Timeseer: Detecting interesting distributions in multiple time series data

Tuan Nhon Dang
University of Illinois at Chicago
tdang@cs.uic.edu

Leland Wilkinson
University of Illinois at Chicago
leland.wilkinson@systat.com

ABSTRACT

Widespread interest in features and trends in time series has generated a need for interactive tools that support discovering unusual events in time series. In this paper, we introduce an application (TimeSeer) for guiding interactive exploration through high-dimensional data. Our application is designed to handle the types of doubly-multivariate data series by working directly on noteworthy features such as density, skewness, shape, outliers, and texture.

ACM Classification Keywords

I.5.2 Pattern recognition: Design Methodology—*Pattern analysis*

Author Keywords

Scagnostics, Time Serie Visualization, Dot Plots

INTRODUCTION

TimeSeer [4] is a platform for the visual analysis of high-dimensional multivariate time series. The data model that TimeSeer is designed to deal with is: 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 design. 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). We normally expect t, p, and n to be large. An traditional approach, of course, would be to examine all individual series. This approach does not scale.

This paper deals with a substantial extension to the TimeSeer model that allows us to examine time series in a dense visual environment. The original model allowed a user to select pairs of time series and analyze relations between them. The current model allows a user to examine all time series in a corpus simultaneously.

VINCI 2012, September 27–28, 2012, Hangzhou, China.

RELATED WORK

Scagnostics

The features we use to process time series are based on Scagnostics. In the mid 1980s, John and Paul Tukey developed an exploratory graphical method to describe a collection of 2D scatterplots through a small number of measures of the pattern of points in these plots [9]. We implemented the original Tukey idea through nine Scagnostics (Outlying, Skewed, Clumpy, Sparse, Striated, Convex, Skinny, Stringy, Monotonic) defined on planar proximity graphs.

We now review the Scagnostic algorithm [15].

Binning

We begin by normalizing the data to the unit interval and then use a 40 by 40 hexagonal grid [2] to aggregate the points in each scatterplot. 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).

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 [5].

Deleting Outliers

We consider an outlier to be a vertex whose adjacent edges in the MST all have a weight (length) greater than F_{inner+}, where

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

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.

Computing Scagnostic Measures

We now present the Scagnostic measures computed on our three geometric graphs: H for the convex hull, A for the alpha shape, and T for the minimum spanning tree. Figure 1 shows an example of the three geometric graphs. We are interested in assessing three aspects of scattered points: *density*, *shape*, and *association*.

Density Measures

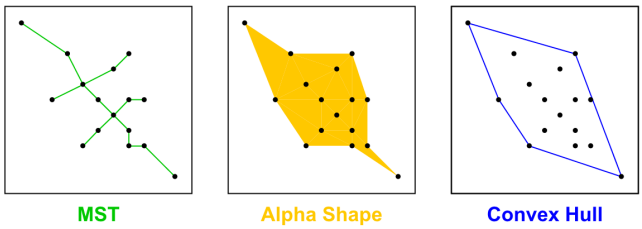


Figure 1. Minimum spanning tree, alpha shape, and convex hull.

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.

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

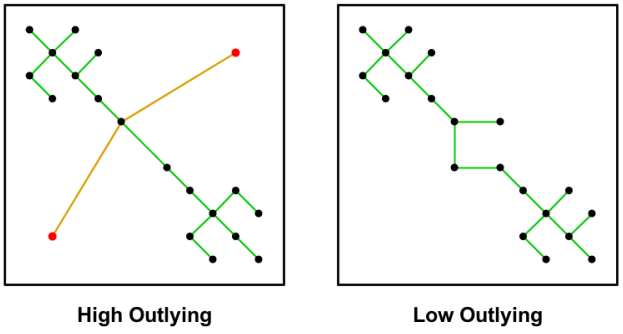


Figure 2. High Outlying and low Outlying distributions (Red vertices are outliers which are not outliers in both projections).

- **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. Figure 3 shows an example of this measure.

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

- **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

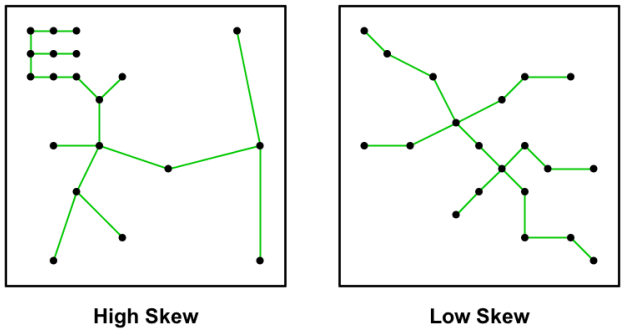


Figure 3. High Skew and low Skew distributions (MST is in green).

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.

$$c_{sparse} = q_{90} \quad (4)$$

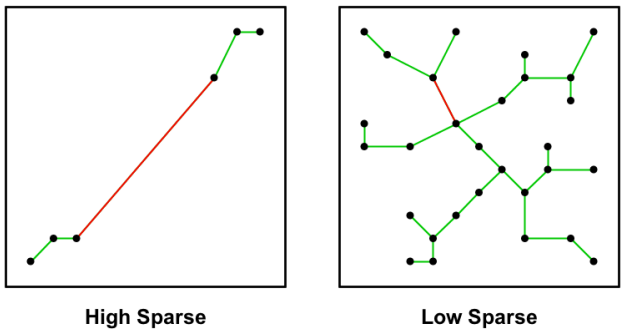


Figure 4. High Sparse and low Sparse distributions (Red edges are q_{90}).

- **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 [8]. 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 [6], we can associate a runt size (r_j) with each edge (e_j) in the MST, as described by [12]. 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.

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

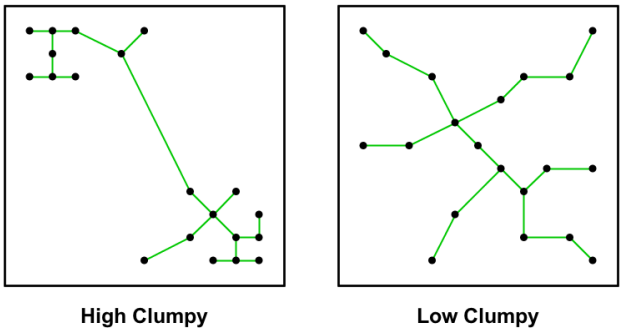


Figure 5. High Clumpy and low Clumpy distributions.

- **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

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

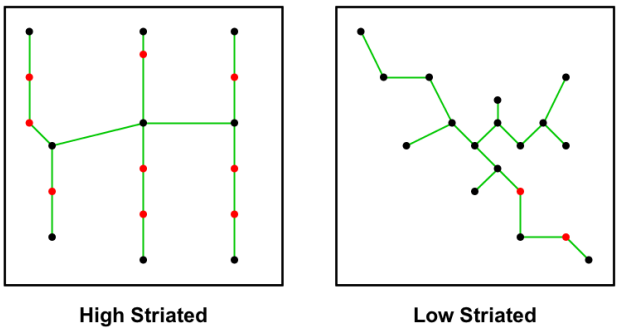


Figure 6. High Striated and low Striated distributions (Red nodes are 2-degree vertices whose cosine is less than -0.75).

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 and the alpha hull.

- **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 (alpha shape) and the convex hull have identical areas.

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

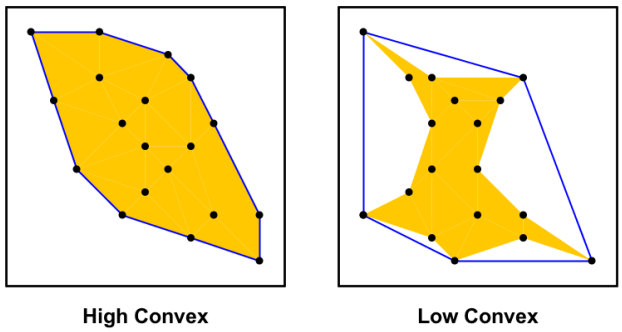


Figure 7. High Convex and low Convex distributions (Alpha shape in yellow and convex hull in blue).

- **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.

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

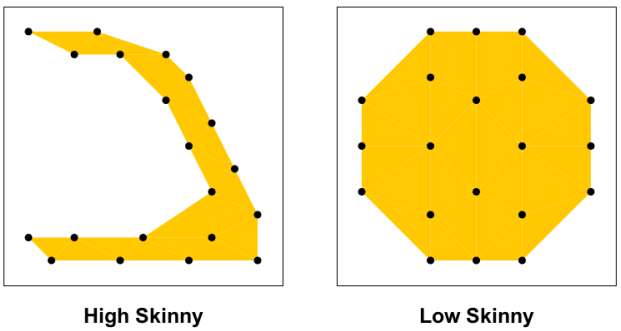


Figure 8. High Skinny and low Skinny distributions (Alpha shape in yellow).

• **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.

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

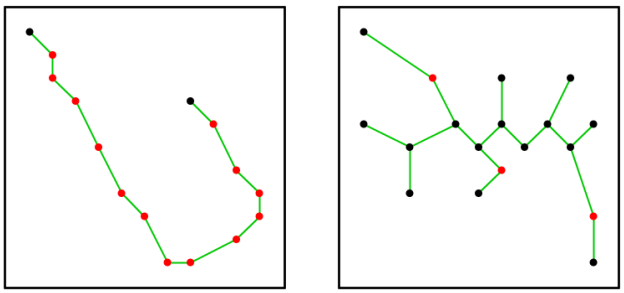


Figure 9. High Stringy and low Stringy distributions(Red nodes are 2-degree vertices in MST).

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.

$$c_{monotonic} = r^2{}_{spearman} \tag{10}$$

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

Visualizing Multivariate Time Series

Some have developed viewers for multivariate time series. Theme River [10] was one of the first applications developed for visualizing multivariate time series. It employed kernel smooths of time series, stacking them in a single display. Based on a similar idea, Wattenberg [14] developed an applet called Name Voyager, which allows one to drill-down to an individual series easily.

Another way to deal with multivariate series is to aggregate across similar series [11, 13, 7]. Aggregation risks concealment of important features, however.

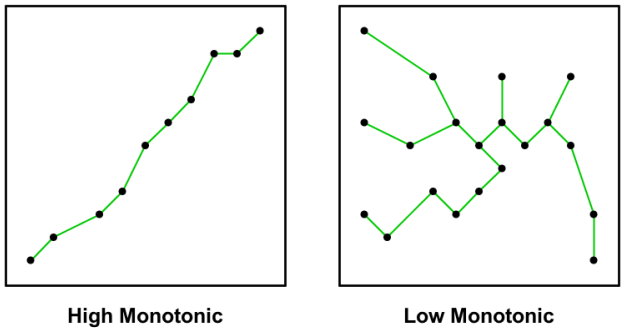


Figure 10. High Monotonic and low Monotonic distributions.

In any case, none of these approaches can deal with the t, p, and n multivariate series that are handled by TimeSeer.

TIMESEER

We will illustrate TimeSeer by using real datasets to show how this visual analytic can be used to detect anomalies and regular patterns.

Data sets

In this section, we use three different datasets to demonstrate the performance of TimeSeer. The first is a series of Weather data, the second is a series of US Employment data, and the third is a series of International Energy data.

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.

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 Govern-

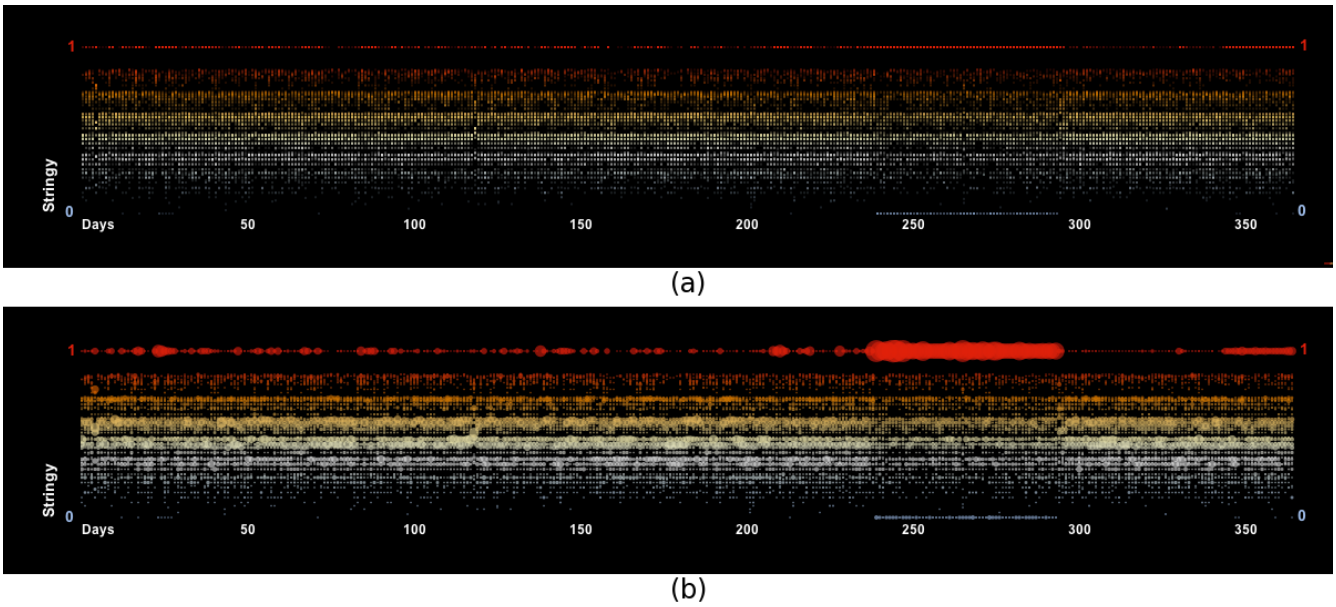


Figure 11. Stringy measure of the Weather data: a) 2D color map b) 2D Dot Plot map.

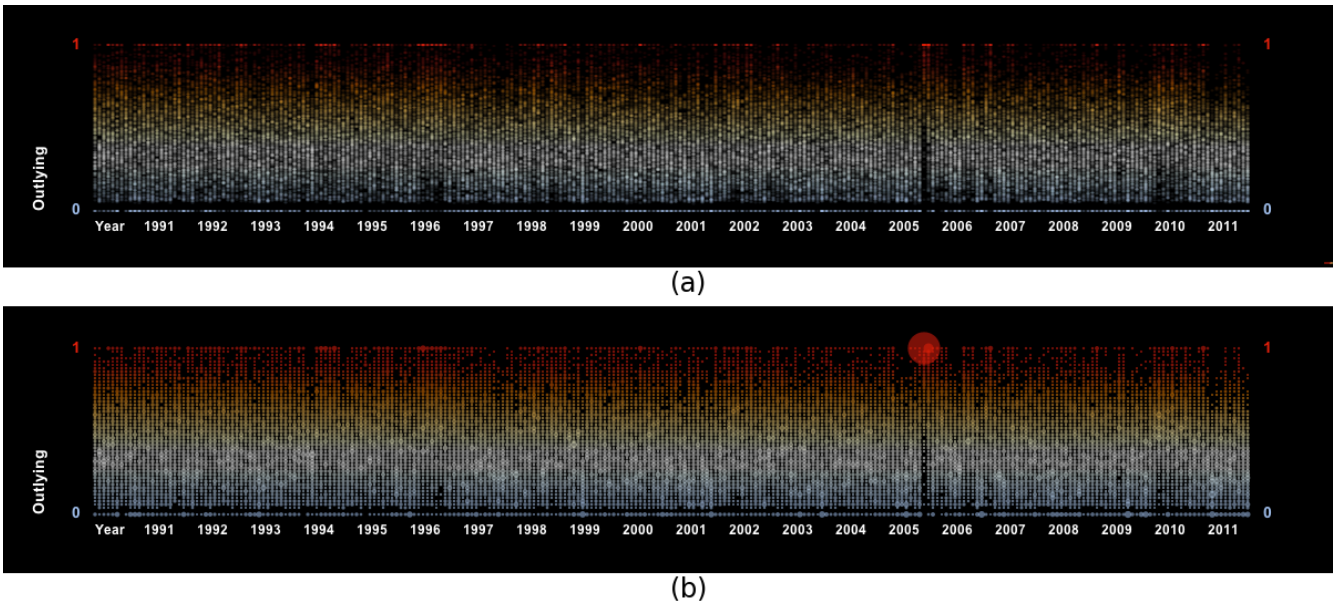


Figure 12. Outlying measure of the US Employment data: a) 2D color map b) 2D Dot Plot map.

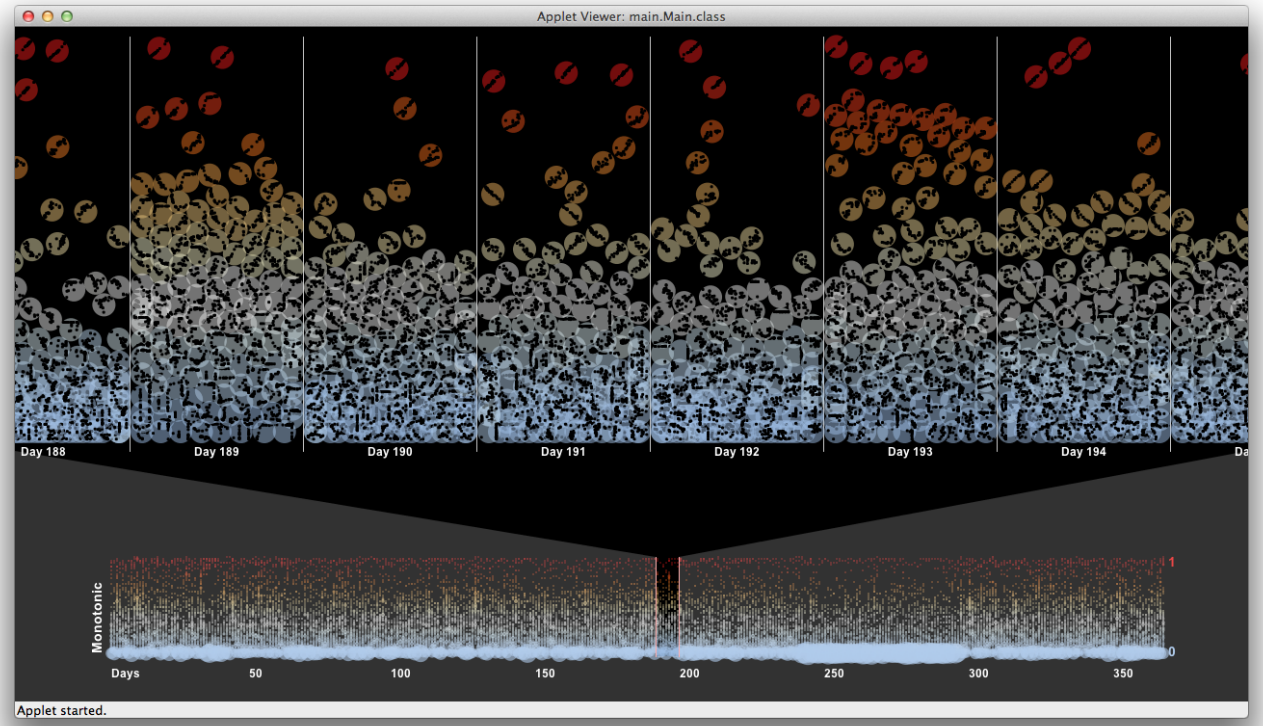


Figure 13. Monotonic measure of the Weather data: Pan and zoom in one week data.

ment, Local Government, and State Employment. For these data, we have 78,600 scatterplots with 50 data points each to examine.

The International Energy data comprise yearly energy statistics for more than 200 countries over 31 years from 1990 to 2010. There are 13 variables represented in the dataset: CO_2 Emissions, Per Capita CO_2 Emissions, Population, Oil Supply, Oil Consumption, Gas Production, Gas Consumption, Coal Production, Coal Consumption, Electric Generation, Electric Consumption, Total Energy Production, and Total Energy Consumption. Data can be found at `http://http://www.eia.gov/countries/data.cfm`. For these data, we have 2,418 scatterplots with 200 data points (200 countries over the world) each to examine.

In the rest of this paper, we describe three basic analysis tasks implemented in TimeSeer: overviewing, panning and zooming, brushing, and drilling-down. These analysis tasks capture people's activities while employing information visualization tools for understanding data [1].

Overview

Users first have to select one of nine measures to visualize. Timeseer generates an overview of the selected measure for all pairs of variables over entire time periods. The horizontal axis shows time. The vertical axis shows the selected measure. We also use the heat color map along vertical axis to highlight value distributions on the selected measure.

Figure 11 shows an example of Weather data. In particular, we use a 2D map to present the overview of 50,000 scatterplots. The horizontal axis contains 365 days in a year. The vertical axis is the Stringy measure.

In Figure 11(a), every scatterplot is presented by a dot in the 2D map. High Stringy scatterplots (plots with points lying on snaky paths) are mapped to red dots, low Stringy scatterplots are mapped to blue dots. The opacity of each dot is used to highlight areas with high occurrences of dots.

Figure 11(b) improves the 2D map by using a bubble symbol. The size of each bubble is determined by the number of scatterplots at the same locations. We use a dot plot algorithm [3] to achieve better location accuracy.

Figure 12 shows another example on US Employment data. In particular, we have selected the Outlying measure for visualization. In Figure 12(b), we can see clearly that there are a lot of scatterplots with outliers in a time point in 2005. In this case, the outliers are Louisiana and Mississippi. Hurricane Katrina wreaked havoc on

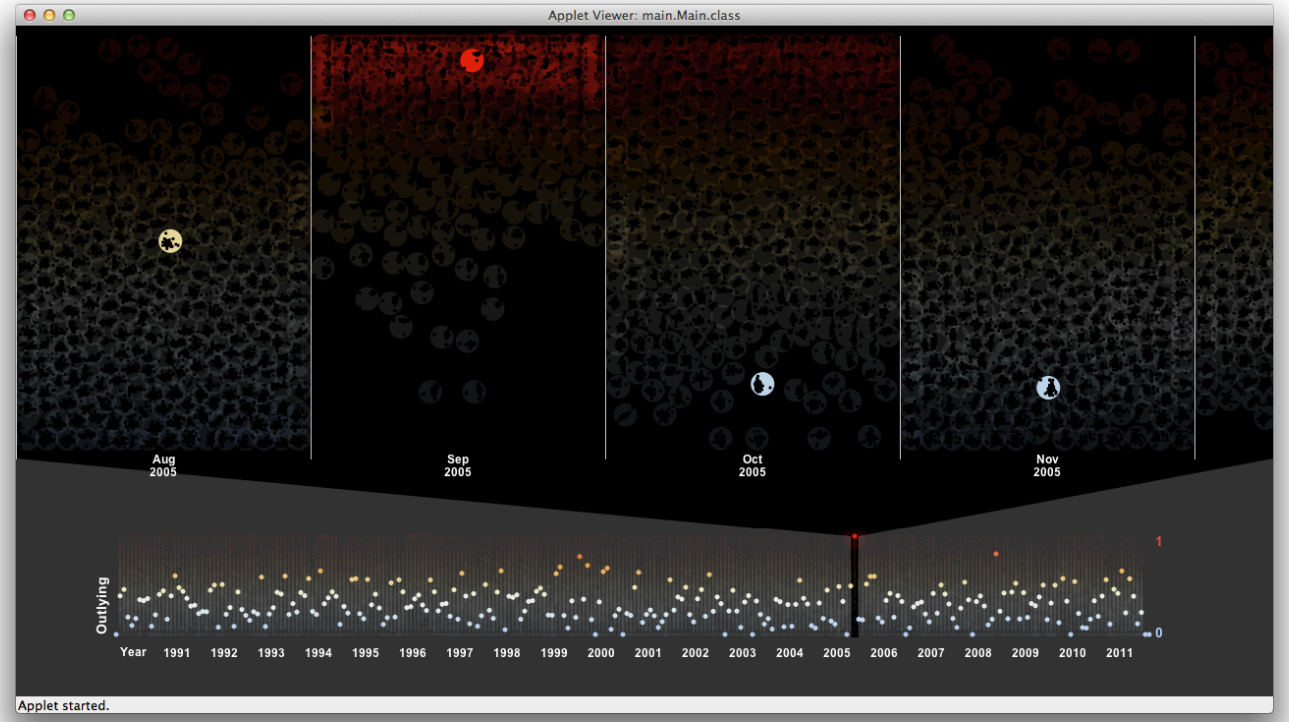


Figure 14. Outlying measure of the US Employment data: Brushing Financial Financial Activities and State Employment.

their employment and productivity figures.

Pan and Zoom

TimeSeer allows one to pan and zoom into a specific region into the overview by using a dragging box. All scatterplots in the dragging box of overview are displayed in the force-directed layout. Since we characterize a scatterplot with orientation-independent features and use them for comparison, we provide an option to view scatterplots in form of circles instead of rectangles. This option makes detecting shapes of data point distribution easier. Moreover, a force-directed layout with circles converges faster than one with rectangles.

Figure 13 shows an example of Pan and Zoom for Weather data. In particular, we zoom into a week from Day 188 to Day 194. All scatterplots in this interval are displayed in the force-directed graph on the top (High Monotonic plots are in red, low Monotonic plots are in blue). Scatterplots move to vertical levels associated to their scagnostic values.

Brushing and Drilling-down

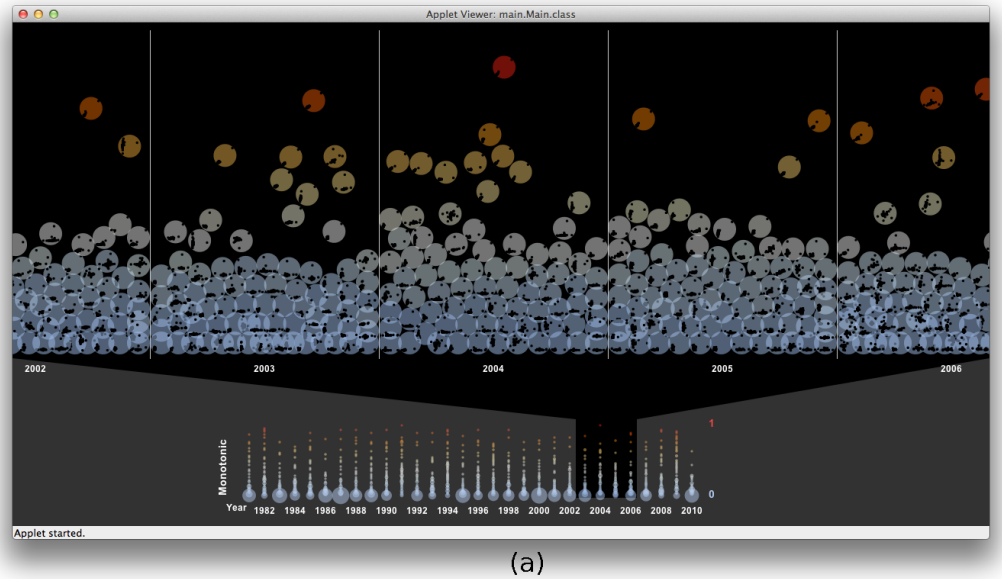
Users can focus on one pair of variables by clicking on any scatterplots in the force-directed graph (the scatterplots of other pairs of variable are faded in both force-directed graph and overview). This is helpful when we want to investigate individual pairs of variables in the entire time series.

Figure 14 shows an example of Brushing for US Employment data. In particular, we have selected one pair of variables (Financial Financial Activities and State Employment). The overview graph shows that there are not many outlying plots, except in September 2005 when Hurricane Katrina happened. In the force-directed graph, scatterplots of the non-selected pairs are faded.

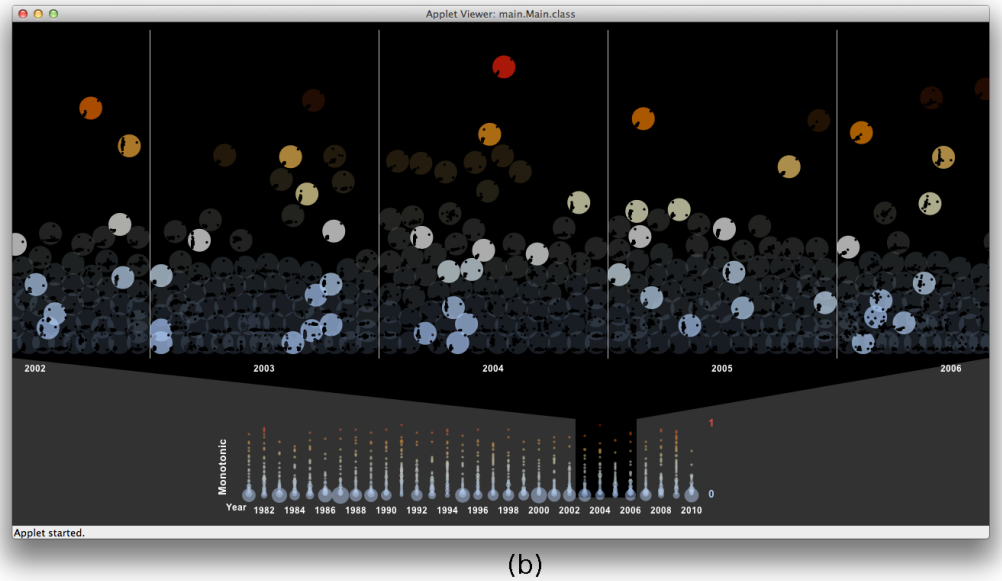
Figure 15 shows another example of Brushing for the International Energy data. We have selected Monotonic to visualize the correlations between variables in the data. Users can focus on a particular variable by selecting only scatterplots containing that variable. For instance, we want to investigate what are the factors that increase CO_2 Emissions (as shown in Figure 15(b)). We then drill down on individual pairs of variables by a simple click on a scatterplot. Figure 15(c) shows an example. In particular, we have selected CO_2 Emissions vs. Total Energy consumption in 2004. In this year, China was the country releasing the most CO_2 and also the country using the most energy.

ACKNOWLEDGMENTS

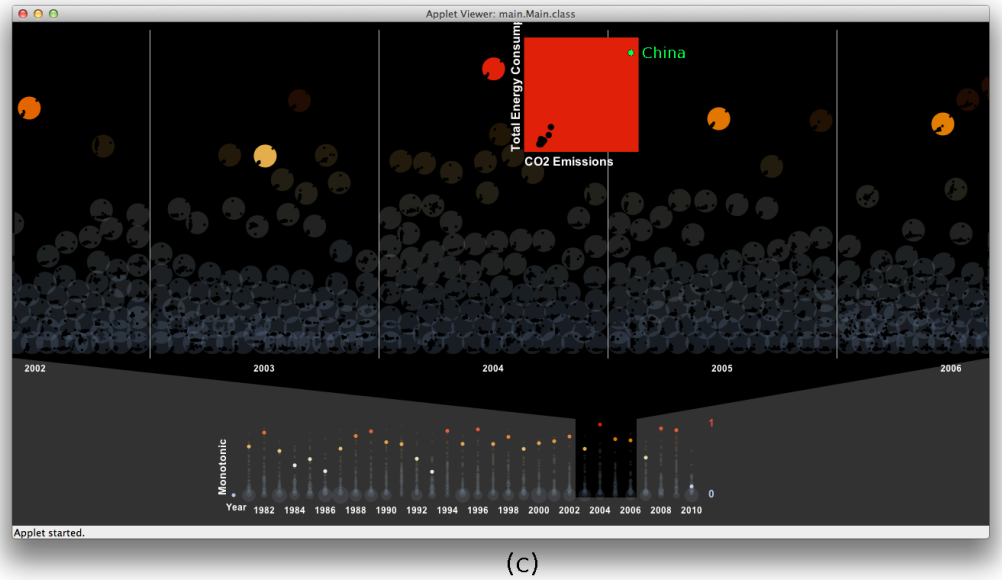
This work was supported by NSF/DHS grant DMS-FODAVA-0808860.



(a)



(b)



(c)

Figure 15. Monotonic measure of the International Energy data: (a) All pairs of variables (b) All pairs containing CO_2 Emissions (c) CO_2 Emissions and Total Energy consumption.

REFERENCES

1. 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.

2. 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.

3. T. Dang, L. Wilkinson, and A. Anand. Stacking graphic elements to avoid over-plotting. In *INFOVIS '10: Proceedings of the IEEE Symposium on Information Visualization (INFOVIS'10)*, Washington, DC, USA, 2010. IEEE Computer Society.

4. T. N. Dang, A. Anand, and L. Wilkinson. Timeseer: Scagnostics for high-dimensional time series. *IEEE Transactions on Visualization and Computer Graphics*, 99(PrePrints), 2012.

5. 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.

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

7. 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.

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

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

10. 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.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.