MatArcs: An Exploratory Data Analysis of Recurring Patterns in Multivariate Time Series

Julia Gaebler, Stephan Spiegel, and Sahin Albayrak

DAI-Lab, Technische Universitt Berlin,
Ernst-Reuter-Platz 7, 10587 Berlin, Germany
{gaebler,spiegel,albayrak}@dai-labor.de
http://www.dai-lab.de/irml/research/

Abstract. The development of algorithms in exploratory data analysis (EDA) requires handling a large amount of data and presenting it in a user-friendly environment. However, the results of EDA are often not self-explanatory. Therefore, we present *MatArcs*, a novel visualization technique for the detection of recurring time series patterns.

Major advantages of MatArcs are the dimensionality reduction of complex multivariate data and the localization of recurring patterns. Geared to the needs of various users which have deviating objectives in finding patterns, we present an algorithm which is able to fullfil the multilayered task of exploratory data analysis; i.e., i) processing multivariate and large time series (even real and noisy data may not pose a problem), ii) the versatility to be applied in fields of theoretical research, iii) extracting the main characteristics of a data set by using significant graphical attributes and iv) identified patterns can be synthesized to new patterns. With regards to the field of visual analytics our proposed approach provides a visualization which extends/increases the retentiveness of the user to get the targeted characteristics, even in multivariate time series.

1 Introduction

Consider the human eye as a high performance camera, which is communicating more impressions to our cerebric than any other organ of perception. However, if it comes to complex time series a clear visualization helps to increase the efficiency of receptivity. In general, any measured data assigned to a time may be termed as a time series e.g., weather data or the population data of China.

In time series analysis regularities in the data can be found to understand the substrate of the data and draw out a tendency (e.g., deducted from the preceding or current state of the atmosphere a weather forecast should be prognosticated). This further guides us to the term pattern, which may be defined as recurring structures in the time series. The identification of patterns is essential when it comes to prediction and forecasting in data sets analysis. A consequence of this is to distinguish trends in the data, which figures out the differences in various *multidimensional* time series.

Not only is finding regularities in the data important but also finding similarities in the structure for the comparison of various time series. For instance, consider data in image processing. There pattern recognition may bring out consistencies and inconsistencies of the image data to enable the image postprocessing or matching of images.

Pattern recognition in general covers a wide range of problems e.g., adornments in arts, action pattern in sociology or the succession of notes in musical theory. It is envisioned that pattern analysis in multivariate time series will gain more focus. This is owing to the fact that most of the data sets in the future will be complex over various dimensions. The basic requirements of the future are, creating an understandable visualisation, its suitability for all types of data and the extraction of the main characteristics. In this work, we propose the *MatArcs* approach to adapt the *multidimensional* data into comprehensible two-dimensional visualization e.g., consider the Figure 1, where recurring patterns are assigned to graphical attributes. We also combine the idea of interactive visualization, which makes the data analysis easy to handle.

We use arcs identifying a coherent pattern and semicircles indicating the time instance for the occurrences of a pattern. One should not neglect the fact that better visualization helps reducing the analysis time. However, it is arguable whether these aspects are intuitively accessible in order to make the prevalent context easy to comprehend and generally intelligible. In the proposed *MatArcs* approach, we aim to provide a visualization technique for depicting the locations of recurrening patterns. In order to calculate recurring patterns in time series, any arbitrary distant measure could be used. However, considering the calculation time, we use the Euclidean distance measure but as well refer to the dynamic time warping method [3].

Our claim of meeting the envisioned requirements comes from the following facts: i) introduction of time line for arc diagrams, ii) the occurrence of pattern may be identified by a specific time instance of the time line, which helps identify the multiple data events of an activity, iii) the differentiation of patterns by the size and color of semicircles, and iv) applying the *Joint Recurrence*, it is necessary and proper to have only *one* visualization instead of multiple visualizations for multiple time series observed coincidently.

The paper is structured as follows. In section 2, we draw a comparison of relevant approaches, in section 3 we elaborate on essential settings and introduce our proposed approach, section 4 presents a case study, section 5 gives an illustrative scenario for different search requests in *Google* and section 6 concludes the paper and discusses the future work.

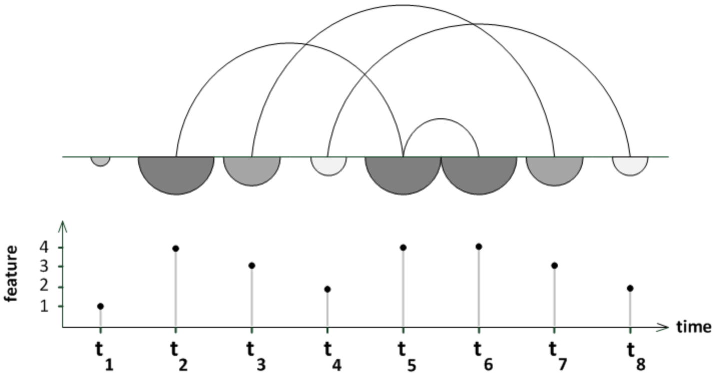


Fig. 1. The proposed approach uses semicircles located by a discrete time instance in a multivariate time series of recurring pattern, whereas arcs are used to clarify the recurrence of the most frequent patterns. Related to an univariate time series (as we use in this figure) containing only one feature the most frequent pattern with length one is the state ”4” which reappears three times.

2 Related Work

In the literature, a wide range of visualization approaches exist e.g., scatter plot, H-Curves, etc. However, many of these approaches lack not only the time reference but are also ambiguous in the visualization [2], [7]. To investigate the performance of our proposed approach against the existing relevant approaches, we compare the proposed MatArcs against the relevant approaches in terms of clarity and understanding of the subjected matter.

Consider scatter plots as methods to visualize the statistical distribution of data and state the frequencies of its features. Note that the multivariate time series has no upper limit on the quantity of variables. The bottleneck in visualizing the multivariate time series using scatter plots is the limitation of dimensions. This limitation is based on the evidence that a visualization is restricted by three coordinates in the phase space and the color spectrum. The scatter plot approach bears resemblance to H-Curves introduced to visualize DNA sequences [2] in a three-dimensional plot. For large data sets, it is difficult to comprehende due to the immense quantity of information in one visualization (since the eyes are limited in the spatial perception).

The mentioned visualization issues may be addressed by an approach like the arc diagram. Arc diagrams were first introduced in musical theory as a technique to visualize coherent structures [6]. They accomplish this by musical procedures for a deeper understanding of complex coherences. Contrary to the tonal analysis, our approach is directed towards finding relations between *multiple* variables. However, literature encamps various approaches using the arc diagram. We believe that the arc diagram algorithm was implemented in the limited scope i.e., for string data as an input [7]. The tonal analysis approach considers subsequences in univariate time series to distinguish a structure in the examined data, which is prior tonal data sets. Authors used a suffix tree to represent the repeating sequences. However, the approach lacks in interpreting the signifiance of the identified patterns and merely states the complexity of a data set. Let us now also discuss the *thread arcs* approach, which considers a compact small-scale visualization [4] and allows for the identification of the contemplated events chronologicaly. The mentioned approach does not meet the requirements of recognizing patterns but an interlinkage or rather a consecution of events.

Owing to increasing complexities in the data sets the extension of the arc diagrams is inevitable, specifically for identifying the patterns. An important aspect to be considered here is that for comprehensible two-dimensional visualization, approaches must be worked out that adapt multidimensional data into two-dimensional visualization. Whereas the two-dimensional visualization should constitute the calculated coherences of multiple variables. The mentioned facts are delt with by the proposed approach. Our approach works out the core pattern for multivariate and specifically large time series. Besides many approaches ([7], [2]) require strings or a sequence of symbols as an input. Another advantage of our approach is all types of time series can be used, even data without a time reference.

3 MatArcs Visualization

In this section we present the proposed visualization framework which is based on Arcs Diagrams. Therefore consider both the following settings and the mathematical consideration. Figure 2 is a flow chart which generalizes the proceeding to calculate the MatArcs.

3.1 Settings

Consider the multivariate time series $T = (x_{ij}) \in \mathbb{R}^{m \times n}$, where x_{ij} corresponds to an element (of ith row and jth column) of matrix T. Owing to the fact that a multivariate time series in the most general sense concerns two or more time series (which are recorded at same time instances), we consider two variables m & n to represent the two different dimensions of the time series. Note that m is the number of measurements and n is the number of discrete time points at which the measurements were taken. Let the pattern

$p = (x_i, x_{i+1}, \ldots, x_{i+k-1})$ of length k starting at time i be a subsequence of k consecutive multivariate states $x_i = (x_{1j}, \ldots, x_{mj}) \in \mathbb{R}^m$ for all $j \in \{1, \ldots, i+k-1\}$.

Suppose that $d : \mathcal{P}_k(T) \times \mathcal{P}_k(T) \to \mathbb{R}_+$ is some dissimilarity function defined n patterns of length k. A recurrent pattern of pattern $p_i \in \mathcal{P}_k(T)$ is any pattern $p_j \in \mathcal{P}_k(T)$ satisfying the following properties:

1. Patterns p_i and p_j are non-overlapping, that is $j + k \leq i$ or $i + k \leq j$,
2. we have $d\,(p_i, p_j) \leq \varepsilon$, where $\varepsilon \geq 0$ is a threshold.

We consider the Euclidean distance $d\,(p_i, p_j) = \|p_i - p_j\|$, though any dissimilarity function can be chosen. For continuous time series the method dynamic time warping may be a suitable choice, for further reading see [3].

Let us now refer to [5], who introduced recurrence plots for visualizing recurring pattern. A recurrence plot is mathematically described by the matrix $R_{ij} = \Theta(\varepsilon - \parallel x_i - x_j \parallel)$, where $x_i, x_j \in \mathbb{R}^m$ are multivariate states at time i and j, resp., and $\Theta(z)$ is the Heaviside function defined by

$$\Theta(z) = \begin{cases} 1 & : \quad z \geq 0 \\ 0 & : \quad z < 0 \end{cases}. \tag{1}$$

The visualization of recurrening patterns in q time series

$$T_1 = \left(x_{ij}^{(1)}\right) \in \mathbb{R}^{n_1 \times m_1}, \ldots, T_q = \left(x_{ij}^{(m)}\right) \in \mathbb{R}^{n_q \times m_q} \tag{2}$$

is based on Joint Recurrence method

$$JR_{ij} = \prod_{r=1}^{q} \Theta\Big(\varepsilon_r - \parallel x_i^{(r)} - x_j^{(r)} \parallel\Big), \tag{3}$$

where $x_i^{(r)} \in \mathbb{R}^{m_r}$ denotes a multivariate state from time series T_r at time i, and ε_r is the r-th threshold for determining recurrence of a state in T_r. The joint recurrence is an extension to the recurrence R_{ij}. Simultanous states in q different dynamic systems are compared i.e., consider weather data containing the humidity and the temperature as two dynamical systems, which are recorded for each hour of one day.

In the following section we carry out multiple steps to process the data for the proposed visualization.

3.2 Proposed Approach

The proposed approach is decomposed into three major functional components namely i) input, ii) solver, and iii) visualization. These components are pictorially depicted in Figure 2. We now elaborate on the functions carried out in each of the mentioned components.

Input component: The existance of this component is mainly the consequence of situations when one is dealing with univariate time series like inputs or multivariate time series with heterogeneous time scales. We address these issues within the input component by merging the univariate time series which means that univariate time series of same length form a matrix. However, when it comes to multivariate series with heterogeneous time scales, we suggest the harmonization of the time scales as a pre-processing step in the input component. It should be noted that the pre-processing is strictly time series dependent. For instance, the univariate time series, $T_1 \ldots T_m$, are pre-processed and merged as T_{mn} in Figure 2. Note that a filtering of the data is not required, even not desirable. This results from the fact that for the calculation of the distance between the time instances outliers may be too far, so that

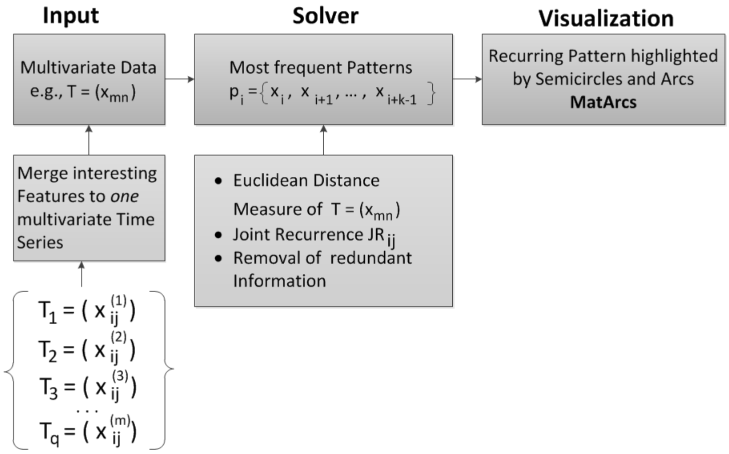

Input
Solver
Visualization
Multivariate Data e.g., T = (x_mn)
Most frequent Patterns p_i = { x_i , x_i+1 , ... , x_i+k-1 }
Recurring Pattern highlighted by Semicircles and Arcs
MatArcs
Merge interesting Features to one multivariate Time Series
Euclidean Distance Measure of T = (x_mn)
Joint Recurrence JR_ij
Removal of redundant Information
T_1 = (x_ij^(1))
T_2 = (x_ij^(2))
T_3 = (x_ij^(3))
...
T_q = (x_ij^(m))

Fig. 2. MatArcs flow chart illustrating the various calculating steps

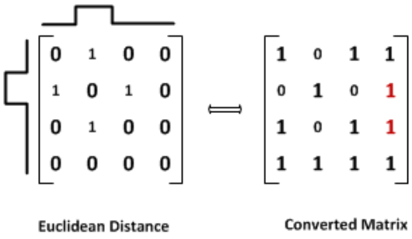
by default they do not form a pattern. Thereby NaN-values will be ignored.

Solver component: This component is responsible for i) measuring the similarity of each component of the univariate time series against others, and ii) measuring the similarities on higher dimensions i.e., for multiple dynamic systems.

In the proposed approach, the earlier functionality (i.e., (i)) is achieved by computing the Euclidean distance. Let us consider the squared Euclidean distance which is defined as a distance $d^2(p_i, p_j)$. Then we propose that for each considered univariate time series, the euclidean distance may be computed seperately. This dictates that a state with homogeneous identity is represented by the distance of Zero, which is mathematically represented as

$$d^2(p_i, p_j) = \begin{cases} 0 & : p_i = p_j \\ non-zero & : p_i \neq p_j \end{cases} \tag{4}$$

The latter functionality (i.e., (ii)) is achieved by computing the joint recurrence in this work. As a first step, identifiers need to be assigned to the components of Euclidean distance. These identifiers play a vital role in identifying the similarities within the multiple dynamic systems e.g., in Figure 3, the matrix on the left hand side corresponds to Euclidean distances. We assign the identifier 1 to the euclidean distance 0 and we assign identifier 0 to any other Eculidean distance value. This results in the matrix on the right hand side.



$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \Longleftrightarrow \begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Fig. 3. Converting of Euclidean distance measure

The joint recurrence of the time series T_{mn} is computed using equation 3. As an example, consider two of the converted matrices seen in Figure 3. By multiplying these matrices containing the assigned identfiers a value of 1 is recieved, if both of the multiplied entities of the matrices are 1. This indicates a time instance for a pattern.

One may choose any suitable threshold value. However, our experience in processing various data advocates that the probalility of finding patterns increases for choosing higher thresholds, see section 4. It should be mentioned that owing to symmetric matrix $R = R^T$ and the calculation of every time instance $t_1, ..., t_n$ for m potential patterns one may expect redundant information (see Definition 1), which should be avoided. In this work, entities containing redundant information are replaced by Zero, as marked in red in Figure 3. When working with extremely large data sets, the computational complexity may turn out to be crucial issue. Thus, we stick to the idea of using the sparse matrices and the upper triangular matrices. Upon carrying out the aforementioned steps within the *solver component*, we get the most frequent patterns e.g., $p = \{x_i x_{i+1} \dots x_k\}$.

Definition 1:*As a result of this, patterns have similar frequencies on at similar time instances. For instance, consider the second and third pattern (second and third row of the converted matrix) in Figure 3. The redundant information corresponds to identical information that exist in the last two time instances.*

Visualization component: The three major functionalities within this component include: i) Computing the radius of the circles $r_i = \frac{nnz(i)}{nnz(JR_{i,j})}$, where nnz corresponds to the number of non-zero elements in one row of the joint recurrence matrix $JR_{i,j} \forall i, j = 1, ..., n$. The radius of a circle is proportional to the frequency of a pattern. ii) Calculating the arcs - the arcs in this work correspond to the distance of elementary entities in one pattern, whose heights are proportional to half the distance between effectively identical entities. iii) Color assignment - for better visualization, each pattern is assigned with a different color.

4 Case Study

In this section, we elaborate on analysing the application of the proposed approach. We reveal a possibility to utilize the MatArcs for any arbitrary data. Let us evaluate the performance of the proposed approach by analysing two examples which are publicly attainable on the internet, namely data of *Google Trends. Trends* is a feature of *Google* which illustrates the number of searches for a specified search request over time and this data can be extracted from the webpage into a *.csv* file. The data is normalized based on the average search traffic of the search term[1]. In particular we consider the search item *renewable energy* in two countries, Germany and the United States.

Within the *input component*, we merge the two terms of same time instances (i.e., January 1, 2004, to March 26, 2012) where each time series contains more than 400 time instances. In the *Solver component*, in addition to computing the Euclidean distances of time instances, we consider the joint recurrences of multiple features (which are then plotted as the values of the vertical axis within the *visualization component*). Furthermore, we define threshold values for both time series as 5% divergence. In general the number of patterns can be reduced via the user defined threshold as long as the quality of the results is not affected. The quality in this work is determined by a drift error e.g., in Figure 5 a threshold of 0.05 is appropriate in order to have a preferable number of patterns and a clear visualization. However, choosing a very small threshold value for a large data set may lead to an immense number of patterns, which in turn will make the pattern calcluation computationally expensive.

[1] http://www.google.com/intl/en/trends/about.html, Chapter 7. How is the data scaled

This motivates us to use and suggest a suitable threshold that helps to avoid the computational complexities. We address the aforementioned issue by providing the users with the *interactive* figure(s). In the considered settings, we take a minimum of three patterns as the most frequent pattern. This is done for attaining a clear visualization of the proposed MatArcs approach.

5 Visual Analysis

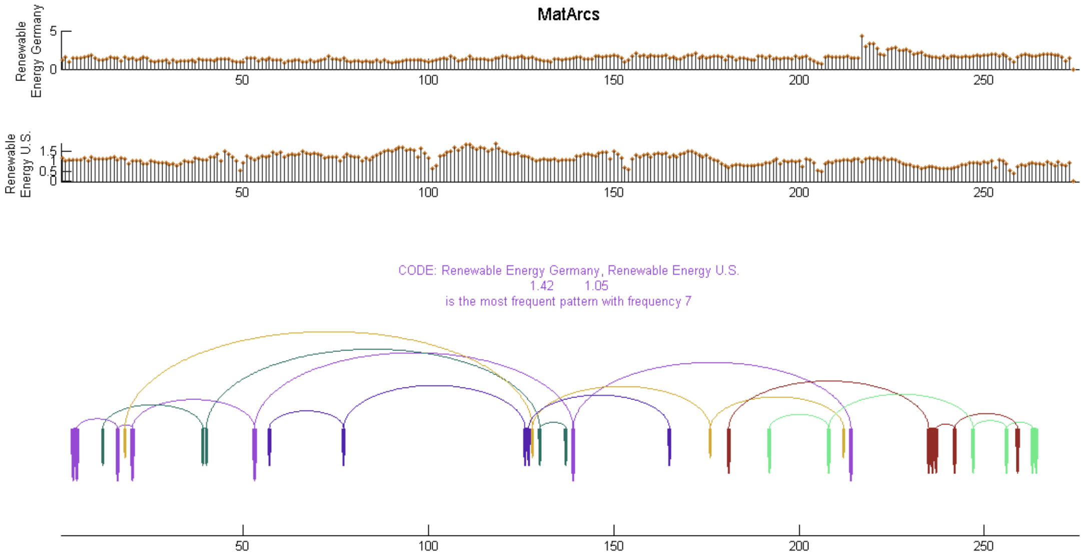


Fig. 4. MatArcs Visualization of the search request *Renewable Energy* which is then subdivided and considered for two counrties (Germany and the United States), Google Trends Data [1], which spans from 2004 to the present. Note that the colour of the middle text coincides with the colour of the most frequent pattern 1.42 1.05 of frequency 7.

Figure 4 presents the visualization of the Google trends (details mentioned earlier). It can be seen in the Figure that the vertial axis represents the number of search requests for the keyword *renewable energy*. A closer look at the Figure 5 reveals that the vertical axis contain various discrete values e.g., 4.36, which means that the value recorded at this time is 4.36 times larger than that of the average recording of all search requests found in *Google*. The horizontal axis represents the number of occurences, which in this case is not continuous. The proposed semicircles indicate the time instances of a given pattern e.g., the most frequent pattern (i.e., 1.42 1.05) advocates that 1.42 corresponds to higher search frequency in Germany compared to the average rate and a 1.05 higher search frequency in the U.S. The connecting acrs of events help identifying similar events which form a pattern.

To examine the results in greater detail, we hereby present a detailed view of Figure 4 in Figure 5 (namely 100 data points of the google trends data containing more than 400 time instances). As can be seen in the figure, the pattern in the range of 190 to 265 is a complete pattern, appearing at the beginning of the year 2011 for the first time (denoted by

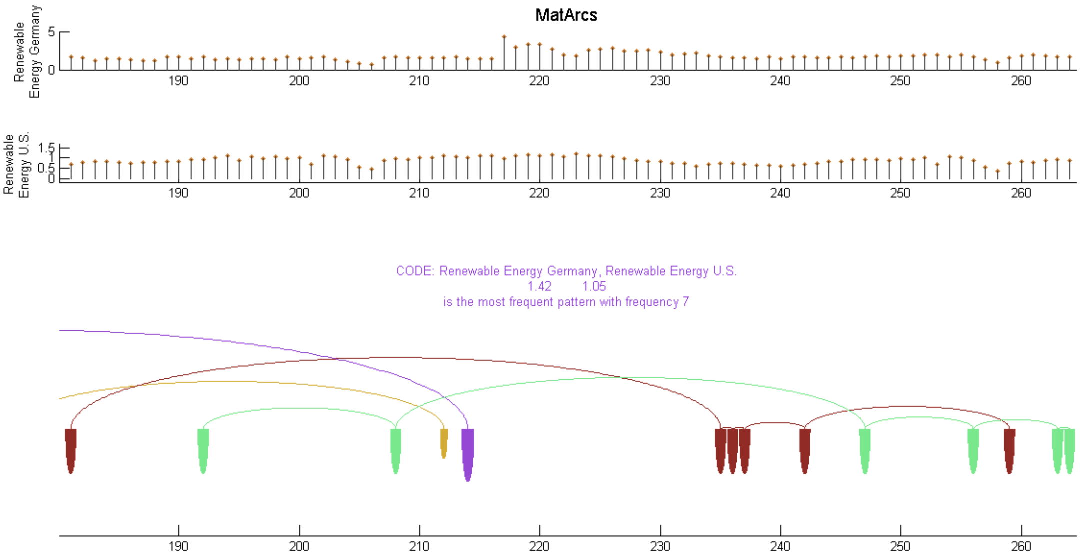


Fig. 5. Detailed Plot: MatArcs Visualization of the search request *Renewable Energy* which is then subdivided and considered for two counrties (Germany and the United States), Google Trends Data [1], which spans from 2004 to the present. Note that the colour of the middle text coincides with the colour of the most frequent pattern 1.42 1.05 of frequency 7.

its slight green colour). Note that 1.0 on the vertical axis corresponds to the average value of all searched terms in Google. Whereas the term *renewable energy* was searched in Germany with a value of 1.64, the said term is below the average value in the United States. The most frequent pattern (1.42 1.05) depicts an above average interest in both countries which might then have been influenced not only by national but also international news.

More precisely arcs provide the readers with the information about the quantity of searching activities for a given time duration. Let us now then refer to some definitions for exploiting the visualization and interpreting the data. When viewing the MatArcs visualization for various time series, particular types of patterns are salient. Patterns that are performed more frequent with regard to the whole period of observation, indicate a long-term consumer motivation of permanent interest (see the purple or yellow pattern in Figure 4). Whereas a short-term interest is shown by a more frequent appearance of patterns that occupy a shorter period of time. However, considering a short period of observation related to the entire observation time (e.g., patterns in red and slight green in Figure 4 and 5). Note that the biggest curve radius reveals an activity which is not executed for a long duration of time. The aforementioned case is conditional to the frequency of a pattern i.e., i) a more frequent pattern containing the biggest curve radius may lead to a long-term interest, and ii) a frequency of 2 for a pattern may show a short-term interest. Additionally to these given data, it is also conceiveable to understand and compare structures by our approach e.g., gene sequences.

Let us now refer to scatter plots to make a comparison between these two approaches. Figure 6 contains the same two variables as seen in Figure 4 for the search term *renewable energy*. Scatter plots are commonly used to reveal a dependency of two variables but can be far to complex to recognize a structure. Merely the agglomeration of the data points may

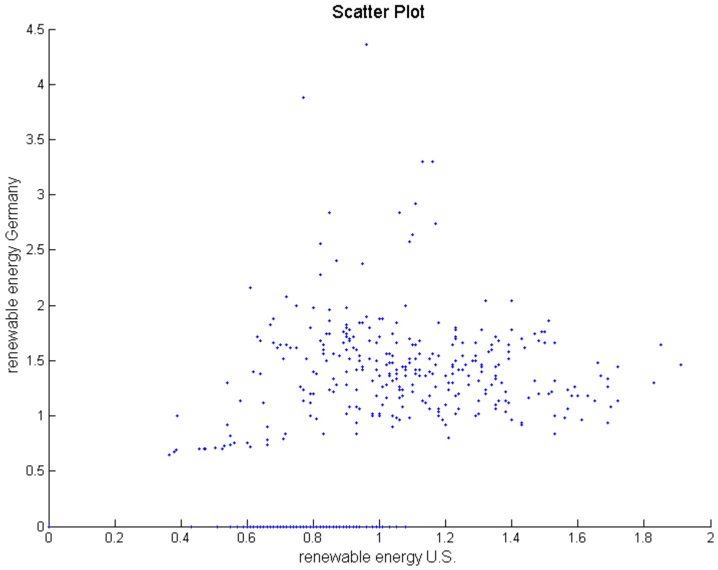


Fig. 6. Scatter plot of the search request *Renewable Energy* which is considered for two counrties (Germany and the United States), Google Trends Data [1].

be detected. Beyond that, multivariate data containing more than three signals may not be visualized by the scatter plot approach. Having observed the performance of the proposed approach, we claim that it helps the comprehensible and easy to interpret evaluation of the complex patterns.

6 Conclusion and Future Work

In this paper we propose a novel interactive visualization approach, namely MatArcs, which enables the user to interactively zoom into the most frequent and common patterns that plainly and obviously express the distinctive characteristics. These characteristics might be unknown or not accessible in advance, but they are identified by our MatArcs approach. We aim to spare additional costs and reduce the time and effort associated with data analysis by identifying structures which may be used for a comparison with other data. The fact that huge data might also be applied and the articulate visualization of our MatArcs evoke a easy comprehensibility and allows an analysis by a pure visualization (we also evaluated the proposed approach for different data sets).

In the future it is not only important to discover and visualize coherences but also to generalise a proceeding. Based on this, the system is able to support the user by means of automatisms. The analysed *Google Trends Data*, up-to-the-minute news to a topical subject may be to an audience. Based on frequent searched terms, news may be personalized in a found rhythm equivalent to the patterns of this terms, which was found. More precisely learnt behaviour provides the basis for assigning a profile to every user. The proposed approach could further be employed to build user profiles and to provide news readers with personalized information.

References

1. Google. Google trends. 2011.
2. E. Hamori and J. Ruskin. H curves, a novel method of representation of nucleotide series especially suited for long dna sequences. *The Journal of Biological Chemistry*, 258(2):1318–1327, 1983.
3. E. Keogh and C. A. Ratanamahatana. Exact indexing of dynamic time warping. *Knowl. Inf. Syst.*, 7(3):358–386, mar 2005.
4. B. Kerr. Thread arcs: an email thread visualization. In *Proceedings of the Ninth annual IEEE conference on Information visualization*, INFOVIS'03, pages 211–218, Washington, DC, USA, 2003. IEEE Computer Society.
5. N. Marwan, M. Carmenromano, M. Thiel, and J. Kurths. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5-6):237–329, jan 2007.
6. M. T. S. of New York State. *Theory and practice: newsletter-journal of the Music Theory Society of New York State*. The Society, 1988.
7. M. Wattenberg. Arc diagrams: visualizing structure in strings. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*, pages 110 – 116, 2002.

Appendix: Explorative Data Analysis

A web framework was developed following the idea of reviewing the context of this paper. For the application of our web tool ***anyTime*** we kindly refer to *http://anytime.dai-labor.de*, which we conceived as a testing ground for any data chosen by the user to explore multivariate time series. Researchers can take advantage of interactive figures for calculating and visualizing common pattern as well as the most frequent pattern.