Visual Features for Multivariate Time Series

Bao Dien Quoc Nguyen
Department of Computer Science
Texas Tech University
Lubbock, Texas, USA
Bao.D.Nguyen@ttu.edu

Rattikorn Hewett
Department of Computer Science
Texas Tech University
Lubbock, Texas, USA
rattikorn.hewett@ttu.edu

Tommy Dang
Department of Computer Science Texas Tech University
Lubbock, Texas, USA
tommy.dang@ttu.edu

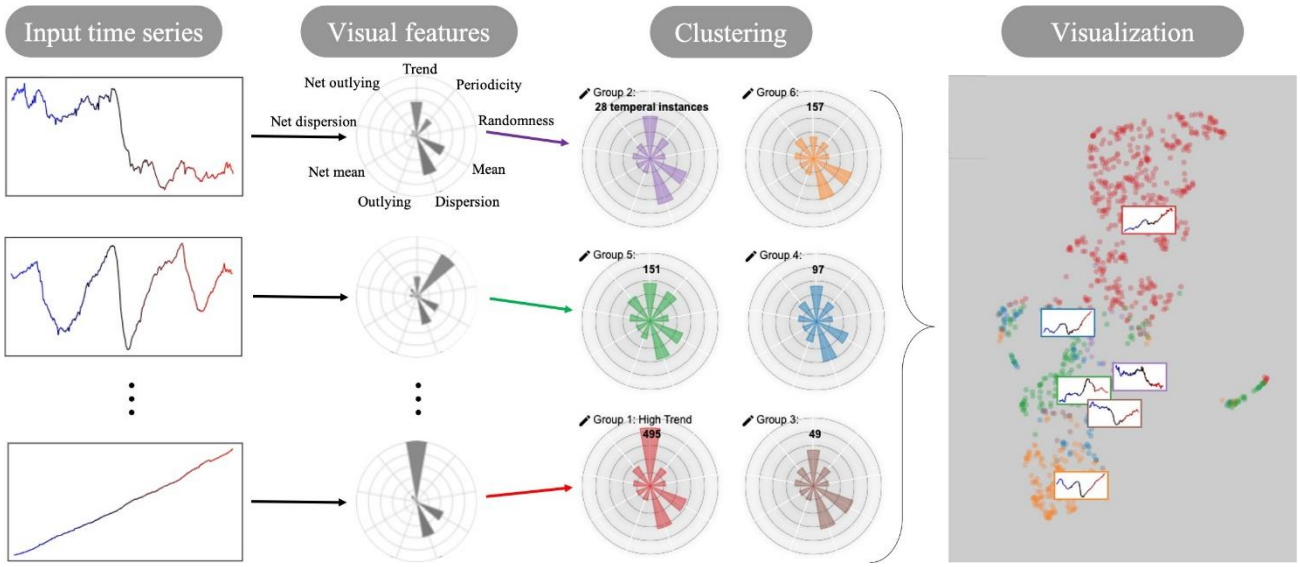

Figure 1: Schematic overview of the paper: visual features extractions for input time series, k-means clustering in visual feature space, and dimension reduction for exploration of the dataset.

ABSTRACT

Visual analytics combines the capabilities of computers and humans to explore the insight of data. It provides coupling interactive visual representations with underlying analytical processes (e.g., visual feature extraction) so that users can utilize their cognitive and reasoning capabilities to perform complex tasks effectively or to make decisions. This paper applies successfulness of visual analytics to multivariate temporal data by proposing an interactive web prototype and an approach that enables users to explore data and detect visual features of interest. A list of nonparametric quantities is proposed to extract visual patterns of time series as well as to compute the similarity between them. The prototype integrates visualization and dimensional reduction techniques to support the exploration processes. Many different temporal datasets are used to justify the effectiveness of this approach, and some remarkable results are presented to show its value.

CCS CONCEPTS

• Human-centered computing → Visual analytics.

KEYWORDS

visual features extraction, clustering method, dimension reduction

ACM Reference Format:
Bao Dien Quoc Nguyen, Rattikorn Hewett, and Tommy Dang. 2020. Visual features for multivariate time series. In *Proceedings of International Conference on Advances in Information Technology (IAIT2020), July 1-3, 2020, Bangkok, Thailand. ACM, New York, NY, USA, 8 pages.* https://doi.org/10.1145/3406601.3406621

1 INTRODUCTION

Multivariate temporal datasets usually contain hundreds or thousands of time series. The time series in these large datasets may have many different characteristics, and their number of time steps may significantly vary from tens to thousands. These complexities make it infeasible for humans to manually analyze each at a time. One typical method for the large temporal datasets is feature-based analysis [23, 34]. Many works have investigated the use of distinctive lists of features of time series for classification or query [16, 28].

IAIT2020, July 1–3, 2020, Bangkok, Thailand

ACM ISBN 978-1-4503-7759-1/20/07…$15.00
https://doi.org/10.1145/3406601.3406621

For instance, one recent work integrates thousands of features into a software tool, namely hctsa [16]. This tool provides a broad and interdisciplinary methodological literature. However, many of these features require specific knowledge of statistics and math to under- stand. Thus, it is difficult for a majority of users who are unfamiliar with some concepts such as entropy or stationary time series to interpret them. In this work, we aim to overcome this issue by using a list of visual characterizations as well as an interactive interface that supports users to navigate and explore large temporal datasets.

Contribution of this work is listed as the following:

- A list of nine metrics is proposed to extract the visual features of the time series. These metrics are discussed and utilized as dimensions for computing similarity between time series. Applying some well-known dimensional reduction tech- niques such as Principal Component Analysis (PCA), t-Distri- buted Stochastic Neighbor Embedding (tSNE), and Uniform Manifold Approximation and Projection (UMAP) to project the temporal dataset from visual feature space to lower di- mensional space for displaying. This approach allows human- centered exploration of the dataset.
- Introduce a web prototype that integrates visualization tech- niques for users to explore the data.

This paper starts with some related works in Section 2, before introducing the list of metrics in Section 3, dimensional reduction techniques in Section 4, and an interactive web prototype in Section
5. After that, some use cases are discussed in Section 6. Finally, summary and conclusion are given in Section 7.

2 RELATED WORKS

In this section, we do not attempt to survey all time series analysis techniques since there are too many. Instead, we focus on the most relevant research on our proposed visual method. We will start with the feature-based time series analysis.

2.1 Feature extraction for time series

One of the first efforts in feature-based analysis of time series is the work of Nanopoulos et al. [34]. They utilized the mean, standard deviation, skewness, and kurtosis of the first-order and second-order features to form the feature vector for each time series for classification purposes. While mean and standard deviation are two well-known statistics of time series, the other two are higher-order statistical metrics. The skewness can indicate time series with many small values and few large values or the contrary visual context with many high values and few low values. Besides, kurtosis describes the 'peakedness' of the probability density function of time series values [13]. Our work ignores these two statistics because the visual aspect of skewness can be illustrated by other metrics such as the mean value. The use of skewness may lead to the overlap of the utilization of different metrics in our proposed list. In the case of kurtosis, it is obliquely to point out its visual information for the time series values so that it may confuse users if we implement this statistic.

The outlier is a crucial characteristic of any dataset because it is able to provide valuable information about the data such as measurement errors, incorrect distributional assumptions, or so on [21]. There are many different definitions of outliers. One classic and commonly accepted definition is the one of Hawkins, which con- siders an outlier as an observation that deviates too much from the others [19]. Tsay [41] discussed and handled distinct types of out- liers such as additive outlier (AO) and innovational outlier (IO), or structure changes like level shift (LS) and temporal change (TC) by ARIMA model. TimeSeer [10] uses scagnostics to explore outliers of large time series data by projecting scatterplots of variables at each time step into a single display. Another popular interest is finding anomalous or surprising patterns in the temporal dataset. A pattern is anomalous if its occurrence frequency "differs substantially from that expected by chance, given some previously seen data" [24]. VizTree detects motif patterns and surprising ones by symbolic aggregate approximation (SAX), which transforms time series into strings [27]. The VizTree requires prior knowledge about the length of the motif, so Li et al. proposed the use of grammar induction to overcome the issue [26]. GrammarViz 2.0 provides a non-parameter detection of anomalies in time-series data [38].

A recent work was carried out a few years ago by Ben D. Fulcher and Nick S. Jones [15]. This work integrates a wide range of time series characteristics, including basic statistics of the distribution of time-series values, linear correlations, stationarity, information-theoretic and entropy/complexity measures, and so on. Their list of features reaches thousands of metrics for time series. This list was then combined into a software toolbox, namely hctsa [16]. The computation of all features is quite expensive, and many of them are similar, so Carl H Lubba et al. have introduced a reduced list, called Catch22 [28]. The reduction from about 7000 features to 22 ones significantly improves the time complexity of the computation while maintains the quality of the classification process. However, many of the features are not human-readable and hence, are difficult for novice users to verify and explore. Kurt. Kang et al. [23] utilized six metrics, which base on Seasonal and Trend de- composition using Loess (STL), to propose an idea of visualizing the effectiveness of proposed forecasting method in M3 dataset [30]. However, the STL decomposition has six parameters, and the choice of seasonal smoothing parameter requires prior knowledge about the dataset [9]. Our work aims to use nonparametric metrics so that users can explore a dataset without any prior knowledge about it.

2.2 Clustering and Dimension reduction techniques on the feature space

There are two main types of classification: supervised and unsupervised clustering. The latter is a crucial element in exploring insight data [22], for it provides interrelationships among data observations. One well-known clustering method is the k-means algorithm in which data is grouped into a given number of clusters [18], but this method is sensitive to the initial partition. ISODATA [3] provides a solution for this drawback by permitting splitting and merging the clusters by comparing the distance between their centers with a pre-specified threshold. Another method is leader algorithm that groups data points, whose similarity is less than a given threshold, into a cluster [18]. ScagExplorer [12] uses this method to classify scatterplots in the space of scagnostics in order to visualize some characteristic patterns of the data.

The combination of dimension reduction and additional visual encodings is able to improve the performance of visual-interactive

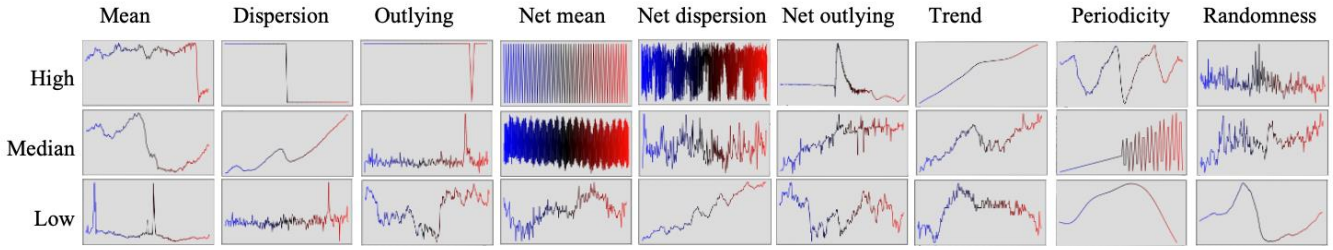

Figure 2: Some example time series that are corresponding to high, medium, and low scores for every metrics

labeling [6]. PCA [46] is a linear approach that reduces high dimensional data to a few most important dimensions. This method was applied to various time series such as financial [25], or finding similarity measure for multivariate time series [47]. T-SNE [29], a nonlinear technique, uses Gaussian distribution to define relationships between data points in high dimensional space, before re- constructing them using Student t-distribution in low dimensional space. The acceleration of this technique using the Barnes-Hut algorithm can reduce the computational and memory complexity from $O(N^2)$ to respectively $O(Nlog(N))$ and $O(N)$ [42]. UMAP is recently introduced as another nonlinear dimensional reduction technique [33]. It is shown to be comparable to the previous method [4]. TimeCluster [1] applies these dimension reduction techniques to long univariate and multi-variate time series to visualize motif and anomalies of the series, and the results show that PCA and UMAP tend to extract the good global structure of the data.

3 CHARACTERISTIC METRICS

This section is going to propose a list of metrics that are utilized to capture visual features for time series, as depicted in Figure 2. In order to compare these scores across different datasets, they are standardized to the unit interval. The input dimensions are also normalized on the unit scale.

3.1 Metrics of time series

Mean: The mean value is a classic statistic for representing the values and distribution of a dataset. Many works have utilized this well-known attribute [11, 34], and it can give visual information about the time series [11]. If a time series has a high mean value, it has many high values and a few low values. In contrast, a time series with many small values and few large values has a low mean score.

$$m_{mean} = \frac{\sum_t x_t}{N} \quad (1)$$

where N is the number of time steps in the time series.

Distribution Dispersion: Another well-known statistic is the standard deviation. This quantity describes the dispersion of the dataset. However, we cannot use this quantity directly because we aim to standardize all metrics to the unit interval for comparison purposes. Let take an example of a binary time series, which is an 'extremely stretch' dataset. Due to the normalization, two values of this series are only 0 or 1, and then, it is easy to get the standard deviation for this dataset that is a half unit. This 'extremely stretch' time series is expected to have a high score for dispersion, not a medium one. Thus, we multiply the standard deviation by two and call this metric as dispersion score.

$$m_{dispersion} = 2\sqrt{\frac{\sum_t (x_t - \bar{x})^2}{N}} \quad (2)$$

where $\bar{x}$ is the mean value of time series x_t over N time steps.

Outlying: Outliers can be detected by several methods. One is the generalized extreme Studentized deviate (ESD) test [37], which is utilized for a univariate dataset. The dataset must follow an approximately normal distribution. Another common approach is the modified Z-score method [20]; however, this score also requires the assumption of a normal distribution dataset. In order to find outliers without prior knowledge of the distribution of the data, the box-plot rule [32, 35] is suitable. This rule determines outliers by quantiles that are not affected by the existence of these outliers. We score this metric as the ratio of the sum of the absolute deviation (AD) from the median of outliers and of all time series values.

$$m_{outlying} = \frac{\sum AD_{outliers}}{\sum AD_{total}} \quad (3)$$

3.2 Metrics of the first different time series

Net mean: Absolute of the first lag difference, $d_t = |x_{t+1} - x_t|$, describes the change rate in values of time series. A representative of the change is the net mean, which is the mean value of the net series.

$$m_{net\ mean} = \frac{\sum_t d_t}{N-1} \quad (4)$$

Net dispersion: Often the first difference is a stationary se- ries [14, 35], so we assume this series, d_t, is the sum of a constant mean and a white noise that follows normal distribution $N(0, \sigma^2)$. Because 99.99% of the data in the normal distribution is within four standard deviations from the mean, we approximate that the range of the first difference equals to four standard deviations (only get the positive tail of the normal distribution). On the other hand, the absolute of the first difference of the normalized data set is between 0 and 1. We synchronize the assumption range with the normalized range to come to a fact that computed standard deviation rarely exceeds 1/4 of the unit. Therefore, we get four times the standard deviation as our net dispersion score.

$$m_{net\ dispersion} = 4 \times \sqrt{\frac{\sum_t (d_t - \bar{d})^2}{N-1}} \quad (5)$$

Next outlying: This score measures whether time series has irregular change rates that are much different from the others. We reuse the Equation 3, but for the first difference series.

$$m_{net\ outlying} = \frac{\sum AD_{net\ outliers}}{\sum AD_{net\ total}} \quad (6)$$

3.3 Metrics for extracting visual features of interest

In time series, three common characteristics are the trend, seasonal, and randomness. This sub-section mentions the metrics for these visual features.

Trend: The global trend of a time series can be scored by a nonparametric Mann-Kendall test [31]. This test compares $N(N-1)/2$ pairs of values of the time series at two distinct time steps and computes their signs, $Sign(x_i - x_j)$. The values of these signs can be 1, 0, or -1 corresponding to three cases: $x_i > x_j$, $x_i = x_j$, and $x_i < x_j$. This score can be derived by summing all signs, then divide the total by the number of compared pairs, and get the absolute value of the quotient.

$$m_{trend} = \left| \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} Sign(x_i - x_j)}{N(N-1)/2} \right| \quad (7)$$

where

$$Sign(x_i - x_j) = \begin{cases} 1 & if\ x_i > x_j \\ 0 & if\ x_i = x_j \\ -1 & if\ x_i < x_j \end{cases} \quad (8)$$

Periodicity: Two conventional approaches for detecting periodicity time series are periodogram and auto-correlation. The latter is difficult to determine periods [44] automatically, so we use the former for this score. The periodogram element is squared of the Fourier coefficient of the series [8, 44]:

$$I_k(f_{k/N}) = \frac{1}{N} \left| \sum_{t=1}^{N} x_t e^{-i2\pi tk/N} \right|^2, \quad k = 0,1, \ldots \left\lceil \frac{N-1}{2} \right\rceil \quad (9)$$

Each element $I_k(f_{k/N})$ depicts the power of signals of frequency k/N in the series, so we score the periodicity measure of a time series as the sum of all peaks in the periodogram, and divide it by the sum of all elements to get standardized value.

$$m_{periodicity} = \frac{\sum I_{peak}}{\sum I_{total}} \quad (10)$$

Randomness: The randomness of time series describes how random the data is. It is related to a noisy series without trend or seasonal pattern. We use the first (lag 1) autocorrelation for this score [35]. Because we do not aim to distinguish negative and positive correlations, the squared Pearson coefficient is utilized.

$$m_{randomness} = 1 - \left[\frac{\sum_{t=2}^{N} (x_t - \bar{x})(x_{t-1} - \bar{x})}{\sum_{t=1}^{N} (x_t - \bar{x})^2} \right]^2 \quad (11)$$

4 DIMENSION REDUCTION TECHNIQUES

The nine metrics in the previous section are computed for every time series in the dataset. Each time series is treated as a point in the space of these metrics. Neighboring points in the nine metrics space are time series that have quite similar values of all metrics and are expected to share common visual features or multi-dimensional shapes. The structure of the dataset in high dimensional space can be preserved and visualized on lower dimensional space (2D in this work) by dimension reduction techniques such as PCA, t-SNE, and UMAP. The reduction process involves humans in the exploration of the dataset. Because the visualization phase depends on the result of the reduction process, the choice of a particular technique is crucial. Many different linear and nonlinear dimension reduction techniques have been reviewed and systematically compared by Laurens Van Der Maaten et al. [43]. In this work, we utilize and compare three popular methods: PCA, t-SNE, and UMAP. The default distance in metric space is Euclidean distance.

PCA is a popular linear technique that preserves as much variance of the dataset in high dimensional space as possible [43]. It transforms the nine metrics in the visual feature space into orthogonal or uncorrelated principal components. The first component has the largest possible variance, and the second component has the second largest one. Therefore, the first two principal components contain most of the information on the dataset. PCA has two main disadvantages [43]. The first one is that PCA might be inappropriate for embedding very high-dimensional space. However, this is not a problem in our work because the visual feature space has only nine dimensions corresponding to nine metrics. For example, it takes less than a second for our web prototype to compute and display the PCA projection of 1727 data points. The second drawback is that PCA focuses on large pairwise distances rather than small ones.

Nonlinear projections can avoid the overlapping problem of distinct clusters [4]. One popular nonlinear technique is t-SNE, which is capable of preserving the local structure of the dataset [29]. UMAP, a new nonlinear projection algorithm, can give better results of the reduction process in comparison to t-SNE [1, 4]. One crucial limitation of t-SNE is computation time, while UMAP is the fastest algorithm for small datasets (less than 100,000 data points) [4]. T-SNE also tends to neglect the global structure of the dataset and spread the low-density areas. Large t-SNE clusters are less dense than the smaller ones. In contrast, the density of UMAP clusters is more uniform [4]. Therefore, UMAP can provide a more meaningful organization of clusters, outliers, and the global structure of the dataset. In other words, UMAP is expected to give more reasonable projection results for the exploration process. Therefore, we adopt UMAP for our framework.

5 WEB PROTOTYPE

Our web prototype provides three main parts for users to look insight the dataset. The first one is the violin charts of the dataset on every metric. The second one is the clustering of all time series in visual features space. The last part is the visualization of the dataset using dimension reduction techniques, which are discussed in the previous section.

5.1 Summary distribution of metric values

This part of the prototype provides a summary view of the dataset on the metric space. Violin charts [cite] are utilized to display the distribution of the dataset on each metric. The input of a violin chart is the set of corresponding metric values of every time series.

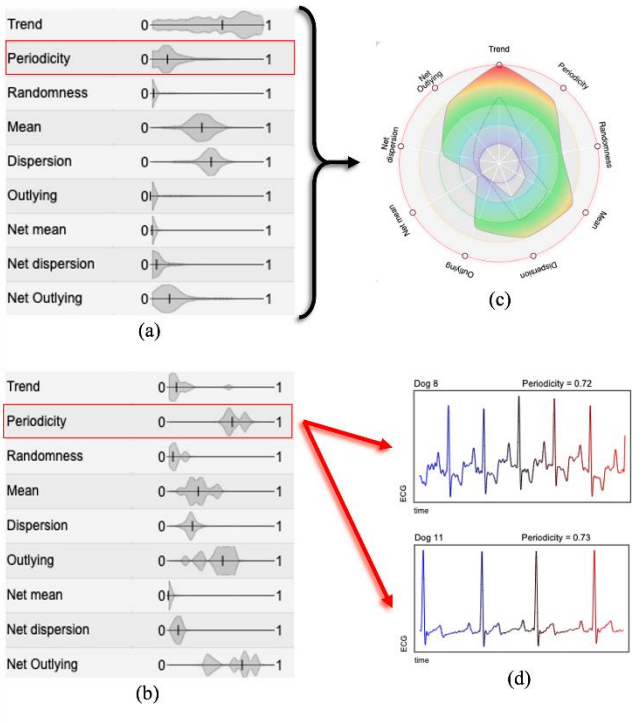

Figure 3: The summary charts in our prototype: (a) The violin charts of the US employment data are used to gain fundamental knowledge about the distribution of the US employment data on the metrics. Besides, (b) the violin charts of ECG signals of 17 dogs are utilized to test the metric *periodicity* (in the red box). (c) The radar chart of the US employment data gives the maximum, the minimum, and the mean values (dash line) of each metric. (d) Two of the time series of the ECG dataset have periodic property.

Because each dataset has different patterns, users can gain some fundamental knowledge about the distribution of the data on the metrics by observing the violin charts. For example, users can know whether most time series in the collection have seasonal patterns by investigating periodicity metric (as highlighted in Figure 3a). Another viable use of these violin charts is checking the effectiveness of proposed metrics. Figure 3.b is an example of the metric periodicity using the ECG signal of dogs [5, 17, 40]. This dataset has 17 time series, which are quite periodic (two of them are depicted in figure 3.d). The comparison between periodicity metric in figure 3.a and 3.b shows that this metric periodicity works well on the ECG dataset. Another chart for summarizing the distribution of metric values is the radar chart (figure 3.c) that gives the maximum, the minimum, and mean values of every metric.

5.2 Clustering of time series via their visual feature space

This part provides two clustering methods: k-means and leader algorithm. The former requires the input of the number of clusters and an appropriate convergence criterion such as no (or minimal) reassignment of patterns to new cluster centers, or minimal decrease in squared error [18, 22]. We use the maximum number of iterations as an input to stop the computation in case the convergence criterion cannot meet. The latter uses an input threshold to assign data points to groups of leaders [18]. This approach de- mands a reasonable number of leaders, for many leaders make computing and visualization too busy while too few leaders tend to over-summary the dataset. Thus, the threshold can be adjusted to get the desired number of leaders [12]. In this prototype, the range of the number of leaders is an input (two numbers that the resultant number of leaders is between). Clustering gives users a first look at the interrelationships among data points, such as how many significant different types of time series are (based on their visual features) in the dataset. The k-means is the default algorithm in this prototype. The default distance for clustering is Euclidean distance, and another option is Manhattan distance. After clustering the dataset, we color points due to their group in the visualization view to help users to identify similar patterns to explore.

5.3 Time series projection via their visual feature space

It is an interactive part of the prototype. By applying a chosen dimension reduction technique (one in three options: PCA, t-SNE, and UMAP), users can observe the visualization of the structure of the dataset in high dimensional space. This part allows users to explore the dataset by recognizing clusters in the low dimensional space (2D in this prototype). Detail discussion of each technique is in the previous section. This part includes many interactive components that allow users to look at the time series of a particular point, all time series of an individual instance, or a variable.

6 USE CASES

In this section, we use some real data to discuss the effectiveness of proposed metrics and the visualization.

6.1 US employment rates

The first data is the number of employees in thousands in 34 different industrial sectors of 53 states in the United States. It is a multivariate time series in which an economic sector of a particular state is one time series. Because some states have less than 34 sectors, the total number of time series in the collection is 1727. The data was recorded monthly over 21 years, from 1999 to 2019. We downloaded the dataset from the website of the Bureau of Labor Statistics: https://www.bls.gov/data/. It was removed seasonal component, or in other words, it is non-seasonal economic data.

There are several reasons why looking at every time series plot in this dataset is ineffective. Firstly, the number of plots is 1727, which is too much for analysts to analyze every graph to find visual features of interest. Secondly, it is also impossible for comparisons between time series plots of a state or of an industrial sector due to the size of the collection. Finally, finding common patterns and anomalous ones in this dataset is difficult. These problems can be overcome by the use of proposed metrics and the projection of the dataset from the visual feature space.

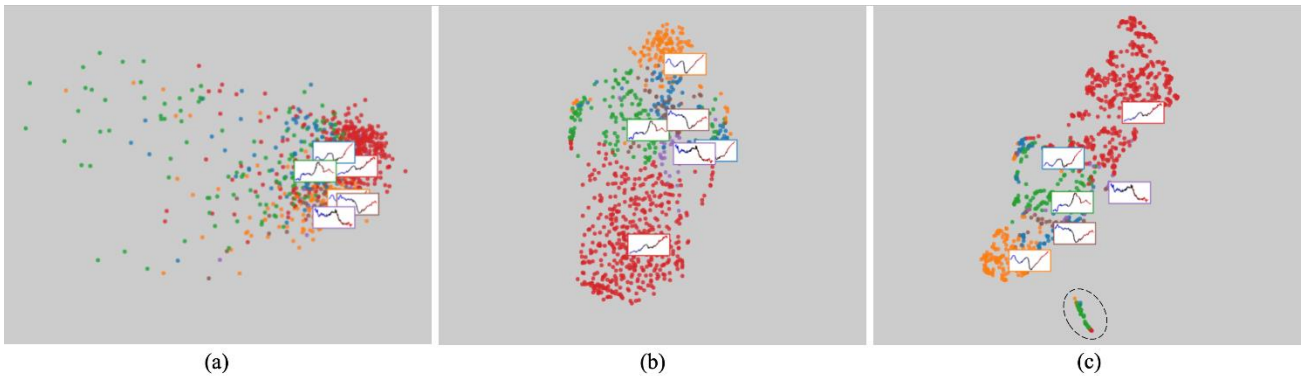

Figure 4: The structure of the dataset US employment in the metric space is visualized in the two-dimensional display using (a) PCA, (b) t-SNE, and (c) UMAP. The black dash cycle highlights the small group of points that stand out the large cluster in UMAP projection. The color indicates six different groups that are computed before the projection process by the k-means algorithm. The six time series are representatives of these groups.

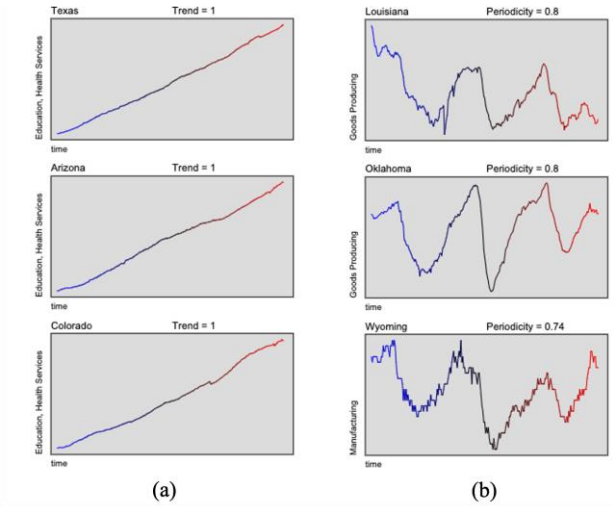

Figure 5: Time series that have (a) global trend and (b) cycles in the US employment data.

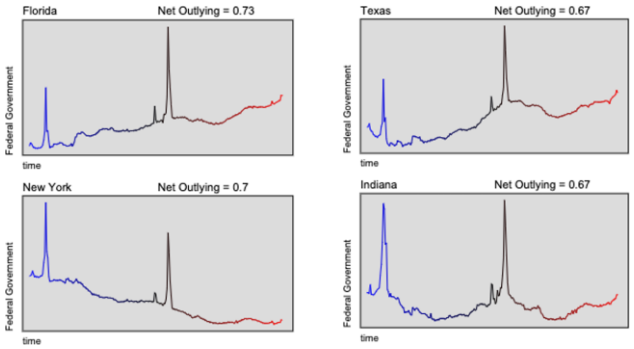

Figure 6: Outliers can be detected in time series of the US employment data by the metric *net outlying*.

Two important visual features of economic time series after removing seasonal components are trend and cycle. These visual features can be extracted respectively by two metrics: trend and periodicity. Sorting time series plots by the values of the metric of interest helps users to gain the corresponding visual feature. For example, figure 5 shows some plots which have high trend scores and high periodicity scores. By sorting the metric net outlying, we can observe outliers in some time series in the collection of US employment data (figure 6).

In terms of the global structure of the dataset, users can look at its visualization in one view after computing the chosen projection technique. Figure 4 depicts this structure on two-dimensional space using PCA, t-SNE, and UMAP. PCA suffers an overcrowding issue which points tend to locate in a small region. Some points, which are much far away from the others in the visual feature space, scatter over less dense areas. A small cluster, which stands out the big group, can be found easily in the figure of UMAP. It may be related to irregular patterns in comparison to others. In contrast, there is no such cluster, which stands out the large one in the t-SNE projection. Because UMAP can distinguish visual groups in the dataset, we will use it for further examples of the exploration of the dataset to represent the meaning of clusters in the projection space.

By interactions with the prototype, users can find that most data points in the small cluster in the UMAP projection (highlighted in figure 4.c) are time series of Federal Government sector. Looking at all time series of this sector, most of them have two extremely high peaks in May 2000 and May 2010 (figure 7.a), except some states such as Alaska, District of Columbia, or Maryland. These irregular increases in the employment data occur in the same years as the decennial censuses. This "coincidence" might point out the impact of the decennial censuses on monthly employment data. The U.S.

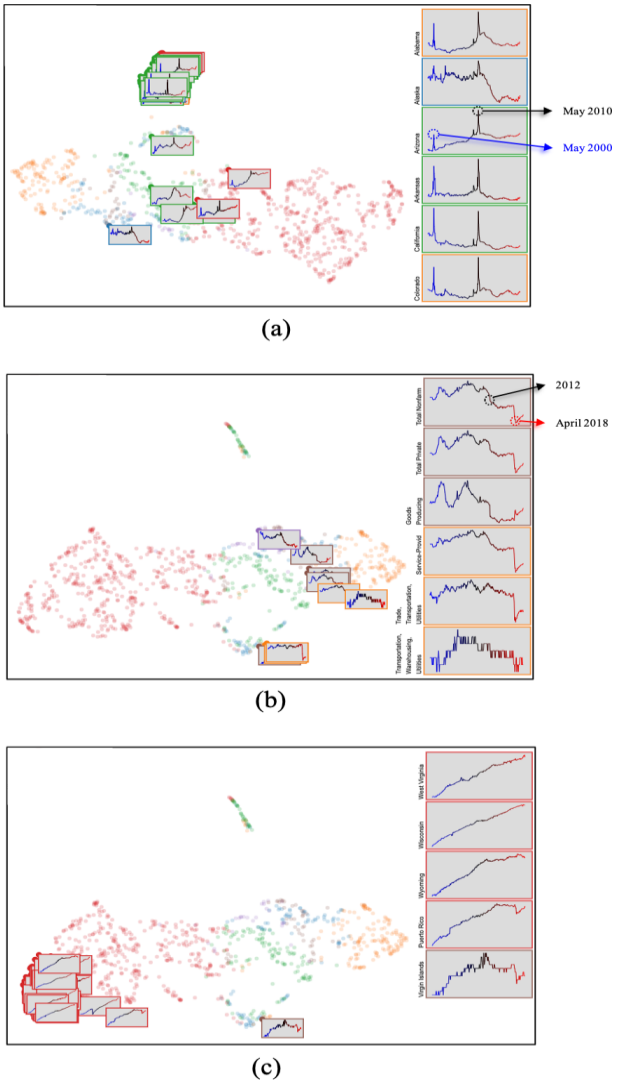

Figure 7: Time series of (a) the sector *Federal Government*, (b) Virgin Islands, and (c) the sector *Education and Health Services*

Bureau of Labor Statistics has confirmed on its website [36] that the Census involves thousands of temporary workers, and the hiring for this process is reflected in the data of the federal government.

Another way for exploration of data is by looking at the relationships of time series of a certain state or a particular industrial sector. By choosing the state (or sector) of interest, analysts can observe clusters of time series over visual features space to gain useful in- formation. For example, Figure 7.b illustrates the time series of the Virgin Islands. Plots of Virgin Islands in Accommodation and Food Services, Leisure and Hospitality, ... tend to focus on a small region, and these plots have a sudden drop in late 2017. This period is the time when the terrible hurricanes Irma and Maria went through the areas [39]. Other plots of this state in sectors like Government or Total Nonfarm have an earlier fall in 2011-2012, which was in the period of a catastrophic recession and the shut down of its largest oil company in 2012 [48]. Another example is time series of the sector Education and Health Services (figure 7.c). They mostly locate in the same region, or in other words, they have similar patterns. Although the period from 1999 to 2019 experienced two economic crises (2001 and 2008-2009) [2], the number of employees in this sector increases quite steadily for almost all states, except the Virgin Islands.

6.2 EEG dataset

This dataset is electroencephalograms that acquire brain signals from facial movements and thinking signals of 10 volunteers [45]. Movements are opening eyes, closing eyes, raising eyebrows, looking left, looking right, and smiling. Thinking signals are about moving forward, moving backward, turning left and turning right. There are five output channels from the Emotiv Insight headset, including AF3, AF4, T7, T8, and O1. Each facial movement and thinking is considered as a data instance, while each channel is a variable. This multivariate dataset has 550 time series. We use this dataset to show that our proposal metrics have the capability of detecting different types of outliers (figure 8).

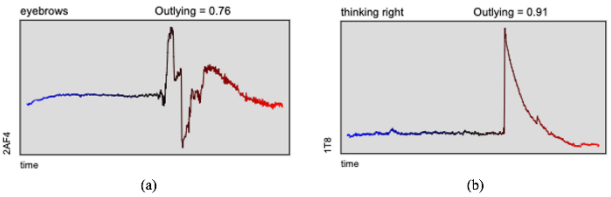

Figure 8: Two common patterns of outliers detected by the metric *outlying* in the EEG dataset.

7 CONCLUSION

This work considers the visual characteristics of the time series with a proposed list of visual features of interest. These visual features can involve users' cognitive and reasoning capabilities in the exploration progress of the large temporal dataset, and hence, make this process easier for the majority of users. The use of them in some real datasets gives noticeable results. Many different types of outliers are effectively detected, such as depicted in Figure 6 of the US employment data by the metric net outlying and in Figure 8 of the EEG data by the metric outlying. Cycles in the US employment dataset are detected (Figure 5.b) by the metric periodicity. Besides, the global structure of datasets is visualized by dimension reduction techniques such as PCA, t-SNE, and UMAP. The result shows UMAP is the most appropriate projection because it gives clear clusters of the dataset.

The visualization of interrelationships of the dataset in the visual feature space allows human-centered exploration of the data. Users can look at time series of points of interest in the two-dimensional display, or observe the set of time series of a particular instance or a variable. This approach helps users to overcome the issue of analyzing plenty of time series in the large collection of thousands of time-series plots.

The interactive interface is implemented as a JavaScript-based web application using D3.js [7]. The demo video, web application, and source codes of our visualization are available on our Github project at https://idatavisualizationlab.github.io/B/timeSeries/.

REFERENCES

[1] Mohammed Ali, Mark W Jones, Xianghua Xie, and Mark Williams. 2019. TimeCluster: dimension reduction applied to temporal data for visual analytics. *The Visual Computer* 35, 6-8 (2019), 1013–1026.

[2] Kimberly Amadeo. 2019. History of Recessions in the United States. website. Retrieved January 30, 2020 from https://www.thebalance.com/the-history-of-recessions-in-the-united-states-3306011.

[3] G.H. Ball and D.J. Hall. 1965. *Isodata, a Novel Method of Data Analysis and Pattern Classification.* Stanford Research Institute. https://books.google.com/books?id=Ti3BGwAACAAJ

[4] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. 2019. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology* 37, 1 (2019), 38.

[5] Joachim A Behar, Aviv A Rosenberg, Ido Weiser-Bitoun, Ori Shemla, Alexan- dra Alexandrovich, Evgene Konyukhov, and Yael Yaniv. 2018. PhysioZoo: a novel open access platform for heart rate variability analysis of mammalian electrocardiographic data. *Frontiers in physiology* 9 (2018), 1390.

[6] J. Bernard, M. Hutter, M. Zeppelzauer, D. Fellner, and M. Sedlmair. 2018. Com- paring Visual-Interactive Labeling with Active Learning: An Experimental Study. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (Jan 2018), 298– 308. https://doi.org/10.1109/TVCG.2017.2744818

[7] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D3 data-driven documents. *IEEE transactions on visualization and computer graphics* 17, 12 (2011), 2301–2309.

[8] Peter J Brockwell and Richard A Davis. 2016. *Introduction to time series and forecasting.* springer.

[9] Robert B Cleveland, William S Cleveland, Jean E McRae, and Irma Terpenning. 1990. STL: A seasonal-trend decomposition. *Journal of official statistics* 6, 1 (1990), 3–73.

[10] Tuan Nhon Dang, Anushka Anand, and Leland Wilkinson. 2013. TimeSeer: Scagnostics for high-dimensional time series. *IEEE Transactions on Visualization and Computer Graphics* 19, 3 (2013), 470–483. https://doi.org/10.1109/TVCG.2012.128

[11] Tuan Nhon Dang and Leland Wilkinson. 2013. TimeExplorer: Similarity search time series by their signatures. In *International Symposium on Visual Computing*. Springer, 280–289.

[12] Tuan Nhon Dang and Leland Wilkinson. 2014. ScagExplorer: Exploring scatter- plots by their scagnostics. *IEEE Pacific Visualization Symposium* (2014), 73–80. https://doi.org/10.1109/PacificVis.2014.42

[13] Jon Danielsson. 2011. *Financial risk forecasting: the theory and practice of forecasting market risk with implementation in R and Matlab.* Vol. 588. John Wiley & Sons.

[14] Samarjit Das. 1994. Time series analysis.

[15] Ben D Fulcher and Nick S Jones. 2014. Highly comparative feature-based time- series classification. *IEEE Transactions on Knowledge and Data Engineering* 26, 12 (2014), 3026–3037.

[16] Ben D Fulcher and Nick S Jones. 2017. hctsa: A computational framework for automated time-series phenotyping using massive feature extraction. *Cell systems* 5, 5 (2017), 527–531.

[17] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101, 23 (2000), e215–e220.

[18] John A. Hartigan. 1975. *Clustering Algorithms* (99th ed.). John Wiley & Sons, Inc., New York, NY, USA.

[19] DM Hawkins. 1980. Monographs on Applied Probability and Statistics.

[20] Boris Iglewicz and David Hoaglin. 1993. Volume 16: how to detect and handle outliers. *The ASQC basic references in quality control: statistical techniques* 16 (1993).

[21] Boris Iglewicz and David Caster Hoaglin. 1993. *How to detect and handle outliers.* Vol. 16. AsqPress.

[22] A. K. Jain, M. N. Murty, and P. J. Flynn. 1999. Data Clustering: A Review. *ACM Comput. Surv.* 31, 3 (Sept. 1999), 264–323. https://doi.org/10.1145/331499.331504

[23] Yanfei Kang, Rob J Hyndman, and Kate Smith-Miles. 2017. Visualising forecasting algorithm performance using time series instance spaces. *International Journal of Forecasting* 33, 2 (2017), 345–358.

[24] Eamonn Keogh, Stefano Lonardi, and Bill'Yuan-chi' Chiu. 2002. Finding surprising patterns in a time series database in linear time and space. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 550–556.

[25] Ragnar H Lesch, Yannick Caillé, and David Lowe. 1999. Component analysis in financial time series. In *Proceedings of the IEEE/IAFE 1999 Conference on Computational Intelligence for Financial Engineering (CIFEr)(IEEE Cat. No. 99TH8408)*. IEEE, 183–190.

[26] Yuan Li, Jessica Lin, and Tim Oates. 2012. Visualizing variable-length time series motifs. In *Proceedings of the 2012 SIAM international conference on data mining*. SIAM, 895–906.

[27] Jessica Lin, Eamonn Keogh, and Stefano Lonardi. 2005. Visualizing and discover- ing non-trivial patterns in large time series databases. *Information visualization* 4, 2 (2005), 61–82.

[28] Carl H Lubba, Sarab S Sethi, Philip Knaute, Simon R Schultz, Ben D Fulcher, and Nick S Jones. 2019. catch22: CAnonical Time-series CHaracteristics. *Data Mining and Knowledge Discovery* 33, 6 (2019), 1821–1852.

[29] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[30] Spyros Makridakis and Michele Hibon. 2000. The M3-Competition: results, conclusions and implications. *International journal of forecasting* 16, 4 (2000), 451–476.

[31] Henry B Mann. 1945. Nonparametric tests against trend. *Econometrica: Journal of the Econometric Society* (1945), 245–259.

[32] Robert McGill, John W Tukey, and Wayne A Larsen. 1978. Variations of box plots. *The American Statistician* 32, 1 (1978), 12–16.

[33] Leland McInnes, John Healy, and James Melville. 2018. UMAP: Uni- form Manifold Approximation and Projection for Dimension Reduction. arXiv:stat.ML/1802.03426

[34] Alex Nanopoulos, Rob Alcock, and Yannis Manolopoulos. 2001. Feature-based classification of time-series data. *International Journal of Computer Research* 10, 3 (2001), 49–61.

[35] NIST/SEMATECH. 2013. e-Handbook of Statistical Methods. e-handbook. Re- trieved January 3, 2020 from https://www.itl.nist.gov/div898/handbook/index. htm.

[36] U.S. Bureau of Labor Statistics. 2019. 2020 Census and CES Employment by State. website. Retrieved January 30, 2020 from https://www.bls.gov/sae/additional-resources/ces-state-and-area-census-2020-workers.htm.

[37] Bernard Rosner. 1983. Percentage points for a generalized ESD many-outlier procedure. *Technometrics* 25, 2 (1983), 165–172.

[38] Pavel Senin, Jessica Lin, Xing Wang, Tim Oates, Sunil Gandhi, Arnold P Boedihardjo, Crystal Chen, Susan Frankenstein, and Manfred Lerner. 2014. Grammarviz 2.0: a tool for grammar-based pattern discovery in time series. In *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 468–472.

[39] National Park Service. 2017. Hurricanes Irma and Maria. website. Retrieved January 30, 2020 from https://www.nps.gov/viis/learn/nature/2017-hurricanes. htm.

[40] Ori Shemla and Joachim Behar. 2019. PhysioZoo - mammalian NSR databases. https://doi.org/10.13026/P63Q-HQ95

[41] Ruey S Tsay. 1988. Outliers, level shifts, and variance changes in time series. *Journal of forecasting* 7, 1 (1988), 1–20.

[42] Laurens Van Der Maaten. 2014. Accelerating t-SNE using tree-based algorithms. *The Journal of Machine Learning Research* 15, 1 (2014), 3221–3245.

[43] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. 2009. Dimensionality reduction: a comparative. *J Mach Learn Res* 10, 66-71 (2009), 13.

[44] Michail Vlachos, Philip Yu, and Vittorio Castelli. 2005. On periodicity detection and structural periodic similarity. In *Proceedings of the 2005 SIAM international conference on data mining*. SIAM, 449–460.

[45] Tien Hoang-Thuy Vo, Tran Luu-Nha Dang, Ngan Vuong-Thuy Nguyen, and Tuan Van Huynh. 2019. Classification Electroencephalography Using Machine Learn- ing. In *2019 19th International Symposium on Communications and Information Technologies (ISCIT)*. IEEE, 14–19.

[46] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 1-3 (1987), 37–52.

[47] Kiyoung Yang and Cyrus Shahabi. 2004. A PCA-based similarity measure for multivariate time series. In *Proceedings of the 2nd ACM international workshop on Multimedia databases*. 65–74.

[48] Matthew Yglesias. 2013. The U.S. Virgin Islands Are in a Catastrophic Recession. website. Retrieved January 30, 2020 from https://slate.com/business/2013/08/virgin-islands-recession.html.