TimeSeer: Scagnostics for High-Dimensional Time Series

Tuan Nhon Dang, Anushka Anand, and Leland Wilkinson

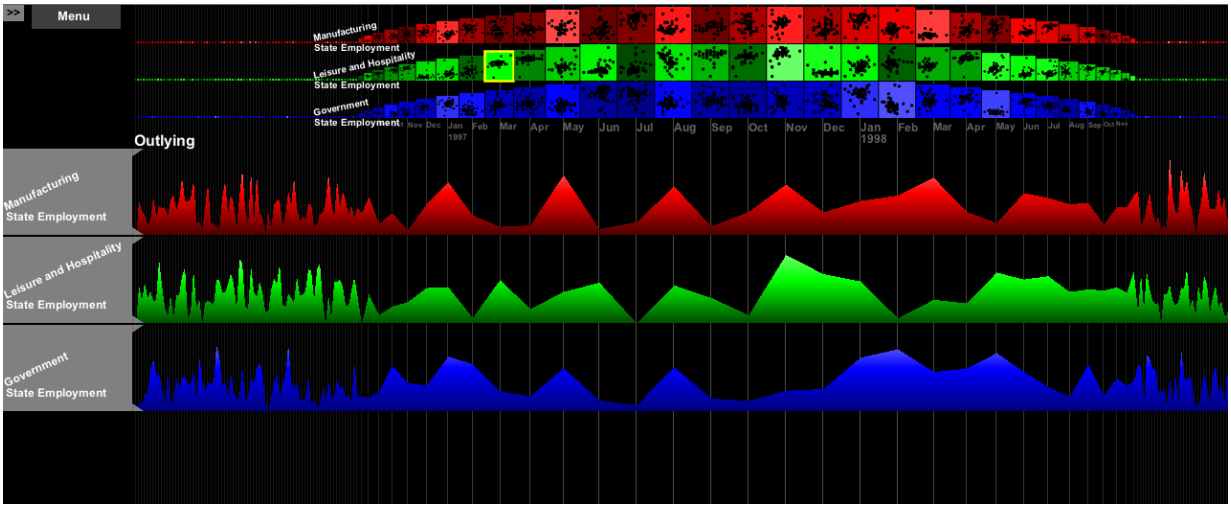


Fig. 1. Visualization of US Employment data highlighting the Outlying Scagnostic feature.

Abstract—We introduce a method (Scagnostic time series) and an application (TimeSeer) for organizing multivariate time series and for guiding interactive exploration through high-dimensional data. The method is based on nine characterizations of the 2D distributions of orthogonal pairwise projections on a set of points in multidimensional Euclidean space. These characterizations include measures, such as, density, skewness, shape, outliers, and texture. Working directly with these Scagnostic measures, we can locate anomalous or interesting sub-series for further analysis. Our application is designed to handle the types of doubly-multivariate data series that are often found in security, financial, social, and other sectors.

Index Terms—Scagnostics, Scatterplot matrix, High-Dimensional Visual Analytics, Multiple Time Series.

1 Introduction

Suppose we have data consisting of many time series over many variables with many time points. Suppose, further, that we want to identify unusual events at individual time points across all the series. If there is only one time series for each variable, the usual analytic approach to this problem is to perform spectral modeling of cross-covariance functions among the series. A visual analytic analog of this approach is to plot pairs of series and highlight noteworthy features in the pairs that appear to be substantially related. We might see, for example, one series showing average fines for auto speeding each week in a US state and another series showing weekly auto accidents in that state. If we see a rise in auto speeding fines preceding by several weeks a reduction in accidents, we might conclude (with appropriate *ceteris-paribus* qualification) that a rise in fines may lead to a reduction in accidents.

This traditional approach will not work if there is more than one time series for each variable (more than one state in our example). An alternative, of course, would be to examine all singletons or pairs of individual series for patterns. This alternative does not scale. Assume we have t time points, p variables, and n series which describe a doubly-multivariate data series. In this case, we have for each pair of variables at each time point a 2D scatterplot of n points. In our US economy example, we would have for a single year's data 52 time points (each week in a year), 10 variables (one for each economic sector), and 50 series (one for each state). In this rather small example (from our perspective), we have 2340 scatterplots with 50 points each to examine. We will consider a much larger example in this paper.

If we can solve the multitude-of-scatterplots problem, we gain some important visual-analytic insights. By looking at scatterplots instead of individual series, we can analyze patterns that would not be evident in other types of multivariate time series visualizations. Suppose, for instance, that we have n individuals (including some potential terrorists) interacting at t time points through p different channels (web sites, text messages, cell conversations, ...). Suppose, also, that we are interested in recognizing time points where one or more subgroups of these individuals begin to cluster together in a communication clique. With the right tools, we should be able to identify scatterplots where these clusters are apparent. Furthermore, we should be able to examine with these tools other features, such as outliers or correlations, that characterize conspiratorial interaction.

An important aspect of our proposal is that it is firmly grounded in time series methodology. We are not simply looking for unusual scatterplots in a large collection of scatterplots. We are looking at time series of *aspects* of scatterplots. For example, we can investigate a time series of Clumpiness (cliques) or Outliers (rogues) or Monotonicity (conspirators). We will see in our examples below that these series have coherent behavior in real data that becomes apparent and revealing when viewed with the right tools. It is noted that multiple views, interactions and analytical components are particularly useful in analyzing time-oriented data [1]. Our system, TimeSeer, not only looks at a data abstraction through aspects of the raw data but also provides multiple views together with filtering, searching and focusing interactions.

This work is a natural extension of our work on *Scagnostics* [48], an idea that allows us to characterize the "shape" of scatterplots. In developing a working platform, however, we discovered that we had to design custom tools to deal with the challenges posed by massive

- *Tuan Nhon Dang is with Department of Computer Science, University of Illinois at Chicago, E-mail: tdang@cs.uic.edu.*
- *Anushka Anand is with Department of Computer Science, University of Illinois at Chicago, E-mail: aanand2@lac.uic.edu.*
- *Leland Wilkinson is with SYSTAT Inc. and Department of Computer Science, University of Illinois at Chicago, E-mail: leland.wilkinson@systat.com.*

time series datasets. We found some clues in related work, but most of what we have developed is, in our understanding, new.

Our contributions in this paper are:

- We devise a framework for applying Scagnostics in the context of time-varying data analysis. This induces data reduction which allows for fast identification of interesting features such as outliers in high-dimensional data sets.
- We propose a dissimilarity measure for scatterplots based on their Scagnostics.
- We design an interactive system for visually mining doubly-multivariate data series using multiple visual metaphors in a novel combination.

The paper is structured as follows: We describe related work in the following section. We describe an overview of our interactive system, TimeSeer, in Section 3. Section 4 illustrates TimeSeer on real datasets. Finally, Section 5 draws conclusions and indicates future developments.

2 Related Work

In reviewing related work, we must keep in mind that some approaches that seem visually similar to ours are fundamentally different and some approaches that seem visually quite different nevertheless provided important guidelines for our own development. We begin with our work on Scagnostics.

2.1 Scagnostics

In the mid 1980s, John and Paul Tukey proposed an exploratory graphical method called *Scagnostics*. The Tukeys intended to characterize a collection of 2D scatterplots through a small number of measures of the pattern of points in these plots. These measures included the area of 2D isolevel kernel density contours, the perimeter length of these contours, a nonlinearity measure of association based on principal curves [23], and other statistics. By using these measures, the Tukeys aimed to detect anomalies in density, shape, association, and other features.

We described Scagnostics in a plenary session at the 2003 InfoVis conference. Seo and Shneiderman followed our general description by using ordinary parametric statistics (mean, standard deviation, correlation coefficient, etc.) instead of the kinds of nonparametric measures proposed by the Tukeys [27]. Consequently, we decided to implement the original Tukey idea through nine Scagnostics defined on planar proximity graphs. We gave these measures ordinary names (Outlying, Skewed, Clumpy, Sparse, Striated, Convex, Skinny, Stringy, Monotonic) and presented a scalable program for computing these new graph-theoretic measures [47]. Following this work, Fu [16] extended Scagnostics to 3D and still others used analogs of the word to describe feature-based descriptions for parallel coordinates and pixel displays[14, 39].

Although the original motivation for Scagnostics was to locate interesting scatterplots in a large scatterplot matrix, we soon realized the idea had more general implications. We have argued [48] that Scagnostics should be regarded as a type of projection that enables us to examine features in Scagnostics space and then make inferences about patterns that would not be apparent in the raw data space. In other words, Scagnostics space can serve as a basis for visual analytics much as the complex plane does for spectral analytics, although the Scagnostics projection is not invertible. Our time series platform rests on this fundamental principle.

We now outline the Scagnostic algorithm.

2.1.1 Binning

We begin by normalizing the data to the unit interval and then use a 40 by 40 hexagonal grid [9] to aggregate the points in each scatterplot. If there are more than 250 nonempty cells, we reduce the bin size by half and rebin. We rebin until there are no more than 250 nonempty cells. The choice of bin size is constrained by efficiency (too many bins slow down calculations of the geometric graphs) and sensitivity (too few bins obscure features in the scatterplots).

We compute all our measures on the binned points using the counts in each bin as weights. The Scagnostics measures depend on proximity graphs that are all subsets of the Delaunay triangulation: the convex hull, the minimum spanning tree (MST), and the alpha complex [15].

2.1.2 Deleting Outliers

Before computing the Scagnostics, we delete outliers to improve robustness. We consider an outlier to be a vertex whose adjacent edges in the MST all have a weight (length) greater than F_{inner+}, where

$$F_{inner+} = q_{75} + 1.5(q_{75} - q_{25}) \tag{1}$$

where q_{75} is the 75th percentile of the MST edge lengths and the expression in the parentheses is the *interquartile range* of the edge lengths.

2.1.3 Computing Scagnostic Measures

We now present the Scagnostic measures computed on our three geometric graphs. In the formulas below, we use H for the convex hull, A for the alpha hull, and T for the minimum spanning tree. We are interested in assessing three aspects of scattered points: *density*, *shape*, and *association*.

Density Measures

The following measures detect different aspects of point densities.

- **Outlying**

The Outlying Scagnostic measures the proportion of the total edge length of the minimum spanning tree accounted for by the total length of edges adjacent to outlying points (as defined above). We do this calculation before deleting outliers for the other measures.

$$c_{outlying} = length(T_{outliers})/length(T) \tag{2}$$

- **Skewed**

We use two other density measures based on MST edge-lengths. The first is a relatively robust measure of skewness in the distribution of edge lengths of the MST.

$$q_{skew} = (q_{90} - q_{50})/(q_{90} - q_{10}) \tag{3}$$

- **Sparse**

The second edge-length statistic, Sparse, measures whether points in a 2D scatterplot are confined to a lattice or a small number of locations on the plane. This can happen, for example, when tuples are produced by the product of categorical variables. It can also happen when the number of points is extremely small. We choose the 90th percentile of the distribution of edge lengths in the MST. This is the same value we use for the α statistic.

$$c_{sparse} = q_{90} \tag{4}$$

- **Clumpy**

An extremely skewed distribution of MST edge lengths does not necessarily indicate clustering of points. For this, we turn to another measure based on the MST: the RUNT statistic [22]. The runt size of a dendrogram node is the smaller of the number of leaves of each of the two subtrees joined at that node. Since there is an isomorphism between a single-linkage dendrogram and the MST [19], we can associate a runt size (r_j) with each edge (e_j) in the MST, as described by [40]. The RUNT graph (R_j) corresponding to each edge is the smaller of the two subsets of edges that are still connected to each of the two

vertices in e_j after deleting edges in the MST with lengths less than $length(e_j)$.

The RUNT-based measure responds to clusters with small maximum intra-cluster distance relative to the length of their nearest-neighbor inter-cluster distance. In the formula below, j runs over all edges in T and k runs over all edges in R_j.

$$c_{clumpy} = \max_j \left[1 - \max_k \left[length(e_k) \right] / length(e_j) \right] \quad (5)$$

- **Striated**

We define coherence in a set of points as the presence of relatively smooth paths in the minimum spanning tree. Smooth algebraic functions, time series, and curves (*e.g.*, spirals) fit this definition. So do points arranged in flows or vector fields. Another common example is the pattern of parallel lines of points produced by the product of categorical and continuous variables.

We use a measure based on the number of adjacent edges in the MST whose cosine is less than -0.75. Let $V^{(2)} \subseteq V$ be the set of all vertices of degree 2 in V and let $I()$ be an indicator function. Then

$$c_{striate} = \frac{1}{|V|} \sum_{v \in V^{(2)}} I(\cos \theta_{e(v,a)e(v,b)} < -.75) \quad (6)$$

SHAPE MEASURES

The shape of a set of scattered points is our next consideration. We want to detect if a set of scattered points on the plane appears to be connected, convex, and so forth. Of course, scattered points are by definition *not* these things, so we need additional machinery (based on geometric graphs) to allow us to make such inferences. In particular, we will measure aspects of the convex hull, the alpha hull, and the minimum spanning tree.

- **Convex**

Our convexity measure is based on the ratio of the area of the alpha hull and the area of the convex hull. This ratio will be 1 if the nonconvex hull and the convex hull have identical areas.

$$c_{convex} = [area(A)/area(H)] \quad (7)$$

- **Skinny**

The ratio of perimeter to area of a polygon measures, roughly, how skinny it is. We use a corrected and normalized ratio so that a circle yields a value of 0, a square yields 0.12 and a skinny polygon yields a value near one.

$$c_{skinny} = 1 - \sqrt{4\pi area(A)}/perimeter(A) \quad (8)$$

- **Stringy**

A stringy shape is a skinny shape with no branches. We count vertices of degree 2 in the minimum spanning tree and compare them to the overall number of vertices minus the number of single-degree vertices.

$$c_{stringy} = \frac{|V^{(2)}|}{|V| - |V^{(1)}|} \quad (9)$$

We cube the Stringy measure to adjust for negative skew in its conditional distribution on n.

ASSOCIATION MEASURE

We are interested in a symmetric and relatively robust measure of association.

- **Monotonic**

We use the squared Spearman correlation coefficient to assess monotonicity in a scatterplot. We square the coefficient to accentuate the large values and to remove the distinction between negative and positive coefficients. We assume investigators are most interested in strong relationships, whether negative or positive.

$$c_{monotonic} = r^2{}_{spearman} \quad (10)$$

This is the only coefficient not based on a subset of the Delaunay graph.

2.2 Time Series Visualization

Visualizing time series has a long history in statistics and geography [30, 17, 20, 6, 11]. Many of the best ideas from centuries-old hand-drawn graphics have been incorporated in modern computer visualizations [4, 43, 44]. Noteworthy recent examples are Spiral Graph [46] and Time Searcher [26].

2.2.1 Visualizing Multivariate Time Series

Some have developed viewers for multivariate time series. Theme River [24] was one of the first; it employed kernel smooths of time series, stacking them in a single display. Theme River can be quite effective for displaying up to 20 time series simultaneously, but it is not as useful for displaying raw series. Theme River trades detail for overall impact. Because it smooths and stacks, the absolute levels of the series are difficult to discern. Cleveland has discussed in more detail the problems involved in stacking time series [12] (see also [7]). Wattenberg [45] developed an applet called Name Voyager that presents interactive stacked graphs of raw series with exploratory widgets that allow the manipulation and visualization of multiple series in a single stacked display. With his tool, it is easy to drill-down to an individual series to investigate details. Name Voyager continues to be one of the more popular visualization sites on the Web, perhaps because it is so engaging and easy to use. Other recent multivariate time series viewers include [3, 41, 25, 34, 28, 5].

2.2.2 Time Series Pattern Search

Long time series cannot be visualized on ordinary or mega-pixel displays. There aren't enough pixels to represent each time point in these series. The problem is especially acute in live feeds or streaming data sources because the feeds are effectively infinite [36]. A common remedy is to pan and zoom into "interesting" segments of the series with lensing or other widgets. How do we identify "interesting" segments, however? One popular method is to search for motifs or anomalous patterns in time series using statistical and data mining algorithms [33, 8, 10, 35, 29]. Time Searcher [26] contains widgets that could be effective when paired with these algorithms. Superimposing similar sub-series can facilitate within-series comparisons.

2.2.3 Aggregation

One way to deal with multivariate series is to aggregate across similar series. If series are already categorized (within states, countries, economic sectors, hospital patients, etc.), then averaging the series is a possibility. Otherwise, one must use cluster analysis to identify clusters of similar series [37, 42, 21]. Aggregation risks concealment of important features, however. One must be sensitive to outlying series and other anomalies that can bias the aggregation.

3 TIMESEER

TimeSeer is a platform for visualizing Scagnostic time series. As we indicated before, our model is fundamentally different from other time series visualizations. It is based on the recognition that synchronized multivariate time series have multivariate point distributions at each time point. Our data model is a multivariate generalization of the series models employed in the papers we have reviewed. We have t time points and p variables, resulting in p-multivariate time series. For each variable, however, we have n series, resulting in a doubly-multivariate

distribution. We have found no visual analytic platform capable of handling this model.

Typical data for this model are: t months, p economic indicators, and n countries; t minutes, p vital signs, and n patients; t trading days, p stock indices, and n markets (exchanges); t seconds, p network protocols, and n nodes. A significant challenge for visual analytics on data like these is scalability. We normally expect t, p, and n to be large. It is not uncommon to find the product of these parameters to be in the tens of thousands. Visualizing them with conventional tools is out of the question.

We will illustrate the features of TimeSeer mainly through examples. In this section, however, we will describe the overall architecture of the system. As we have indicated, our solution to the overall problem is to regard simultaneous time points in this multivariate system as collections of point sets. Characterizing those point sets will allow us to discern patterns that we could not see with conventional time series analytics or statistics. Our system leverages juxtaposition and explicit encoding of relationships in data (through the Scagnostics) as visual comparison strategies [18].

The most obvious benefit of our parameterization is to reduce n to 1. That is, if we can characterize a scatterplot with a single measure (monotonicity, clumpiness, etc.), then we can use methods designed for ordinary multivariate time series. The tradeoff here is, of course, that we might lose detail at a given time point. Our remedy for that tradeoff is to devise a display that incorporates a pixel-scale scatterplot at each time point. An additional way we ameliorate this problem is to provide selection tools to switch easily between different Scagnostics. Our display changes almost instantly when a different Scagnostic is selected for analysis. This feature allows an analyst to focus on a particular aspect of scatterplots without excluding other possibilities.

We confine our model at this point to 2D scatterplots. There is nothing preventing us from computing most Scagnostics in higher dimensions, but display issues come into play as the dimensionality increases. We believe that analysts are more familiar with 2D scatterplots than with more exotic displays, but that is a belief that requires testing in the future.

3.1 The TimeSeer GUI

The TimeSeer GUI incorporates two major systems: Variable Selection SPLOM and Time Series Viewer. The first enables the analyst to select Scagnostics and then variables. We employ a scatterplot matrix and a novel lensing tool to navigate through the matrix and select cells. The selections made in the SPLOM system direct the visualization in the Time Series Viewer. Figure 5 shows instances of this tool. The top panel shows an implementation of Table Lens [38], in which a row/column is enlarged and the remaining rows/columns are reduced. The lower two panels show our implementation, which involves a smooth lens so that distant rows/columns are reduced proportionally.

Figure 2 depicts the basic difference between Table Lens and our lensing method. In Table Lens, the Degree Of Interest (DOI) of rows/columns outside the lensing area is uniformly small. In our lensing method, we implement a smooth transition in DOI proportional to the distance of a frame to the lensing frame. This lensing technique is similar to Cartesian Fisheye View [31].

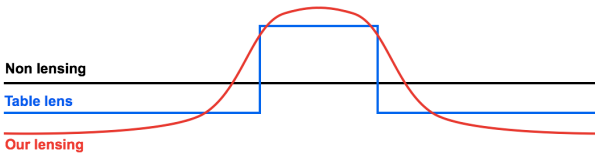


Fig. 2. DOI function maps from cell address to interest level.

The following algorithm shows how we extend Table Lens to achieve smooth lensing. For simplicity, we will explain the algorithm in one dimension, say X-axis, and one side of lensing area, say on the right of lensing area. The two-dimensional smooth lensing can be achieved by applying the same algorithm for both sides and both dimensions.

1. As with Table Lens, we increase the width of the lensed column W_{max}. Let k be the number of columns to the right of the lensed column, the widths of these columns are reduced to $W_1 = W_2 = ... = W_k = W_{small}$.
2. Now we compute a lensing factor $s = (W_{small} - W_{min})/((k-1)/2)$ where W_{small} is the smallest width (of the farthest column) that we want to lens.
3. The width of columns are recomputed by $W_i = W_i + s((k+1)/2 - i)$ for $i = 1,...,k$

Figure 3 shows an example of our lensing method on X-axis where $W_{max} = 98$, $W_{small} = 34$, $W_{min} = 10$, $k = 7$, and $s = 8$. All widths are in pixels.

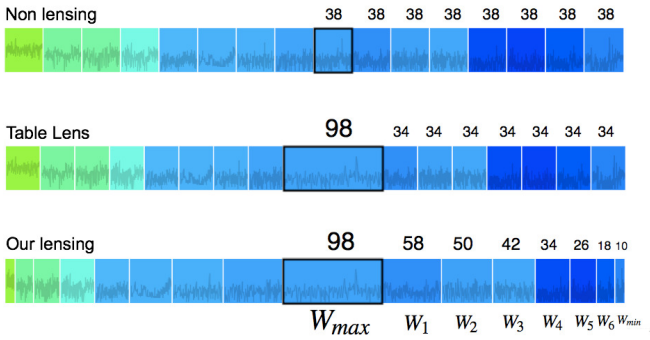


Fig. 3. Horizontal lensing.

After we have selected pairs in the SPLOM system, we go to the Time Series Viewer where time series in selected cells are expanded in a full window as depicted in Figure 7. This main window contains multiple Scagnostic series at the bottom and scatterplots at each timepoint. A variety of pan-and-zoom tools facilitate smooth navigation throughout this window. As depicted in Figure 7(b) and Figure 7(c), the same smooth lensing technique is applied on the X-axis.

Simple buttons offer alternate views and filtering. The Time Series Viewer is discussed in greater details in Section 4.3.

We could elaborate on details in this section, but it is easier to understand the workings of this rather large application by looking at real data examples. We will explain technical details as we cover various capabilities of the program.

4 Examples

In this section, we use two different datasets to demonstrate the performance of TimeSeer. The first is a series of US Employment data and the second is a series of US Weather data. The US Employment data comprise monthly employment statistics for 50 states over 22 years from 1990 to 2011. The data were retrieved from `http://www.bls.gov/`. There are 25 variables in the collected data: Total Nonfarm, Construction, Manufacturing, Non-Durable Goods, Trade and Transportation, Wholesale Trade, Retail Trade, Transportation and Utilities, Financial Activities, Real Estate and Leasing, Professional and Business, Scientific and Technical, Administrative and Support, Education and Health, Educational Services, Social Assistance, Leisure and Hospitality, Arts and Entertainment, Accommodation and Food, Other Services, Government, Federal Government, State Government, Local Government, and State Employment. For these data, we have 78,600 scatterplots with 50 data points each to examine.

The Weather data comprise hourly meteorological measurements over a year from the Gulf of Maine in 2008. There are 17 variables represented in the dataset: current speed, current direction, temperature, East Current Velocity, North Current Velocity , significant

wave height, dominant wave period, air temperature, wind speed, wind gust, wind direction, visibility, barometric pressure, water temperature, salinity, sigma-T, and conductivity. Data and variable descriptions can be found at `http://gyre.umeoce.maine.edu/buoyhome.php`. For these data, we have 50,000 scatterplots with 24 data points (24 hours in a day) each to examine.

We begin with the US Employment data.

4.1 Variable Selection SPLOM

With p variables, there are $p(p-1)/2$ pairs of variables. To do our analysis, we need to: 1) select the Scagnostic of interest: Outlying, Monotonic, Stringy, Skinny, Sparse, Striated, Convex, Clumpy, Skewed, 2) select a criterion to order variables in SPLOM: mean or variance of the Scagnostic series, and 3) select a subset of the scatterplots, either by picking individual frames in the scatterplot matrix or by picking all frames corresponding to a single variable.

The mean and variance of a Scagnostic time series (a pair of variables) is computed by averaging that Scagnostic measure over time series as shown in Equation 11 and Equation 12 where T is the number of data points in time series, p and q are two variables.

$$Mean(p,q) = \frac{\sum_{i=1}^{T} X_i}{T} \tag{11}$$

$$Variance(p,q) = \frac{\sum_{i=1}^{T} (X_i - Mean(p,q))^2}{T} \tag{12}$$

The Scagnostic mean and variance of a variable p is computed by averaging all pairs of variables containing p as an element as shown in Equation 13 and Equation 14 where V is the number of variables. The mean or variance of variables (depending on which one we have selected) is used to order variables in SPLOM.

$$Mean(p) = \frac{\sum_{q=1,q\neq p}^{V} Mean(p,q)}{V} \tag{13}$$

$$Variance(p) = \frac{\sum_{q=1,q\neq p}^{V} Variance(p,q)}{V} \tag{14}$$

We offer both mean and variance for ordering Scagnostics series because each captures a different aspect of the Scagnostic process that might interest an analyst. The mean selection offers the opportunity to pick series with extremely high or low series means on a Scagnostic. The variance selection ranks by variability, so that single peaks and valleys in the Scagnostic time series will be more discernable in the main time series window.

Figure 4(a) shows the scatterplot matrix for 25 variables in the US Employment data. We have selected the Outlying measure and sorted the variables by their means. In particular, each plot (each pair of variables) is colored by its mean of the selected Scagnostic time series; the embedded small graph shows a thumbnail of the actual Scagnostic time series. On the top of Figure 4(a) is the color legend for the mean of Outlying Scagnostic time series. We use a Kelvin color temperature scale [32] to encode the range of all possible Outlying mean values with red corresponding to high values of means and green corresponding to low values of means. This range (always within the 0 and 1 interval) is different when we select a different Scagnostic feature. TimeSeer sorts the variables so that low Outlying series are at the bottom and high Outlying series are at the top. Notice that we also color variable names to differentiate and group them by categories and subcategories.

Single *plot* selection is depicted in Figure 4(a). This mode is invoked by clicking on any of the panes in the scatterplot matrix. This selection mode allows the analyst to investigate specific Scagnostic series that show interesting patterns of behavior among the two featured variables. Single *variable* selection mode is depicted in Figure 4(b) This mode is invoked by clicking on the angled variable names to the right of the scatterplot matrix diagonal. The figure shows Total Nonfarm selected. Black rectangles are used to denote selected plots. This selection mode allows the analyst to examine all variables paired with a specific variable of interest.

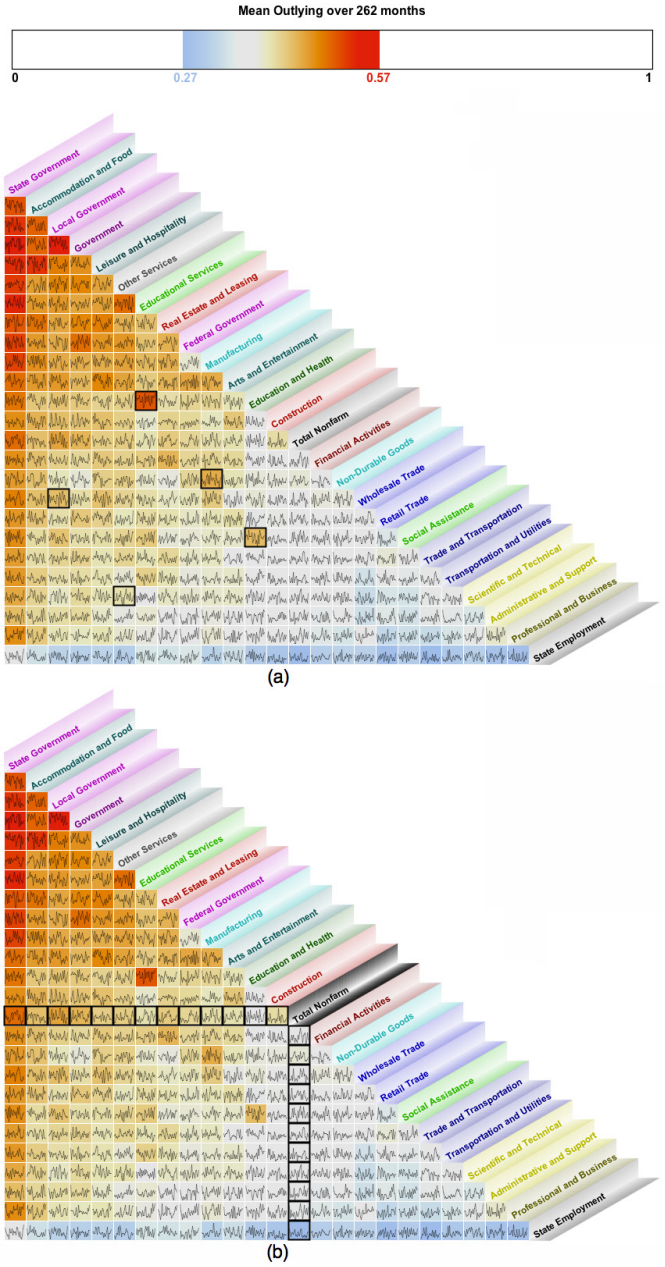


Fig. 4. Plot selection in scatterplot matrix of 25 sorted variables in the US Employment data.

4.2 Lensing

For many more variables, the Scagnostic time series will be difficult to discern inside the scatterplot matrix. Consequently, we added zooming and lensing to the matrix. If one hovers over a plot the graph inside is enlarged. There are two types of lensing. Figure 5(a) shows an implementation of Table Lens; Figure 5(b) and Figure 5(c) show smooth lensing on the top and in the corner. Unlike the standard implementation of Table Lens, our smooth lensing offers a smooth transition when we move the mouse over different plots. Therefore, it is easier to keep track of the whole context (SPLOM) and the focused area.

After a user has selected variables from the scatterplot, all pairwise combinations are displayed in the controller window, as depicted in Figure 6 (8 pairs of variables are selected). Nine other Scagnostic measures on each pair of variables are presented on the same row as the selected pair but they are faded. Each plot is colored by the mean of that feature over a year. Additionally, all pairs are reordered by

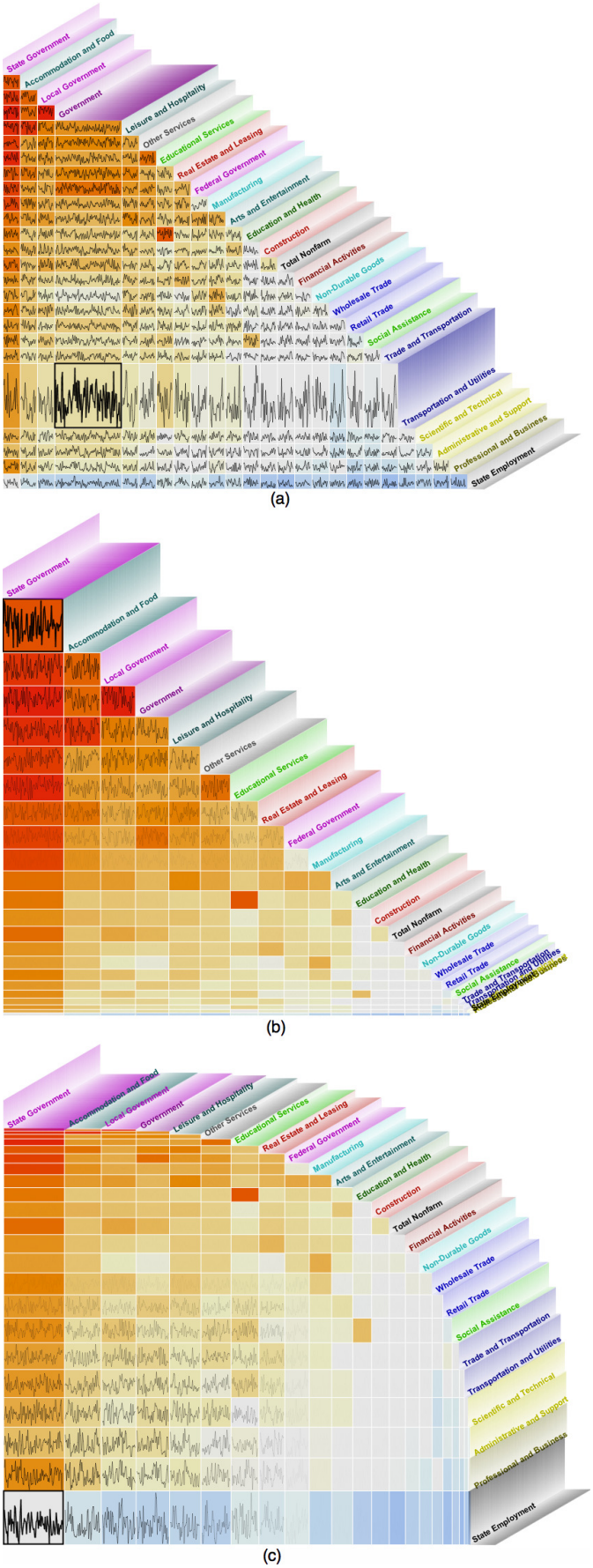


Fig. 5. Lensing in scatterplot matrix of 25 sorted variables in the US Employment data.

the selected criterion: the mean of the selected Outlying feature. In Figure 6, the third pair is brushed, and other features of the third pair are highlighted. We can keep track the position of the brushing pair in the scatterplot matrix by highlighting the graph inside this selection display. Both views are linked. Moreover, variable names are colored in the same as they appear in SPLOM. This helps in locating variables in both views.

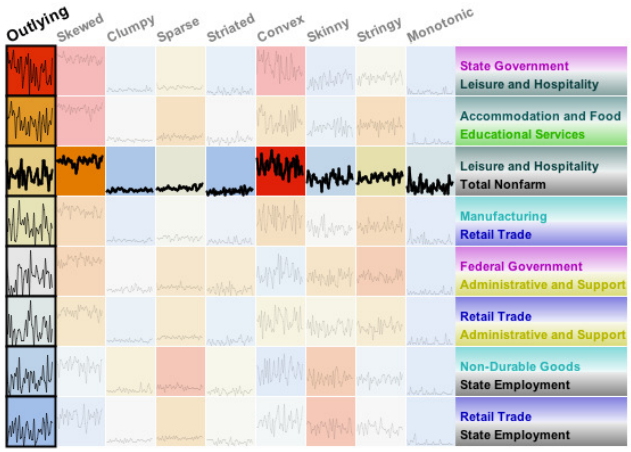


Fig. 6. Pairwise variable selecion: 8 pairs of varibables.

4.3 Time Series Viewer

After we have selected pairs, we go to the Time Series Viewer. There are several ways to visualize multiple time series: small multiples or multiples superimposed with or without lensing. Figure 7 is an example. The series are built from ones we selected in the controller based on the US Employment data. We selected Monotonic as the Scagnostic for this example, and we chose 9 pairs of variables sorted by their means. Notice the slanted orientation of the second variable in each pair. This device helps the viewer to understand which variable is on the X axis and which variable is on the Y axis.

Figure 7(a) shows 9 small multiples corresponding to 9 pairs. The lensing in the Y-dimension is applied to the first three series, colored red, green, and blue respectively; the other pairs are colored gray and greatly reduced in size. We also employ a gradient on the lensed series to make the profiles more discernible and to coordinate highlighting with the scatterplots at the top of the window. The larger the series value, the lighter the coloring (like snow on mountains). This use of brightness also facilitates the highlighting of the scatterplots at the top. Each scatterplot corresponds to the appropriate colored Scagnostic series directly below it, and is highlighted with the same brightness as the point on the series.

We can change the number of pairs in the lensing area or replace them by other pairs in the non-lensing area by a simple click and drag. Moreover, we can resize the Y-dimension of a Scagnostic time series (both lensing and non lensing area) by a simple mouse scroll. This helps to accommodate different numbers of Scagnostic time series into the fixed height of application window, as users can always go back to the SPLOM to select different pairs of variables.

For the arrangement we have selected, we notice that the Monotonic Scagnostic shows a distinct seasonal pattern with an annual cycle. This is consistent with what we would expect for variables related to farming.

In the US employment data, there are 262 data points on each Scagnostic time series. As we can see in Figure 7(a), however, the data point are crowded enough so that we cannot read any details in a season. The common solution in this case is displaying a selected season or interval. One limitation of this method is that when we select a season or interval, we lost the overall context of the time series. As a remedy, we chose to implement smooth lensing for the X-dimension. When we lens a season, we still can see what is going on in the other

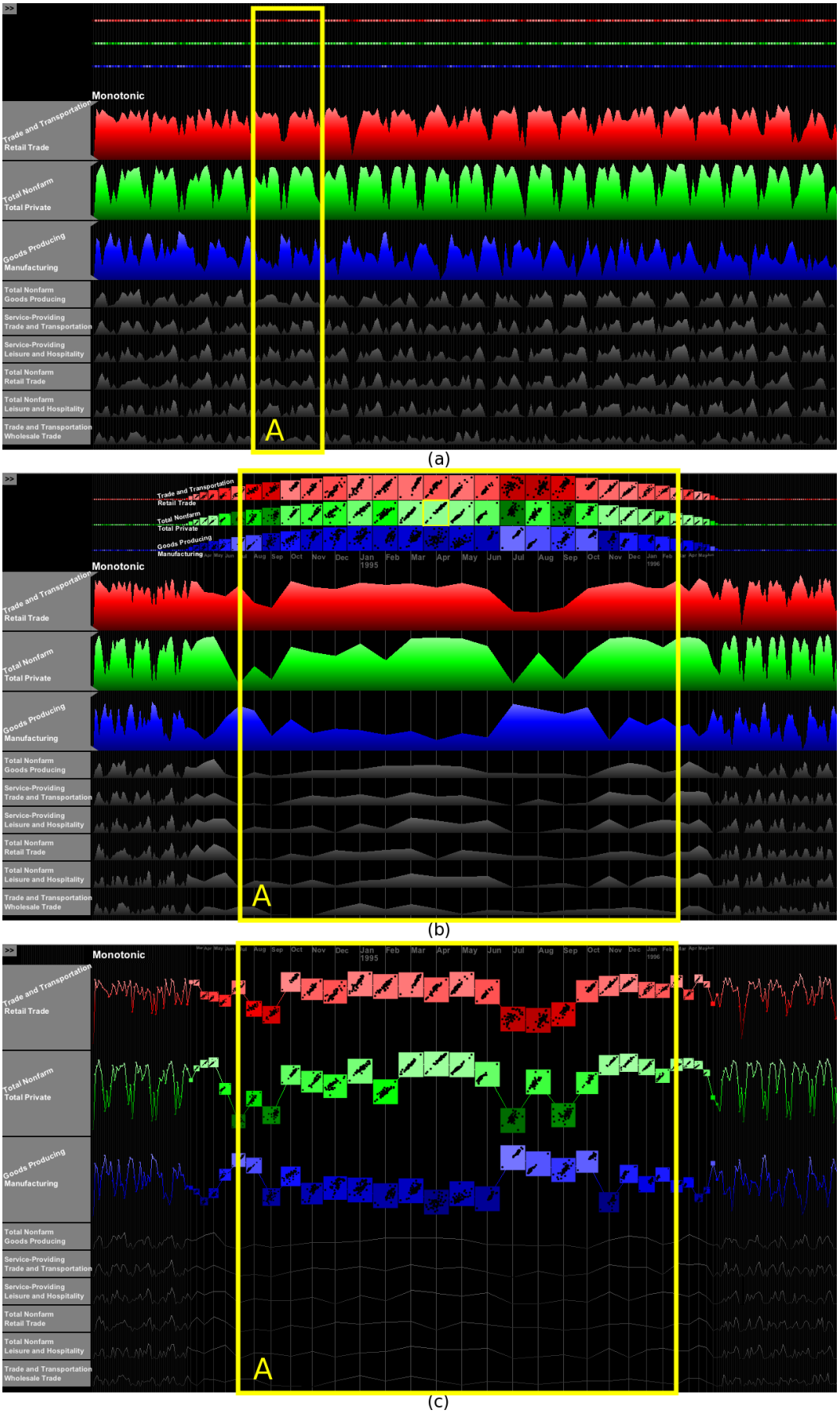


Fig. 7. Visualization of US Employment data: a) Small multiples Overview with lensing on Y-dimension b) Small multiples with lensing on both X and Y-dimension c) Line graphs superimposed by plots with lensing on both X and Y-dimension

seasons throughout the entire time series. We can simply move the mouse to enlarge a different season. Moreover, smooth lensing allows continuous transition as we move the mouse.

Figure 7(b) shows a lens applied to the Scagnostic series. Vertical lensing is applied to three first pairs of variables and horizontal lensing are applied to two seasons (highlighted in Box A). The lensing works over the series as well as the scatterplots, so we are able to investigate individual scatterplots to see the configuration of points that led to the value of the Scagnostic shown in the series.

Figure 7(c) shows an alternate view of the same series. This view superimposes scatterplots on line graphs of the Scagnostics series. Such an arrangement allows investigation of individual scatterplots without anchoring or reference to a row of scatterplots elsewhere in the window. We believe this layout is useful once interesting segments are found in the series. In any case, toggling between views can be done in an instant.

4.4 Filtering, Brushing, and Drill-down

Information visualization systems should allow one to perform analysis tasks that largely capture people's activities while employing information visualization tools for understanding data [2]. In the rest of this section, we describe four basic analysis tasks implemented in TimeSeer: filtering, brushing, drill-down, and searching.

4.4.1 Filtering

We employ a gradient on the Scagnostic series and the scatterplots at the top to help users locate scatterplots with high Scagnostic values. However, it is not possible for users to filter only scatterplots with selected Scagnostic values in a specific interval (for example, with Monotonicity from 0.6 to 0.8). The range sliders on the left of each Scagnostic time series allow users to do that. Figure 8(c) shows an example of filters applied to the Outlying series for the three pairs in the lensing area. We are looking for outliers. When the user moves a range slider, a number is displayed to show the current value on the range slider. The filtered parts of time series are faded. In the pairwise view area, the filtered distributions are faded so viewers can focus on data distributions with high numbers of outliers.

4.4.2 Brushing

Looking at an interesting distribution, we may want to check out the data point for further details. For example, we may want to see which state is the outlier in an Outlying distribution. Or, we may ask where is New York in the overall picture for 2001. Or, we may want to compare Illinois and California in 2011. We implement a brushing tool allowing users to do these things. Figure 8(d) illustrates the use of a brush. When we brush a state (a data point in a scatterplot), the state name is displayed in a tool tip, the same state is highlighted in other plots and a line appears connecting adjacent plots. This reveals, in effect, a spatial-temporal series. We can see the changes in the orientation of the state in scatterplots over time. This kind of detail view provides information that cannot be discerned in the original raw time series.

4.4.3 Drill-down

Figure 8 shows an additional view invoked by a simple user action. The Scagnostic is Outlying, and the pairs of variables selected involve State Employment against {Accommodation and Food, Leisure and Hospitality, Education and Health, Construction, Retail Trade, and Trade and Transportation}.

In Figure 8(a), we see a peak in several Scagnostic series. This suggests a time point in 2005 (highlighted in Box A) in which we would expect to see outliers in the relevant scatterplots. We lens this region in Figure 8(b). We see that a period in the Fall has an unusually high peak. This is the precise point where we expect to find an outlier. In Figure 8(c), filtering out low outlying plots allows us to focus on outliers. We now can see clearly the high outlying plots in all selected time series in September of 2005.

We get our details on demand, as shown in Figure 8(d), by clicking on the red scatterplots. Subsequently, the raw series of State Employment and Accommodation and Food (which are the 1-month net change in employment rate) are displayed in the lower graph on two separated scales (in cyan and pink). By brushing the outlier in the scatterplot of September of 2005, we see the actual line graph for that brushed state in yellow. In this case, the outlier is Louisiana. Hurricane Katrina wreaked havoc on their employment and productivity figures (note the sudden drop in Louisiana employment rate and many industries). Notice that Louisiana is also the outlier in the scatterplot of December of 2005, even as it recovered.

In Figure 8(e), we brush another outlier of State Employment and Accommodation and Food scatterplot which is Mississippi. Similarly, Mississippi is also an outlier in the scatterplot of December of 2005. However, Mississippi situation is different (Mississippi got another even more serious drop in Accommodation and Food in December of 2005). We can obtain that information by only looking at the scatterplots. Further details can be found in raw data time series.

4.5 Searching for Similar Patterns or Interesting Distributions

Upon finding an interesting distribution, one may want to look for similar ones. For example, one may wonder if there is another month having a similar distribution to the Katrina example in State Employment vs. Accommodation and Food in September of 2005. Other may want to see if other pairs of variables have similar distributions to the Katrina example in State Employment vs. Accommodation and Food in September of 2005.

TimeSeer offers several methods for discovering similar patterns in the Scagnostic series. The dissimilarity of two scatterplot (S and P) is computed by the following equation:

$$Dissimilarity(S,P) = \sqrt{\sum_{i=1}^{9}(S_i - P_i)^2} \tag{15}$$

where S and P are two arrays of nine Scagnostics of the two scatterplots.

4.5.1 Automatic Search for Similar Distributions

In Figure 9, we show the result from the user selecting a plot from the main screen and requesting a search. TimeSeer searches and plots the top 5 most similar scatterplots, as characterized by the selected Scagnostic (Outlying in this case). In this example, taken from the US employment data, we have selected a plot with a high outlier on State Employment against Accommodation and Food in September of 2005. In the lower panel, the first plot on the left is the plot we selected, highlighted by a yellow rectangle. The five most similar plots are ordered over the nine Scagnostics (the smaller the index, the more similar). Again, the background of a plot is colored by the time series containing that plot; saturation encodes the value of interested feature (the brighter the shade, the more salient the Scagnostic).

We also can see the time (on the top of each plot) for when the data distributions happen to be similar. The Scagnostics of the selected plot and top five plots are also grouped and ordered appropriately. From Figure 9, we note that it is interesting that Louisiana and Mississippi are outliers in all six plots. Additionally, four out of five similar plots are in the same month (September of 2005). This tells us that Hurricane Katrina affected Louisiana employment in the selected economy sectors.

The last similar plot is another month which has similar Scagnostics as State Employment against Accommodation and Food in September of 2005. However, the situation in December of 2005 is different. While Louisiana had recovered in both Employment Rate and Accommodation and Food, Mississippi was still struggling.

Figure 10 shows a similar result for the Weather data. We have selected a plot with a high Stringy and Skinny Scagnostic value on barometric pressure vs. air temperature and searched for similar distributions in the same time series. This is a rather fascinating example of an unusual relation between variables that would not be evident in summary statistics such as the Pearson correlation. It is well-known that air temperature and barometric pressure are related, but

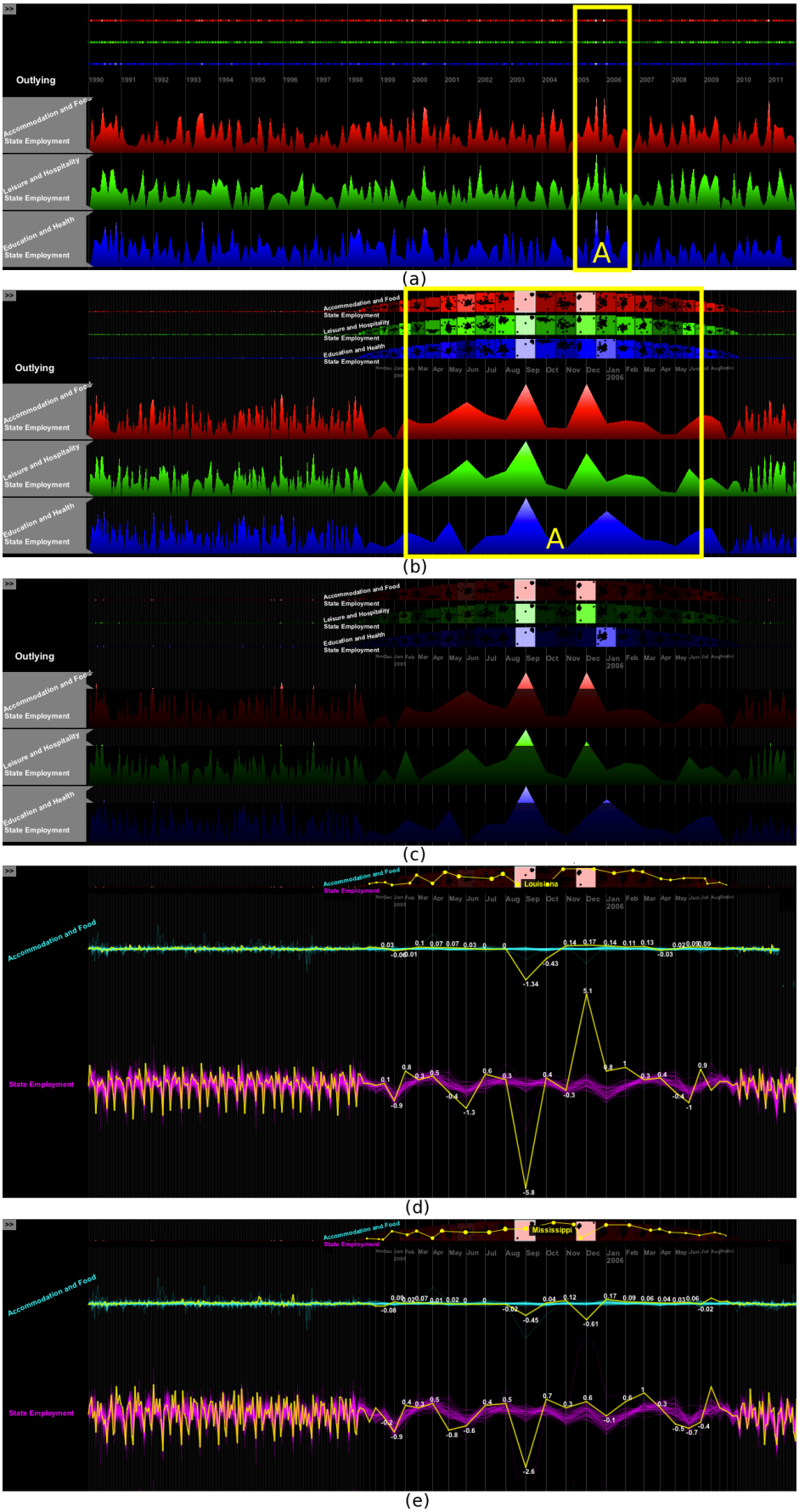


Fig. 8. Details on demand: a) Overview b) Lensing in interested area c) Filtering: Outlying ≥ 0.7 d) Requested actual data plots of the first outlier (Louisiana) e) Requested actual data plots of the second outlier (Mississippi).

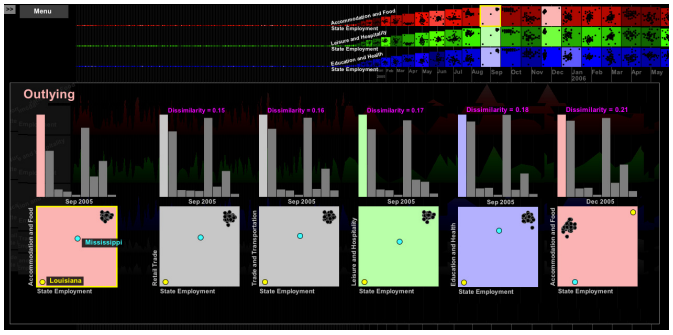


Fig. 9. US Employment data, searching for distributions which are similar to State Employment against Accommodation and Food in September of 2005.

these plots make clear that it is not a simple functional relationship. By searching for Stringy Scagnostics, we see that this dynamic relationship between barometric pressure and air temperature has little error (the strings/paths are quite smooth) but is highly nonlinear (they wind around instead of following a straight line).

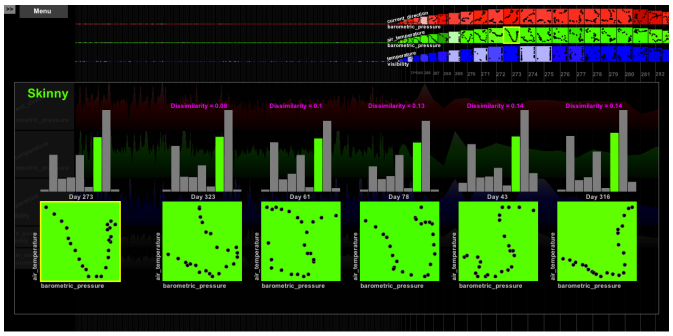


Fig. 10. Weather data, searching for distributions which are similar to barometric pressure and air temperature on day 273.

4.5.2 Manual Search for Similar Distributions

We have devised an annotation that allows a user to search for similar plots manually. The user selects a plot from the main screen and TimeSeer computes the Scagnostic dissimilarity of each plot compared to the selected plot. It then displays this dissimilarity underneath the series. Figure 11(a) shows an example. We have selected an Outlying Scagnostic. We also selected the scatterplot in State Employment against Accommodation and Food in September of 2005. The similarity at each time point compared to the selected scatterplot is presented by the saturation (in purple) of the bar under it; the higher the saturation, the more similar the scatterplots. The slider at the bottom is used to filter similarity. Above this slider is the dot histogram showing the similarity distribution of all plots in the 6 selected time series colored accordingly. Notably, as we have selected a high outlying plot, the brighter plots tend to appear in the front of the slider and vice versa.

The user can filter these plots to see only the most similar ones. Figure 11(b) shows an example. All plots with a dissimilarity greater than 0.5 have been filtered (by using the slider). The user can brush on the remaining dissimilarity purple bars or on the dot histogram to check the dissimilarity. When the mouse is over a purple bar or a plot in the dot histogram, a small window appears right below the purple bar. On this window, a new scatterplot and its Scagnostics is plotted next to the selected plot. In the example in Figure 11(b), we can see that the distribution of State Employment against Education and Health in September of 2005 (and its Scagnostic histogram) is similar to the distribution of State Employment against Accommodation and Food (and its Scagnostic histogram) in the same time.

Figure 12 shows a similar result for the Weather data. We have selected a plot with a high Striated and Skinny Scagnostic value on temperature vs. salinity on day 64 and searched for similar distributions in the selected time series.

We should note that the time for searching similar distribution does not depend on the number of data points. The time for comparing two scatterplots is $O(1)$ compared to $O(n)$ because we are searching on nine Scagnostics, not on individual points. Therefore, the time to search for distributions similar to one in a selected plot is $O(t * p^2)$ instead of $O(t * p^2 * n)$

In a real-time application with a huge time series involving multiple variables, and many data points at each time point, we can cache Scagnostics at each time point. When we need to find plots having a distribution similar to a target plot, we can exploit our cache. The time saved in this approach is considerable.

5 Conclusions

TimeSeer is a visual analytic tool for analyzing a doubly-multivariate time series. It highlights the strength of visual analytics itself because statistical modeling of this type of series is problematic. There are no off-the-shelf algorithms for dealing with the doubly-multivariate time series design, even in advanced statistical packages like SAS or R and even in existing visual analytics platforms.

It should be clear that TimeSeer is not a simple application designed for non-technical users. To leverage its capabilities, a user needs to become familiar with scatterplots, scatterplot matrices, Scagnostics, and multivariate time series. Consequently, we have focused on giving this relatively sophisticated class of analysts a set of tools that enables searches for structure in very high dimensional time series spaces. There are surely ways these tools can be improved, and a study of user interactions can help with this task. Nevertheless, our first challenge has been to devise an interactive platform that can handle huge multivariate collections like the BLS and weather datasets without running out of memory, time, or screen resolution.

One might wonder whether data fitting in this model are rare and whether the model itself is esoteric. We believe the opposite is true. We have cited examples in economics, security, medicine, and other fields that indicate how prevalent these types of data are. An exploratory tool that provides an integrated analytic environment for these types of series can make it possible for the first time to examine real-world datasets that arise from massive data collection systems and sensor networks. The use of Scagnostics also provides an ordinary-language descriptor for distinctive patterns in time series. We see the power of this descriptive language when we compare the plots in Figure 9 and Figure 10. It is appropriate and obvious to characterize the first as Outlying and the second as Stringy.

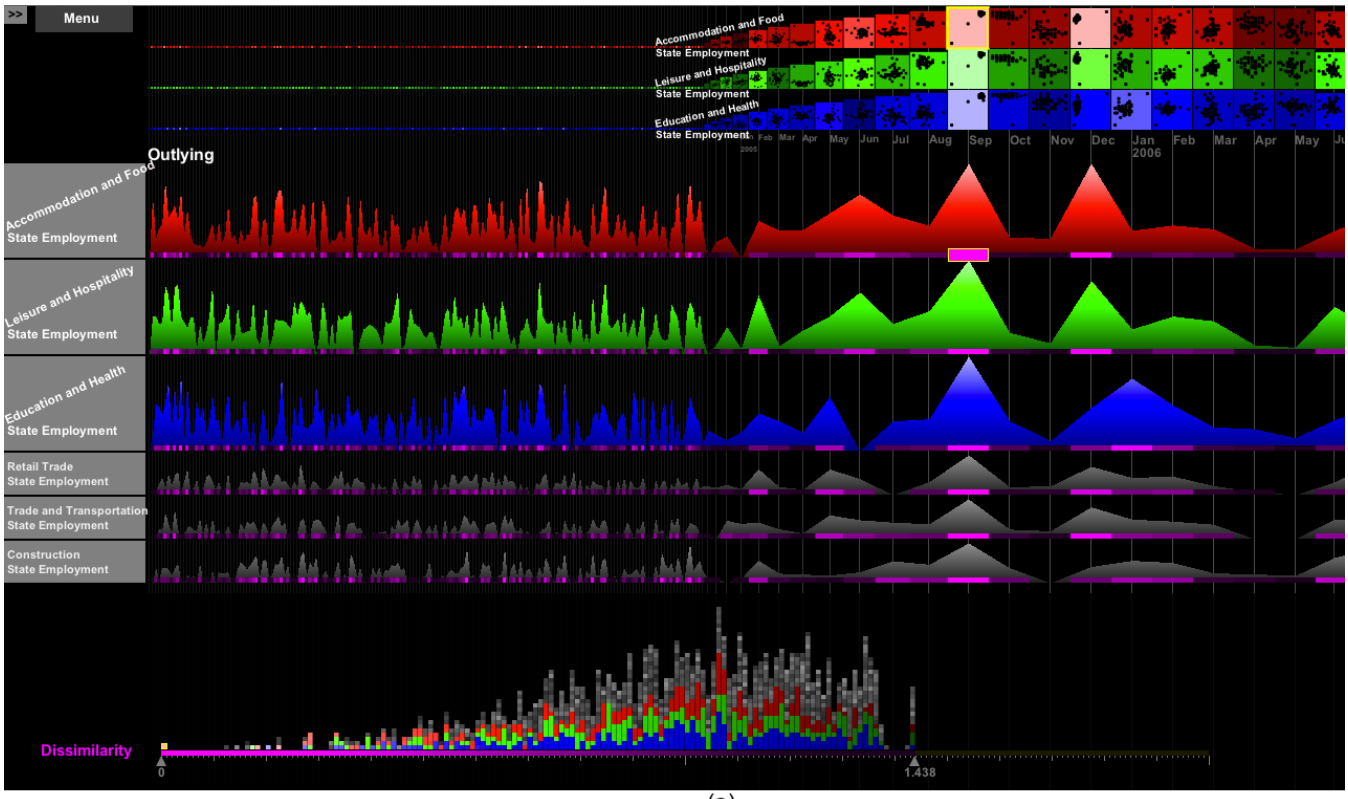
TimeSeer is sensitive to how data is normalized. We normalize each variable independently to have the range 0 to 1. Variables that change the same amount in the normalized scale even with vastly different relative values, produce scatterplots that appear similar and Scagnostics do not help detect these changes especially if the changes in absolute values are small.

The enormous compression we achieve by collapsing n to 1 through the use of Scagnostics provides TimeSeer with the scalability to handle huge datasets. Subcomponents of the system can deal with thumbnails rather than multiple raw series.

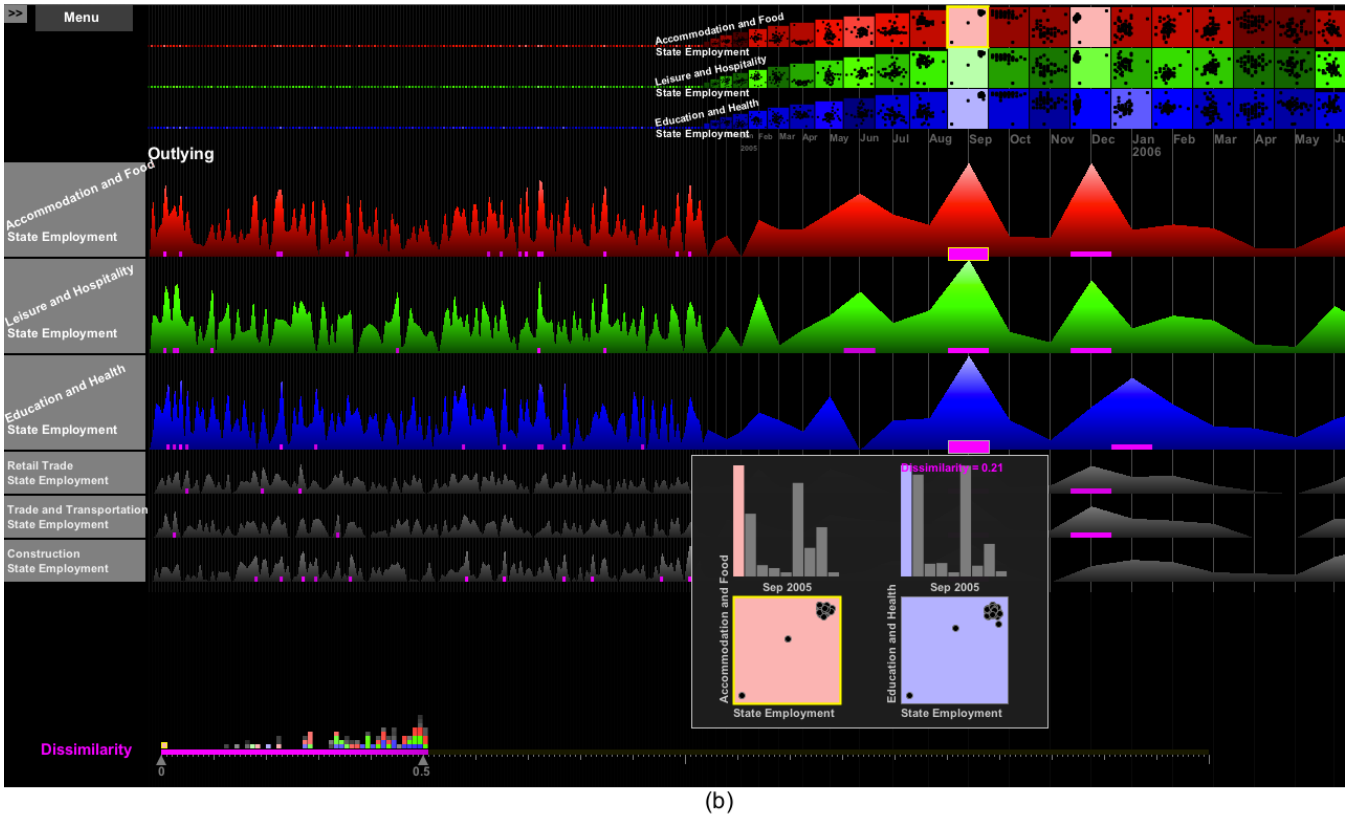
Finally, we plan to investigate the use of TimeSeer on large security databases to assess the gains we claim for its performance. In addition, we expect to investigate how TimeSeer can be extended to spatial data. Time and space have similar statistical issues when modeling [13], so extending time series analytics to spatial analytics makes sense.

Acknowledgments

This work was supported by NSF/DHS grant DMS-FODAVA-0808860.

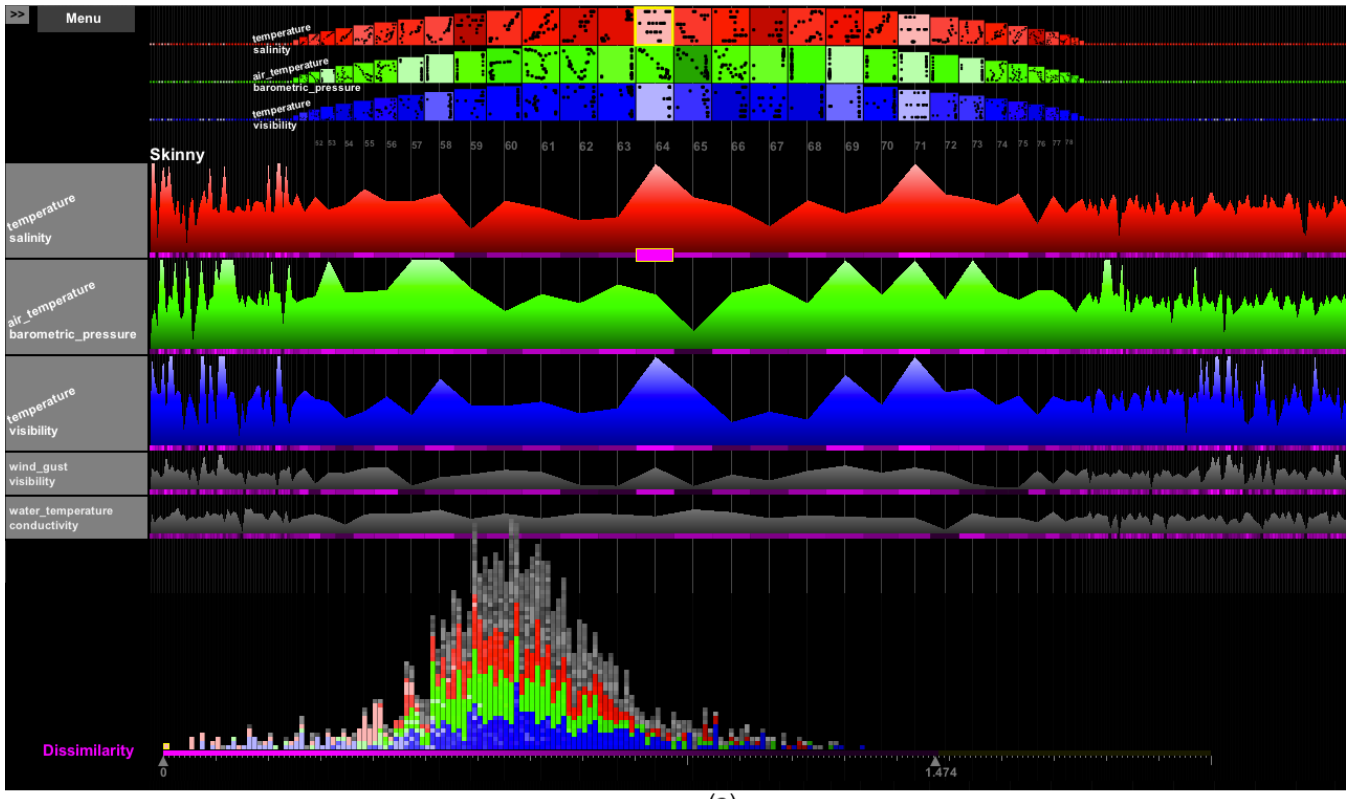


(a)

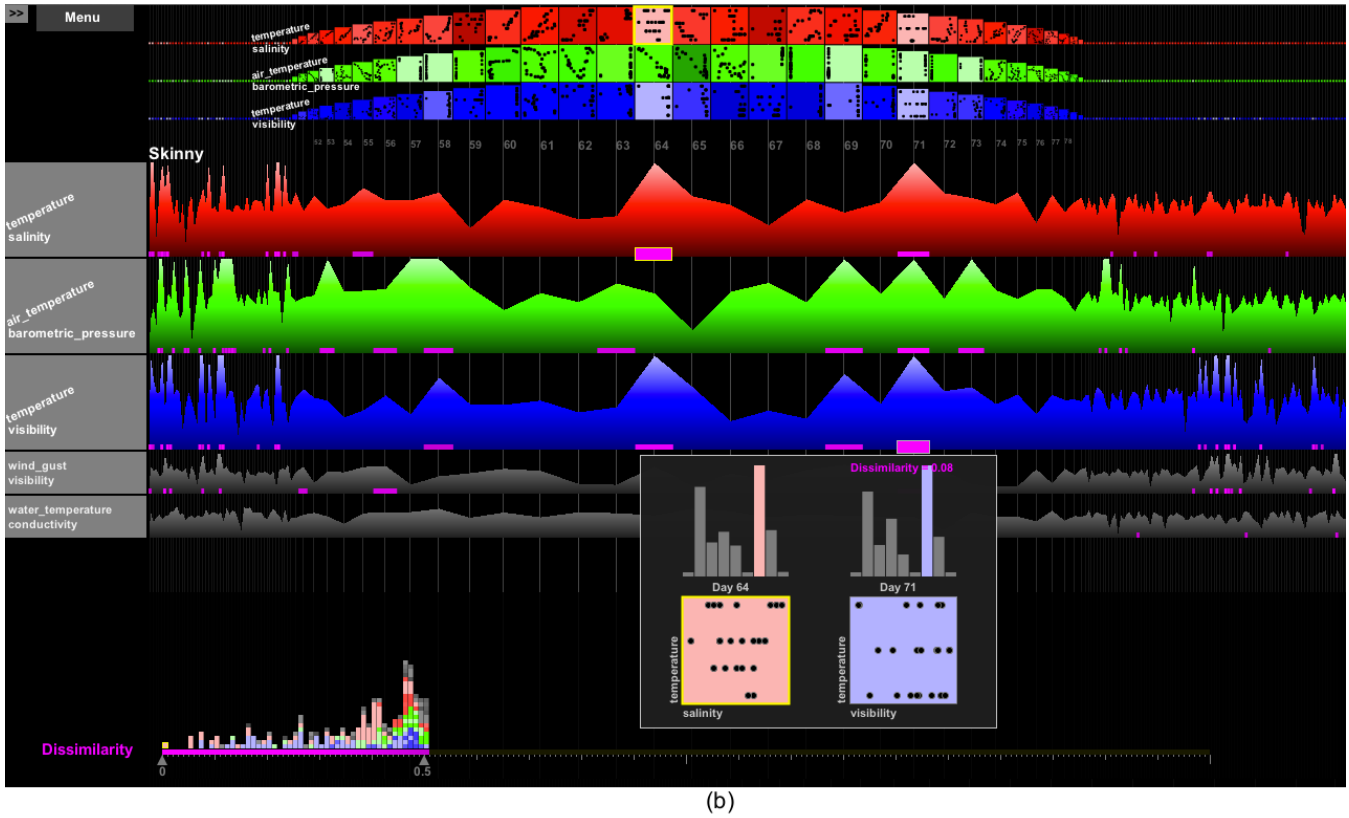


(b)

Fig. 11. US Employment data, searching for distributions which are similar to State Employment against Accommodation and Food in September of 2005: a) Dissimilarity index for all selected pairs b) Filter applied: dissimilarity ≤ 0.5 .



(a)



(b)

Fig. 12. Weather data, searching for distributions which are similar to temperature vs. salinity on day 64: a) Dissimilarity index for all selected pairs b) Filter applied: dissimilarity ≤ 0.5 .

REFERENCES

[1] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. Visualizing time-oriented data-a systematic view. *Computer Graph.*, 31(3):401–409, 2007.

[2] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *Proc. of the IEEE Symposium on Information Visualization*, pages 15–24, 2005.

[3] K. Beard, H. Deese, and N. Pettigrew. A framework for visualization and exploration of events. *Information Visualization*, 7:133–151, April 2008.

[4] J. Beniger and D. Robyn. Quantitative graphics in statistics: A brief history. *The American Statistician*, 32:1–11, 1978.

[5] J. Blaas, C. Botha, and F. Post. Extensions of parallel coordinates for interactive exploration of large multi-timepoint data sets. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1436–1451, Nov. 2008.

[6] G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control (rev. ed.)*. Holden-Day, Oakland, CA, 1976.

[7] L. Byron and M. Wattenberg. Stacked graphs – geometry & aesthetics. *IEEE Transactions on Visualization and Computer Graphics*, 14:1245–1252, November 2008.

[8] J. Caraça-Valente and I. López-Chavarrías. Discovering similar patterns in time series. In *Proc. of the ACM SIGKDD*, KDD '00, pages 497–505, New York, NY, USA, 2000. ACM.

[9] D. B. Carr, R. J. Littlefield, W. L. Nicholson, and J. S. Littlefield. Scatterplot matrix techniques for large n. *Journal of the American Statistical Association*, 82:424–436, 1987.

[10] B. Chiu, E. Keogh, and S. Lonardi. Probabilistic discovery of time series motifs. In *Proc. of the ACM SIGKDD*, KDD '03, pages 493–498, New York, NY, USA, 2003. ACM.

[11] W. S. Cleveland. *The Elements of Graphing Data*. Hobart Press, Summit, NJ, 1985.

[12] W. S. Cleveland. *Visualizing Information*. Hobart Press, New Jersey, 1993.

[13] N. Cressie. *Statistics for spatial data*. John Wiley & Sons, New York, 1991.

[14] A. Dasgupta and R. Kosara. Pargnostics: Screen-space metrics for parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 16:1017–2626, 2010.

[15] H. Edelsbrunner, D. G. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29:551–559, 1983.

[16] L. Fu. Implementation of three-dimensional scagnostics. Master's thesis, University of Waterloo, Department of Mathematics, 2009.

[17] H. Funkhouser. Historical development of the graphical representation of statistical data. *Osiris*, 3:269–404, 1937.

[18] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization, 2011.

[19] J. C. Gower and G. J. S. Ross. Minimal spanning trees and single linkage cluster analysis. *Applied Statistics*, 18:54–64, 1969.

[20] A. Grafton and D. Rosenberg. *Cartographies of Time: A History of the Timeline*. Princeton Architectural Press, Princeton, NJ, 2010.

[21] R. L. Grossman, M. Sabala, A. Aanand, S. Eick, L. Wilkinson, P. Zhang, J. Chaves, S. Vejcik, J. Dillenburg, P. Nelson, D. Rorem, J. Alimohideen, J. Leigh, M. Papka, and R. Stevens. Real time change detection and alerts from highway traffic data. In *ACM/IEEE SC 2005 Conference (SC '05)*, 2005.

[22] J. A. Hartigan and S. Mohanty. The runt test for multimodality. *Journal of Classification*, 9:63–70, 1992.

[23] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.

[24] S. Havre, B. Hetzler, and L. Nowell. Themeriver: Visualizing theme changes over time. In *Proceedings of the 2000 IEEE Symposium on Information Visualization*, pages 115–123, Washington, DC, USA, 2000. IEEE Computer Society.

[25] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *Proc. of ACM SIGCHI*, CHI '09, pages 1303–1312, New York, NY, USA, 2009. ACM.

[26] H. Hochheiser and B. Shneiderman. Dynamic query tools for time series data sets: Timebox widgets for interactive exploration. *Information Visualization*, 3:1–18, March 2004.

[27] S. J. and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4:96–113, March 2005.

[28] J. Johansson, P. Ljung, and M. Cooper. Depth cues and density in temporal parallel coordinates. In K. Museth, T. Mller, and A. Ynnerman, editors, *EuroVis*, pages 35–42. Eurographics Association, 2007.

[29] E. Keogh, S. Lonardi, and B. Chiu. Finding surprising patterns in a time series database in linear time and space. In *Proc. of the ACM SIGKDD*, KDD '02, pages 550–556, New York, NY, USA, 2002. ACM.

[30] J. Klein. *Statistical Visions in Time: A History of Time Series Analysis, 1662-1938*. Cambridge University Press, Cambridge, UK, 1997.

[31] Y. K. Leung and M. D. Apperley. A review and taxonomy of distortion-oriented presentation techniques. *ACM Trans. Comput.-Hum. Interact.*, 1:126–160, June 1994.

[32] H. Levkowitz. *Color Theory and Modeling for Computer Graphics, Visualization, and Multimedia Applications*. Kluwer Academic Publishers, Boston, MA, 1997.

[33] J. Lin, E. Keogh, S. Lonardi, J. Lankford, and D. Nystrom. Visually mining and monitoring massive time series. In *Proc. of the ACM SIGKDD*, KDD '04, pages 460–469, New York, NY, USA, 2004. ACM.

[34] P. McLachlan, T. Munzner, E. Koutsofios, and S. North. Liverac: interactive visual exploration of system management time-series data. In *Proc. of the ACM SIGCHI*, CHI '08, pages 1483–1492, New York, NY, USA, 2008. ACM.

[35] F. Mörchen and A. Ultsch. Efficient mining of understandable patterns from multivariate interval time series. *Data Mining and Knowledge Discovery*, 15:181–215, October 2007.

[36] A. Norton, M. Rubin, and L. Wilkinson. Streaming graphics. *Statistical Computing and Graphics Newsletter*, 12(1):11–14, 2001.

[37] T. Oates. Identifying distinctive subsequences in multivariate time series by clustering. In *Proc. of the ACM SIGKDD*, KDD '99, pages 322–326, New York, NY, USA, 1999. ACM.

[38] R. Rao and S. K. Card. The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*, CHI '94, pages 318–322, New York, NY, USA, 1994. ACM.

[39] J. Schneidewind, M. Sips, and D. Keim. Pixnostics: Towards measuring the value of visualization. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 199–206, Baltimore, MD, 2006.

[40] W. Stuetzle. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of Classification*, 20:25–47, 2003.

[41] S. Thakur and T.-M. Rhyne. Data vases: 2d and 3d plots for visualizing multiple time series. In *Proc. of the International Symposium on Advances in Visual Computing: Part II*, ISVC '09, pages 929–938, Berlin, Heidelberg, 2009. Springer-Verlag.

[42] J. Van Wijk and E. Van Selow. Cluster and calendar based visualization of time series data. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 4–10, Washington, DC, USA, 1999. IEEE Computer Society.

[43] H. Wainer. *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot*. Lawrence Erlbaum Associates, Mahwah, NJ, 2000.

[44] H. Wainer. *Graphic Discovery: A Trout in the Milk and Other Visual Adventures*. Princeton University Press, Princeton, NJ, 2004.

[45] M. Wattenberg. Baby names, visualization, and social data analysis. In *Proceedings of the 2005 IEEE Symposium on Information Visualization*, pages 1–7, Washington, DC, USA, 2005. IEEE Computer Society.

[46] M. Weber, M. Alexa, and W. Müller. Visualizing time-series on spirals. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 7–17, Washington, DC, USA, 2001. IEEE Computer Society.

[47] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *Proceedings of the IEEE Information Visualization 2005*, pages 157–164. IEEE Computer Society Press, 2005.

[48] L. Wilkinson, A. Anand, and R. Grossman. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1363–1372, 2006.

Tuan Nhon Dang is a Ph.D. student in Computer Science at the University of Illinois at Chicago. He received his B.Sc. degree in Computer Science from the University of Technology, Ho Chi Minh City, Vietnam in 2006. In 2009, he received his M.Sc. degree in Computer Science at University of Illinois at Chicago and a M.Sc. degree in Computing Systems at Politecnico di Milano, Italy. His research interests are in visual analytics and computer animation.

Anushka Anand is a Ph.D. student in Computer Science at the University of Illinois at Chicago. She received her M.Sc. degree in Computer Science from the University of Illinois at Chicago and her B.Sc. degree in Computer Science from the American University of Sharjah, United Arab Emirates. She currently serves as a student member of the Board of Trustees of the Anita Borg Institute and as Chair of the Chicago Chapter of the ACM. Her research interests are in visual analytics and data mining.

Leland Wilkinson is Executive VP of SYSTAT Software Inc. and Adjunct Professor of Computer Science at the University of Illinois Chicago. He received an A.B. degree from Harvard in 1966, an S.T.B. degree from Harvard Divinity School in 1969, and a Ph.D. from Yale in 1975. Wilkinson wrote the SYSTAT statistical package and founded SYSTAT Inc. in 1984. Wilkinson is a Fellow of the American Statistical Association, an elected member of the International Statistical Institute, and a Fellow of the American Association for the Advancement of Science. He has served on the Committee on Applied and Theoretical Statistics of the National Research Council and has been Vice Chair of the Board of the National Institute of Statistical Sciences (NISS). In addition to authoring journal articles, the original SYSTAT computer program and manuals, and patents in visualization and distributed analytic computing, Wilkinson is the author (with Grant Blank and Chris Gruber) of *Desktop Data Analysis with SYSTAT* and *The Grammar of Graphics*.