Exploring High-D Spaces with Multiform Matrices and Small Multiples

Alan MacEachren[1], Xiping Dai[1], Frank Hardisty[1], Diansheng Guo[1], Gene Lengerich[2]
[1]GeoVISTA Center, Department of Geography, The Pennsylvania State University, University Park, PA 16802
[2]Department of Health Evaluation Sciences, The Pennsylvania State University, Hershey, PA 17033
{maceachren, xpdai, fhardisty, dguo, elengerich}@psu.edu

Abstract

We introduce an approach to visual analysis of multivariate data that integrates several methods from information visualization, exploratory data analysis (EDA), and geovisualization. The approach leverages the component-based architecture implemented in GeoVISTA *Studio* to construct a flexible, multiview, tightly (but generically) coordinated, EDA toolkit. This toolkit builds upon traditional ideas behind both small multiples and scatterplot matrices in three fundamental ways. First, we develop a general, *MultiForm, Bivariate Matrix* and a complementary *MultiForm, Bivariate Small Multiple* plot in which different bivariate representation forms can be used in combination. We demonstrate the flexibility of this approach with matrices and small multiples that depict multivariate data through combinations of: scatterplots, bivariate maps, and space-filling displays. Second, we apply a measure of *conditional entropy* to (a) identify variables from a high-dimensional data set that are likely to display interesting relationships and (b) generate a default order of these variables in the matrix or small multiple display. Third, we add *conditioning*, a kind of dynamic query/filtering in which supplementary (undisplayed) variables are used to constrain the view onto variables that are displayed. Conditioning allows the effects of one or more well understood variables to be removed from the analysis, making relationships among remaining variables easier to explore. We illustrate the individual and combined functionality enabled by this approach through application to analysis of cancer diagnosis and mortality data and their associated covariates and risk factors.

CR Categories and Subject Descriptors: H.2.8 Database Management: Database Applications--data mining, spatial databases and GIS; I.5.2 Pattern Recognition: Design Methodology--feature evaluation and selection; I.5.3 Pattern Recognition: Implementation--interactive systems; J.3 Life and Medical Sciences: Health

Additional Keywords: geovisualization, EDA, scatterplot matrix, bivariate map, space-filling visualization, conditional entropy, small multiples, conditioning, GeoVISTA *Studio*

1. Introduction

Data often become information when methods and tools enable multivariate relationships to be uncovered. Identifying and understanding complex, multivariate relationships requires examination of those relationships from multiple perspectives. Here, we present an approach to multivariate data analysis that extends from and integrates several well-known visual display and analysis methods used in information visualization, exploratory data analysis (EDA), and geovisualization. A suite of methods are incorporated into a pair of dynamic, computationally-assisted, visual analysis tools. These tools emphasize juxtaposition and sorting of multiple, bivariate representations that: (a) use interactive statistical analysis of bivariate association to generate default ordering in the display and to select data subspaces with variables of potential interest and (b) can be "conditioned" by filtering on non-displayed variables.

Our work builds on concepts and methods from several perspectives on visual information analysis; key aspects of this past work are highlighted in section 2. Section 3 provides a brief overview of GeoVISTA *Studio*, our application-building environment that supports flexible, component-based application design and implementation. Then, we introduce a suite of MultiForm, multivariate, visual data analysis tools, outline a method for selecting and ordering variables based on conditional entropy, and discuss addition of data conditioning (section 4). Throughout Section 4 we demonstrate the use of the methods and tools through application to georeferenced cancer and risk factor data. Section 5 provides conclusions and plans for future work.

2. Related Work

The work reported here is informed by past research on five general categories of multivariate information analysis methods: information sorting (permutation matrices and correlation ordering), small multiples (with emphasis on user manipulation), bivariate dynamically linked matrices (brushable scatterplot matrices and their extensions), space-filling visualizations (tree-maps, mosaic plots, and pixel-oriented displays), and dynamic scope control (including dynamic query and conditioning).

2.1 Sorting

Sorting is a fundamental information organization and analysis method incorporated in database, spreadsheet, and related software. The ability to sort a table or spreadsheet on selected attributes allows a user to understand relationships between one variable and many or (with nested sorting) among multiple variables. The power of visually-facilitated sorting as a method to uncover multivariate relationships has been explored by several authors. Many ground their work in Bertin's concept of a permutation matrix, e.g., (Gluck et al., 1999; Siirtola, 1999; Wilkinson, 1979). A *permutation matrix* is essentially a graphical table for which: (a) cell values are replaced by a graphic depiction of the value and (b) rows and columns can be sorted by an agent (human, computational, or both) to search for groupings of related entities.

Several authors have applied automated sorting routines to derive a useful matrix depiction. Schmid and Hinterberger (1994) advocate completely automated sorting to reduce user bias. Mäkinen and Siirtola (2000), in contrast, demonstrate that at least some sub-problems underlying the permutation matrix approach are

NP-complete. As a result, they advocate user controllable, computationally-assisted sorting. Their "reorderable matrix." directly implements Bertin's concept of a matrix with attributes (depicted by the graphic variable size) along the x-axis and objects along the y-axis, while balancing the task of sorting for interesting patterns between the system and the user. In complementary work, Gluck (2001) has extended the basic idea, under the label of *augmented seriation*. His extensions include dynamic linking with small multiples of univariate and bivariate maps and application of redundant graphic variables (size, color saturation), as well as sonic variables (pitch, loudness), to enhance pattern identification and hypothesis generation.

In contrast to the permutation matrix/seriation method (which is applied to univariate entries in a table), (Friendly, 2002b) proposes the concept of *correlation ordering*, which is applied to bivariate relationships in correlation matrices. The goal is to put "similar" variables (in terms of correlations) together in the matrix to enhance detection of trends and anomalies. Ankerst et al. (Ankerst et al., 1998) have proposed a related computational approach to sorting that uses a heuristic method to cluster similar dimensions. Below, we present an alternative approach for sorting variables based on conditional entropy as the measure of bivariate association and minimum spanning trees as the sorting method.

2.2 Small multiples

A second information visualization method that underlies the work presented here is the *small multiple* – a set of juxtaposed data representations that together support understanding of multivariate information. Although Tufte (1983) popularized this method (and is often credited with the term), the idea was introduced as a data analysis technique by Bertin in 1967 (called simply *collections*); and presented in English editions of his books in the early 1980s (Bertin, 1981, 1983).

A key difference between Bertin's concept and that labeled small multiples by Tufte is that Bertin emphasized the potential for insight about patterns and relationships that can be gained by providing users with the ability to sort the collections of related views (thus to reorder the small multiples). We have previously considered the potential of interactive "extensions" to the small multiple method in the context of geovisualization (DiBiase et al., 1994; MacEachren, 1995, 2001). More recently, Pinnel (2000) implemented a version of manipulable, map-based small multiples that includes both temporal map representations and difference maps to show change at each time point (in relation to a base time). The method has also been applied recently, by Iizuka and colleagues (1998), to visual comparison of tabular data using graphic table displays that are similar in appearance to permutation matrices. In this work, they apply computational methods to sort the tables into a similarity-based small-multiple arrangement.

2.3 Linked bivariate matrices

A third core visual analysis method on which our work builds is the *scatterplot matrix*, a method with roots in statistical EDA (Chambers et al., 1983). While small multiples have most often been used for display and analysis of multiple, univariate representations, the scatterplot matrix by design focuses on multiple, bivariate relationships. Many extensions have been made to the basic method over the years, see: (Becker & Cleveland, 1987), (Reed et al., 1995), (Rousseeuw et al., 1999; Zani et al., 1998). Several authors have used dynamic linking to combine the scatterplot matrix with other display forms such as maps (Carr et al., 1987; Cook et al., 1997; Monmonier, 1989) and alternative multivariate depictions (Schmid & Hinterberger, 1994). Friendly (1999), in work particularly relevant to our own, adapted the basic matrix form to produce both a conditional scatterplot matrix and a bivariate mosaic plot matrix. In the conditional scatterplot matrix, each panel depicts the partial correlation between row and column variable, given the remaining variables. The mosaic matrix is the categorical data analog of a scatterplot matrix, showing selected aspects of the bivariate relation between all pairs of variables.

2.4 Space-filling visualization

A mosaic plot (used for each panel in a mosaic matrix) is one of several forms of *space-filling visualization* developed to represent categorical data. The term space-filling seems to have been coined by Johnson and Shneiderman (1991) in a paper "introducing" tree maps. However, the basic idea of space-filling visualization can be traced to the early origins of information graphics in late 17th century Europe (Friendly, 2002a) and a close precursor to modern tree maps was proposed in 1934 by the cartographer Raisz, as an abstract form of cartogram (Raisz, 1934). There are a number of other possible variants on the space-filling idea, each of which imposes different constraints on the display. Stasko and colleagues (2000), for example, consider their sunburst method to be space-filling, but relax the constraint that the space is rectangular. They cite empirical evidence that their concentric representation of nested classification has user performance advantages over a tree map representation.

We consider the pixel-oriented methods developed by Keim (Keim & Kriegel, 1996; Keim, 2000) to be part of the family of space-filling visualizations, since these methods also fill the entire space by subdividing it. The primary differences between pixel-oriented methods and tree-maps or mosaic plots are that the pixel-oriented methods constrain all entities to representation by equal size (generally square) rectangles, with order of the rectangles in the space and color attributes assigned to them used to communicate categories or quantities. In addition, pixel-oriented methods were developed for depicting very large data sets and Keim and Kriegel (1996) proposed that they can be effective when using only one pixel per data value. Below, we introduce a hybrid space-filling visualization that draws most heavily on the pixel-oriented method.

2.5 Dynamic scope control

Dynamic scope control can be a powerful complement to the multivariate visual display and analysis methods reviewed above. Scope control, as defined by Goldstein and Roth (1994), is any user-initiated operation that constrains the data entities displayed to those meeting specified parameters on one or more variables (displayed or not). *Dynamic query* is a common form of scope control used in information visualization. It involves rapid adjustment of multiple query parameters, often through user adjustable sliders (Ahlberg et al., 1992). The method is applied in a wide range of visualization tools for an equally wide range of applications (Ahlberg & Shneiderman, 1994; Goldstein & Roth, 1994; Kumar et al., 1997; Tweedie, 1997).

Conditioning is a variant on dynamic query in which the view on the displayed data is constrained by controlling the scope of one or more secondary variables that are not displayed (Carr et al., 2000). Conditioning is applied when the goal is to control (or condition) for the effects of one variable while exploring others. For example, to explore risk factors for lung cancer, we might condition the display on data for smoking (a known risk factor) while we explore potential (but less well understood) risk factors.

The method is derived from the *prosection* technique first introduced by Furnas and Buja (1994) and has been successfully applied to bivariate matrix displays (Brunsdon, 2001; Tweedie et al., 1996) and to map-based displays (Brunsdon, 2001; Carr et al., 2000; Carr et al., submitted).

3. GeoVISTA *Studio*

A primary contribution of the work presented here is to develop and implement methods that enable integration of ideas from the diverse set of InfoVis/EDA/geovisualization developments discussed above. We do this by leveraging GeoVISTA *Studio*, an open source visual programming environment for building Java applications and applets from JavaBeans (Figure 1). The *Studio* engine provides a generic environment to support construction of applications in any domain. Components available with the current *Studio* distribution (http://sourceforge.net/) include ones for visualization, computational analysis (e.g., high-dimensional clustering, inductive learning, K-means classification) and statistical analysis (e.g., outlier detection, local spatial autocorrelation). Our focus in this paper is on integration of a suite of visualization components that we have specifically implemented and that support multiple kinds of coordinated behavior applied to multivariate data exploration. For more complete discussion of the *Studio* environment, see (Gahegan et al., 2002; MacEachren et al., 2003)

Studio uses Java's introspection capabilities to discover each component's available methods, allowing application designers to combine components that were not originally developed to work together. This mechanism, allowed us to add an attribute *animator* that supports an animated grand tour (by variable and subcategory – see accompanying video). This was accomplished in about one hour, even though none of the components had been specifically designed to support animation. Figure 2 illustrates a typical *Studio* design (for an applications discussed below).

4. Multiform represention

The focus of the work presented here is on integrating methods and tools for exploratory analysis of multivariate data. Specifically, the work emphasizes juxtaposition, ordering, manipulation, and linking of multiple bivariate views. We begin by defining three classes of display types, and their subclasses.

1. UniForm Univariate Matrix – matrices building on Bertin's univariate permutation matrix ("Form" is a generic label for a kind of representation; here it can be point symbols that vary in size, cell fills that vary in color value, a map that varies in content, etc.). This category includes both permutation matrices and sortable small multiples.
2. Bivariate Small Multiples – arrangements of bivariate representation forms that obtain data assignments from columns and rows:
   a. UniForm Bivariate Small Multiple – each row uses the same representation form and depicts a different;
   b. Multiform Bivariate Small Multiple – each row uses a different representation form (with the same variable applied to all rows).
3. Bivariate Matrix – matrices extending the scatterplot metaphor:
   a. UniForm Bivariate Matrix – matrix with one representation form (e.g., all scatterplots);
   b. MultiForm Bivariate Matrix – matrix with two (or more) representation forms (e.g., maps + scatterplots, bagplots).

Our focus here is on the MultiForm Bivariate displays (2b and 3b). Implementation of each is described and illustrated below.

4.1 MultiForm Bivariate Small Multiple

The MultiForm Bivariate Small Multiple display (henceforth shortened to MultiForm Small Multiple or Small Multiple) supports analysis of multiple (potentially) related bivariate relationships through the perspectives of different bivariate representation forms. In the version currently implemented, each row uses a different representation form to display bivariate representations that relate a single common variable (assigned to rows) with any number of additional variables (one per column). Our Small Multiple is designed as a generic JavaBean component. The component provides a Java interface through which any bivariate representation form (instantiated as a JavaBean) can communicate

Figure 3 provides the simplest case, with two representation forms and two columns showing data for 156 counties making up the Pennsylvania, West Virginia, and Kentucky portion of the Appalachia Cancer Network. The common variable is per capita income and the columns contain small multiples depicting age-adjusted, cervical and breast cancer mortality. Coloring of dots on the scatterplot provides a legend through which the bivariate maps can be interpreted (the bivariate legend in map panel two is part of another component in the working system and has been superimposed here, along with the labels in the bottom panels to make the

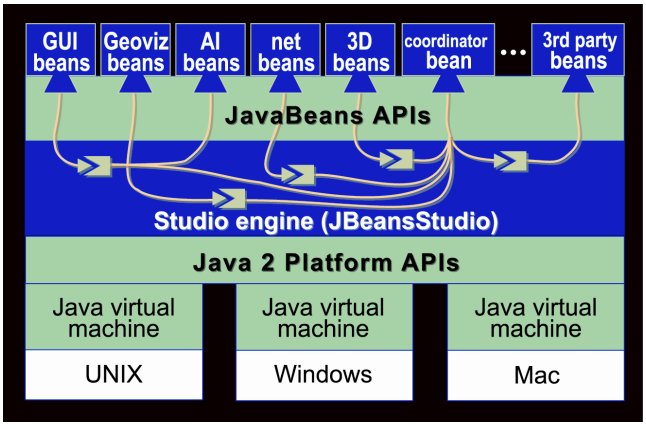

Figure 1. GeoVISTA Studio component-based architecture.

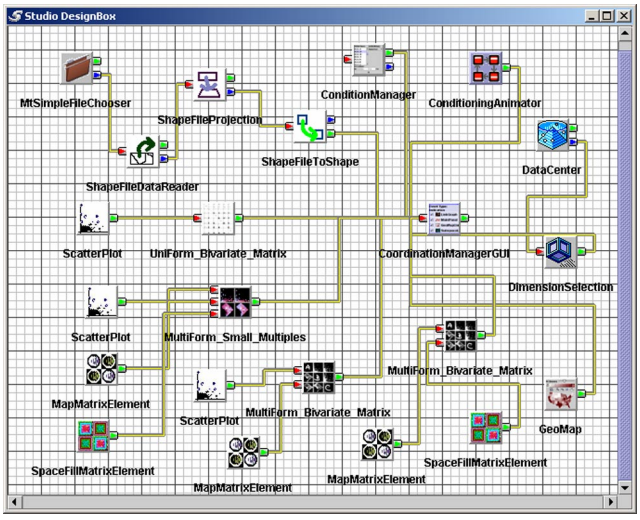

Figure 2. GeoVISTA *Studio* application design.

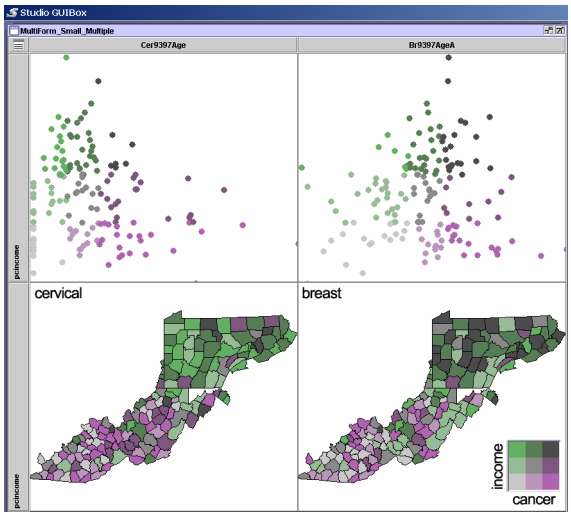

Figure 3. MultiForm Small Multiple with scatterplot and map.

print figure understandable). The color scheme applied uses sequential value (lightness) ranges for complementary hues to depict each variable, resulting in a range from light to dark grey along the diagonal (light grey = low-low, medium grey = medium-medium, and dark grey = high-high). The hue combinations used are ones developed by Brewer (1994) to be both logical and robust (e.g., they work for people with color vision deficiencies).

One interesting pattern is evident in Pennsylvania, where the bright green counties for cervical cancer mortality in the left map panel indicate relatively high income but low cancer rates while the dark grey counties in the right map panel indicate high breast cancer corresponding to high income. A hypothesis to explain the difference is that, low income women do not receive the same level of health care and this increases the chance for cervical cancer (since it is a cancer that can be prevented if early signs are detected). However, for breast cancer, the relationship is somewhat reversed because having children late in life is a risk factor and higher income women tend to delay having children.

Exploration is supported in the Small Multiples by allowing users to manually sort columns to juxtapose variable pairs of interest or to sort the entire set based on some ordered characteristic. This ability to sort by column (or row) is similar to functionality in other interactive small multiple implementations. The differences here are that: (a) the juxtaposition of different representation forms allows users to integrate perspectives (e.g., the relationship between geographic clusters and clusters in attribute space) and (b) we have implemented computational detection of interesting variable subsets along with sorting of the subsets to generate potentially meaningful default orders (see section 4.3 below).

4.2 Multiform Bivariate Matrix

As with a standard scatterplot matrix, the Multiform Bivariate Matrix (referred to henceforth as the MultiForm Matrix or simply the Matrix) allows users to select interesting variables to be displayed in juxtaposed, elemental, bivariate representations (with variables depicted in each representation defined by row-column position). Our Matrix is designed as a generic JavaBean component. As above, the Matrix is implemented as a JavaBean with a Java interface for communicating with any bivariate component. Representation forms that work with the matrix thus far include scatterplots, bivariate maps, and the spacefill display. We are in the process of adding treemaps (with size and color fill controlled by column and row variables), bagplots (bivariate boxplots), and mosaic plots as additional bivariate forms. The same or different representation forms can appear above and below the diagonal (with scatterplots in both places, the result is a standard scatterplot matrix). The central diagonal accepts any univariate representation form (also as a JavaBean).

Any number of variables can be selected for display within the Matrix; selecting five variables would produce a set of 20 bivariate displays. The selection of variables can be achieved via a pull-down menu, as well as through coordinated events generated by other components (as discussed in section 4.3). As with the Small Multiple plot, sorting of representations is supported in the Matrix (both through direct user manipulation or other system events – again, details of our computational sorting methods are discussed in section 4.3).

Figure 4 shows a Matrix containing scatterplots paired with maps; here, histograms are used for each individual variable along the diagonal. This Matrix is part of a typical application that includes a dynamically linked univariate map and conditioning controls (described in section 4.5). In each Matrix panel, users can access a detailed version of the representation form with additional controls on functionality. Changes made in the detailed view (e.g., constraining the data range depicted for a variable in a scatterplot) are propagated to other cells for which the change is relevant.

Our Matrix implementation (and the Small Multiple plot) provides flexible support for coordinating user and data events across components by a separate *coordinator component*. This component extends the traditional idea of linked brushing considerably. In addition to standard brushing (including AND and OR selection operations), users can selectively establish dynamic connections among components for the following events: Background Color, Classification, Color Array, Color Classifier, Conditioning, Data Set, Indication, Selection, Spatial Extent, Subspace.

The Matrix and linked bivariate map in figure 4 shows the relationship between four variables for counties in Pennsylvania, West Virginia, and Kentucky that are part of the Appalachia Cancer Network (in the Matrix: age-adjusted breast cancer mortality rate/100,000, percent of cancer diagnoses at local stage in the body, and percent of women age 54-65 who have had mammograms in the past 2 years; in the univariate map: number of screening mammography facilities/1000 population). In general, an increase in cancer screening leads to more early detection (thus a positive relationship with percent local stage diagnosis – the lower right scatterplot). There is, however, a group of counties with a high percent of women having had a recent mammogram but with low percent early detection. Using the tools capability to do multiple additive selections, these counties have been selected in the scatterplot and highlighted in all scatterplots and maps. The linked univariate map of screening facility accessibility shows that this pattern occurs mostly in counties where accessibility to screening is low. A possible explanation for the observed pattern is that, although women in these counties have been screened recently, lack of screening facilities may have limited their past screening, leading to detection of cancer at later stages.

4.3 Selecting and sorting variables

As discussed above, sorting is a fundamental tool of information visualization. The tools described here include components that support: (a) automatic or interactive selection of potentially interesting subspaces of variables and (b) generation of an initial viewing order for variables in the display components. While we pre-

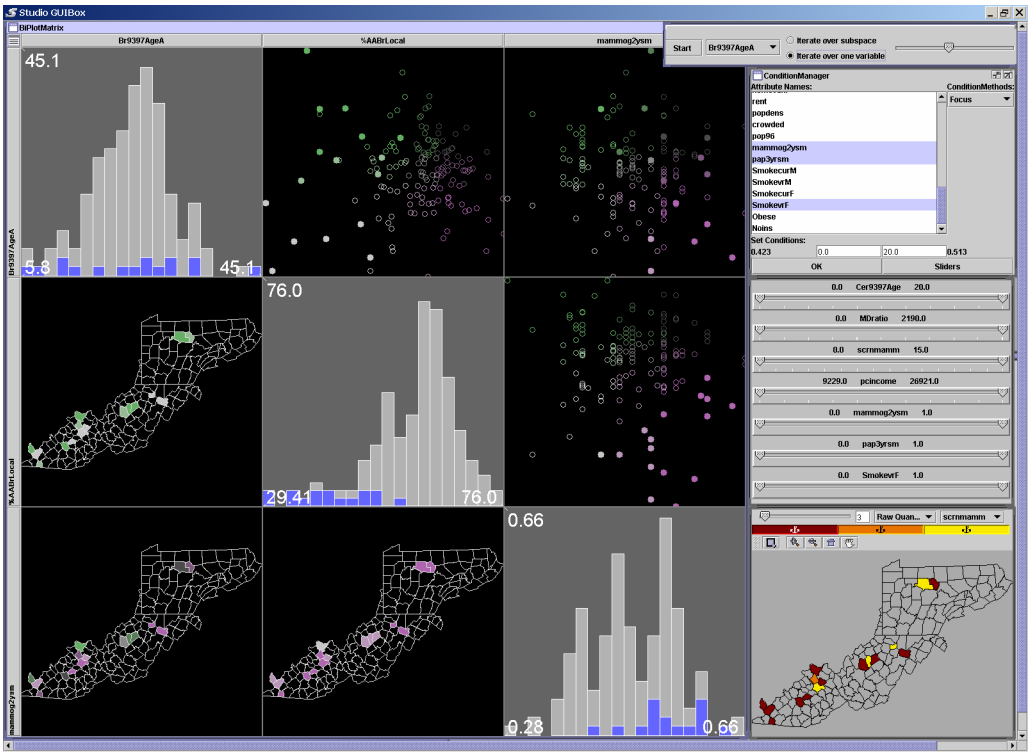

Figure 4. Typical display. MultiForm Bivariate Matrix with map and scatterplot (left), univariate map (lower right), conditioning picker and sliders generated by user selection (middle right), and conditioning animator (top right). An additive selection has been made in lower right scatterplot.

sent one method for variable selection and sorting (detailed below), our component-oriented approach supports simple substitution of alternative methods.

In contrast to Friendly (2002b), we display bivariate correlation (for reference) but sort on conditional entropy. Figure 5 shows a matrix in which correlation (independent of sign) is displayed above the diagonal and conditional entropy is displayed below (in both cases, the bright cells, thus white, represent strong relationships). Variables are ordered on their paired entropy. Given an entropy threshold, interesting subspaces can be automatically identified. The user can also pick variables to form a subspace based on her domain knowledge and interest.

The dimension selection tool (Figure 5) shows a subset from a larger set of 60 variables (several cancers and a wide range of risk factors). The tool is used to select clusters of variables from the active subset, here from a subset that includes breast and/or cervical cancer mortality rates, success in early diagnosis, and several variables indicating access to (and ability to afford) screening for cancer and subsequent treatment. Highlighted in the matrix shown is a cluster of three variables that have strong bivariate relationships (based on entropy). This cluster contains percent of age-adjusted, breast cancer diagnoses made at local (early) disease stage (%AABrLocal) along with the related risk factors of doctors/100,000 (MDratio) and per capita income. These selected variables are broadcast to other components, along with a default initial ordering for display calculated, as discussed in section 4.3.2 (see Figure 6 for the result in one component). Below, we discuss (briefly) the calculation of conditional entropy (and its advantages over other measures, e.g., correlation) as a measure of relationship in an exploratory environment and the generation of a reasonable sorting of variables overall and variables in selected subspaces.

4.3.1 How to calculate maximum conditional entropy

Let S_2 be a 2D subspace comprised of dimensions X and Y. To calculate the maximum conditional entropy of S_2, both X and Y must be discretitized into ξ intervals, which partition S_2 into a matrix of grid cells. The discretization of each dimension is a critical step, which can dramatically influence the final measure. We use a nested-means approach instead of the traditional equal-interval approach to partition each dimension. After the discretization of X and Y, let χ be the set of grid cells (including empty ones) for a column X_i ($i = 1..\xi$) in the matrix, and $d(x)$ be the density of a cell $x \in \chi$, i.e., the number of points in x divided by the total number of points in X_i. Then the entropy of this column X_i is

Figure 5. Dimension selection tool. Conditional entropy values (below diagonal) and correlation values (above diagonal).

calculated using the following equation:

$H(X_i) = - \sum_{x \in \chi} [d(x) \log d(x)] / \log|\chi|$.

Conditional entropy $(Y|X)$ is a weighted sum of $H(X_i)$, $i = 1..\xi$. Conditional entropy $(X|Y)$ can be calculated similarly using rows, Y_j ($j = 1..\xi$), instead of columns. Then the larger value of these two conditional entropies, maximum conditional entropy, is taken as the final entropy value for subspace S_2. The most prominent advantage of this conditional-entropy-based measure is that it can capture the existence of more complex patterns (clusters) than either the linear (or logarithmic, or exponential) relationships that correlation can detect or the sequential patterns measured by similarity functions presented in (Ankerst et al., 1998). A detailed introduction and evaluation of this part of work can be found in (Guo et al., 2003).

4.3.2 How to generate an ordering of dimensions

Let $S = a_1 \times a_2 \times \ldots \times a_d$ be a d-dimensional data space. Let $S_2 = \{a_i \times a_j \mid i=1..d, j = 1..d, i < j\}$ be the set of all unique 2D subspaces from S. The conditional entropy values of these 2D subspaces can form a matrix, which is a complete graph with each attribute as a vertex. Between any two attributes there is an "edge", whose length is the conditional entropy value between the attributes. To present a better display of the entropy matrix (and hence a better visualization of a high-dimensional dataset), it is important to decide an ordering of all attributes. Ankerst and colleagues (1998) view such an arrangement of dimensions as a global optimization problem and introduce a heuristic algorithm to solve it. Here we argue that, since irrelevant or noisy attributes are common in high-dimensional data, a global optimal solution may not be the best ordering to detect subspace patterns (which involve only subsets of original dimensions).

We use a hierarchical clustering approach to generate an ordering of all dimensions so that the more correlated two attributes are (in terms of a low conditional entropy value), the closer they should be in the ordering. The method adopted here first builds a minimum spanning tree (MST) from the graph and then derives an ordering of all attributes (Guo et al., 2002). The ordering preserves all hierarchical clusters and extra proximity information among attributes. The method regards a cluster (connected graph) as a chain of points and each chain has two end points. At the beginning, each cluster (or chain) contains a single point (attribute) and its two end points are the same point. When merging two clusters into one with an edge, the method will connect the closest two ends (each from a distinct chain) in the new chain.

Visualized in the order derived above, those dimensions of a multidimensional subspace with "good" clusters are likely to neighbor each other. Each cell is colored according to its entropy values—lower entropy values are assigned brighter colors. Striking bright blocks indicate subspaces with good clusters.

4.3.3 How to select subspaces

Given a threshold of conditional entropy e, a *maximum subspace* $S_{max}(e)$ satisfies two conditions: (1) the conditional entropy value of any 2-D subspace from $S_{max}(e)$ is smaller than e; and (2) adding any new dimension that is not in $S_{max}(e)$ will violate the first condition. The rationale for selecting multidimensional subspaces based on this matrix (graph) is: *if an L-dimensional ($2 < L < d$) subspace S_L has good clusters, all possible 2-D subspaces of S_L should have low conditional entropy values*. This rationale is similar to the monotonicity lemma used by CLIQUE: *if a collection of points P is a cluster in a k-dimensional space, then P is also part of a cluster in any (k-1)-dimensional projections of this space* (Agrawal et al., 1998). Based on the conditional entropy matrix, the user can also interactively form a subspace.

Once a subspace is selected (either interactively formed or automatically identified), the dimensions within this subspace (not all dimensions) will be re-ordered using the approach introduced in the previous section.

4.4 Grid-based space-filling display

As noted above, one representation "form" implemented for use in both MultiForm displays (as well as independently) is a space-filling visualization tool. It is a variant of pixel-oriented techniques (Keim & Kriegel, 1996). As noted above, pixel oriented techniques map data observations onto pixels, using color and arrangement to make patterns visible. Research has shown that these techniques can effectively present a large quantity of information to the user (Keim, 2000). One of the main advantages of this technique is the prevention of over plotting. A scatterplot with thousands of observations will obscure many of them; combining small scatterplots in a matrix exacerbates the problem.

When used in a Bivariate Small Multiple or Matrix, our *spacefill* component matches its ordering to the variable defined by the column and its color to the variable defined by the row. The component implements the following space-filling orders: spiral, scan-line, boustrophedon, Peano curves, or Morton ordering.

Unique aspects of our pixel-oriented display include:

- ability to function in a small multiple display or a matrix
- coordination with other representations
- linked exploration
- symbol representations

First, our grid-based component can be arranged in a matrix, varying the "color by" and "arrange by" variables simultaneously. Pixel-oriented displays have been organized in matrix form previously (Keim et al., 2002), but not in a bivariate matrix in which the rows and columns provide the variables to depict. *Second*, our grid-based component can appear in a matrix with an arbitrary number of other components, such as a scatterplot or map, using the same variables that are reflected in the grid-based component. This enables the user to have immediate access to a familiar representation such as a scatterplot, which may be of particular importance to users unfamiliar with grid-based (pixel-oriented) displays. Alternatively, using the map together with the grid-based component, users can rapidly compare geographic patterns with patterns shown in the grid-based display. *Third*, our grid-based component is linked to many other information visualization and geographic visualization components via our coordinator component. This allows colors, selections, and other user driven display changes to be automatically propagated and shared amongst display elements. *Fourth*, we have implemented the ability to place other symbols on top of the grid cells of the display (see video).

Figure 6 shows the three variables selected as a related subset using the entropy-based methods discussed above, arranged in the order broadcast by the component. As above, the application domain to which we apply our grid-based spacefill visualization is population data for 156 counties in the Appalachian Cancer Network. With this relatively small number of observations, we have projected the pixel-oriented display into larger spaces (cells occupy multiple pixels). Those counties having a doctor ratio of

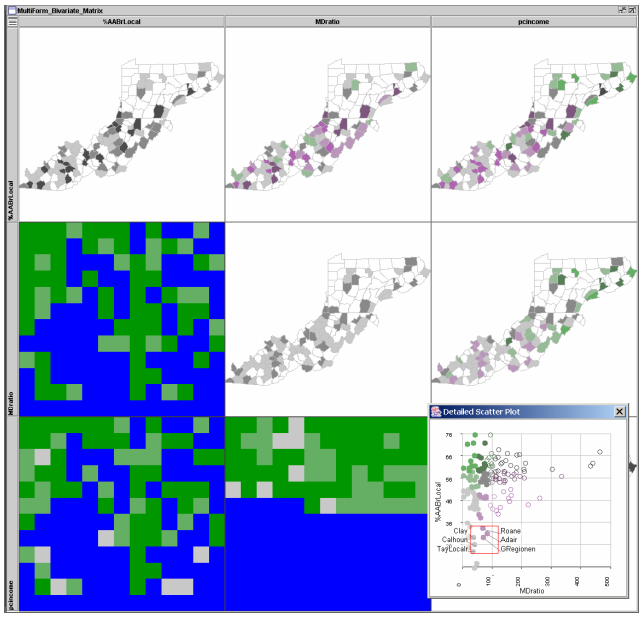

Figure 6. Matrix with bivariate map and spacefill components (univariate maps along the diagonal). Counties with a doctor ratio under about 80/1000 are highlighted in the popup scatterplot – with highlighting propagated to other views.

about 80/1000 or less have been highlighted in a popup scatterplot. The spacefill component makes clear that these counties represent about one half of the total (middle column). The maps illustrate that low doctor ratios are more prevalent in West Virginia and Kentucky than in Pennsylvania, and that Kentucky has large regions with low doctor ratios and low early detection.

The scatterplot in Figure 6 also demonstrates use of Feteke and Plaisant's (1999) excentric labeling method. We downloaded this component from www.lri.fr/~fekete/InfovisToolkit and easily linked with our own components using GeoVISTA *Studio* and our own coordinator component.

4.5 Conditioning

Our current implementation supports conditioning on any variable or variables. For example, we could condition on percent poor (removing from consideration places with a high percent poor, thus places with the greatest potential cost of health care to be a factor in cancer). This allows attention to be focused on remaining places for which some factor other than financial resources might be responsible for high rates of cancer mortality. When the user selects variables to condition on, the appropriate number of conditioning sliders is generated (see figure 4).

The conditioning tool gives the user a choice of suppressing data entities that are either outside or inside the range specified. It is also possible to activate a *conditioning animator* to explore any systematic relationships between the one well known risk factor and others. To do so, the analyst sets an initial conditioning range. The conditioning animator then steps through each data value (adding the next highest and dropping the lowest from the display). This animator component, like the data animator component discussed above, was easily added to our application design (even though components had not been built to support it). Conditioning and the conditioning animator are illustrated in the accompanying video.

5. Conclusions and future work

This paper presents an integrated perspective on a set of multivariate visual analysis methods that emphasize bivariate relationships as the starting point for discussion. This integrated perspective begins by considering links across a diverse Information Visualization, EDA, and geovisualization literature, then it addresses the integration of related methods in a set of MultiForm, computationally assisted visual tools that enable simultaneous application of multiple perspectives on a data space.

The methods and tools described are being developed within the context of an interdisciplinary project to achieve integrated visual, statistical, and computational methods for application to interpreting, testing, and validating patterns of cancer incidence and mortality in the United States. Particular attention is directed to significant cancers for which prevention and screening have been shown to be effective. As part of this process, next steps in our research include a user task analysis based on in-depth interviews with potential tool users, a set of focus group assessments of our methods and tools, and subsequent controlled experiments to test selected aspects of those tools, particularly as prompts for problem formulation and hypothesis generation.

Acknowledgement:

Research presented in this paper is supported in part by the U.S. National Science Foundation, grant #EIA-9983451 and by grant CA95949 from the National Cancer Institute.

References

Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. 1998, Automatic subspace clustering of high dimensional data for data mining applications. *ACM SIGMOD International Conference on Management of Data*, Seattle, WA USA, pp. 94-105.

Ahlberg, C. & Shneiderman, B. 1994, Visual information seeking: tight coupling of dynamic query filters with starfield displays. *Human Factors in Computing Systems (CHI '94)*, Boston, MA, April 24-28, 1994, pp. 313-317.

Ahlberg, C., Williamson, C., & Shneiderman, B. 1992, Dynamic Queries for Information Exploration: An Implementation and Evaluation. *Proceedings of CHI'92 Human Factors in Computing Systems*, Monterey, CA, May 3-7, 1992, pp. 619-626.

Ankerst, M., Berchtold, S., & Keim, D. A. 1998, Similarity clustering of dimensions for an enhanced visualization of multidimensional data. *Proceedings, Information Visualization '98*, Raleigh-Durham, NC, Oct. 19-20, 1998, pp. 52-60, 153.

Becker, R. A. & Cleveland, W. S. 1987, Brushing Scatterplots. *Technometrics, 29*, 127-142.

Bertin, J. 1981, *Graphics and Graphic Information Processing* (W. Berg, Trans.). Berlin: Walter de Gruyter.

Bertin, J. 1983, *Semiology of Graphics: Diagrams, Networks, Maps* (W. Berg, Trans. (translation from French 1967 edition) ed.). Madison, WI: University of Wisconsin Press.

Brewer, C. A. 1994, Color use guidelines for mapping and visualization. In A. M. MacEachren & D. R. F. Taylor (Eds.), *Visualization in Modern Cartography* (pp. 123-147). Oxford, UK: Pergamon.

Brunsdon, C. 2001, The Comap: exploring spatial pattern via conditional distributions. *Computers, Environment and Urban Systems, 25*, 1, 53-68.

Carr, D., Wallin, J. F., & Carr, A. 2000, Two new templates for epidemilogy applications: linked micromap plots and conditioned choroplth maps. *Statistics in Medicine, 19*, 2521-2538.

Carr, D. B., Littlefield, R. J., Nicholson, W. L., & Littlefield, J. S. 1987, Scatterplot matrix techniques for large N. *Journal of the American Stiatistical Association, 82*, 398, 424-436.

Carr, D. B., MacPherson, D., White, D., & MacEachren, A. M. submitted, Conditioned choropleth maps and hypothesis generation. *Annals of the Association of American Geographers*.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. 1983, *Graphical Methods for Data Analysis*. Monterey, CA: Wadsworth.

Cook, D., Symanzik, J., Majure, J. J., & Cressier, N. 1997, Dynamic graphics in a GIS: more examples using linked software. *Computers & Geosciences, special issue on Exploratory Cartographic Visualization, 23*, 4, 371-386.

DiBiase, D., Reeves, C., Krygier, J., MacEachren, A. M., von Weiss, M., Sloan, J., & Detweiller, M. 1994, Multivariate display of geographic data: Applications in earth system science. In A. M. MacEachren & D. R. F. Taylor (Eds.), *Visualization in Modern Cartography* (pp. 287-312). Oxford, UK: Pergamon.

Fekete, J.-D. & Plaisant, C. 1999, Excentric labeling: dynamic neighborhood labeling for data visualization. *Proceeding of the CHI 99 conference on Human factors in computing systems: the CHI is the limit*, Pittsburgh, PA, May 15 - 20, 1999, pp. 512-519.

Friendly, M. 1999, Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics, 8*, 3, 373-395.

Friendly, M. 2002a, A brief history of the mosaic display. *Journal of Computational and Graphical Statistics, 11*, 1, 89-107.

Friendly, M. 2002b, Corrgrams: Exploratory displays for correlation matrices. *The American Statistician, 19*, 316-325.

Furnas, G. & Buja, A. 1994, Prosection Views: Dimensional Inference Through Sections and Projections. Journal of Computational and Graphical Statistics and Computing, 3, 4, 323-385.

Gahegan, M., Takatsuka, M., Wheeler, M., & Hardisty, F. 2002, Introducing GeoVISTA Studio: an integrated suite of visualization and computational methods for exploration and knowledge construction in geography. *Computers, Environment and Urban Systems, 26*, 4, 267-292.

Gluck, M. 2001, Multimedia Exploratory Data Analysis for Geospatial Data Mining: The Case for Augmented Seriation. *Journal of the American Society for Information Science and technology, 52*, 8, 686-696.

Gluck, M., Yu, L., Ju, B., Jeong, W.-S., & Chang, C.-T. 1999, Augmented seriation: usability of a visual and auditory tool for geographic pattern discovery with risk perception data. *GeoComputation 99*, Mary Washington College , Fredericksburg, Virginia, USA , 25-28 July, 1999, pp. (CD, no page numbers).

Goldstein, J. & Roth, S. F. 1994, Using aggregation and dynamic queries for exploring large data sets. *Proc. of the CHI'94 (Conference on Human Factors and Computing Systems)*, Boston, MA, pp. 23-29.

Guo, D., Gahegan, M., Peuquet, D., & MacEachren, A. 2003, Breaking Down Dimensionality: An Effective Feature Selection Method for High-Dimensional Clustering. *Workshop on Clustering High Dimensional Data and its Applications, the Third SIAM International Conference on Data Mining*, San Francisco, CA, May 1-3.

Guo, D., Peuquet, D., & Gahegan, M. 2002, Opening the Black Box: Interactive Hierarchical Clustering for Multivariate Spatial Patterns. *The Tenth ACM International Symposium on Advances in Geographic Information Systems*, McLean, VA, November 8-9, 2002, pp. 131-136.

Iizuka, Y., Shiohara, H., Iizuka, T., & Isobe, S. 1998, Automatic visualization method for visual data mining, *Research and Development in Knowledge Discovery and Data Mining* (Vol. 1394, pp. 174-185). Berlin: Springer-Verlag Berlin.

Johnson, B. & Shneiderman, B. 1991, Tree-maps: A space-filling approach to the visualization of hierarchical information structures. *Proc. IEEE Visualization'91*, Piscataway, NJ, pp. 284-291.

Keim, D. & Kriegel, H.-P. 1996, Visualization techniques for mining large databases: a comparison. *IEEE Transactions on Knowledge and Data Engineering (Special Issue on Data Mining), 8* ,6, 923-938.

Keim, D. A. 2000, Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics, 6*, 1, 59-78.

Keim, D. A., Hao, M. C., Dayal, U., & Hsu, M. 2002, Pixel bar charts: a visualization technique for very large multi-attribute data sets. *Information Visualization, 1*, 1, 20-34.

Kumar, H. P., Plaisant, C., & Shneiderman, B. 1997, Browsing hierarchical data with multi-level dynamic queries and pruning. *International Journal of Human-Computer Studies, 46*, 1, 105-126.

MacEachren, A. M. 1995, How Maps Work: Representation, Visualization and Design. New York: Guilford Press.

MacEachren, A. M. 2001, An evolving cognitive-semiotic approach to geographic visualization and knowledge construction. *Information Design Journal, 10*, 1, 26-36.

MacEachren, A. M., Hardisty, F., Dai, X. P., & Pickle, L. 2003, Supporting visual analysis of federal geospatial statistics. *Communications of the ACM, 46*, 1, 59-60.

Mäkinen, E. & Siirtola, H. 2000, Reordering the Reorderable Matrix as an Algorithmic Problem. *Theory and Application of Diagrams, Diagrams 2000, Lecture Notes in Artificial Intelligence 1889*, Edinburgh, Scotland, September 2000, pp. 453-467.

Monmonier, M. 1989, Geographic brushing: Enhancing exploratory analysis of the scatterplot matrix. *Geographical Analysis, 21*, 1, 81-84.

Pinnel, L. D., Brush, A. J. B., & Borning, A. 2000, Manipulable Small Multiple Displays. *Symposium on User Interface Software and Technology (UIST) 2000*, May 2000.

Raisz, E. 1934, The Rectangular Statistical Cartogram. *Geographical Review, 24*, 292-296.

Reed, D. A., Shields, K. A., Scullin, W. H., Tavera, L. F., & Elford, C. L. 1995, Virtual Reality and Parallel Systems Performance Analysis. *Computer, 28*, 11, 57-67.

Rousseeuw, P. J., Ruts, I., & Tukey, J. W. 1999, The bagplot: A bivariate boxplot. *The American Statistician, 53*, 4, 382-387.

Schmid, C. & Hinterberger, H. 1994, Comparative Multivariate Visualization Across Conceptually Different Graphic Displays. *Proceedings of the Seventh International Working Conference on Statistical and Scientific Database Management, SSDBM 94*, Charlottesville, Virginia, September 28 - 30.

Siirtola, H. 1999, Interaction with the Reorderable Matrix. *Proceedings International Conference on Information Visualization IV '99*, July 1999, pp. 272 -277.

Stasko, J., Catrambone, R., Guzdial, M., & McDonald, K. 2000, An evaluation of space-filling information visualizations for depicting hierarchical structures. *International Journal of Human-Computer Studies, 53*, 5, 663-694.

Tufte, E. R. 1983, *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.

Tweedie, L. 1997, Characterizing interactive externalizations. *Proceedings of the SIGCHI conference on Human factors in computing systems (CHI97)*, Atlanta, Georgia, pp.375-382.

Tweedie, L., Spence, R., Dawkes, H., & Su, H. 1996, Externalising Abstract Mathematical Models. *CHI'96: Conference on Human Factors in Computing Systems*, Vancouver, BC, pp. 406-412.

Wilkinson, L. 1979, Permuting a matrix to a simple pattern. Proceedings of the Statistical and Computing section of the American Statistical Association, 409-412.

Zani, S., Riani, M., & Corbellini, A. 1998, Robust bivariate boxplots and multiple outlier detection. *Computational Statistics and Data Analysis, 28*, 257-270.