Visual Inquiry Toolkit – An Integrated Approach for Exploring and Interpreting Space-Time, Multivariate Patterns

Jin Chen[1], Alan M. MacEachren[1] , Diansheng Guo[2]

[1]GeoVISTA Center and Department of Geography
Pennsylvania State University
302 Walker Building, University Park, PA16802
Email [jxc93, maceachren]@psu.edu
Phone (814-865-1633) **Fax** (814-863-7943)

[2]Department of Geography
University of South Carolina
709 Bull Street, Columbia, SC 29208
Email [guod]@sc.edu

ABSTRACT

While many datasets carry geographic and temporal references, our ability to analyze these datasets lags behind our ability to collect them because of the challenges posed by both data complexity and scalability issues. This study develops a visual analytics approach that integrates human knowledge and judgments with visual, computational, and cartographic methods to support the application of visual analytics to relatively large spatio-temporal, multivariate datasets. Specifically, a variety of methods are employed for data clustering, pattern searching, information visualization and synthesis. By combining both human and machine strengths, this approach has a better chance to discover novel, relevant and potentially useful information that is difficult to detect by any method used in isolation. We demonstrate the effectiveness of the approach by applying the Visual Inquiry Toolkit we developed to analysis of a dataset containing geographically referenced, time-varying and multivariate data for U.S. technology industries.

1 INTRODUCTION

Exploring and analyzing large spatio-temporal multivariate datasets is challenging because of data complexity (i.e. potential interactions among space, time, and attributes) and scalability issues (i.e. both data volume and high dimension). Existing approaches, from entirely computational to visually-led methods, are limited in analyzing complex patterns that include space, time and attribute components together. Moreover, traditional information visualization methods do not support analysis of large data sets. Data mining and other computational methods have been developed explicitly to deal with large and high-dimensional data sets, but typically do not provide ways to incorporate both space and time nor do they leverage the power of human vision and cognition to help analysts notice and quickly interpret patterns in complex data. One goal for visual analytics is to bridge this gap by developing analytic methods that couple visual and computational methods in productive ways. Specifically, we introduce a Visual Inquiry Toolkit that integrates visual, computational, and cartographic methods to support an *overview+detail* strategy for visual analytics of multivariate spatio-temporal datasets of relatively large size. Our Toolkit also emphasizes flexible interaction strategies designed to enable human knowledge and judgment to be coupled productively with computational pattern finding methods. The presented research was initiated for the IEEE InfoVis 2005 contest, the focus of which was to analyze the changing characteristics of U.S. technology industries and companies over time.

The reminder of paper is organized as follows. Section 2 reviews related literature. Section 3 discusses our strategy and methodologies, with a focus on representation issues. Section 4 demonstrates an interactive visual analytics approach for mining spatio-temporal multivariate patterns. Section 5 discusses the advantages and limitations of the approach and possible further work.

2 RELATED WORK

We first discuss visualizing multivariate data. The commonly used data representations for multivariate visualization include tables and scatter plots; more sophisticated methods include scatterplot matrices (1972), parallel coordinate plots (Inselberg 1985), matrix permutation (Bertin 1981; Mäkinen and Siirtola 2000), and multivariate glyphs (Pickett, Grinstein et al. 1995). A comprehensive review of the methods can be found in a paper by Keim, et. al.(2005).

All of these methods have difficulty representing large data sets. As the number of data entities/variables (dimensions) goes up, the potential for over plotting on displays goes up as well. Two major solutions are proposed to address this problem. One is to reduce data size being displayed by grouping individual data records into subsets (e.g. aggregation or clustering), in this case, collective characteristics of the grouped data are visualized and investigated (Johansson, Treloar et al. 2004; Ward 2004; Guo, Gahegan et al. 2005). The other solution is data selection which allows zooming, filtering, and focusing on a subset or individual data (Keim, Panse et al. 2005). This research takes a combination of both approaches as discussed in section 3.1.

Visualization of spatio-temporal, multivariate data is challenging because traditional single 2D or 3D views do not provide enough dimensional space to simultaneously display all space, time, and multivariate attribute components. A widely-adopted method for space-time data is to represent these data in a three-dimensional view where time data is visualized in the third dimension over the two-dimensional map (Kwan 2000; Lodha and Verma 2000; Gatalsky, Andrienko et al. 2004; Kapler and Wright 2004). But the method has severe limitations for visualizing multivariate data with even modest size (e.g., hundreds of records for more than 2-3 variables). Some other systems use animation to display time, presenting sequential representations of spatial information at a moment of time (Stojanovic, Djordjevic-Kajan et al. 1991; Oberholzer and Hurni 2000; Slocum, Yoder et al. 2000). However, this technique imposes burdens on human short memory to retain temporal changes thus not suitable for complex, large datasets. Two approaches that show some potential to address these issues are: (1) small multiple adjacent views (MacEachren, Dai et al. 2003); (2) linked views (MacEachren, Wachowicz et al. 1999; Andrienko and Andrienko 2001; Robinson, Chen et al. 2005). We extend both approaches, combining with computational clustering methods.

Successful analysis of large, space-time-attribute datasets requires more than advances in visual representation methods. Human interaction also plays an important role in detecting and interpreting complex patterns. Considerable efforts have been made in interactively mining multivariate patterns (Seo and Shneiderman 2002; Harri 2004), temporal patterns (Carlis and Konstan 1998; Paolo, Aleks et al. 2005), and spatio-multivariate patterns (Guo, Gahegan et al. 2005). Some efforts have been made in visually mining spatio-temporal patterns, with a focus respectively on spatial distribution of temporal behaviors (Andrienko and Andrienko 2005), on events (Gatalsky, Andrienko et al. 2004), and on activity patterns (Kwan 2000). However, few approaches have been developed to interactively search for patterns using strategies that consider all aspects of space, time and multivariate attribute components.

3 VISUAL INQUIRY TOOLKIT– AN INTEGRATED APPROACH

We first discuss tasks and our strategy. Questions posed by spatio-temporal data analysis usually involve three components: *where* (space), *when* (time), and *what* (attribute/thematic objects) (Peuquet 1994). Drawing upon Peuquet's ideas, Andrienko, et al (2003) discussed three basic analysis tasks in detail as they relate to exploratory visualization: *when + where* → *what; when + what* → *where; where + what* → *when.* The tasks follows a general question scheme $A+B \rightarrow X$, where A and B denote known and X stands for unknown information. Based on Bertin's concept of *levels of reading* (elementary, intermediate, and overall), Andrienko, et al (2003) introduced two "search levels" to the analysis tasks: (1) elementary level on which a task deals with individual objects (such as a time, a place or a characteristic) ; (2) general level on which a task considers a set of object as general situations. A sample elementary level task related to our industry analysis is: What were the industry sales in the year t for the state l? A question of this style can be easily answered by an interactive data query. The more challenging questions are the general level ones like: What industrial characteristics (e.g. industry combinations) have been developed in U.S. during the past decade? What geographical areas have similar or unusual characteristics, and what are they? These questions are hard to answer through interactive queries alone because: (1) exploratory goals are initially vague and we do not know what data components (what, when, where) to query; (2) these questions are too general to be stated as formal queries. Furthermore, a system that forces users to query data iteratively and view and act on a partial result at each iteration is time-consuming, error-prone and often does not produce the desired results (Kapler and Wright 2004). Hence, analysts working with large and complex data sets need to first gain an overview of the entire dataset to quickly understand the scope of their data set and discriminate between the interesting and uninteresting content (Greene, Marchionini et al. 2000); then focus on a subset of data with more viable patterns in detail views.

Our toolkit employs the *overview + detail* strategy. Moreover, it enables human knowledge and judgment to be coupled productively with computational and visual methods for incrementally filtering and searching for novel and relevant patterns. Specifically, we employ a self-organizing map (SOM) (Kohonen 2001) to cluster multivariate data, then encode the clusters with a 2D diverging-diverging cartographic color scheme. The colored clusters are visualized in a Space-Time matrix (a recognizable, graphical tabular view). Supported by some hierarchical clustering methods, the matrix reorders layout of the rows (clusters), thus presents an overview of coarse grained patterns and exposes major explicit patterns by grouping similar entities. A parallel coordinate plot, linked to the matrix, serves as a legend for interpreting the multivariate patterns in a detail view. A matrix of small multiple geographic maps supports the perception of both spatial distribution of multivariate patterns and changes in that distribution over time. Finally, a *Pattern Basket* (a place in which an analyst can store interesting fragments of information during an extended analysis process) supports pattern synthesis. The proposed approach and tools outlined above are implemented in a proof-of-concept software application– the *Visual Inquiry Toolkit.*

reference: place + time

18 attributes: industries composition

588 records:
49 places *12 years

State	Year	AUT	...	TEL	TRN
Alabama	**1992**	0.022	...	0.02	0.01
...	**...**				
Alabama	**2003**	0.005	...	0.027	0.015
Arizona	**1992**	0.005	...	0.045	0.13
...	**...**				
Wyoming	**2003**	0.1	...	0.069	0.04

Figure 1: Aggregated data model in a tabular view. It has 588 records (49 places * 12 years) and 20 columns. The state and year columns constitute a "reference column" and 18 attribute columns represent 18 industries respectively. Hence, each record has a state-year as reference and 18 attribute values, each of which is an industry's percentage of total sales. For each record, the 18 attribute values depict the industry composition for the state in the year, and sum of the value is equal to 1.

We demonstrate our research and the *Visual Inquiry Toolkit* through an application to a benchmark dataset, provided for the IEEE InfoVis 2005 contest(Grinstein, Cvek et al. 2005) . The full dataset has around 563,000 records. The focus is on the change in geographic patterns for national industry composition over time. Hence, the data are aggregated by state and year as shown in Figure 1. The 18 industries to be analyzed are: factory automation (AUT), biotechnology (BIO), chemicals (CHE), computer hardware (COM), defense (DEF), energy (ENR), environmental (ENV), manufacturing equipment (MAN), advanced materials (MAT), medical (MED), not- primarily-high-tech (NON), pharmaceuticals (PHA), Photonics (PHO), computer software (SOF), Test & Measurement (TAM), telecommunications and internet (TEL), transportation (TRN) and subassemblies and components (SUB).

3.1 Visualization of Multivariate Patterns

The Parallel Coordinate Plot (PCP) method (Inselberg 1985) is one of the widely used techniques for visualizing multivariate data. As one component in our Visual Inquiry Toolkit, we have extended the method and implemented an enhanced, dynamic parallel coordinate plot to serve as a “legend” for multivariate patterns.

A well-known problem with PCPs is overplotting – data patterns become illegible with increasingly amount of data entities displayed. This problem has been addressed primarily from two directions (Edsall 2003; Ward 2004; Andrienko and Andrienko 2005; Guo, Gahegan et al. 2005): (1) computationally grouping the data (e.g. aggregating or clustering) to achieve an overview with fewer data groups displayed; (2) data selection (zooming, focusing, filtering) to investigate individual data record or a subset of data in a detail view. Both methods are adopted in our research.

Grouping data can be achieved by a wide array of computational clustering methods (Hastie, Tibshirani et al. 2001). Among them, the Self-Organizing Map (SOM) has proved to be effective method for multivariate clustering (Kohonen 1997; Vesanto and Alhoniemi 2000). Basically, a SOM projects a set of n-dimensional data vectors to an array of nodes on a 2D space. As a result, the entire dataset is divided into a group of non-overlapping subsets; and a cluster contains a subset of similar multivariate vectors. Johansson (2004) demonstrated an integration of a SOM with a PCP for the exploration of large multivariate data. Guo,et al (2005) also adopts the similar approach, enhancing interpretation of cluster results by applying a 2-D cartographic color scheme so that similar clusters are assigned similar colors while different clusters are assigned distinct colors. More importantly, **a color is used to uniquely encode a multivariate pattern.** This approach is adopted in this research.

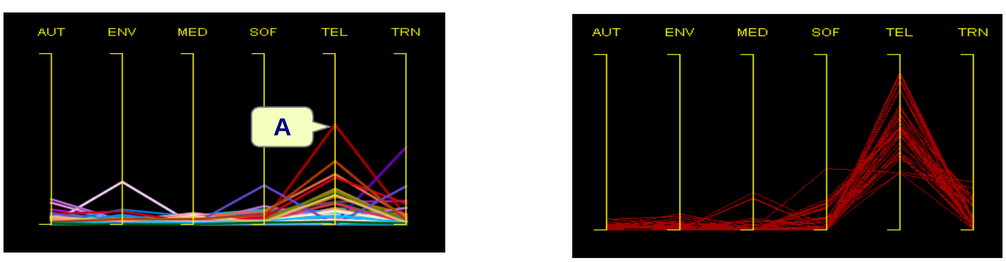

Figure 2: Six industries are displayed in the PCP, each represented by an axis. Each axis is scaled to a range from 0 to 1, a value on the axis represents an industry’s proportional contribution to the total for a place and time. We select the salient red string marked as A in the *overview*, and switch the PCP to *detail view* mode which displays the data items belong to the cluster. The outline of the strings depicts a pattern of the cluster: it has a high percentage of TEL industry, some TRN, SOF) and MED industries.

Our PCP implementation allows switching between *overview* mode (to display data groups) and *detail view* mode (to display individual data items). In the *overview* mode (Figure 2, left plot), a string represents a cluster (with it median vector values) and depicts a multivariate pattern for the cluster. In the *detail view* mode (Figure 2, right plot), a string represents an individual data item; and these strings together represent a cluster of data items. The outline of the strings depicts the pattern of the cluster. The *overview* implementation in the PCP alleviates the overplotting problem while also reducing the "noise" generated by individual variations, exposing multivariate patterns in a more legible manner.

3.2 Visualization of Space-Time and Multivariate Patterns

To visualize distribution of mutlivariate patterns across places over time, a space-time matrix is employed and coupled to the PCP to achieve a holistic view of space, time, attribute components of the data. Specifically in the matrix (Figure 4, top-left), rows represent states, columns represent years. A cell reflects via its color a coarse grained multivariate pattern in a place for a year. The pattern, uniquely encoded by a color, roughly depicts an industry composition. The detail of the pattern is visualized in the PCP. Simply put, the matrix provides an overview of patterns carrying all the space, time, and attribute components. The coupled PCP serves as a legend for illustrating the multivariate pattern in detail.

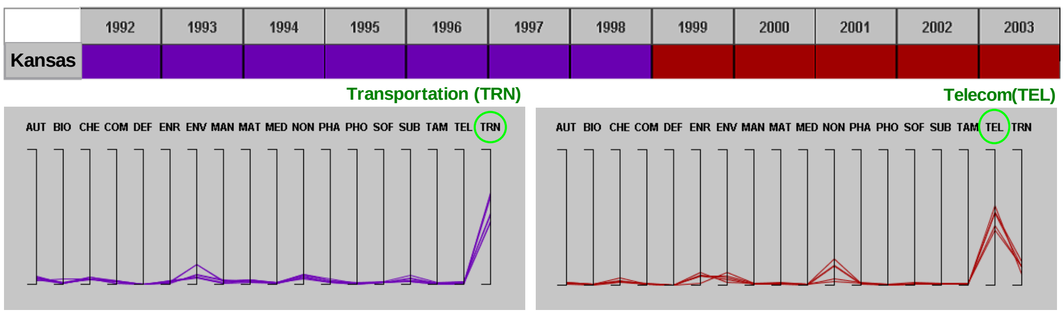

Figure 3: The matrix row representing Kansas (shown independently here) has two distinct colors (thus two patterns): purple and red. Each snapshot of the PCP (in *detail-view* mode) depicts a pattern: purple strings depict an industry mix for Kansas in 1992-98 dominated by Transportation, while red strings depict a Kansas industry mix in 1999-03 dominated by the Telecom industry. In both cases, Environmental is the most apparent secondary industry. Hence, Kansas switched focus from Transportation (purple) to Telecom (red) in 1999, while maintaining a secondary focus on Environmental through the period.

The PCP and matrix are coupled on two levels: a dynamic link and a "static" link. A dynamic link(Becker and Cleveland 1987; Buja, McDonald et al. 1991) typically means simultaneously highlighting visual elements that represent a single data record in all the views so that various aspects of the data can be investigated concurrently. "Static" link, a term introduced by Andrienko and Andrienko (2005) refers to visual connections in a static graphic having multiple views divided into a number of classes. Each class of elements is assigned unique symbol or color to so that they are known to visualize the same "thing" in the multiple views. Figure 3 illustrates the "static" link implemented in our toolkit: The group of purple strings (or red strings) is (visually) linked to the group of purple cells (or red cells) in the matrix, depicting a pattern (on industry mix) for the state and years. A string in the PCP is linked to a matrix cell in the same color. The link is visualized dynamically via highlighting operations.

In addition, a map matrix (Figure 4, top-right) is coupled to the space-time Matrix and PCP in the same way. It complements the space-time matrix by displaying the distribution of multivariate patterns on small multiple geographic map views that are ordered by years. While each individual map is useful to display patterns across space, the multiple yearly-ordered maps are particular useful to uncover how geographic patterns change across space over time (see Figure 4, D), especially when a subset of data are displayed (see detail in section 4.2).

The space-time matrix, map matrix and PCP complement each other to construct a holistic overview of complex patterns from spatial, temporal and thematic perspectives. The space-time matrix displays the salient patterns as distinct color regions (e.g. (A), (B), (C) in Figure 4). Each of the patterns is depicted in detail in the PCP. For example, the green region (A) reflects an industry composition dominated by NON industry as depicted by a green string in the PCP (in overview mode). The map matrix, as noted above, supports interpretation of spatial distribution of multivariate patterns over time. Obtaining such an initial overview of the entire dataset is critical for exploring large datasets, especially when the exploration goal is vague initially; because it can help determine what subsets of the data may potentially contains interesting patterns and thus be worth further investigation. Iteratively querying the entire dataset without such guidance can be time-consuming, error-prone and usually does not produce desired results. In the Visual Inquiry Toolkit, the linked views provide a complementary human interaction interface for exploratory analysis that allows analysts to move back and forth between overview and detail. We discuss it in section 4.2.

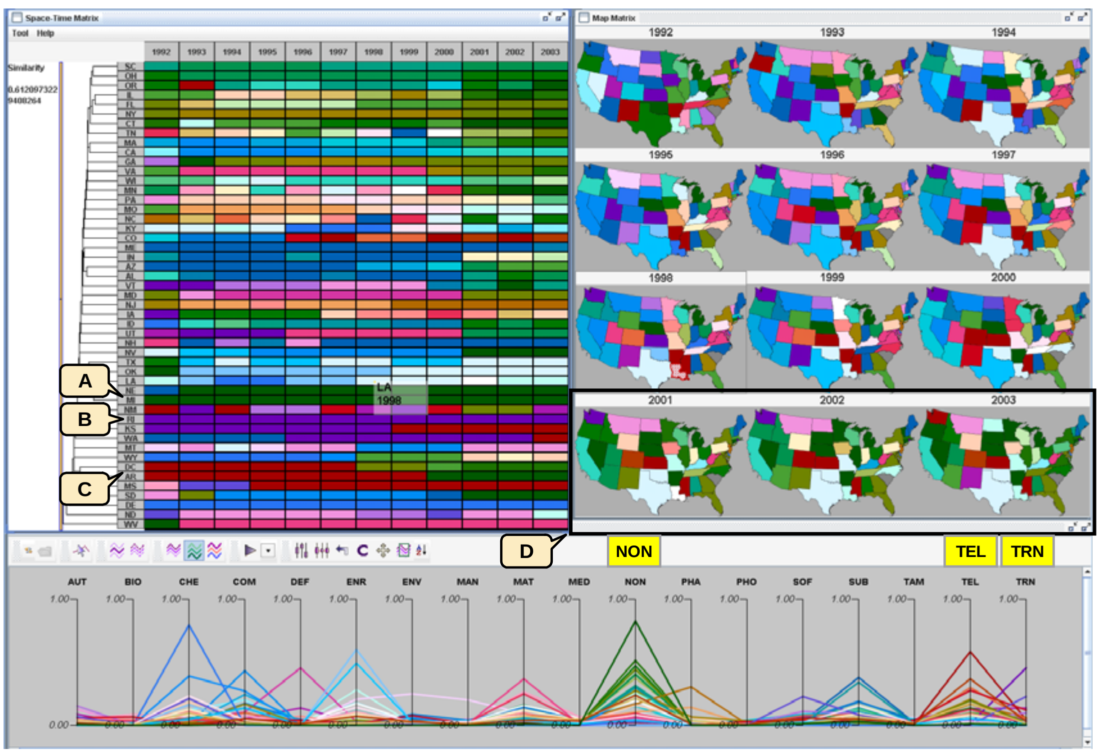

Figure 4: An overall view of spatio-temporal multivariate patterns. Here, the PCP is in *overview* mode.

(A) The green clusters statically linked to green strings in PCP, which have NON as the dominant industry. (B) Purple clusters have TRN as the dominant industry. (C) Red clusters have TEL as the dominant industry. (D) In the last row (year 2001-2003) of the map matrix, many states change to green, indicating a shift in relative importance of non-primary high tech industry.

Next, we discuss how the patterns (that appear as similarly colored regions) are computationally constructed within the space-time matrix. An agglomerative hierarchical clustering method is employed to reorder the matrix rows and to group rows (states) with similar industry compositions (represented by similar colors) together. The general approach has been found to be effective for pattern identification (Bar-Joseph, Gifford et al. 2001; Seo and Shneiderman 2002). Basically, the method includes following three basic steps: (1) initially each data item is treated as its own cluster, (2) the most "similar" pair of clusters (e.g. A and B in Figure 5) is found and merged into a new parent cluster. (3) Repeat step 1, 2 until all the clusters are agglomerated in one root cluster. Based on the definition of "similar", various algorithms have been developed (Jain and Dubes 1988). Generally, no single algorithm alone is best suited for all cases. Hence, we implement Single-Link and Complete-Link algorithms for comparison purpose and other clustering methods can be easily added.

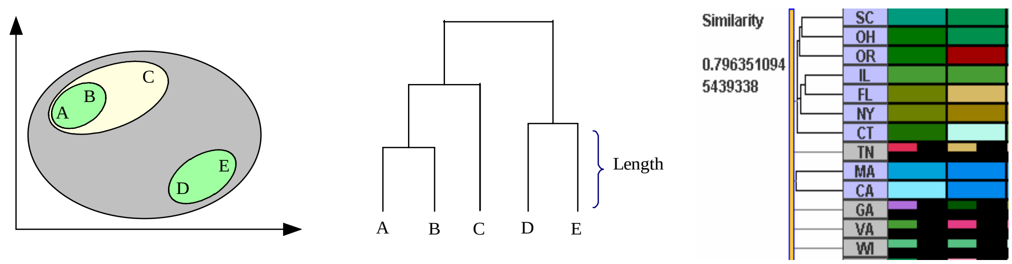

Figure 5. A dataset {A, B, C, D, E} is agglomeratively clustered and visualized in a dendrogram (middle). In this research, a dendrogram (oriented horizontally) is attached to the matrix (right), with similarity value ranges from 0 (means not similar) to 1 (means same). States with similarity value less the threshold value (currently 0.79635 as specified by the vertical bar) are filtered out (e.g. TN, GA, VA, WI).

Besides, the clustering outcome – the clusters can be conveniently visualized in a dendrogram. Basically, a dendrogram(Jain and Dubes 1988) is a binary tree in which each branch (connecting two sub-branches or leaves) is a cluster and the leaves are individual data items. The length of branches expresses a distance between two clusters/elements (e.g. the length in Figure 5means the distance between D and E). A shorter length means a shorter distance (thus more similar) between two clusters/elements. With our implementation of a dendrogram (Figure 5, right), an analyst can dynamically adjust the similarity threshold value by horizontally dragging the vertical bar across the dendrogram. Rows with similarity values less than the threshold are filtered out and appear as deselected (the cells are shrunken to a quarter of their original size). The remaining rows (in full size) are what attention should be focused on at the moment. The detail application of dendrogram is illustrated in section 4.1.

4 EXPLORATORY PATTERNS ANALYSIS

Our exploratory data analysis process follows three steps of the so-called "information-seeking mantra" (Shneiderman 1996; Keim, Panse et al. 2004): (1) Overview- examine the representation of a summary of the entire data, which presents a context from all space, time, attribute perspectives;(2) Focus and filter - select interesting patterns revealed by the holistic view or by previous processes; (3) Details on demand - focus on the patterns identified in the previous step, inspecting details from various perspectives. In addition, we add one more step: information synthesis - export novel, relevant patterns and reorganize them to yield more useful information. The concept of synthesis was introduced by DiBiase (1990) and has been discussed repeatedly as a core stage in the geovisualization process by MacEachren and colleagues (1994; 1997; 2004), but little progress has been made toward tools that support synthesis. The four steps form a cyclic process to incrementally search, identify and analyze patterns, and eventually synthesize useful information out of the patterns for knowledge construction and decision making.

4.1 Pattern detecting, filtering and synthesizing

Initially, the space-time matrix displays an overview of patterns in which matrix rows are ordered by a computational method alone (Figure 4). Patterns seen in this overview are not fully satisfying because all kinds of patterns – relevant/irrelevant, explicit/implicit ones - are put together on the matrix view. The mixture of patterns can distract human attention and hinder an analyst in identifying the most important patterns as well as any hidden ones. Hence, we need to go through step 2 (as mentioned at beginning of section 3) -*Focus and filter* -to find the most interesting and relevant patterns. A key tool for achieving this is the interactive dendrogram that is attached to the space-time matrix.

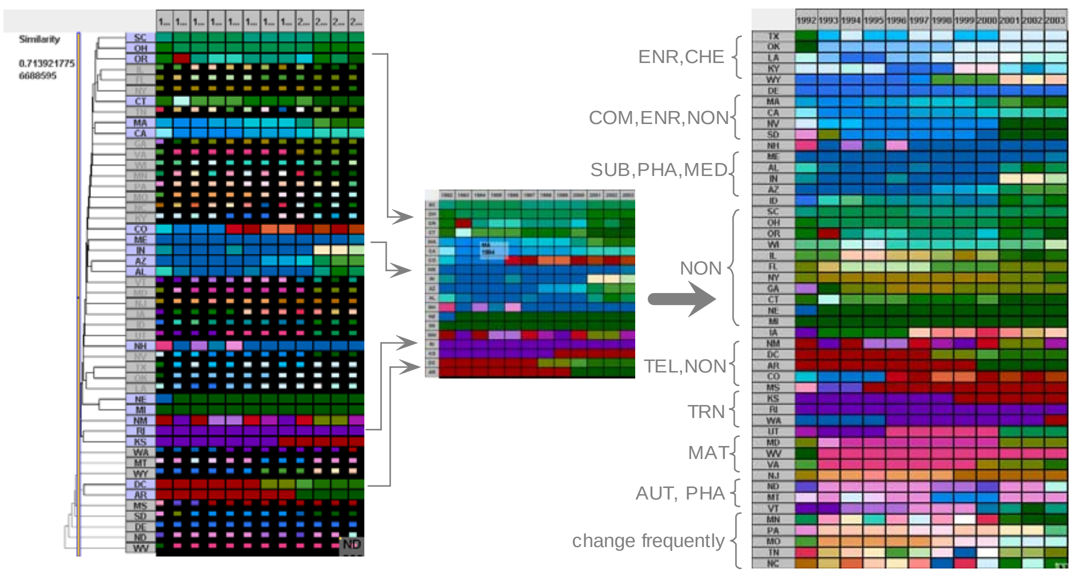

Figure 6: Left - pattern filtering: novel, relevant and important patterns are picked and export to a Pattern Basket (middle) through which interpretations can be captured and later retrieved. Well known, irrelevant and unimportant rows can be filtered out. Right – pattern synthesis: rows of Pattern basket are manually adjusted and some implicit patterns are exposed (e.g. TEL region shows DC, AR switched away from TEL while CO, MS, KS switched to TEL in recent years).

Analysts can gradually adjust the threshold value (by dragging the bar) and select the most interesting rows, which usually are ones having high similarity values. However, high similarity values do not necessarily mean being relevant and potentially useful, hence an analyst must investigate the detail patterns in the PCP and apply domain knowledge to determine what patterns are novel, relevant and important. Rows that carry irrelevant, known or unimportant patterns can be manually disabled (thus filtered out). The rows with interesting patterns can be exported. By "export", we mean copy the selected rows from the space-time matrix to the Pattern Basket (Figure 6 middle) so that the patterns can be saved for further analysis.

A Pattern Basket is a variation of the space-time matrix. The motivation for introducing the component is: human working memory is limited and unable hold all the patterns found during the exploratory analysis process; the Pattern Basket enables the analyst to expanding the memory capacity by offloading found patterns to external memory – the basket. Moreover, row order in the pattern baskets can also be manually adjusted based on domain knowledge to achieve a better view of patterns (Figure 6 middle) and to expose implicit patterns. For example, the TEL region on the matrix (right) shows how some states switched industry focus to/away from telecommunications over time. The space-time matrix, PCP, and map matrix, when combined with Pattern baskets, provides a platform for a spiral and incremental processes of pattern searching, filtering, analyzing and synthesizing. Eventually with more rows exported, we achieve an overview with patterns thematically organized (Figure 6, right).

4.2 Interactive pattern examining

Having found some interesting patterns via step 2 - filtering, we can go to step 3 – examine the patterns in detail. We demonstrate the process via an analysis case.

We select the blue region (Figure 6, in the middle of the matrix). It contains four states - Maine (ME), Indiana (IN), Arizona (AZ), Alabama (AL). The strings in the PCP approximately depict an industry composition pattern that is dominated by SUB industry (Figure 7 left). This is a *where+when*→ *what* process in which *where* and *when* are specified in the space-time matrix and *what* is found in PCP. We do an incremental brushing along the SUB axis. It is a *what* → *where+when* process and two more states - NH (New Hampshire) and WA (Washington) – are found to focus on the SUB industry. The six rows are exported into Pattern Basket and manually reordered for better interpretation (Figure 7 right).

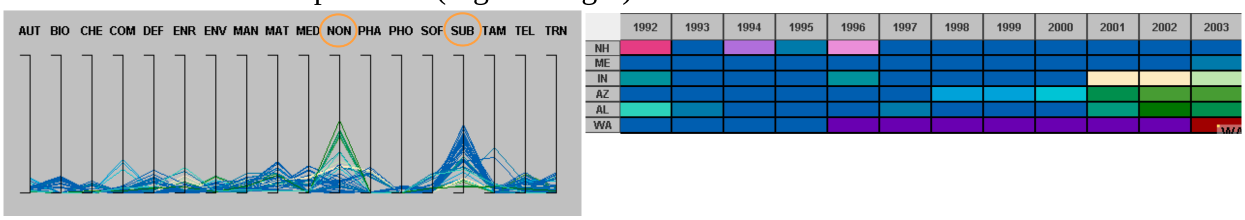

Figure 7. Left - PCP (in *detail view* mode) shows industry mix for four states: ME, IN, AZ, AL. All axes scaled from 0 as the min value to 1 as the max value – representing the proportion of all technology industry in that place and time represented by the specific industry. The industry mix is dominated by SUB industry with considerable amount of NON industry. Right - Six states are found to focus/ had focused on an industry mix dominated by SUB industry, which is represented in blue.

In the map matrix (Figure 8, top plot), we notice that most of the states were in dark blues in early years; more recently, from 1998 to 2000, Arizona changed to bright blue and some dramatic changes in color occurred from 2001 to 2000. To determine what this means, we make four drill-down selections according to the four major color themes in the map matrix. The outcomes are displayed in all three views: space-time matrix, PCP and map matrix. Because of the space limitation of this paper, we take four PCP screenshots for the four selection operations, and they "share" one map matrix screenshot. Our complementary linked views effectively visualize how multivariate patterns (i.e. A, B, C in Figure 8) change across spaces (i.e. the six states) over time (for details, see figure 8). In addition, the views allow interactive exploring from all space, time, and attribute perspectives; and compound selection operations can be made to achieve any combination of *where*, *when* and *what* for a query schema of either *A+B*→ *C* or *A*→ *B+C*. The queries can be made either on a group of entities to address general level questions (explained in section 3) or on an individual entity to address elementary level questions. For example as shown in Figure 7, a selection on the space-time matrix makes a query of "*where+when* → *what*", with *what* displayed in the PCP. Conversely, a brushing on the PCP makes a query of *what*→ *where* + *when*, with the outcome displayed in the matrix. In summary, being coupled together, the space-time matrix, map matrix and PCP, with their underlying computational clustering and sorting methods, complement each other to support the *overview* → *detail* strategy for an exploratory data analysis process.

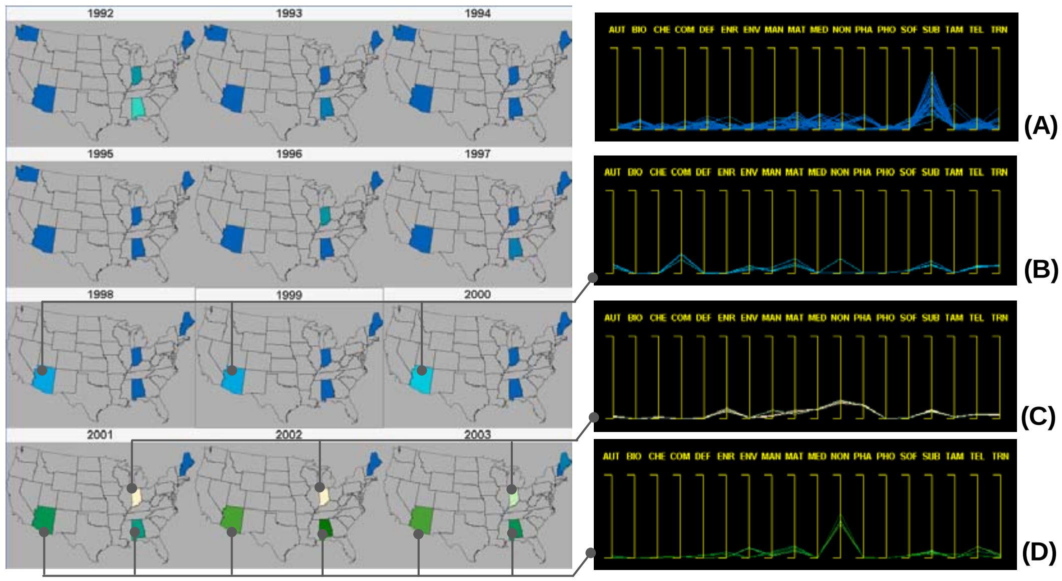

Figure 8. A map matrix screen shot is "shared" by four screen shots (A, B, C, D) which display the outcome of four selecting operations. (A) Almost all six states except NH focused on SUB in the early years; (B) From 1998 to 2000, AZ switched focus to COM, NON, MAT, but still kept some SUB. (C) From 2001 to 2003, IN switched focus to NON and PHA, while still kept some SUB. (D) From 2001 to 2003, AZ and AL switched focus to NON. We conclude SUB expanded from ME to NH while it "died" in the other states after 2001.

5 CONCLUSION

The proposed exploratory analysis approach for spatio-temporal multivariate data provides two major advantages: (1) by integrating visual, computational, and cartographic methods, we achieve a holistic overview of the entire data set as well as detailed views on a particular aspect/a subset of data. The views effectively support our *overview+detail* exploratory analysis strategy; (2) by productively coupling human knowledge with computational, our toolkit supports incremental pattern searching, filtering, and synthesizing processes, thus has a better chance to find novel, relevant and potentially useful information. One of the future efforts of this research will be to address the scalability issue. Currently, our methods work well for modest numbers of places and times (data aggregated to states and to years). Additional insights are likely if methods can be extended to deal effectively with multiple levels of drill down – for the current case study to monthly data by zip code. In general, the goal is to support visual analytics activities that move smoothly across scales of analysis as suggested by patterns and relationships that are uncovered.

ACKNOWLEDGEMENT

This study was supported and monitored by the Advanced Research and Development Activity (ARDA) and the Department of Defense. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the National Geospatial- Intelligence Agency or the U.S. Government. The research is also supported, in part, by grant CA95949 from the National Cancer Institute and by the Penn State Regional Visualization and Analytics Center, which is part of a National Consortium led by the National Visualization and Analytics Center at the Pacific Northwest National Laboratory.

REFERENCES

Andrews, D. F. (1972). "Plots of High-Dimensional Data." Biometrics **29**: 125-136.

Andrienko, G. and N. Andrienko (2001). Constructing Parallel Coordinates Plot for Problem Solving. 1st International Symposium on Smart Graphics, Hawthorne, New York, USA.

Andrienko, G. and N. Andrienko (2005). "Blending Aggregation and Selection: Adapting Parallel Coordinates for the Visualization of Large Datasets." The Cartographic Journal **42**(1): 49-60.

Andrienko, G. and N. Andrienko (2005). Visual exploration of the spatial distribution of temporal behaviors.

Andrienko, N. and G. Andrienko (2005). Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach, Springer.

Andrienko, N., G. Andrienko, et al. (2003). "Exploratory spatio-temporal visualization: an analytical review." Journal of Visual Languages & Computing **14**: 503-541.

Bar-Joseph, Z., D. K. Gifford, et al. (2001). "Fast optimal leaf ordering for hierarchical clustering." Bioinformatics **17**(Suppl. 1): S22-S29.

Becker, R. A. and W. S. Cleveland (1987). Brushing scatterplots, American Society for Quality Control and American Statistical Association.

Bertin, J. (1981). Graphics and Graphic InformationProcessing. translated by W. J. Berg and P. Scott.

Buja, A., J. A. McDonald, et al. (1991). Interactive data visualization using focusing and linking. Visualization, 1991. Visualization '91, Proceedings., IEEE Conference.

Carlis, J. V. and J. A. Konstan (1998). Interactive visualization of serial periodic data. Proceedings of the 11th annual ACM symposium on User interface software and technology. San Francisco, California, United States, ACM Press.

DiBiase, D. (1990). "Visualization in the Earth Sciences." Earth and Mineral Sciences, Bulletin of the College of Earth and Mineral Sciences, Penn State University **59**(2): 13-18.

Edsall, R. M. (2003). "The parallel coordinate plot in action: design and use for geographic visualization." Computational Statistics & Data Analysis **43**(4): 605-619.

Gatalsky, P., N. Andrienko, et al. (2004). Interactive analysis of event data using space-time cube.

Gatalsky, P., N. Andrienko, et al. (2004). Interactive Analysis of Event Data Using Space-Time Cube. Eighth International Conference on Information Visualisation (IV'04), London, England.

Greene, S., G. Marchionini, et al. (2000). "Previews and Overviews in Digital Libraries: Designing Surrogates to Support Visual Information-Seeking " Journal of the American Society for Information Science **51**(3): 380-393.

Grinstein, G., U. Cvek, et al. (2005). "IEEE InfoVis 2005 Contest, Technology Data in the US." Retrieved April 14, 2005, from http://ivpr.cs.uml.edu/infovis05.

Guo, D., M. Gahegan, et al. (2005). "Multivariate Analysis and Geovisualization with an Integrated Geographic Knowledge Discovery Approach." Cartography and Geographic Information Science **32**(2): 113-132.

Harri, S. (2004). Interactive Cluster Analysis. Proceedings of the Information Visualisation, Eighth International Conference on (IV'04) - Volume 00, IEEE Computer Society.

Hastie, T., R. Tibshirani, et al. (2001). The Elements of Statistical Learning, Springer.

Inselberg, A. (1985). " The plane with parallel coordinates." The Visual Computer **1**: 69-97.

Jain, A. K. and R. C. Dubes (1988). Algorithms for clustering data. Englewood Cliffs, NJ, Prentice Hall.

Johansson, J., R. Treloar, et al. (2004). "Integration of Unsupervised Clustering, Interaction and Parallel Coordinates for the Exploration of Large Multivariate Data." Proceedings of the Eighth International Conference on Information Visualisation (IV'04).

Kapler, T. and W. Wright (2004). GeoTime Information Visualization.

Keim, D. A., C. Panse, et al. (2005). Information Visualization: Scope, Techniques and Opportunities for Geovisualization. Exploring Geovisualization. J. Dykes, A. M. MacEachren and M.-J. Kraak, Amsterdam: Elsevier**:** 23-52.

Keim, D. A., C. Panse, et al. (2004). "Visual data mining in large geospatial point sets." Computer Graphics and Applications, IEEE **24**(5): 36-44.

Kohonen, T. (1997). Self-organizing maps, Springer.

Kohonen, T. (2001). Self-organizing maps, Berlin ; New York : Springer.

Kwan, M. P. (2000). "Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set." Transportation Research Part C-Emerging Technologies **8**(1-6): 185-203.

Lodha, S. K. and A. K. Verma (2000). Spatio-Temporal Visualization of Urban Crimes on a GIS Grid. Proceedings of the 8th ACM Symposium on GIS, Washington D.C., ACM Press.

MacEacchren, A. M. (1994). Visualization in modern cartography: Setting the Agenda. Visualization in Modern Cartography. A. M. MacEachren and D. R. F. Taylor. , Oxford, UK, Pergamon 1-12.

MacEachren, A., X. Dai, et al. (2003). Exploring high-D spaces with multiform matrices and small multiples.

MacEachren, A. M., M. Gahegan, et al. (2004). "Geovisualization for knowledge construction and decision support." Ieee Computer Graphics and Applications **24**(1): 13-17.

MacEachren, A. M. and M.-J. Kraak (1997). "Exploratory cartographic visualization: advancing the agenda." Comput. Geosci. **23**(4): 335-343.

MacEachren, A. M., M. Wachowicz, et al. (1999). "Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in database methods." International Journal of Geographical Information Science **13**(4): 311-334.

Mäkinen, E. and H. Siirtola (2000). Reordering the Reorderable Matrix as an Algorithmic Problem. Theory and Application of Diagrams, Diagrams 2000, Lecture Notes in Artificial Intelligence 1889, Edinburgh, Scotland, Springer-Verlag.

Oberholzer, C. and L. Hurni (2000). "Visualization of change in the interactive multimedia atlas of Switzerland." Computers & Geosciences: 423-435.

Paolo, B., A. Aleks, et al. (2005). Interactive pattern search in time series, SPIE.

Pickett, R. M., G. Grinstein, et al. (1995). Harnessing preattentive perceptual processes in visualization. Perceptual Issues in Visualization. G. Grinstein and H. Levkowitz. New York, Springer**:** 33-45.

Robinson, A. C., J. Chen, et al. (2005). "Combining Usability Techniques to Design Geovisualization Tools for Epidemiology." Cartography and Geographic Information Science **32**(4): 243-255.

Seo, J. and B. Shneiderman (2002). "Interactively exploring hierarchical clustering results." Computer **35**(7): 80-+.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations.
Slocum, T., S. Yoder, et al. (2000). "MapTime: software for exploring spatiotemporal data associated with point locations." Cartographica 15-31.
Stojanovic, D., S. Djordjevic-Kajan, et al. (1991). Cartographic visualization and animation of the dynamic geographic processes and phenomena. Proceedings of 19th International Cartographic Conference, Ottawa, Canada,.
Vesanto, J. and E. Alhoniemi (2000). "Clustering of the self-organizing map." Neural Networks, IEEE Transactions on **11**(3): 586-600.
Ward, M. O. (2004). "Finding Needles in Large-Scale Multivariate Data Haystacks." Computer Graphics and Applications **24**(5): 16-19.