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. In this paper, we specify the problem, review our measures on point clouds, describe our approach, and present application results based on a real world datasets.

Categories and Subject Descriptors

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

General Terms

Design

Keywords

Scagnostics, Scatterplot Matrix, Time Serie Visualization

1. INTRODUCTION

TimeSeer [5] 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.

Scatterplots are used to analyze the relationship between two variables. Figure 1 shows an example. The two variables are Birth Rate and Death Rate for 200 countries in 2011 ($p = 2$, $n = 200$, and $t = 1$).

As p increases, a scatterplot matrix accommodates multiple scatterplots in a single display. Figure 2 shows an example from The World Bank Data. The data contains 10 health-related variables for 200 countries in 2010 ($p = 10$, $n = 200$, and $t = 1$).

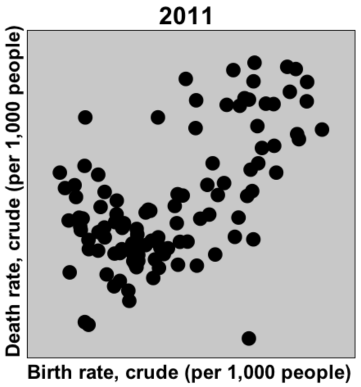


Figure 1: Birth Rate and Death Rate of 200 countries in 2011.

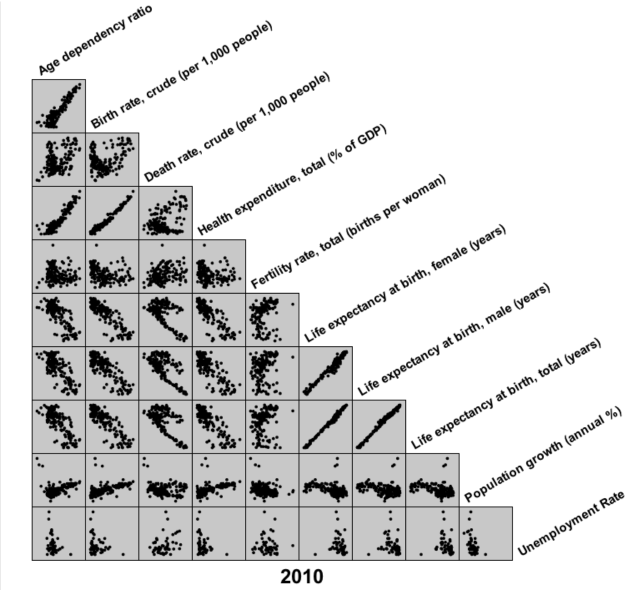


Figure 2: Scatterplot Matrix of 10 health-related variables from The World Bank Data in 2010.

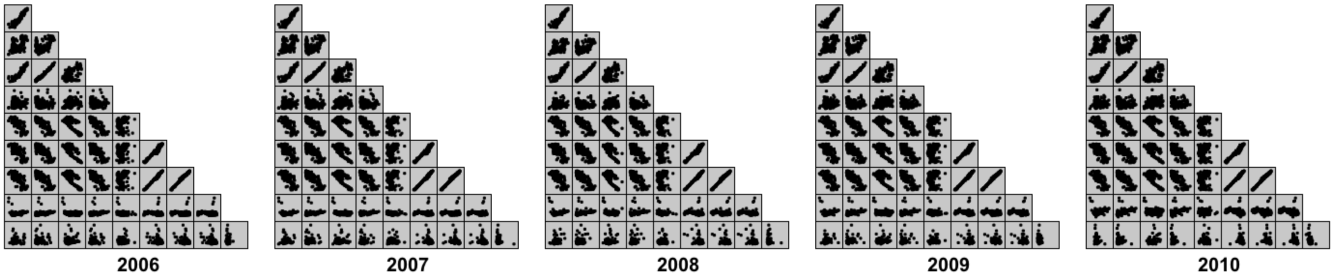


Figure 3: Scatterplot Matrix of 10 health-related variables from The World Bank Data from 2006 to 2010.

TimeSeer is designed to deal with a much larger problem: time series of scatterplot matrices. For example, we will examine The World Bank Data which contains 10 health-related variables of 200 countries over 52 years from 1950 to 2011. Figure 3 conceptually shows this data model. We have, in effect, t scatterplot matrices, one at each time point.

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.

2. RELATED WORK

2.1 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].

2.1.1 Binning

We begin by normalizing the data to the unit interval and then use a 40 by 40 hexagonal grid [3] 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 [6].

2.1.2 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

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

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: H for the convex hull, A for the alpha shape, and T for the minimum spanning tree. Figure 4 shows an example of the three geometric graphs. We are interested in assessing three aspects of scattered points: *density*, *shape*, and *association.*

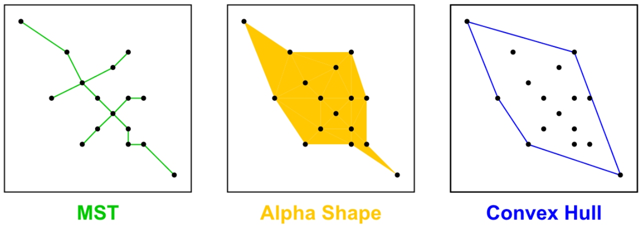


Figure 4: Minimum spanning tree, alpha shape, and convex hull.

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. Figure 6 shows an example of this measure.

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

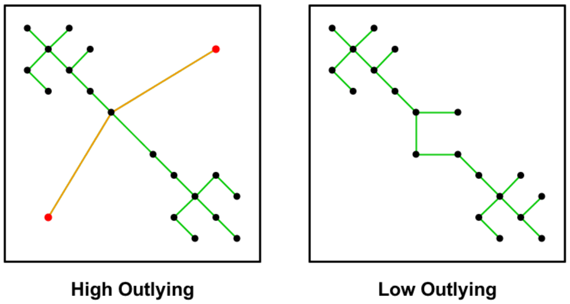


Figure 5: High Outlying and low Outlying distributions (Red vertices are outliers which are not outliers in both projections).

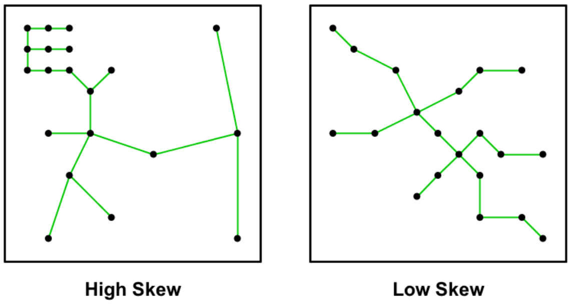


Figure 6: High Skew and low Skew distributions (MST is in green).

- **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.

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

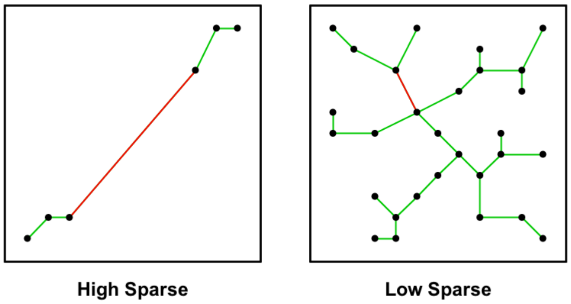


Figure 7: High Sparse and low Sparse distributions (Red edges are q_{90}).

- **Clumpy**

An extremely skewed distribution of MST edge lengths does not necessarily indicate clustering of points. For this purpose, we need another measure based on the MST: the RUNT statistic [8]. The RUNT graph (R_j) corresponding to an 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.

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

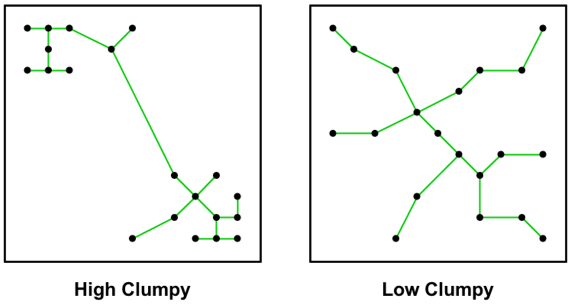


Figure 8: High Clumpy and low Clumpy distributions.

- **Striated**

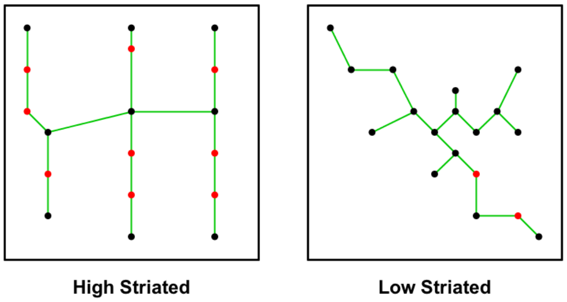
We define coherence in a set of points as the presence of relatively smooth paths in the minimum spanning tree. Smooth algebraic functions, time series, and curves (*e.g.*, spirals) fit this definition. So do points arranged in flows or vector fields. Another common example is the pattern of parallel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

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

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 and the alpha hull.



High Striated

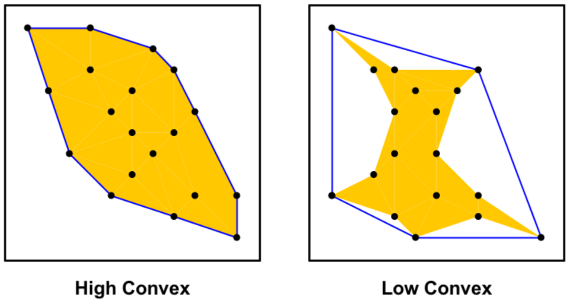
Low Striated

Figure 9: High Striated and low Striated distributions (Red nodes are 2-degree vertices whose cosine is less than -0.75).

- **Convex**

The 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.

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



High Convex

Low Convex

Figure 10: High Convex and low Convex distributions (Alpha shape in yellow and convex hull in blue).

- **Skinny**

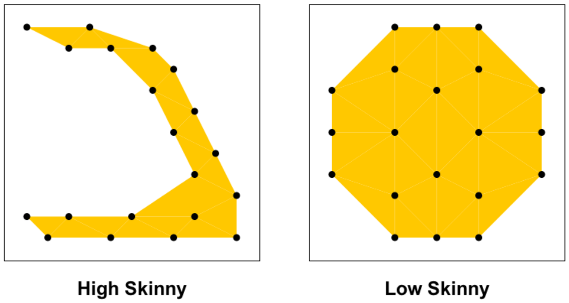
The ratio of perimeter to area of a polygon measures, roughly, how skinny it is. We use a corrected and normalized ratio so that a circle yields a value of 0, a square yields 0.12 and a skinny polygon yields a value near one.

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

- **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.

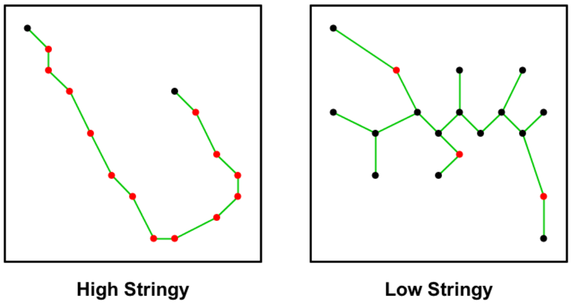
$$c_{stringy} = \frac{|V^{(2)}|}{|V| - |V^{(1)}|} \tag{9}$$



High Skinny

Low Skinny

Figure 11: High Skinny and low Skinny distributions (Alpha shape in yellow).



High Stringy

Low Stringy

Figure 12: 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.

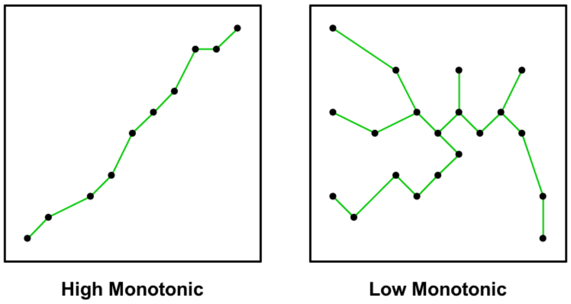


Figure 13: High Monotonic and low Monotonic distributions.

2.2 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, 12, 7]. Aggregation risks concealment of important features, however.

Landesberger [13] monitors 2D point cloud developments by tracing hull and mid point of the point cloud. This approach is not scalable. In particular, tracing the development of 10 point clouds in a single display can clutter the visualization. Timeseer can be used to analyze up to a hundred of point clouds at the same time. Moreover, we do not just trace how the hull of a point cloud changes over time, we inspect features of the point cloud. The sudden changes in these features suggest a special event happens at a time point where we can drill-down for further details.

3. TIMESEER OVERVIEW

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

3.1 Quality Metrics

A quality metric captures properties useful to the extraction of meaningful information about the data. Quality metrics for detecting interesting patterns in high-dimensional data are described in [2]. Now, we describe the quality metrics used in TimeSeer.

What is measured. *Clustering* metrics measure the extent to which the data contain groupings. Our Clumpy measure captures this feature in a scatterplot. *Correlation* captures the extent to which systematic changes to one dimension are accompanied by changes in other dimensions. Our Monotonic measure belongs to this category. *Outlier* metrics capture the extent to which the data segment under inspection contains elements that behave differently from the large majority of the data. Our outlying measure is used to detect outliers in a scatterplot. *Complex patterns* metrics capture aspects that cannot be easily categorized as any of the classes described above. Other scagnostic measures, such as Stringy, Striated, and Skinny, belong to this category.

Where it is measured. Our quality metrics can be calculated in *data space* or *image space*. In case of small-size data (less than 200 data point in each scatterplot), scagnostics can be computed directly on normalized data. For large-size data, scagnostics can be computed on binning data. This makes our quality metrics scalable.

Purpose. Purpose describes the main reason for using quality metrics, that is, what is the goal to be achieved with the metric. ScagExplorer can be used for the following purposes. *Projection* aims at finding 2D projections in which interesting patterns reside. For example, Skinny and Stringy measures are used to identify unusual correlations of variables in data. *Abstraction* aims at providing an overview on scatterplot groupings in entire dataset (each scatterplot is presented as a ball). Then, one can sample on each scatterplot cluster to see how they confine their quality metrics (see Section 4.1). *Visual mapping* aims at coloring scatterplots by their measures. This helps viewer to discern patterns of scatterplot distributions (see Section 4.2).

3.2 Data sets

In this section, we use three different datasets to demonstrate the performance of TimeSeer. The first is a series of US Employment data, and the second is a series of The World Bank 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 14 variables in the collected data: Total Nonfarm, Manufacturing, Trade and Transportation, Retail Trade, Financial Activities, Professional and Business, Education and Health, Leisure and Hospitality, Accommodation and Food, Other Services, Government, State Government, Local Government, and State Employment. For these data, we have 24,024 scatterplots with 50 data points each to examine.

The World Bank Data comprise annual statistics for 200 countries over 52 years from 1960 to 2011. The data were retrieved from `http://www.worldbank.org/`. We have collected 10 health-related variables: Age dependency ratio, Birth rate (per 1,000 people), Death rate (per 1,000 people), Health expenditure (% of GDP), Fertility rate (births per woman), Life expectancy at birth for female (years), Life expectancy at birth for male (years), Life expectancy at birth for both (years), Population growth, Unemployment rate. For these data, we have 2,340 scatterplots with 200 data points each to examine.

4. TIMESEER COMPONENTS

In the remainder of this paper, we describe four basic analytic tasks implemented in TimeSeer: overviewing, panning and zooming, brushing, and drilling-down. These tasks capture an analyst's activities when employing information visualization tools for understanding data [1].

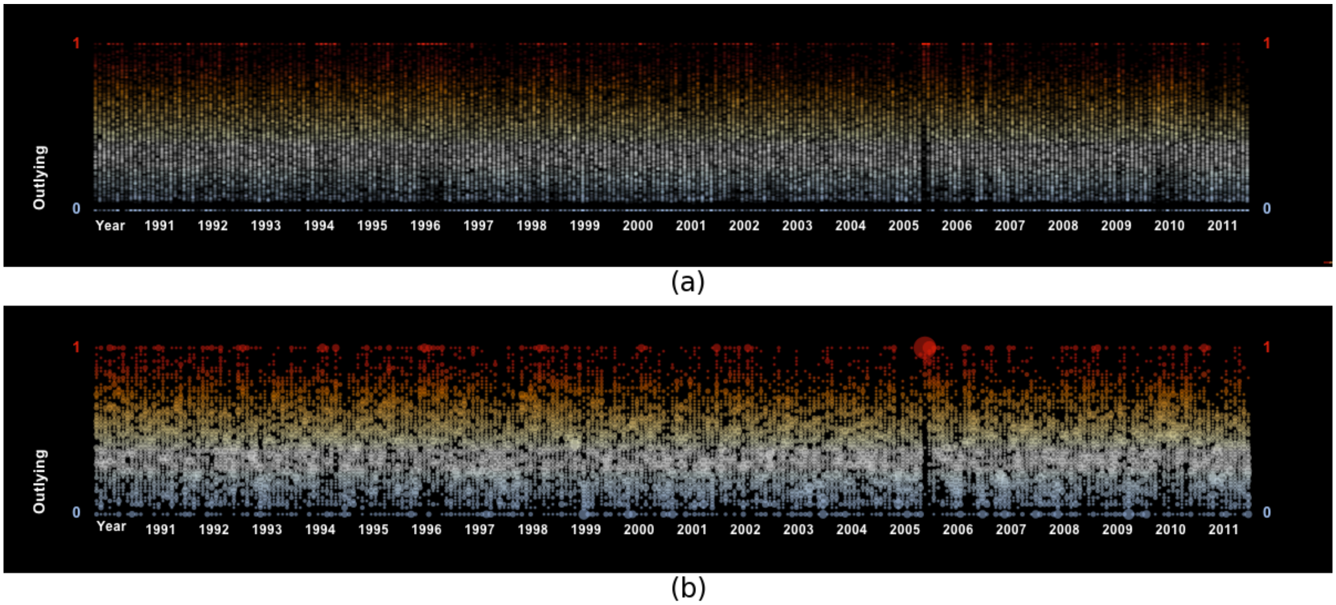


Figure 14: Outlying measure of the US Employment data: a) 2D color map b) 2D Dot Plot map.

4.1 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 the vertical axis to highlight value distributions on the selected measure.

Figure 14 shows an example on the US Employment data. In particular, we use a 2D map to present the overview of 24,024 scatterplots. The horizontal axis contains 264 months in 22 years. The vertical axis is the Outlying measure.

In Figure 14(a), every scatterplot is presented by a dot in the 2D map. High Outlying scatterplots (plots with points lying on snaky paths) are mapped to red dots, low Outlying scatterplots are mapped to blue dots. The opacity of each dot is used to highlight areas with high occurrences of dots.

Figure 14(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 [4] to achieve better location accuracy. We can see clearly that there are many scatterplots with outliers in a time point in 2005.

4.2 Pan and Zoom

After identifying an interesting time point, we want to inspect that time point. TimeSeer allows one to pan and zoom into a specific region of the overview by using a dragging box. All scatterplots in the dragging box on the overview are displayed on the top.

Figure 15 shows an example of Pan and Zoom in a time point in 2005 of the US Employment data. In particular, we zoom into August, September, and October of 2005. The scatterplot matrices in this interval are displayed on the top (high Outlying plots are in red, low Outlying plots are in blue).

Figure 16 provides a different view on the same data. In stead of scatterplot matrices, the scatterplots at each time point are displayed in a forced-directed graph. Scatterplots move to vertical levels associated to their scagnostic values. This view conforms to the overview map: most of scatterplots are high Outlying in September 2005. In this case, the outliers are Louisiana and Mississippi. Hurricane Katrina wreaked havoc on their employment and productivity figures.

4.3 Brushing

Figure 17 shows an example of Pan and Zoom into 4 years of the World Bank Data (from 1997 to 2000). Users can focus on one pair of variables by clicking on any scatterplots (the scatterplots of other pairs of variable are faded in both scatterplot matrices and the overview map). This is helpful when we want to investigate individual pairs of variables in the entire time series.

Figure 18 shows an example of brushing a pair of variables. We have selected life expectancy of male vs. female. The overview map indicates that life expectancy of male and female are highly correlated over entire time series.

Figure 19 inspects Outlying measure on the same pair of variables. The overview map indicates that there are a few intervals where we can find Outlying plots of life expectancy of male vs. female. We zoom into one of the highest outlying intervals (from 1982 to 1986). The scatterplots on the top are life expectancy of male vs. female of these years. We can check the details of each data plaint by simply clicking on it. This reveals the 3 outliers: Iran in green, Iraq in blue, and El salvador in red. The Iran-Iraq war (First Persian Gulf War) lowered the life expectancy of males because men

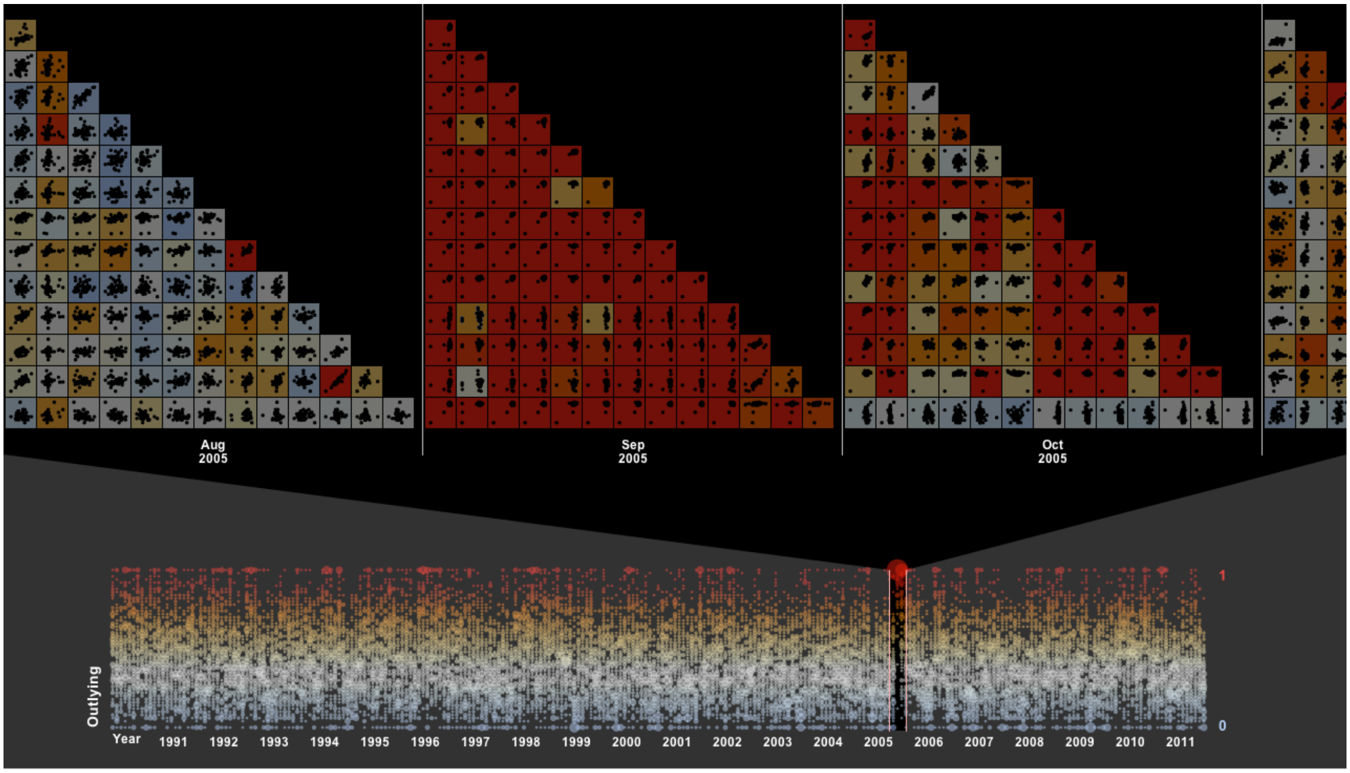


Figure 15: Outlying measure of the US Employment data: Pan and zoom in 3 months.

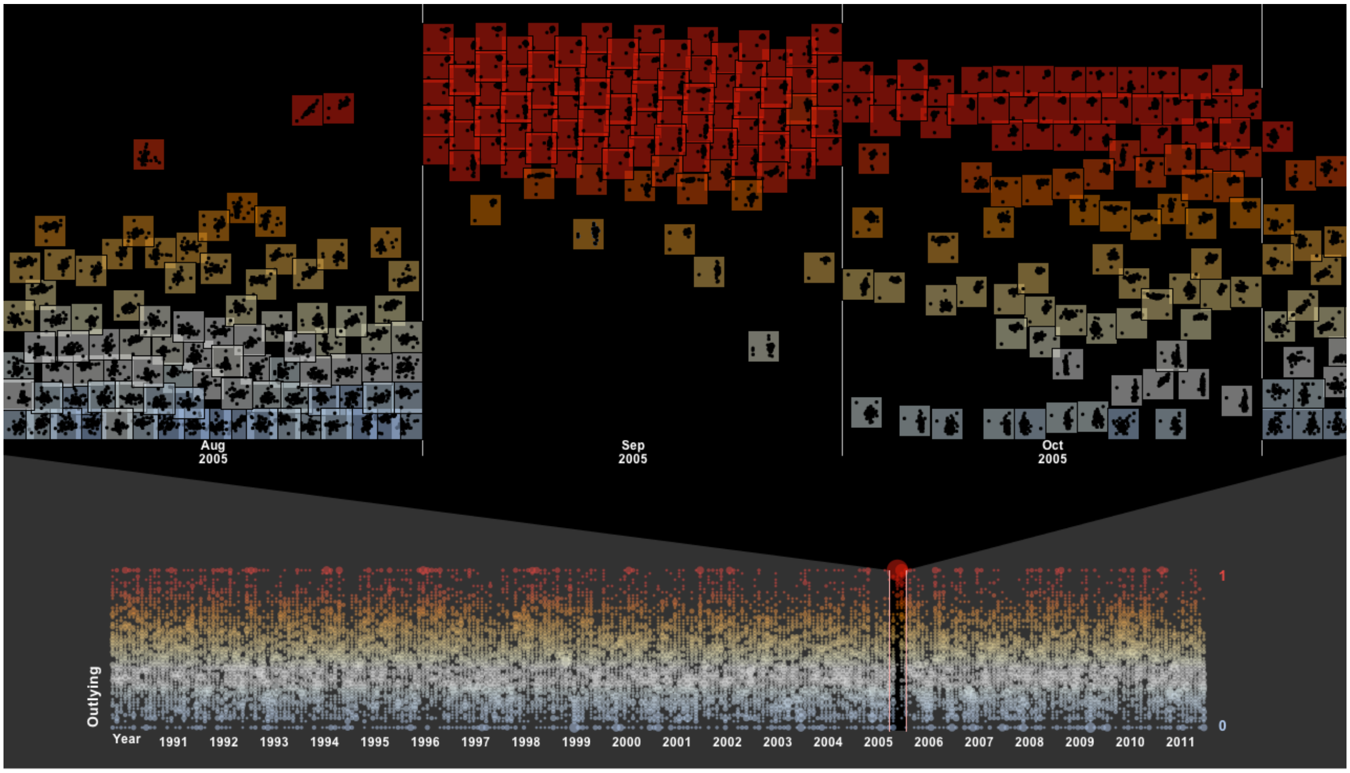


Figure 16: Outlying measure of the US Employment data: Forced-directed layout.

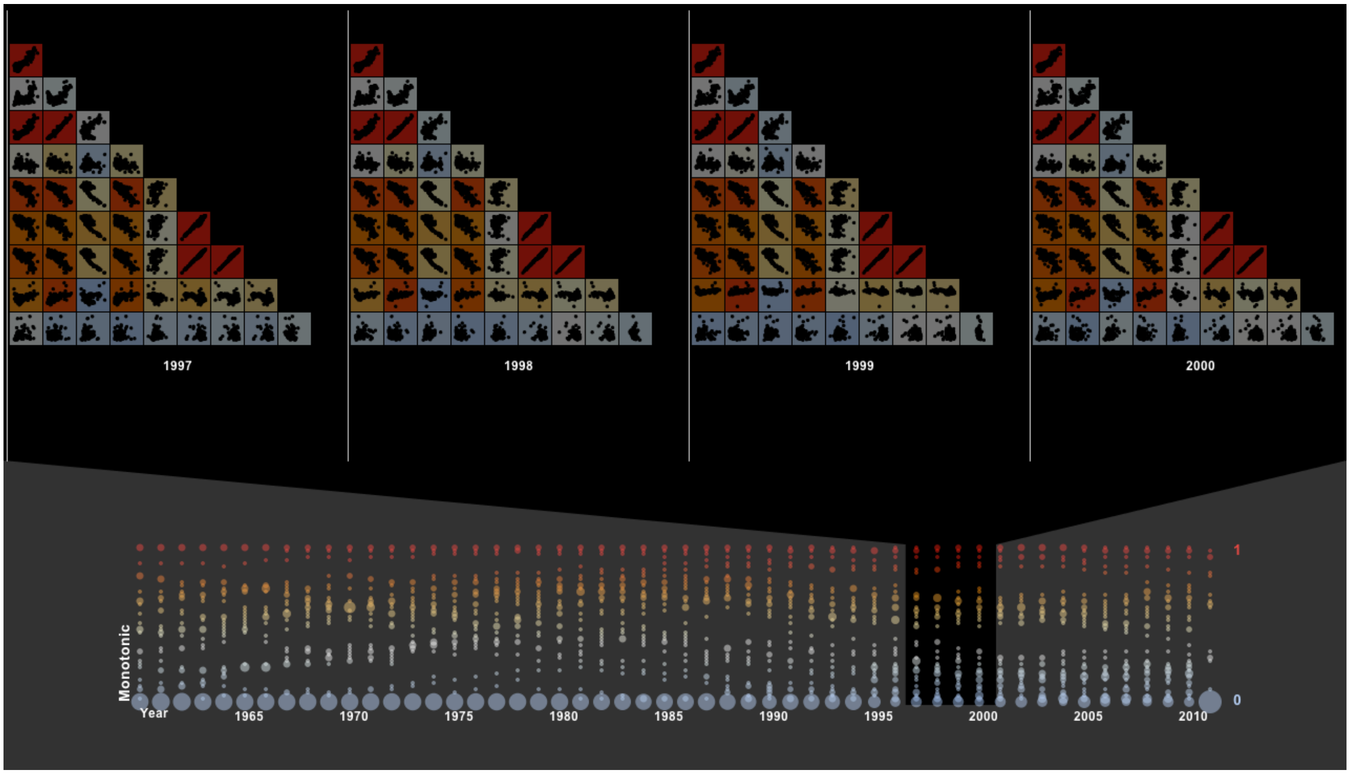


Figure 17: Monotonic measure of the World Bank Data.

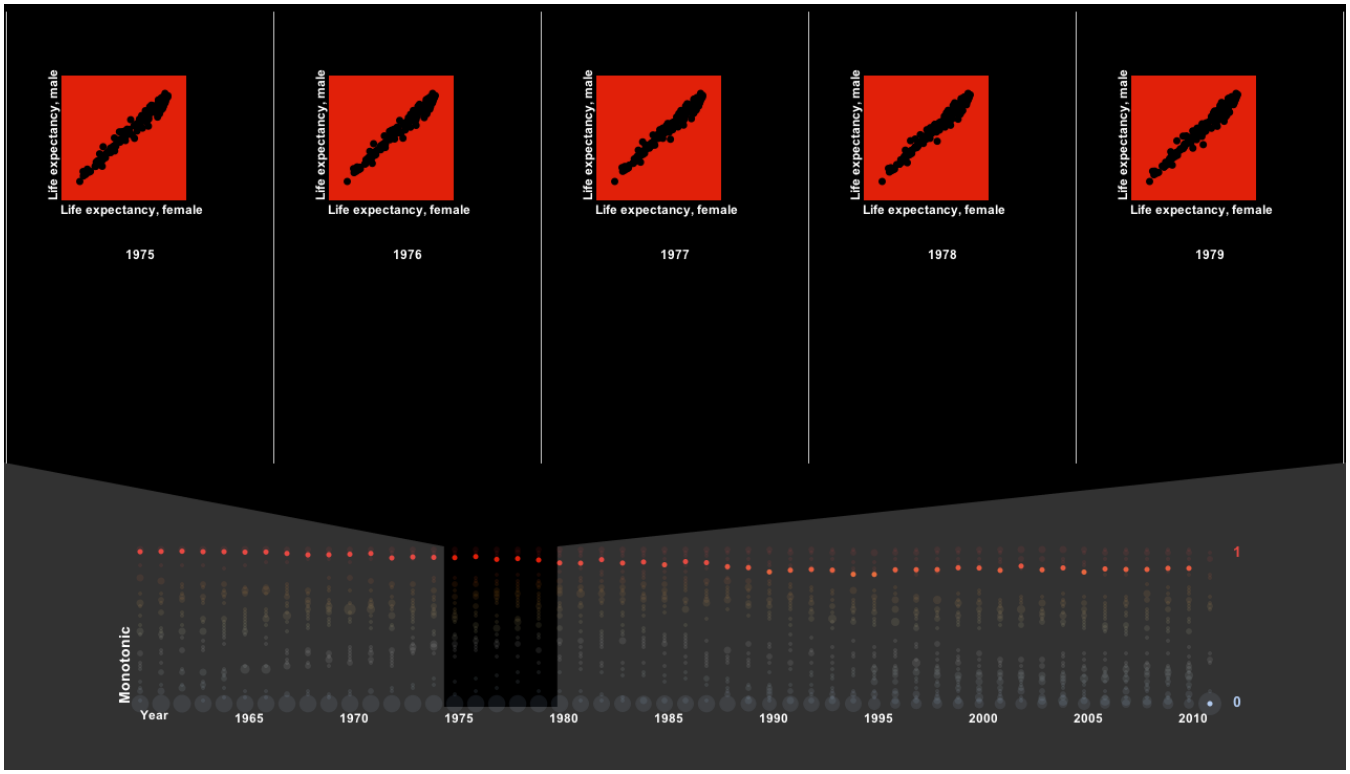


Figure 18: Monotonic measure of The World Bank Data: Selecting a pair of variables.

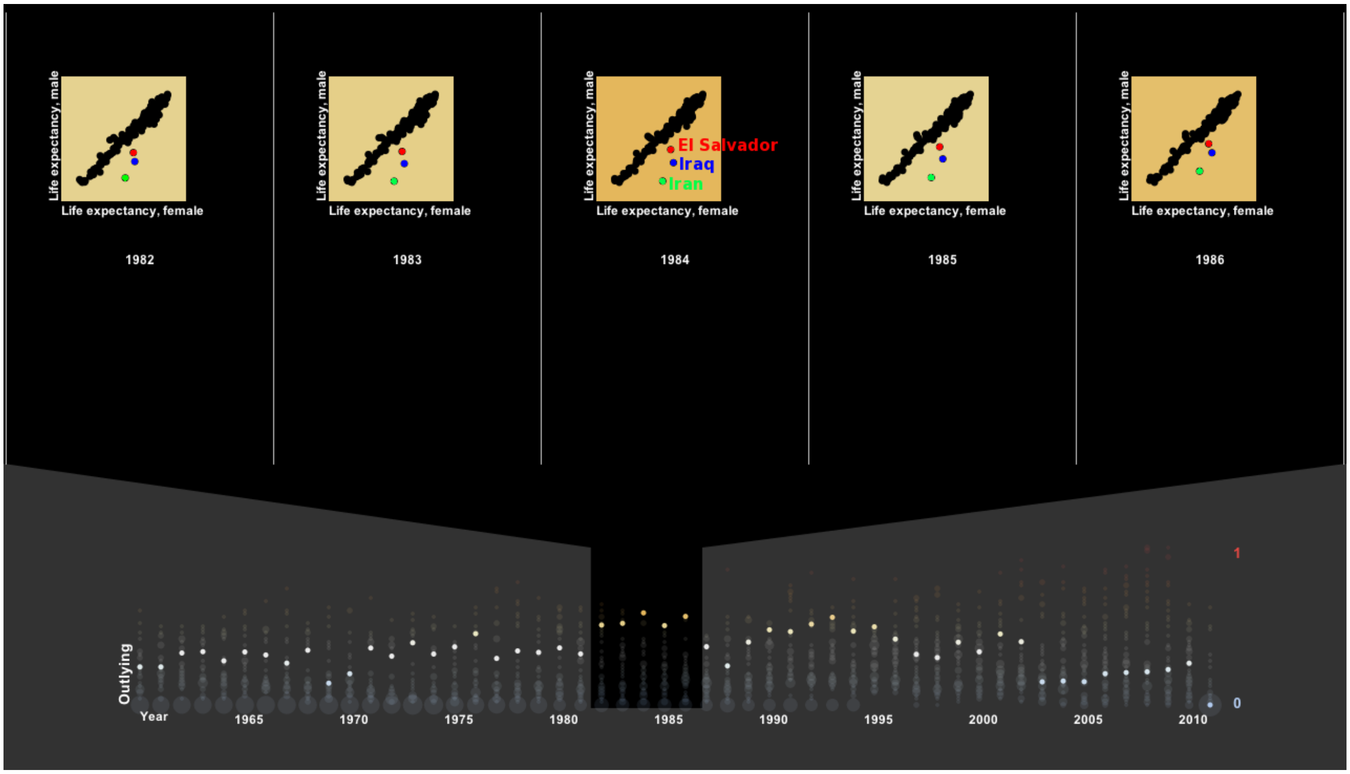


Figure 19: Outlying measure of the World Bank Data: Selecting a pair of variables.

were needed for the war. However, the life expectancy of Iranian males was much lower than that of Iraqi males. A similar situation happened to El salvador because this was the time period inside the Salvadoran Civil War (1979-1992).

4.4 Drilling-down

Now we inspect another pair of variables: Life expectancy vs. Population growth as depicted in Figure 20. The overview map indicates that we can find high Outlying plots at the first half of 1990s. We zoom into this time period (from 1990 to 1994). Then we can find a very high Outlying plot in 1991 with 3 outliers: Croatia in green, Jordan in blue, and Rwanda in red. In 1991, Croatia and Jordan both had high life expectancy. While Croatia got a drop in population growth of -6% (The Croatian War of Independence was fought from 1991), Jordan got a pump in population growth of nearly 12%. Rwanda has a very different situation. The prelude to genocide from 1990 to 1994 created a drop in both Life expectancy and Population growth of Rwanda. The Rwandan Genocide was the 1994 mass murder of an estimated 800,000 people in the East African state of Rwanda.

Figure 21 shows the raw time series for Life expectancy. Figure 22 shows the raw time series for Population growth. We use the same colors of Figure 20 to encode Rwandan, Jordan, and Croatia.

5. CONCLUSIONS

Timeseer is a visual tool for analyzing high-dimensional time series. Timeseer is designed to handle the data models with t time points and p variables. For each variable, we have n series. It should be clear that Timeseer does not inspect individual raw time series. Timeseer inspects the correlations of time series. We use scagnostics to monitor these correlations.

By working on doubly-multivariate data series, Timeseer can be use to find outliers which can be very difficult to be detected by traditional approaches for time series. For example, the three outliers in the scatterplots of Figure 19 are not outliers in any 1D projections on life expectancy of male or female.

6. ACKNOWLEDGMENTS

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

7. 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] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2203–2212, Dec. 2011.

[3] 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.

[4] T. Dang, L. Wilkinson, and A. Anand. Stacking

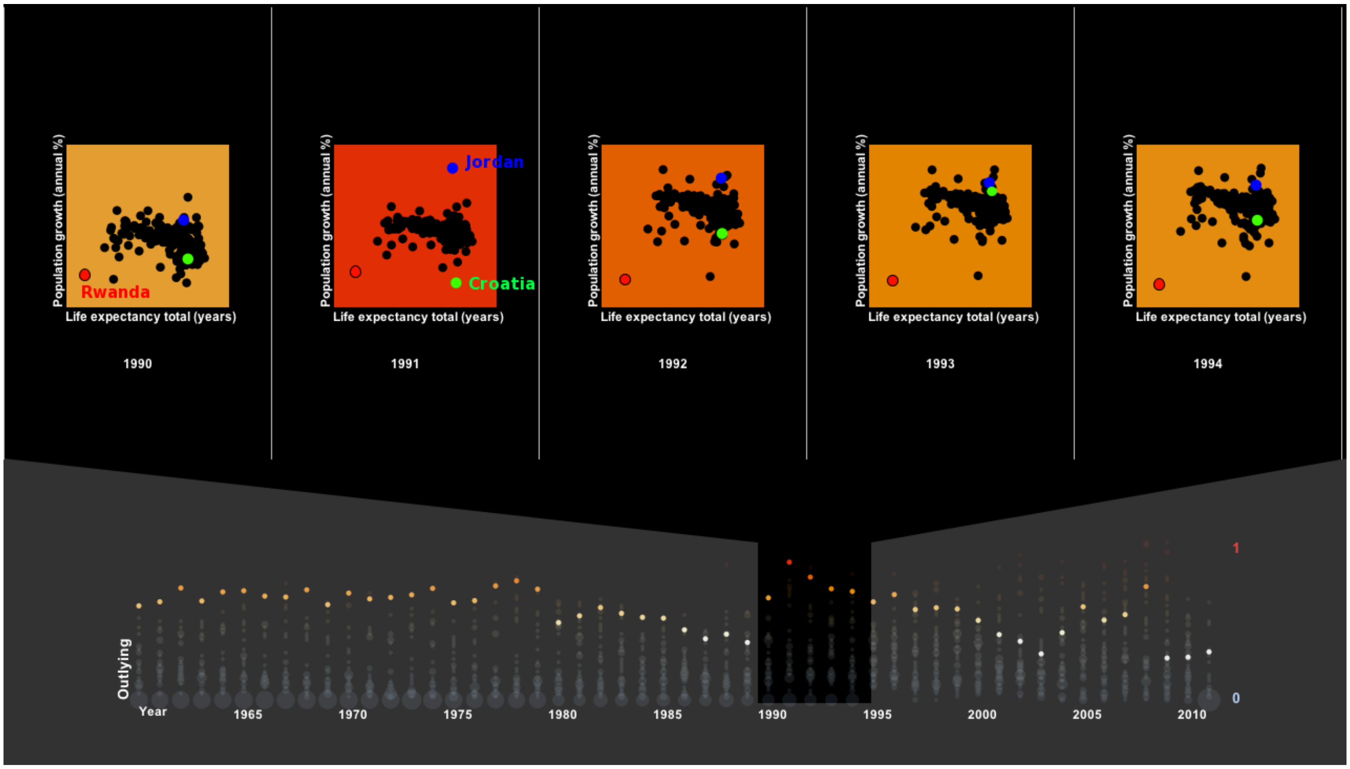


Figure 20: Outlying measure of the World Bank Data: Selecting Life expectancy vs. Population growth.

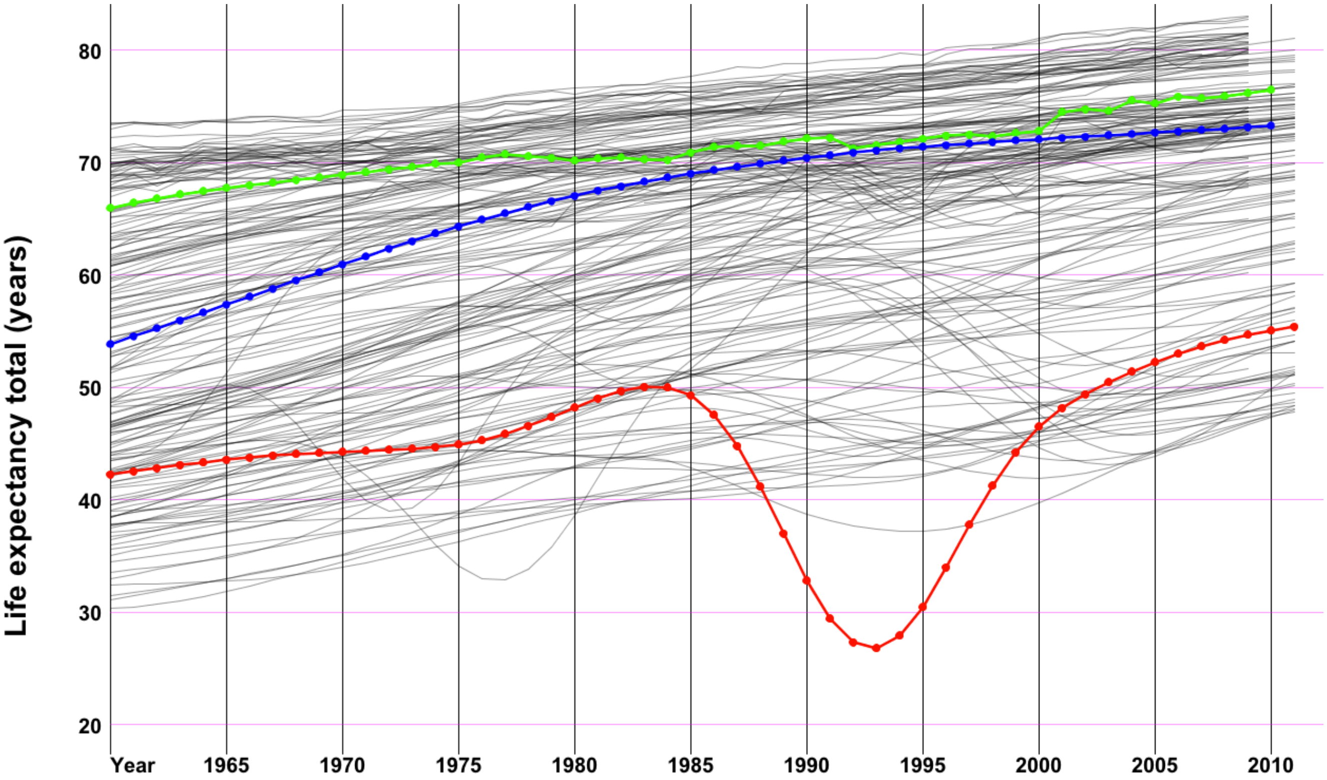


Figure 21: Raw time series for Life expectancy: Croatia in green, Jordan in blue, and Rwanda in red.

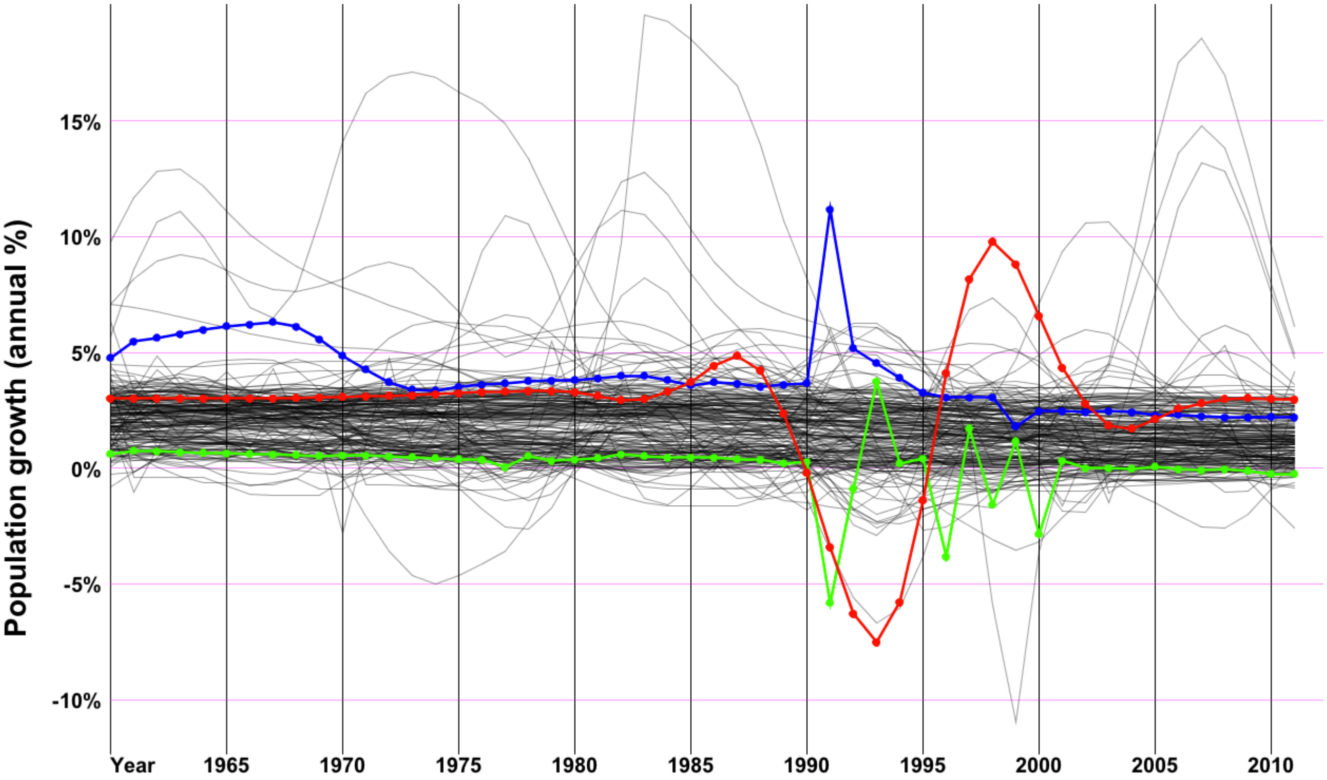


Figure 22: Raw time series for Population growth: Croatia in green, Jordan in blue, and Rwanda in red.

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.

[5] 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.

[6] 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.

[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] 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.

[13] T. von Landesberger, S. Bremm, P. Rezaei, and T. Schreck. Visual analytics of time dependent 2d point clouds. In *Proceedings of the 2009 Computer Graphics International Conference*, CGI '09, pages 97–101, New York, NY, USA, 2009. ACM.

[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.