EuroVis Workshop on Visual Analytics (2020)
K. Vrotsou and C. Turkay (Editors)

Congnostics: Visual features for doubly time series plots

Bao Nguyen[1] , Rattikorn Hewett[1], and Tommy Dang[1]
[1]Department of Computer Science, Texas Tech University

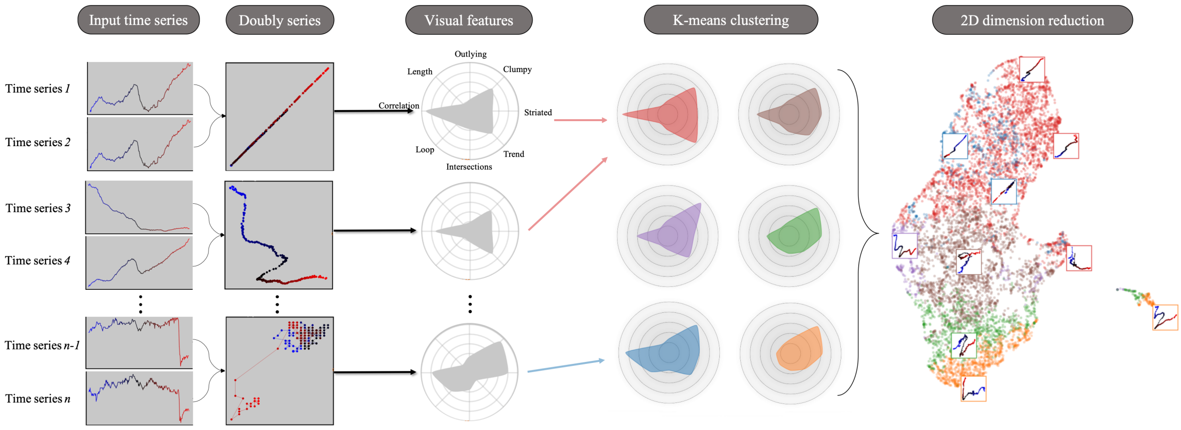


Figure 1: *Major stages in our Congnostics system: Visual feature extraction for doubly time series, K-means clustering on the visual space, and dimensional reduction using UMAP.*

Abstract

In this paper, we propose an analytical approach to automatically extract visual features from doubly time series capturing the unusual associations which are not otherwise possible by investigating individual time series alone. We have extended the visual measures for 2D scatterplots, incorporated univariate time series analysis, and proposed new visual features for doubly time series plots. These measures are discussed and demonstrated via visual examples to clarify their implications and their effectiveness. The results show that distributions, trend, shape, noise, among other characteristics, can be used to uncover the latent features and events in temporal datasets.

1. Introduction

Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces [CT05]. This analytical science provides the reasoning framework upon which tools and techniques are built based on cognitive and perceptual principles [PT04]. In other words, the human is the center of the analytical reasoning and decision-making processes. An example of such visual analytical reasoning is the Scagnostics [WAG05], which aims to capture the shape, density, correlation, and texture of data point distributions in a 2D space. These visual characterizations are computed based on the proximity graphs that are all subsets of the Delaunay triangulation: the minimum spanning tree (MST), the alpha complex, and the convex hull. These proximity graphs can be considered as the backbone analytical processes for producing visual features that capture users' interests in analyzing the 2D scatterplots. Scagnostics measures are designed to handle multivariate data series that are often found in financial sectors, social behaviors, biological experiments, among other application domains [DW14b].

Scatterplots present the data distributions in 2D space but discarding the temporal information of data points. Time series visualizations (such as line charts [DW13] or area chart [JME10]) focus on representing the temporal values of individual variables while their associations are not trivially discerned (except the obvious correlations). Doubly time series uses connected scatterplots to combine both information into a single view where consecutive data points are connected by lines [HKF16]. It is important to automatically capture the features of these connected lines for large and high-dimensional data analysis. This paper fills this gap

DOI: 10.2312/eurova.20201086

by extending the original scatterplot visual features implemented by Wilkinson et al. [WAG06] and adapts additional time-series features in the context of 2D projections. Moreover, we extend higher-level features, which are closer to users' interests. These higher-level features could be the combinations of the basic time series and/or scatterplot features. We will demonstrate these higher-level features through visual examples.

The contributions of this paper are three-fold:

- We propose an approach to characterize the visual aspects of connected scatterplots, which allow the user to embed the temporal information into the standard 2D data projections.
- We design an interactive system, called Congnostics, for visually investigating doubly multivariate data series. The visualization prototype provides summary views of the data and allows users to narrow down the event of interest by filtering our proposed visual features.
- We demonstrate our interactive interface on various contexts of time-varying data analysis. The visual examples in this paper allow users to highlight the strength of analyzing doubly time series rather than investigating these time series alone.

This paper is organized as follows: The next Section presents related research in visual features as well as popular techniques for time series analysis. Section 3 presents our visual characterization in detail. Section 4 provides the architecture of our Congnostics interactive prototype for visually mining the doubly time series data. Use cases of Congnostics are demonstrated in Section 5. Lastly, Section 6 concludes the paper and presents future direction.

2. Related Work

2.1. Time series analysis

A time series can be characterized by different measures, such as trend, seasonality, and noise. Trend, especially monotonic trend, can be detected effectively by nonparametric tests like Mann-Kandall Test [Con71, Gil87]. Seasonality can be modeled by the seasonal ARIMA model [PJB16], and there are many smoothing methods for handling noise [MMB19]. These basic univariate features are useful to detect latent events in large time series, which has application in various domains of social and scientific applications [HS04, DW13]. A common method for analyzing the relations between two-time series is the cross-correlation. This method computes the correlation coefficient, which is usually the Pearson correlation, between a pair of time series at a certain time lags. The time lag, or offset, indicates a delay in responding to the follower series under the impacts of the leader series. The true-time lag is the one at which the cross-correlation coefficient reaches its highest value [Cha96], and the two time series are most synchronized at that time. To find the best alignment between two time series, there is a typical technique, namely Dynamic time warping, which computes the minimum distance between two sequences [Mul07].

2.2. Multivariate analysis and feature extraction

Multivariate analysis has a long and extensive history, including statistical feature extraction [WWW07] or dimensional reduction methods, such as Principal Component Analysis [WEG87] and subspace analysis [PHL04, BGS07]. The Rank-by-feature technique [SS05] standard statistical summaries, such as means, standard deviations, correlations, etc., on bivariate distributions and allows the user to explore the input data in 2D clustered projections. ScagExplorer [DW14a] provides a comprehensive overview of high dimensional data set by presenting typical pair-wise distributions after applying k-means clustering using Euclidean distance on their visual space. Nguyen et al. [NPTS17] propose Extended Frobenius norm [YS04] to replace Euclidean distance for multivariate time series. A thorough review of feature extraction techniques for multivariate data can be found in more recent research [BTK11, SA15, AEM11].

2.3. Scagnostics

Scagnostics (or Scatterplot Diagnostics) were designed to discern meaningful patterns of data points in large collections of pairwise projections. The nine *Scagnostics* measures are named *Outlying*, *Skewed*, *Clumpy*, *Dense*, *Striated*, *Convex*, *Skinny*, *Stringy*, and *Monotonic*. These features are computed based on the proximity graphs that are all subsets of the Delaunay triangulation: the minimum spanning tree (MST), the alpha complex, and the convex hull. Figure 2 shows some example scatterplots and their *Scagnostics*. In particular, each column contains three example scatterplots associated to a *Scagnostics* measure with different scores from low to high. All *Scagnostics* are standardized on the unit interval. The implementation of *Scagnostics* are described in detail in [WAG06].

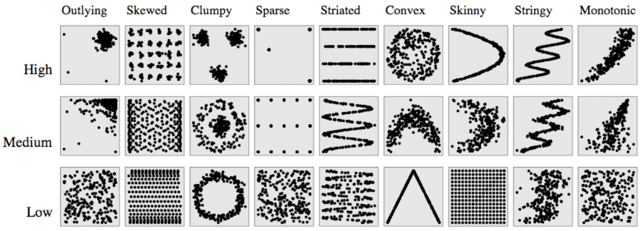


Figure 2: *Sample scatterplots and their Scagnostics measures: Scatterplots with high scores on the associated feartures are on the first row while scatterplots with low scores are at the bottom row.*

3. Visual features for doubly time-series plot

We start with normalizing the input data and then compute our proposed visual features on the 2D normalized space. In Congnostics features, four are extended directed from the most applicable scatterplot measures [DW14b] while others are extracted and extended from the time series literature. The measures are standardized on the unit scale for convenient comparisons across measures/series. The coordinates (x_t, y_t) of doubly time series at each time point are considered as vertices while the link between consecutive vertices is considered as edges in our computations as follows.

- **Outlying**: The first visual feature that we concern in our list is Outlying. Outliers are determined by box-plot rule [MTL78, NIS13]. Then, the Outlying score is computed as the ratio of the sum of all Outlying edge lengths vs. the total edge lengths [WAG06] of the doubly time series.

- **Clumpy**: Similar to the 2D scatterplot, a doubly time series can form temporal clusters. These clusters indicate a sharp change in at least one variable after a stable period. The relative positions of the clusters point out which variable suddenly drops or rises, and also provide an estimation of the changes. This measure is constructed by taking into account two runt sets derived from each edge in the series. A runt graph (R_i) is the longer of two subsets of shorter edges, in comparison to edge e_i, that are still connected to e_i. In the next equation, j denotes the edge e_j in the runt graph of edge e_i.

$$c_{clumpy} = \max_i \left\{ 1 - \max_j \{ length(e_j) \} / length(e_i) \right\}$$

- **Striated**: Striated measures the relative smoothness [WW08] in the time series. N is the number of time points.

$$c_{striated} = \frac{1}{N-2} \sum_{i=1}^{N-2} I(\cos\theta_{e_i,e_{i+1}} < -0.75)$$

- **Correlation**: We use the absolute value of the Pearson coefficient for scoring the correlations between two time series.
- **Intersection**: Edge crossing is an interesting feature of the doubly time series, which shows that both variables repeat to their earlier values. This may translate as at least one series in the doubly plot has clear noise. The *Intersections* gives a normalized score for the number of crossing ($N_{intersection}$) and depicts the randomness in the doubly series.

$$c_{intersection} = 1 - e^{-\frac{N_{intersection}}{N-1}}$$

- **Circular**: This is the measure that we had a lot of difficulties to compute since there is no algorithms in literature to formally define/score the circular patterns in doubly time series. We proposer to score this measure as follows.

$$c_{circular} = \max_k \left\{ \frac{O_k}{C_k} \times \frac{A_{c,k}}{A_{s,k}} \right\} \times n_c$$

where O_k is the number of obtuse angles in the k^{th} circle whose the number of points is C_k and area is $A_{c,k}$. The $A_{s,k}$ is the area of the smallest square that can cover the corresponding circle, and n_c is the number of circles in the plot.

- **Trend**: Trend can be detected by an extension of the Mann Kendall test. Every pair of vertices form a vector whose direction indicates the temporal order. These vectors can be classified by their directions and are divided into four groups corresponding to four quadrants. If the plot has a clear trend, most vectors tend to have similar direction, and as a result, there will be one group whose size is much larger than the others. The standardized score for the trend is given the following formula. N_k is the number of vectors in group k^{th}.

$$c_{trend} = \frac{4}{3} \frac{\max\{N_k\}}{N(N-1)/2} - \frac{1}{3}$$

- **Length**: This measure is the mean edge length, and it estimates the amplitude of the change rates in values of two time series.

To sum up, Figure 3 gives some examples of the high, the medium, and the low values of each measure. The plots are from real datasets, discussed later in Section 5.

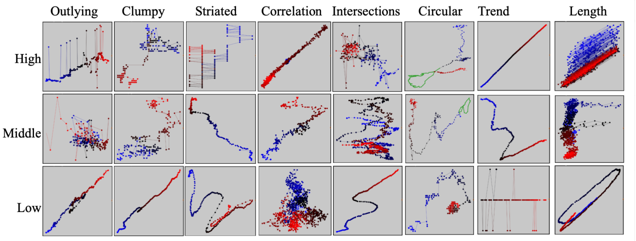


Figure 3: *Some examples of high values, medium values, and low values for the proposed measures. Color depicts temporal order: from blue to red. Green loops highlight circular patterns in the plot.*

4. Visualization process

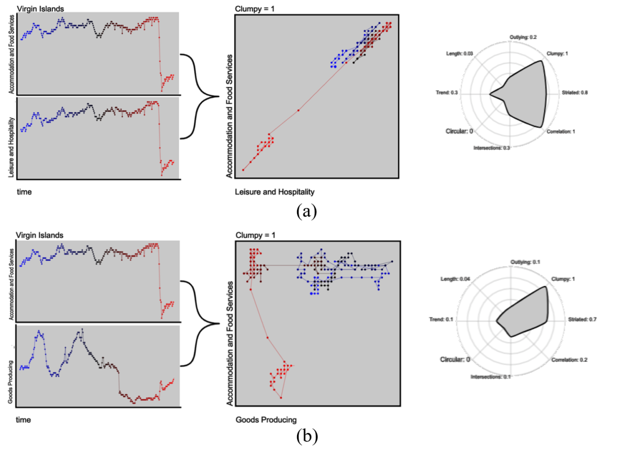
Figure 1 depicts the schematic overview of our Congnostics visualization. After computing the visual measures for every doubly time series, we apply the k-means algorithm [Har75] to cluster these doubly plots on their feature space into a pre-defined number of groups [JMF99]. In our web prototype, the maximum number of iterations is set as a user input to stop the clustering process. Classification helps to estimate the major patterns of timed data points in the doubly series. Finally, UMAP [MHM18] dimension reduction technique is utilized to project the dataset to two-dimensional space so that a large dataset can be visualized in only one view. We choose this technique because it can help to reconstruct the global structure of datasets in the visual features space [BMH*19, AJXW19].

5. Use cases

5.1. US employment data

In this section, we invest in the monthly US employees in 20 different economic sectors for each state over 21 years, from 1999 to 2019. This dataset has 8,570 doubly time series due to the missing time series (for a few states). The data was downloaded from the Bureau of Labor Statistics website: https://www.bls.gov/data/.

Clumpy: Figure 4 shows two high *Clumpy* plots, both belong to Virgin Islands. The transition between these clusters indicates a significant change in one or both series. We can infer the changes based on the relative position of these clusters regarding to temporal order. In Figure 4(a), the first cluster is on the top-right corner of the plot, and the second (red) one is at the bottom-left area. The relative position between them hints sharp drops in both economy sectors: *Accommodation and Food Services* vs. *Leisure and Hospitality*. The employment drop was due to the impact of the hurricanes Irma and Maria in 2017. In Figure 4(b), the first two clusters (one in the top-right corner and another is in the top-left corner) implicate that only the number of employment in Goods Producing decreased dramatically while *Accommodation and Food Services* maintains its numbers. By investigating the input time series on the left, we can see that the drops happened in 2011 when this state experienced a catastrophic recession, and its largest oil company shut down in 2012 [Ygl13]. Accommodation and Food Services were not affected by the recession, but it plummeted in late 2017 due to the hurricanes mentioned above, so there is a small cluster in the bottom-left of the plot. The relative position between the cluster in the top-left region and the one in the bottom-left area implies that the hurricanes did not affect the *Goods Producing* sector.



(a)

(b)

Figure 4: *Clusters in doubly time series of Virgin Islands in sectors: (a) Accommodation and Food Services vs. Leisure and Hospitality, (b) Accommodation and Food Services vs. Goods Producing.*

Circular pattern: A circular pattern in doubly time series indicate an offset (or time shift) between peaks or valleys of the pair of variables. Figure 5 shows an example of seasonal pattern in a plot, highlighted by green and orange, in *Total Private* vs. *Financial Activities* of Massachusetts. The green one relates to the 2001 recession [NBE10], which started in March and ended in November, while the orange one depicts the Great Recession of 2008 [**?**]. Because the crisis in 2008 initially affected financial companies, the drop in the number of employees in this sector occurred several months before a similar situation happened in *Total Private*. The time shift between the drops indicates the relation of two sectors, and in this particular case, the crisis in finance had an impact on the activities of other private companies. This offset also points out how fast the crisis in *Financial activities* affected the *Total Private*. This interesting relation can not be captured by analyzing individual time series.

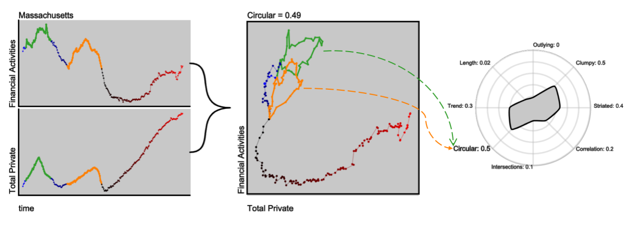
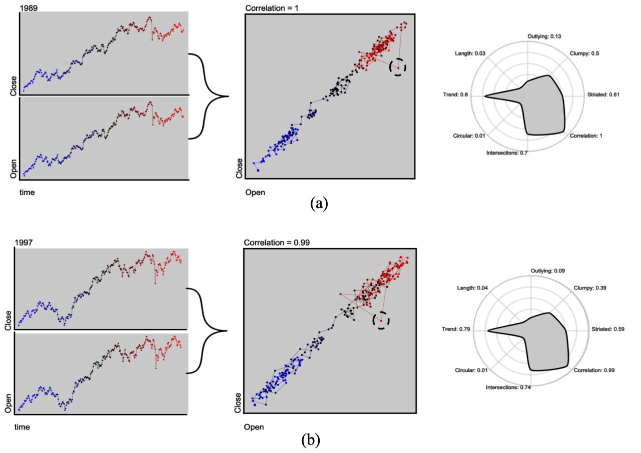


Figure 5: *Two circular patterns are detected in the doubly time series of Massachusetts: Total Private vs. Financial Activities.*

5.2. SP 500 data

This use case is the stock data from 1980 to 2018. We focus on the yearly time series of *open* and *close* values of SP 500. Each time series has around 247 data points (no stock data on Saturday and Sunday). Figure 6 displays the doubly series of daily open price and closing price of the SP 500 index in 1989 and 1997. The two variables are strongly correlated, as shown on the *Trend* dimension of the radar chart. In particular, most data points located near the diagonal. The outliers, which are highlighted within dash circles, indicate the date where SP 500 *open* and *close* value are significantly different. These outliers correspond to mini-crashes in the 13^{th} of October 1989 and the 27^{th} of October in 1997. The former crash corresponds to leveraged buyout deal for UAL Corporation [IDM02] while the latter caused by an economic crisis in Aisa [SC99]. Notice that these unusual data points are not evident in the marginal distributions on the left of Figure 6. Due to the dynamic nature of stock data, many line crossings are visible in the doubly time series plots. These noisy features can be captured effectively by our *Intersection* measure, as depicted on the radar chart on the right. The scores are 0.7 and 0.74 correspondingly.



(a)

(b)

Figure 6: *Correlation between open and close index of SP 500 data in (a) 1989 and (b) 1997. The dashed circles highlight outlying dates in the open vs. close correlations.*

6. Conclusion

This paper has proposed a set of visual features for doubly time series. These features aim to automatically characterize pair-wise relations of variables in multivariate time series. Two real-world datasets have been used to demonstrate our capability of capturing the visual patterns (such as loops), which are difficult or even impossible by investigating marginal time series. Although there are some drawbacks of our feature descriptors when the data is too noisy, providing a holistic overview of multivariate time series through the visual feature space of doubly time series is an interesting research direction. We will investigate the approaches to stabilize our feature descriptors, such as using Convolution Neural Network [MTW*20] with multiple layers [LBB*98]. Congnostics is implemented as a JavaScript-based web application using D3.js [BOH11]. The demo video, web application, and source codes of our visualization are available on our Github project at https://idatavisualizationlab.github.io/B/congnostics/.

References

[AEM11] ALBUQUERQUE G., EISEMANN M., MAGNOR M.: Perception-based visual quality measures. In *IEEE VAST* (2011), pp. 13–20. 2

[AJXW19] ALI M., JONES M. W., XIE X., WILLIAMS M.: Timecluster: dimension reduction applied to temporal data for visual analytics. *The Visual Computer 35*, 6-8 (2019), 1013–1026. 3

[BGS07] BERTINI E., GIROLAMO A. D., SANTUCCI G.: See What You Know: Analyzing Data Distribution to Improve Density Map Visualization. In *Eurographics/ IEEE-VGTC Symposium on Visualization* (2007), Museth K., Moeller T., Ynnerman A., (Eds.), The Eurographics Association. 2

[BMH*19] BECHT E., MCINNES L., HEALY J., DUTERTRE C.-A., KWOK I. W., NG L. G., GINHOUX F., NEWELL E. W.: Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology 37*, 1 (2019), 38. 3

[BOH11] BOSTOCK M., OGIEVETSKY V., HEER J.: D^3 data-driven documents. *IEEE Transactions on Visualization & Computer Graphics*, 12 (2011), 2301–2309. 4

[BTK11] BERTINI E., TATU A., KEIM D.: Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2203–2212. 2

[Cha96] CHATFIELD C.: *The analysis of time series*. Chapman & Hall, 1996. 2

[Con71] CONOVER W.: *Practical Nonparametric Statistics*. Wiley, 1971. 2

[CT05] COOK K., THOMAS J.: *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, vol. 54. 05 2005. 1

[DW13] DANG T. N., WILKINSON L.: Timeexplorer: Similarity search time series by their signatures. In *International Symposium on Visual Computing* (2013), Springer, pp. 280–289. 1, 2

[DW14a] DANG T. N., WILKINSON L.: Scagexplorer: Exploring scatterplots by their scagnostics. In *2014 IEEE Pacific Visualization Symposium* (March 2014), pp. 73–80. 2

[DW14b] DANG T. N., WILKINSON L.: Transforming scagnostics to reveal hidden features. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec 2014), 1624–1632. 1, 2

[Gil87] GILBERT R. O.: *Statistical Methods for Environmental Pollution Monitoring*. Van Nostrand Reinhold, New York, 1987. 2

[Har75] HARTIGAN J. A.: *Clustering Algorithms*, 99th ed. John Wiley & Sons, Inc., New York, NY, USA, 1975. 3

[HKF16] HAROZ S., KOSARA R., FRANCONERI S. L.: The connected scatterplot for presenting paired time series. *IEEE Transactions on Visualization and Computer Graphics 22*, 9 (Sep. 2016), 2174–2186. 1

[HS04] HOCHHEISER H., SHNEIDERMAN B.: Dynamic query tools for time series data sets: Timebox widgets for interactive exploration. *Information Visualization 3*, 1 (Mar. 2004), 1–18. 2

[IDM02] ISLAM R., DJANKOV S., MCLEISH C.: *The right to tell: the role of mass media in economic development*. The World Bank, 2002. 4

[JME10] JAVED W., MCDONNEL B., ELMQVIST N.: Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics 16*, 6 (Nov. 2010), 927–934. 1

[JMF99] JAIN A. K., MURTY M. N., FLYNN P. J.: Data clustering: A review. *ACM Comput. Surv. 31*, 3 (Sept. 1999), 264–323. 3

[LBB*98] LECUN Y., BOTTOU L., BENGIO Y., HAFFNER P., ET AL.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324. 4

[MHM18] MCINNES L., HEALY J., MELVILLE J.: Umap: Uniform manifold approximation and projection for dimension reduction, 2018. 3

[MMB19] MCDOWALL D., MCCLEARY R., BARTOS B. J.: *Interrupted time series analysis*. Oxford University Press, 2019. 2

[MTL78] MCGILL R., TUKEY J. W., LARSEN W. A.: Variations of box plots. *The American Statistician 32*, 1 (1978), 12–16. 2

[MTW*20] MA Y., TUNG A. K. H., WANG W., GAO X., PAN Z., CHEN W.: Scatternet: A deep subjective similarity model for visual analysis of scatterplots. *IEEE Transactions on Visualization and Computer Graphics 26*, 3 (March 2020), 1562–1576. 4

[Mul07] *Dynamic Time Warping*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 69–84. 2

[NBE10] NBER: Us business cycle expansions and contractions. report, September 2010. Retrieved March 2, 2020 from https://www.nber.org/cycles.html. 4

[NIS13] NIST/SEMATECH: e-handbook of statistical methods. e-handbook, October 2013. Retrieved January 3, 2020 from https://www.itl.nist.gov/div898/handbook/index.htm. 2

[NPTS17] NGUYEN M., PURUSHOTHAM S., TO H., SHAHABI C.: m-tsne: A framework for visualizing high-dimensional multivariate time series. *arXiv preprint arXiv:1708.07942* (2017). 2

[PHL04] PARSONS L., HAQUE E., LIU H.: Subspace clustering for high dimensional data: a review. *SIGKDD Explor. Newsl. 6*, 1 (June 2004), 90–105. 2

[PJB16] PETER J. BROCKWELL R. A. D.: *Introduction to Time Series and Forecasting*. Springer, 2016. 2

[PT04] PAK CHUNG WONG, THOMAS J.: Visual analytics. *IEEE Computer Graphics and Applications 24*, 5 (Sep. 2004), 20–21. 1

[SA15] SEDLMAIR M., AUPETIT M.: Data-driven Evaluation of Visual Quality Measures. *Computer Graphics Forum* (2015). 2

[SC99] SECURITIES U., COMMISSION E.: Trading analysis of october 27 and 28, 1997. article, July 1999. Retrieved March 2, 2020 from https://www.sec.gov/news/studies/tradrep.htm. 4

[SS05] SEO J., SHNEIDERMAN B.: A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization 4*, 2 (July 2005), 96–113. 2

[WAG05] WILKINSON L., ANAND A., GROSSMAN R.: Graph-theoretic scagnostics. 1

[WAG06] WILKINSON L., ANAND A., GROSSMAN R.: High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Transactions on Visualization and Computer Graphics 12*, 6 (2006), 1363–1372. 2

[WEG87] WOLD S., ESBENSEN K., GELADI P.: Principal component analysis. *Chemometrics and intelligent laboratory systems 2*, 1-3 (1987), 37–52. 2

[WW08] WILKINSON L., WILLS G.: Scagnostics distributions. *Journal of Computational and Graphical Statistics 17*, 2 (2008), 473–491. 3

[WWW07] WANG X., WIRTH A., WANG L.: Structure-based statistical features and multivariate time series clustering. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)* (Oct 2007), pp. 351–360. 2

[Ygl13] YGLESIAS M.: The u.s. virgin islands are in a catastrophic recession. website, August 2013. Retrieved January 30, 2020 from https://slate.com/business/2013/08/virgin-islands-recession.html. 3

[YS04] YANG K., SHAHABI C.: A pca-based similarity measure for multivariate time series. In *Proceedings of the 2nd ACM international workshop on Multimedia databases* (2004), pp. 65–74. 2