Clustering Trend Data Time-Series through Segmentation of FFT-decomposed Signal Constituents

Daniyal Kazempour, Anna Beer, Oliver Schrüfer, and Thomas Seidl

Ludwig-Maximilians-Universität München, Munich, Germany
{kazempour,beer,seidl}@dbs.ifi.lmu.de, oliver.schruefer@campus.lmu.de

Abstract. When we are given trend data for different keywords, scientists may want to cluster them in order to detect specific terms which exhibit a similar trending. For this purpose the periodic regression on each of the time-series can be performed. We ask in this work: What if we not simply cluster the regression models of each time-series, but the periodic signal constituents? The impact of such an approach is twofold: first we would see at a regression level how similar or dissimilar two time-series are regarding their periodic models, and secondly we would be able to see similarities based on single signal constituents between different time-series, containing the semantic that although time-series may be different on a regression level, they may be similar on an constituent level, reflecting other periodic influences. The results of this approach reveal commonalities between time series on a constituent level that are not visible in first place, by looking at their plain regression models.

Keywords: Clustering, Time-series, FFT, Signal, decomposition, constituents

1 Introduction

Whenever data is collected over a period of time we have time-series. Let it be time-series in the life science, economy, health, social science and many other research fields, we are faced with time-series containing information on attributes changing over time. In context of time-series the detection of periodicity is of special interest, such as the periodicity in the predator-prey behavior among animals, power-consumption cycles, or trends of certain terms in search engines etc. As a motivation we introduce in this work the following use-case: Suppose there is a journal which publishes four times a year. Its content encompasses topics ranging from health, over society, education and many more. The editor in chief decides now to include topics in each of the issues which are most trending among the readers. The question at this point which concerns the editor in chief

Copyright ©2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-2454/paper_64.pdf

is: what is most trending? And how can such information be obtained? Or in other terms: How can the editorial stuff decide which of the most trending topics belong together in an issue? For this purpose one of the editors comes up with the idea to mine the Google trends platform[1]. From an editors-compiled list of topics, the trend data over the past years are collected from the trends platform. The data scientists in the journal obtain first for the keyword "turkey" (the poultry) the time-series and perform a periodic regression on it. We elaborate on periodic regression in Section 3 in more detail. The data scientists, however, recognize that the regression approach neglects many information of the original time-series as it can be seen in Figure 1 (left). Then one of the data scientists comes up with the idea to perform a Fast Fourier Transformation (FFT) by which the time-series for "turkey" is decomposed through the frequency domain into its so called constituents as seen in Figure 1 (right). In fact, the sum over all the constituents resembles the original signal. What the data scientists can do now is to cluster the keywords by using the constituents information instead of clustering the regressions of the time-series. We shall see in the experiments section of this work that the clustering on basis of signal constituents yields meaningful clusters compared to the case where we just use the regression of time-series. By the constituent-based approach the editors can identify the topics for each calendar quarter of the year, providing articles in their issues which satisfy the (seasonal) interests of readers.

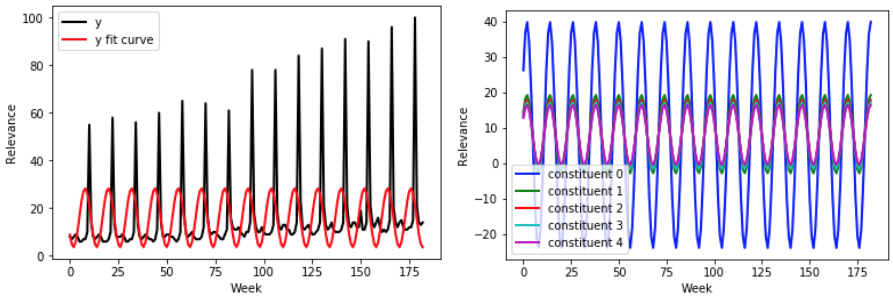

Fig. 1. Left: Google trends data from 2004 until now on the keyword 'turkey' with a periodic regression, Right: top-5 constituents from the turkey time-series using FFT.

Apart from our journal example, in [1] the authors list 25 applications of clustering on time-series data from 11 different domains (aviation, energy, finance etc.). Each of the clustering applications comes with references to at least one paper, emphasizing that the topic of time-series clustering is undoubtedly of significance and broad range regarding its application. Concluding this introductory part our contributions in this work are as follows: (1) Providing a representation of decomposed signal constituents from a periodic signal as parameter-space vec-

[1] https://trends.google.de/trends/

tors. (2) A clustering method on regression and on constituent level of time-series applied on periodic parameter space.

2 Related Work

Most related work can be found in context of expression patterns of genes, since many of them are associated with "circannual (yearly), circadian (daily), cell-cycle and other periodic biological processes [that] are known to be rhythmic" [7]. Early methods like, e.g., [3] use the standard correlation coefficient between the normalized vectors representing the correlations, since it captures similarity in the shape of periodically correlated datasets well. However, it does not fulfill the triangle inequality (and is thus no metric), plus magnitude and shift of time series are neglected here, which can lead to undesired results. [3] apply a pairwise average-linkage cluster analysis and focus on organizing and graphically displaying data in a way to allow users to explore data in an intuitive manner. RAGE [7] is phase independent and uses a true distance metric based on the undirected Hausdorff distance. Nevertheless, it is based on finding the periodic representation of each time series by fitting it to an ideal synthetic sinusoid or another already known (gene expression) profile. CorrCos [4] first generates over 100000 synthetic models and then matches each time series to one of those models using cross correlation. In a more recent work, "a periodic covariance function based on a projection of the Matern covariance" [5] is used. They regard noise and allow time series to share only a periodic component, structuring them in a hierarchical manner. However, they need some information like the phase-length beforehand, which may be popular in this field, but not generally accessible for new datasets. The representation of time series in form of FFT and other approaches is quite common in signal processing as discussed in [9], which also gives a brief overview on some of the most common techniques. With regards to the aspect of time-series distance measures there is a wealth of literature, such as in a more recent work of [8]. To the best of our knowledge, we are not aware of any algorithm regarding the singular components of each periodically correlated time series, but only such working on simple periodical correlations.

3 Clustering via Signal Constituents

In this section we elaborate on our approach of clustering periodic time series by clustering signal constituents. For this purpose we first describe the periodic feature vectors and parameter space and proceed on periodic regression. We contrast the regression approach by describing on how we perform signal decomposition using FFT. Having obtained the signal constituents, we cluster them, which requires specific handling, being elaborated on in the last subsection.

3.1 Periodic Feature Vectors and Parameter Space

A periodic function can be represented as follows:

Definition 1 (Periodic Function) *Given an object* $(x, y) \in \mathbb{R}^2$*, a periodic sinusoidal function is defined as:*

$$y = a \cdot sin(f \cdot x - p) + v$$

Where a *is the amplitude,* f *the frequency,* p *the phase-shift and* v *the vertical shift. As such a periodic feature vector* φ *is defined as:*

$$\varphi = (a, f, p, v)^T$$

The periodic feature vector φ can be regarded as a model which describes a periodicity within a given time series. At this point we recapitulate that we do not aim at comparing residue-by-residue multiple time-series with each other, but their respective models. For this purpose we define the periodic parameter space $\mathbb{P}$ as a feature space being spanned by the set of parameters of a periodic function $\{a, f, p, d\}$ with $\varphi \in \mathbb{P}$.

3.2 The FFT and Periodic Regression

The most common usage of the Fourier transform is to convert a given signal from the so-called time spectrum to a frequency spectrum. One core aspect on which the Fourier transform relies, is that every non-linear function can be represented as a sum of sine functions. A Fourier transform decomposes the time signal and reveals information about the frequency of all involved sine waves that generate the original signal. For time-series of evenly spaced values, such as time series of having data from every second, hour, week etc. the so called Discrete Fourier Transform (DFT) is defined as follows:

Definition 2 (Discrete Fourier Transformation)

$$X_k = \sum_{n=0}^{N-1} x_n e^{-2\pi i k n / N}$$

where N *denotes the number of samples,* n *the current sample,* x_n *the value of a signal at the sample point* n*,* k *the current frequency and* X_k *the result of the DFT encapsulating amplitude and phase.*

Since the computation of the DFT as shown in Definition 2 has a quadratic runtime complexity of $\mathcal{O}(n^2)$, we use in this work the so called Fast Fourier Transform (FFT)[2] which reduces the runtime to $\mathcal{O}(Nlog(N))$ by recursively dividing a given DFT into smaller DFTs .

A periodic regression is defined as:

Definition 3 (Periodic Regression) *Given a set DB of objects from e.g. a time series* (x_i, y_i) *of variables (dependent and independent), the Damped Least Squares (DLS, also known as Levenberg-Marquardt algorithm) aims to find the*

parameters $\boldsymbol{\beta}$ of a model curve $\tau(x, \boldsymbol{\beta})$ s.t. the sum of squares of the deviation $S(\boldsymbol{\beta})$ is minimized:

$$S(\boldsymbol{\beta}) =_{\beta} \sum_{i=1}^{m} (y_i - \tau(x_i, \boldsymbol{\beta}))^2$$

A periodic regression R is defined as the DLS of a periodic function

$$y = a \cdot sin(f \cdot x - p) + v$$

with an initial guess $p_0 = (a_0, f_0, p_0, v_0)$ which is obtained from a FFT taking the frequency, phase-shift and vertical shift with the highest amplitude.

For a periodic regression we perform first an FFT on a given time-series. We take the peak frequency in the frequency domain from the FFT and discard all other frequencies. The peak refers to that particular frequency with the highest energy (=amplitude). Intuitively it can be compared to a principal component analysis (PCA), where for dimension reduction purposes we keep for example only that principal component with the strongest Eigenvalue. The amplitude is determined by computing the absolute value of the complex conjugate from an FFT. Having the frequency with the highest amplitude (=peak) we now need to determine the phase-shift which is easily obtained by computing the angle of the complex conjugate result. With the amplitude, frequency and phase we can now compute the missing vertical shift. The FFT-obtained frequency, phase-shift, vertical-shift and amplitude are taken as an initial guess for a DLS as stated in Definition 3. We got now the periodic feature vector φ which represents the regression of a given periodic time series.

However, the regression approach comes with the drawback that we may lose information on the constituents which actually contribute to the detailed "shape" of a time-series. This is an issue which we shall approach in the upcoming subsection.

3.3 Signal Constituent Tensors through Signal Decomposition

What are the most determinant constituents of an original time-series signal? Or, to express it rather in terms of a principal component analysis: what are the most determinant principal components? For this purpose we perform an FFT as described in the previous section and keep the k strongest signals, where the signal strength is determined by its amplitude. The other constituents are discarded. A possible approach by which k can be chosen is, e.g., by using the elbow-method[6]. As a result we obtain a tensor of the following structure:

Definition 4 (Signal Constituent Tensor) *Given the top k signal constituents $\sigma_i = (a_i, f_i, p_i, v_i)^T$ from an FFT on a time-series τ, the resulting signal constituent tensor is defined as:*

$$\Lambda(\tau) = (\sigma_0^T, ..., \sigma_{k-1}^T)^T$$

Where the single constituents are ordered in a way such that it holds:

$$a_i > a_{i+1} > ... > a_{k-1}$$

The signal constituent tensor from Definition 4 provides us now a representation of the top k strongest signal constituents. They pose the very basis for computing similarities between different time-series among their respective constituent signals.

3.4 Similarity and Clustering of Signal Constituents

Having now obtained the signal constituent tensors, we use them as a base for performing a clustering. As a design decision we use here hierarchical clustering with average linkage. In order to perform hierarchical clustering we need to have a distance matrix. For such a distance matrix the question of how to actually compute the distance of the signal constituent tensors between two time-series arises. For this purpose we like to introduce the MinCT-dist distance in this work.

Definition 5 (Minimum Constituent Tensor Distance (MinCT-dist)) *Given two signal constituent tensors, each of them from one time-series $\Lambda(\tau_1)$ and $\Lambda(\tau_2)$. The minimum constituent tensor distance is defined as:*

$$\begin{aligned} &d_{MinCT}(\Lambda(\tau_1), \Lambda(\tau_2)) \\ &\quad = min\{(\sigma_{0_{\Lambda(\tau_1)}}, \sigma_{0_{\Lambda(\tau_2)}}), ..., (\sigma_{k-1_{\Lambda(\tau_1)}}, \sigma_{k-1_{\Lambda(\tau_2)}})\} \end{aligned}$$

where $(\sigma_{i_{\Lambda(\tau_1)}}, \sigma_{j_{\Lambda(\tau_2)}}) \in \Lambda(\tau_1) \times \Lambda(\tau_2)$

The intuition behind Definition 5 is that we take between two constituent tensors the minimum of two constituents as the distance. Through this approach we link those time-series which have a small distance on a specific constituent. By that definition we also have some sort of a lower bound, since the distance between two time-series can not be smaller than any of the constituents.

Such an approach as provided by MinCT-dist would work well on coarse resolutions (such as one measure per week), but will fail on finer temporal resolutions, since we neglect with MinCT-dist repetitive patterns on daily, or weekly level. To mitigate this effect, we provide a different distance measure taking into account the different phase-shifts for each time-series. This leads to a first definition of our next distance measure:

Definition 6 (Phase-shift based distance) *Given two time-series τ_i, τ_j where each of the time-series has a common series of different frequency intervals like $f_\tau = [f_1, f_2, ..., f_i]$ where e.g. $f_1 = [23h, 35h]$. For each time-series τ a FFT is computed, obtaining the respective amplitudes a and phase-shifts p among the frequency f. The distance between two time-series τ_i and τ_j is defined as:*

$$d_{phase}(\tau_i, \tau_j) = \sum_{k=0}^{n-1} |p_{\tau_{i_{f_k}}} - p_{\tau_{j_{f_k}}}| \tag{1}$$

Definition 6 also relies on taking the top-k impactful frequencies for each of the frequency intervals. However, this phase-shift-centric approach comes with the major drawback that all frequency intervals have the same impact being considered as 'equally important'. Looking for a keyword like e.g. 'Beer garden' (Biergarten in German) the daily cycles may be more important since people query their search engine of choice for the nearest beer garden in the afternoon. Looking at an annual resolution, the summer may be of more relevance since people would look for a beer garden to spend their time in summer rather then in winter. To counteract this effect, we multiply each of the absolute phase-shift distances with their associated amplitude, yielding the next stage of our distance definition:

Definition 7 (Amplitude-weighted phase-shift based distance) *Given two time-series τ_i, τ_j. The amplitude-enhanced variant of the phase-shift based distance is defined as:*

$$d_{aw-phase}(\tau_i, \tau_j) = \sum_{k=0}^{n-1} a_{\tau_{i_{f_k}}} \cdot a_{\tau_{j_{f_k}}} \cdot |p_{\tau_{i_{f_k}}} - p_{\tau_{j_{f_k}}}| \tag{2}$$

By including the amplitudes of both time series as weighting factors like in Definition 7, frequency intervals with small impact (small amplitude) have a smaller impact on the whole distance. In the scenario that in both time series we have high amplitudes in the same frequency intervals, the phase-shift distance is also meaningful for the overall distance. However this approach will fail given the case that both time-series do not share impactful frequencies. This would result in small distances rendering it impossible to tell if they have similar phase shifts or if the distance itself is just insignificant. This effect is however mitigated by adding up the multiplied amplitudes without the phase shifts and normalizing Definition 7 by it.

Definition 8 (Normalized amplitude-weighted phase-shift based distance) *Given two time-series τ_i, τ_j. The normalized amplitude-enhanced variant of the phase-shift based distance is defined as:*

$$d_{naw-phase}(\tau_i, \tau_j) = \frac{1}{\sum_{k=0}^{n-1} a_{\tau_{i_{f_k}}} \cdot a_{\tau_{j_{f_k}}}} \sum_{k=0}^{n-1} a_{\tau_{i_{f_k}}} \cdot a_{\tau_{j_{f_k}}} \cdot |p_{\tau_{i_{f_k}}} - p_{\tau_{j_{f_k}}}| \tag{3}$$

With Definition 8 we have now a distance function which computes the distance with respect to (a) the phase-shift, (b) the amplitudes and thus the impact and (c) different frequency intervals. We shall see in the experiment section the effects of our defined distance function.

4 Experiments

In order to re-connect to the initial journal use-case from the introduction section, we take for the conducted experiments 10 time-series from Google trends[2],

[2] http://trends.google.com

namely: college dorm, fireworks, fitness, flu, healthy eating, holidays, i have a dream, sunburn, superbowl and turkey from a time frame between 2004 and today in the region of the United States. We have the relevance for each of these key terms which denote the vertical axis and each sample represents one month. This yields a total of 183 samples per key term. We used the FFT from the numpy library on which we wrote routines to extract the amplitude, phase-shift and vertical shift of each constituent, ranking them by amplitude and encapsulating them into a constituent tensor. We then performed a hierarchical clustering using first a periodic regression and then all three of our distance measures. Computing first the clustering for the regression models for each of the keywords we obtain the following result as seen in Figure 2.

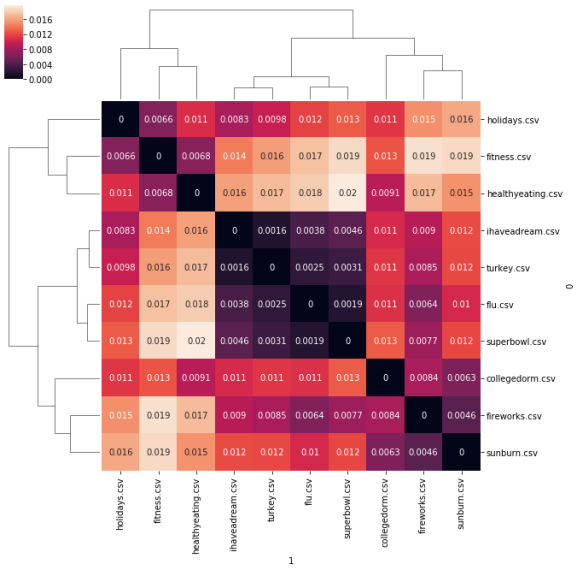

Fig. 2. Clustermap based on the periodic regression models

In the periodic regression model clustering we can observe one cluster block on the diagonal with low intra-cluster distance (center) and two with high intra-cluster distance. The clustering however does not seem to make that much of sense, since the Martin Luther King day (I have a dream) and turkey as well as flu or super bowl do not share any temporal closeness. They all occur annual (except flu maybe) but do not have the same time of the year (phase).

Comparing the regression result against the minimum constituent tensor distance, we obtain the cluster map as shown in Figure 3. It shows, that our approach yields a more meaningful result. We have two clusters with very low intra-cluster distances. The bottom-right cluster encompasses the keywords healthy eating, I have a dream, fitness and holidays. Taking a closer look at the keywords reveals that healthy eating and fitness are terms of high relevance during

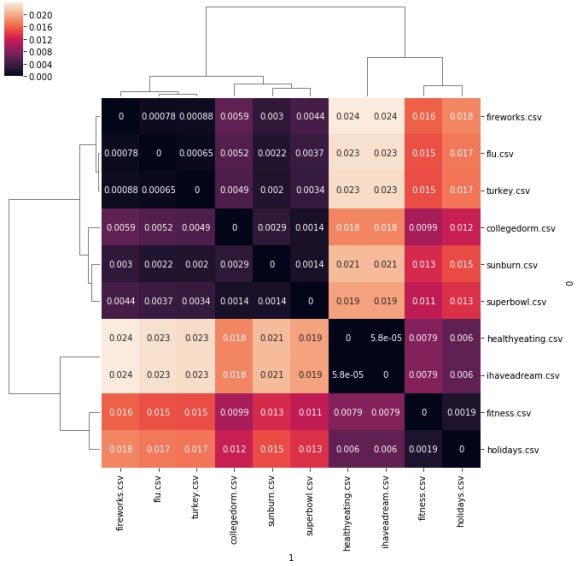

Fig. 3. Clustermap based on the minimum constituent tensor distance model

the beginning of the year (mostly new years resolutions). Also the Martin Luther King day is in January (20.01). In the top-left cluster we can see fireworks, flu, turkey, college dorm, sunburn and super bowl. Fireworks has its peak around the independence day (04.07) and sunburn also matches well, since we are in the summer time. Flu take some kind of special role, since the people seem to look more for this keyword in the search engine during October which is also close to thanksgiving (turkey). College dorm is close to the summer vacations and as such also suitable in the cluster. Just the term super bowl does not seem to fit in the pattern, since it is in February. Compared to the simple regression model, a clustering based on decomposed signal constituents seems to provide a much more intuitive result.

However, in this first experiment the time-series data was obtained on a weekly-basis. What happens if we take different keywords on different temporal resolutions? The frequency intervals encompass 8, 12, 24, 84, 168, 420, 735, 802, 981, 1103, 1471, 2207, 2943, 3924, 4415, 7064, 8830 and 11773 hours representing, working day, weeks, months, seasons etc. For each of the keywords for all of the frequency intervals their respective amplitudes and phase-shifts are computed through FFT. As keywords for this second data set we chose weather (Wetter), outdoor pool (Freibad), winter tires (Winterreifen), bakery (Bäcker), MVV (Munich local public transport), beer garden (Biergarten), sunburn (Sonnenbrand), tv program (Fernsehprogramm), Sunrise (Sonnenaufgang), Twitter, Club, Ikea, Christmas (Weihnachten), Ski and Firework (Feuerwerk). Computing clusters on this trend data with MinCT-dist yields the clustermap as seen in Figure 4. While club and twitter are together in one cluster, most of the rest of the data

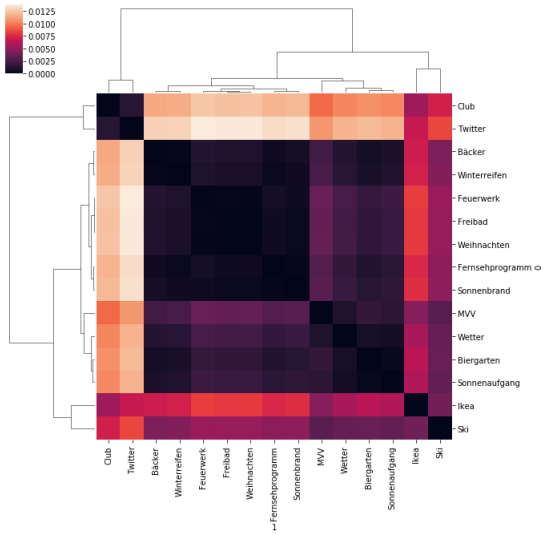

Fig. 4. Clustermap based on the minimum constituent tensor distance model using a second data set.

set is put into one massive cluster where Christmas and outdoor pool are put together. The result seems arbitrary.

If we look at the clustermap yielded by the normalized amplitude-weighted phase-shift based distance, one can see that more fine-grained clusters are revealed as seen in Figure 5. Weather and outdoor pool are put together into one cluster which makes sense, since people would go to an outdoor pool only if the weather is suitable. Further sunburn, beer garden, outdoor pool and MVV fall into the same cluster.

Taking a look at the time-series on a 30-days interval reveals no common periodic pattern as seen in Figure 6.

However if we change from the 30-days interval to a 7-days interval one can observe that all keywords have a high(er) number of queries between 8am and 10pm as seen in Figure 7. Individually the MVV peeks at around 8am and 9pm which corresponds roughly to the morning time people getting to work and the evening time returning from evening events. Outdoor pool peeks at around lunch time, which reflects the time people seeking to go swimming in the afternoon. In the late afternoon beer garden peeks which is due to the effect that people look for beer gardens to spend their evening at. Lastly we have as an interesting insight an increased query for sunburn around 9pm-10pm which may be speculated that people have got a sunburn and look for ways to heal or alleviate the pain from it.

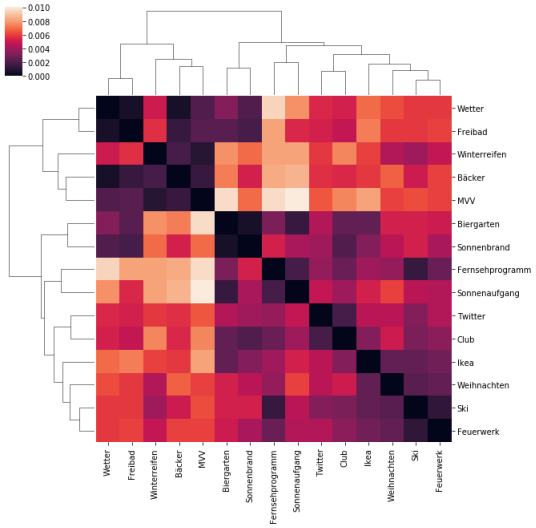

Fig. 5. Clustermap based on the normalized amplitude-weighted phase-shift based distance measure using a second data set.

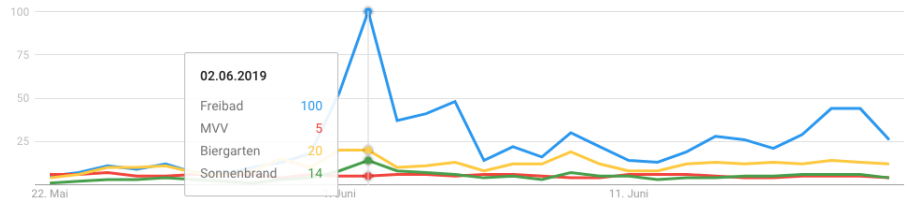

Fig. 6. Time-series for the keywords beer garden, sunburn, outdoor pool and MVV on a 30-days interval

5 Conclusion and Future Work

In this work-in-progress we have presented two distance functions for clustering time-series based on their signal constituents which emerged from an FFT. The first experiments on real-world data showed that our method reveals similarities between time-series on the constituent models which would not have been visible by simply clustering the regression models. While our first approach goes beyond a simple periodic regression, it fails when considering different frequency intervals with their different phase shifts. For this purpose we developed the normalized amplitude-weighted phase-shift based distance, which provides a better reflection of distances by putting phase-shift and amplitudes on different frequency intervals into account. As this work may seem like 'yet another distance measure for temporal data' it aims to explore the computation of distances based on single constituents, respecting the influences of phase-shift, amplitude and frequency intervals. We hope that this work leads to new discoveries on time-series

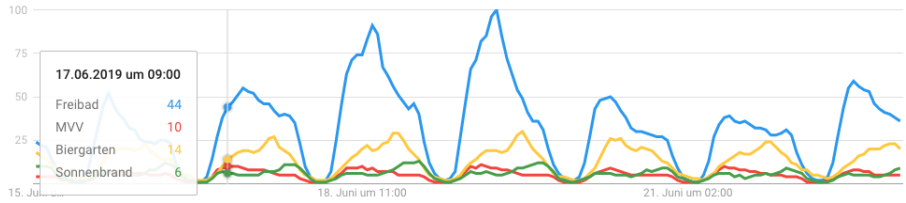

Fig. 7. Time-series for the keywords beer garden, sunburn, outdoor pool and MVV on a 7-days interval

data and the team in our fictional journal to more time-specific topics for their readers.

Acknowledgement

This work has been funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibilities for its content.

References

1. Aghabozorgi, S., Shirkhorshidi, A.S., Wah, T.Y.: Time-series clustering–a decade review. Information Systems **53**, 16–38 (2015)
2. Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complex fourier series. Mathematics of computation **19**(90), 297–301 (1965)
3. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences **95**(25), 14863–14868 (1998)
4. Harmer, S.L., Hogenesch, J.B., Straume, M., Chang, H.S., Han, B., Zhu, T., Wang, X., Kreps, J.A., Kay, S.A.: Orchestrated transcription of key pathways in arabidopsis by the circadian clock. Science **290**(5499), 2110–2113 (2000)
5. Hensman, J., Rattray, M., Lawrence, N.D.: Fast nonparametric clustering of structured time-series. IEEE transactions on pattern analysis and machine intelligence **37**(2), 383–393 (2015)
6. Ketchen, D.J., Shook, C.L.: The application of cluster analysis in strategic management research: an analysis and critique. Strategic management journal **17**(6), 441–458 (1996)
7. Langmead, C.J., Yan, A.K., McClung, C.R., Donald, B.R.: Phase-independent rhythmic analysis of genome-wide expression patterns. Journal of computational biology **10**(3-4), 521–536 (2003)
8. Lucas, B., Shifaz, A., Pelletier, C., ONeill, L., Zaidi, N., Goethals, B., Petitjean, F., Webb, G.I.: Proximity forest: an effective and scalable distance-based classifier for time series. Data Mining and Knowledge Discovery **33**(3), 607–635 (2019)
9. Popivanov, I., Miller, R.J.: Similarity search over time-series data using wavelets. In: Proceedings 18th international conference on data engineering. pp. 212–221. IEEE (2002)