journal of the American Statistical Association* **85**(411): 664-675. 1990.
[4] Edsall, R. "Development of Interactive Tools for the Exploration of Large Geographic Databases." 19th International Cartographic Conference, Ottawa, University of Victoria. 34B. 1999.