Robust Periodicity Detection Algorithms*

Srinivasan Parthasarathy
Department of Computer Science and Engineering
Ohio State University

Sameep Mehta
Department of Computer Science and Engineering
Ohio State University

Soundararajan Srinivasan
Biomedical Engineering Department
Ohio State University

ABSTRACT

In this article, we present an algorithm for detecting periodicity in time series datasets. The algorithm leverages the frequency characterization and autocorrelation structure inherent in a time series to estimate its periodicity. We extend the methods to handle non-stationary time series by tracking the candidate periods using a Kalman filtering technique. We also address the interesting problem of finding multiple interlaced periodicities. We show the effectiveness of our approach on many publicly available real datasets and demonstrate the robustness of our approach in the presence of noise and missing data values.

1. INTRODUCTION

Periodicity is an interesting property of many time series datasets. A period can be informally defined as a self repeating pattern. This pattern provides useful information about the inherent structure in cyclic datasets. Cyclic processes are ubiquitous, e.g. rotation and revolution of earth, solar activity, climate and temperature changes, power consumption in urban areas, human heart beat and respiration processes and number of hits on websites. The importance of periodicity can be gauged by the fact that basic concepts such as the number of hours in a day and number of days in a year are motivated by the periodic nature of earth's movement. The human respiration pattern is another examples of an important periodic process. Deviations from normal periodic behavior are observed in many diseases. Periodicity can be used to derive the signature of normal breathing patterns and thereby facilitating abnormality detection. The power consumption in urban areas also tends to demonstrate a periodic behavior which allows energy companies to analyze the power consumption patterns to facilitate proper planning. Similarly, operating systems and database systems often use periodicities to detect the periods of low activities to schedule maintenance tasks. Many day-to-day activities of human beings also show periodic behavior. For example, a person walking, running, waving the hand and jumping are remarkably periodic processes. Cutler and Davis [3] showed the usefulness of periodicity for object classification and discriminating among different motion types. Periodicity not only helps to understand the properties of a single time series, but can also capture complex relationships among multiple time series. For example, our analysis shows that heart rate, chest volume and blood oxygen concentration can be related through their periodic patterns. Recently, Vlachos et al. [19] also demonstrated the use of structure based descriptors for clustering of time series. The authors showed that a periodicity based feature vector often outperforms shape based descriptors for a variety of signals.

To reiterate, if detected robustly, periodicity can aid in various time series tasks like classification, clustering, compression, abnormality detection and discovery of hidden relationship among attributes. The periodicity detection task is not difficult in very simple cases such as when the signal is a sinusoidal wave -*a single period for the whole data*. Unfortunately, such datasets are rarely found in real life. Most of the time series datasets exhibit one or more of the following properties: i) non-stationarity, ii) interlaced cyclic patterns and iii) data contamination.

Non-stationarity is the property attributed to the change in the underlying distribution of the data generating process. This change can result in different periodicities at different times. Many datasets also exhibit more than one periodicity in the same part of the data, i.e. the series is composed of multiple periods. One such example, *the SunSpot series*, was reported by Kanjilal et al. [9]. The authors reported presence of three periods at 10, 11 and 12 years. *Please note that multiple periodicities in the non-stationary case should not be confused with multiple periodicities in the interlaced case. In the former scenario, the periodicity is gradually changing due to change in the data and only one period is observed in each portion of data, while in the latter case multiple periods are present at every time instant.* Data contamination often occurs due to measurement errors in data acquisition methods resulting in noisy datasets. Incomplete datasets or missing values further compound to the problem. Pearson et al. [14] reported that it is common for many datasets to have as much as 30% of missing values.

In this article we present algorithms that take into account the above-mentioned properties of the dataset in detecting their period(s). The proposed techniques find the periodicity in a stationary time series by using a combination of time-frequency and autocorrelation analysis. Non-stationarity is handled by using sliding window techniques with Kalman filtering to compute the periodicities detected in each window. Multiple interlaced periods are found by repetitively removing the high energy components of the data and finding the periodicity on the remaining signal. We use a comb filter to efficiently accomplish this task. Missing data is handled by treating the data as an unevenly sampled signal. The time-frequency and autocorrelation analysis lends itself easily to the analysis of unevenly spaced datasets. Finally, we demonstrate that our design of using the infor-

*Contact email: srini@cse.ohio-state.edu

mation simultaneously from time-frequency and autocorrelation analysis makes the algorithms robust to the presence of noise. To reiterate, the proposed algorithm is able to handle non-stationary, multi-periodicity and contaminated datasets. We empirically evaluate performance of our algorithm on various publicly available datasets originating from various domains such as astronomy, meteorology, medicine, mathematics, automotive, video surveillance and geography. We also present results describing how this approach can be used to discover hidden and complex relationships among attributes of multi-dimensional time series.

The rest of the article is organized as follows: In Section 2, we present background and related work. Section 3 presents the details of our algorithms, along with motivating examples. Results are given in Section 4, followed by discussion and conclusions in Section 5.

2. BACKGROUND AND RELATED WORK

In this section, we define the period of a random process provide a brief survey of existing algorithms for periodicity detection.

Definition 1. A wide-sense stationary random process $z(t)$ is said to be be mean-square periodic, if $\exists\ T > 0$ such that

$$R_{zz}(\tau) = R_{zz}(\tau + T), \forall\ \tau,$$

where $R_{zz}(\tau)$ is the autocorrelation fucntion corresponding to $z(t)$. The period of this process is then defined to be the smallest such T.

It can be shown [16] that a mean-square periodic process also obeys our intuitive definition of periodicity:

$$z(t) = z(t + T). \tag{1}$$

Existing periodicity detection algorithms can be broadly classified in two groups: time domain methods and frequency domain methods. Time domain methods make use of autocorrelation functions while frequency domain methods make use of spectral density functions.

The primary motivation of using time domain methods [2] (also known as autocorrelation based methods) stems from the observation that if $z\,(t)$ is periodic with period T, then $R_{zz}(\tau)$ will also exhibit a period of T. The peak of $R_{zz}(\tau)$ occurs when $\tau = 0$. Depending on how "strongly periodic" $z(t)$ is, the peak value may also obtained at the values of τ that correspond to the period T and multiples of T. Time domain methods are useful and efficient in detecting periodicities when the time series is approximately characterized by a sinusoid uncontaminated by noise. However, for other signals, the performance degrades rapidly.

Spectral or frequency domain methods decompose a signal into it's constituent frequencies. The main motivation behind the use of frequency domain methods is that the power spectral density of $z(t)$ is a line spectra consisting of impulses located at multiples of $\frac{2\pi}{T}$ with heights (areas) determined by the corresponding Fourier coefficients of $R_{zz}(\tau)$. Recall that $R_{zz}(\tau)$ is the autocorrelation with time lag of τ corresponding to $z\,(t)$. If, the Fourier co-efficients are extracted directly from the signal, the resulting decomposition is known as the periodogram [17]. Although, frequency based methods mitigate some the drawbacks of time domain methods, power loss of the impulsive frequencies due to spectral leakage pose a serious problem [12].

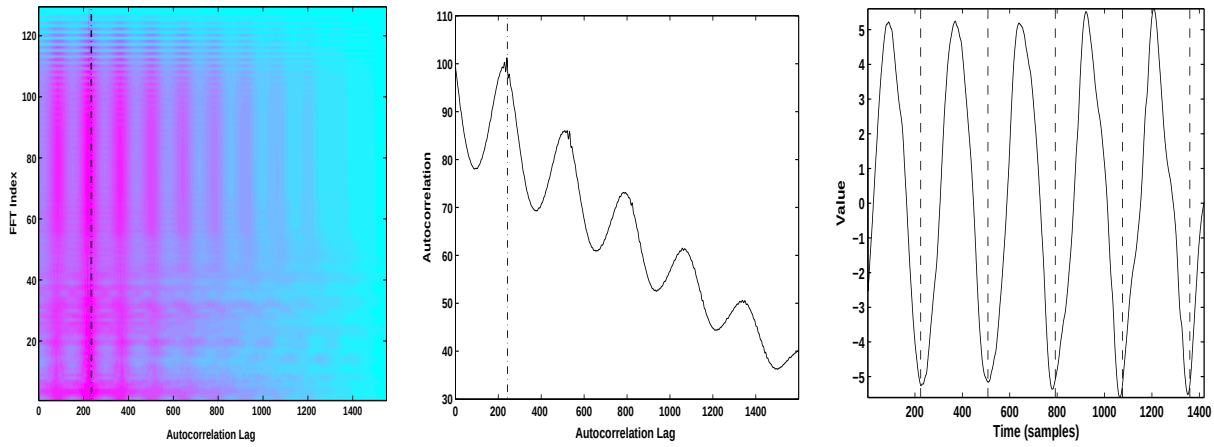
Recently, Vlachos et al. [19] identified some of these issues and presented an algorithm which sequentially uses both time and frequency domain methods for periodicity detection. The authors first detect potential candidates of periodicity by using the periodogram. In the next stage, all candidate periods are evaluated against the auto-correlation function (ACF). The periods which do-not lie on the "hill" of ACF are discarded. The rest of the periods are refined further, if needed. However, the issues of non-stationarity, contaminated datasets and multiple periodicities are not addressed by the authors. A combination of time and frequency domain methods have also been proposed in the context of pitch detection for speech signal processing (see e.g. [11, 20]). Typically, these methods filter the speech signal using a bank of bandpass filters that simulate the human cochlea. Each filter output then undergoes a nonlinear compression, followed by the computation of the autocorrelation. The autocorrelations are then summed across different filters and the largest peak with a certain range in the resulting sum is deemed to be the pitch period [11]. However, the algorithms are not generalizable across domains and often find applicability in the pitch estimation problem only.

Pearson et al. [14] proposed a spectral estimation technique for periodicity detection in non ideal datasets. However, no discussion of non-stationarity and multiple periodicities was provided. Kanjilal et al. [9] used eigen value analysis instead of Fourier transform to perform spectral decomposition. Their algorithm performed well in detecting multiple interlaced periods. However, their technique needs a separate training phase that utilizes a part of the data. Moreover, the authors make an unrealistic assumption that this training data is noise-free.

Another body of related work stems from the sequence mining literature. Ozden et al. [13] presented an algorithm to find periodic patterns in transactional datasets. A time stamp is associated with each transaction. The primary goal is to find association rules which repeat themselves throughout the dataset. Han et al. [6] discussed an algorithm for mining partial periodic patterns. Yang et al. [21] extended the algorithm to mine partial and asynchronous patterns. Finally, Yang and Lee [22] proposed algorithms to mine non-redundant partial periods. However, all these algorithms are developed for transactional datasets, which represent discrete time sequence. In the context of continuous time series, some form of binning algorithm has to be employed as a pre-processing step. The final results will therefore be extremely sensitive to the number of bins and the discretization scheme.

3. ALGORITHMS

In this section we first present an algorithm for detecting the periodicity in a stationary time series. The algorithm yields robust estimates of the period by analyzing the data in the joint frequency-autocorrelation domain. We then discuss the shortcoming of this algorithm when it is applied to non-stationary datasets. For such datasets, we propose a sliding-window approach that independently estimates a period within each window. These estimates are then smoothed using a fixed-interval Kalman smoother. Finally, we discuss an extension to handle time series that contain multiple interlaced periodic processes within. The key intuition is to iteratively estimate the period in the data



(a) (b) (c)

Figure 1: Determination of the period of a wideband signal. (a) Spectrogram of a motor current signal (best viewed in color). (b) The corresponding summed autocorrelation. (c) A pulse train corresponding to the detected period overlaid on the original signal.

by successively subtracting the harmonic components corresponding to the detected period.

3.1 Periodicity in Stationary time series

We first propose an algorithm to detect periodicity in a stationary time series data. The algorithm finds the periodicity by combining time-frequency and autocorrelation analysis. The signal is first decomposed into a short-time frequency representation by using a spectrogram analysis. Specifically, a 256-point discrete Fourier transform (DFT) is computed on successive 256 length portions of the signal generated by applying a running rectangular window to the signal. Consecutive windows are shifted by 1 sample. The outputs from the short-time Fourier transform analysis are then used to compute the corresponding short-time magnitude spectra. Spectrograms provide a powerful tool to analyze a signal in time and frequency domain simultaneously. They monitor the evolution of spectral components across time. A spectrogram plot for the motor current signal from [5] is shown in Figure 1(a). The x-axis is time in increasing order and y-axis is frequency in increasing order. The lighter the color in the spectrogram, the more the signal power (magnitude) at that frequency. Notice that the signal is wideband and contains harmonics across frequencies. However, there is also considerable smearing of the harmonic powers. This could be mitigated to a certain extent by using either more points in the DFT analysis or by a window designed to minimize smearing (e.g. Kaiser) [12]. However, this comes at the cost of assuming either that the signal has the same period for the extended analysis length or certain *a priori* properties of the signal that would enable the optimal window design [12].

It can also be seen from Figure 1(a) that the fine-shift along the time dimension causes the spectrogram to retain the periodicity property along its time axis too. Note that this occurs across different frequencies. To exploit this property, the evolution of each DFT coefficient across time is considered to form a time series. However, each DFT coefficient cannot be analyzed independently due to the smearing problem described above. Hence, we use the discrete cosine transform of the DFT magnitude to partially decorrelate [12] the different spectral time series. Then, for each such series the discrete-time autocorrelation is computed as

$$R_{ZZ}[m] = \sum_{n=0}^{N-m-1} (Z[n]Z[n+m]), \qquad (2)$$

where $Z[n]$ is the $n^{'th}$ order decorrelated coefficient. The resulting autocorrelations are summed across different orders. Figure 1(b) shows the plot of the summed autocorrelations for the motor current data. Note that the resulting autocorrelation is periodic. The various peaks in Figure 1(b) denote the integer multiples of the period. We calculate the gradient at each point of this curve and pick the point that has highest gradient. The corresponding autocorrelation lag index gives the period of a signal. For example, the first peak in Figure 1(b) is the dominant period for the motor current case. For verification, Figure 1(c) shows the original signal overlaid with a pulse train at the detected periods given by the dotted lines.

Aside from combining the individual advantages of frequency decomposition and autocorrelation, the joint time and Fourier analysis improves the robustness of periodicity detection. For example, if the original broadband data were to be corrupted by narrowband noise, the periodic structure in the noise-free frequency components would still allow for accurate estimation. Similarly, addition of time-limited broadband noise is handled by integrating the information spectral information across across time.

3.2 Periodicity in Non-stationary time series

Definition 2. A random process $z(t)$ is non-stationary, if $\exists\ n \in \mathbb{Z}^+$ such that the n^{th} order probability distribution function changes across time.

In this paper, we consider the particular case of non-stationary random processes that possess time-varying periods. The

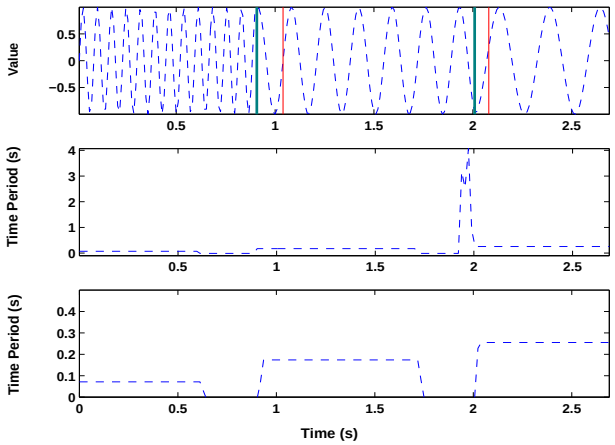


Figure 2: Periodicity Detection for the non-stationary case.

top graph in Figure 2 shows an example of one such non-stationary time series. The dataset has three different frequencies, which can be easily seen to occur at distinct time intervals. Our approach to handle such signals is motivated by the observation that even though the periodicity varies globally but it remains unchanged locally. Specifically, we apply a sliding rectangular window of length M with an overlap between adjacent windows of length o. Within each window we use the stationary version of the algorithm, as explained in Section 3.2, to find the periodicity. Since the point where the change in period occurs is unknown, we will get some windows within which the data generated will encompass different periods. One solution to this problem is to use a large value of o. In the limit, o can be as large as $M-1$. However, this incurs a large computational cost. In practice, one may trade-off the computational burden with the desired time accuracy. The middle graph in Figure 2 shows the result of applying stationary version on the windowed signal. Note that the stationary version does not perform well in the regions where the signal changes its periods. There is a lot of noise in results (notice the big peak). The red lines in top graph indicates the time instances where the period changes are detected. Notice that the results contain small errors.

We smooth the results using a Kalman Filter [1]. The time-variation of periods is modeled as an auto-regressive (AR) process. For simplicity, we have assumed that this variation can be modeled as a second-order AR process. The state space model of this system is

$$x(m) = F(m)\,x(m-1) + Gv(m), \tag{3}$$

$$y(m) = Hx(m) + w(m). \tag{4}$$

In the above, $y(m)$ is the estimated period within a particular window as described above; m is the index of that window.

$$F(m) = \begin{bmatrix} a_1(m) & a_2(m) \\ 1 & 0 \end{bmatrix}, \tag{5}$$

where $a_1(m)$ and $a_2(m)$ correspond to estimated AR coefficients at window m. We let $G = [1\ 0]^T$ and $H = [1\ 0]$. The system Gaussian noise $v(m)$ and the observation Gaussian noise $w(m)$ are zero mean with variances $vv(m)$ and variance $vw(m)$. The conditional state mean $x(m|m-1)$ that corresponds to the desired period sequence and the error covariance $V(n|n-1)$ are predicted as:

$$x(m|m-1) = F(m-1)\,x(m-1|m-1), \tag{6}$$

$$V(m|m-1) = F(m-1)\,V(m-1|m-1)\,F^T(m-1) + GQ(m-1)\,G^T; \tag{7}$$

and tracked by the Kalman filter as:

$$x(m|m) = x(m|m-1) + K(m)\,(y(m) - Hx(m|m-1)), \tag{8}$$

$$V(m|m) = (I - K(m)\,H)\,V(m|m-1), \tag{9}$$

where $K(n)$ is the Kalman gain computed as:

$$K(m) = V(m|m-1)\,H^T\left(HV(m|m-1)\,H^T + vw(m)\right)^{-1}. \tag{10}$$

Due to the overlap between consecutive windows, we expect a smooth change in the detected period. Hence, the observations noise related variance $vw(m)$, is set to be the absolute time-difference of $x(m-1|0\ \ldots\ m-1)$. The parameters of the AR model a_1, a_2 and vv are estimated in each window using the Yule-Walker method [17]:

$$\begin{bmatrix} r_{xx}(0) & r_{xx}(-1) & r_{xx}(-2) \\ r_{xx}(1) & r_{xx}(0) & r_{xx}(-1) \\ r_{xx}(2) & r_{xx}(1) & r_{xx}(0) \end{bmatrix} \begin{bmatrix} 1 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} vv \\ 0 \\ 0 \end{bmatrix}, \tag{11}$$

where $r_{xx}(k)$ is the autocorrelation sequence corresponding to $x(m-1|0\ \ldots\ m-1)$.

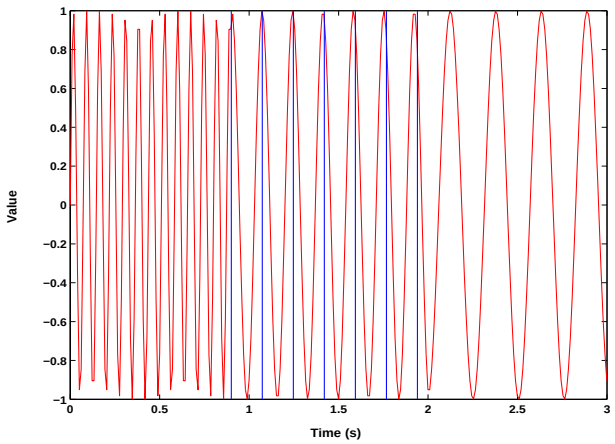


Figure 3: Periodicity Detected in the non-stationary case after Kalman Smoothing. Periods shown only for a selected portion.

Finally, the filtered periods are smoothed by the following fixed-interval Kalman smoother.

$$A(m-1) = V(m-1|m-1)\, F^T(m-1)\, V^{-1}(m|m-1), \tag{12}$$

$$x(m-1|J) = x(m-1|m-1) + A(m-1)\left(x(m|J) - x(m|m-1)\right), \tag{13}$$

$$V(m-1|J) = V(m-1|m-1) + A(m-1)\left(V(m|J) - V(m|m-1)\right) A^T(m-1), \tag{14}$$

where J is the number of sliding windows applied to the signal. The bottom graph in Figure 2 shows the final result after Kalman smoothing. Notice that the three different frequencies are correctly identified. The vertical bold (green) line in top graph shows the time instances which our algorithm detects as the places of changes in signal periods. Note also that the three periods detected also closely match the actual values and we are able to segment the signal using periodicities. Figure 3 shows the original signal with detected periods (represented by vertical line segments).

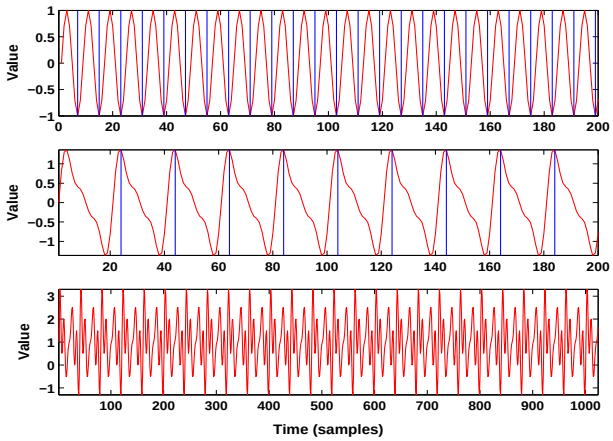


Figure 4: Time series with multiple interlaced periods.

3.3 Detection of Multiple Interlaced Periodicities

Thus far, we have presented algorithms which perform when there is only one period in whole or parts of a time series. However, many data sets have interlaced periods occurring simultaneously. For example, consider the time series in the bottom row of Figure 4. The data has two alternating frequencies. To handle such cases, we adopt a sequential detection and pruning approach. We use the existing algorithm to find the primary period. This period is used to design a comb filter [8]. A comb filter, in general, is a filter that resonates a "selected" frequency and all harmonics of that frequency. We set the "selected" frequency to be the inverse of the detected period, thereby constructing a filter which will only let pass the harmonics corresponding to the detected period.

The transfer function of a comb filter is given by

$$P(l) = \frac{1 - l^{-e}}{1 - (1/e)\, l^{-e}}, \tag{15}$$

where e is the primary period detected. This filter is used to "select" those harmonic components of the signal that occur at integer multiples of $1/e$. The is accomplished by:

$$S(l) = z \otimes P(l) \tag{16}$$

We then subtract the comb filtered version $S(l)$ from the original signal, $z(t)$ and re-run the algorithm on resulting data. This process is continued iteratively until the comb filtered signal contributes to no more that 10% to the original signal power. The bottom graph in Figure 4 shows the original signal. The middle part shows the period corresponding to first primary frequency. The top part shows the signal after the primary frequency is removed. We super impose the period detected in the remaining signal.

3.4 Robustness to Missing-data

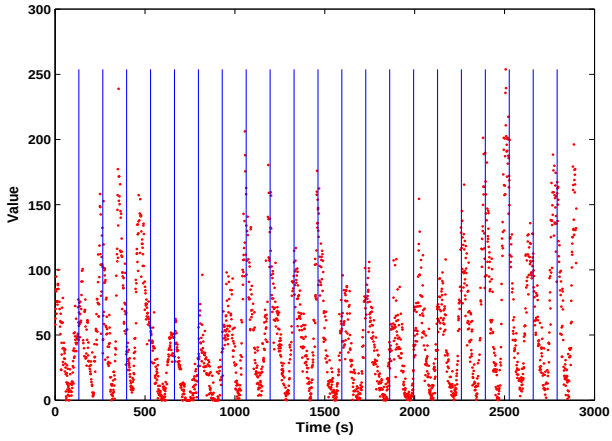


Figure 5: Periodicity detection for an incomplete dataset.

Incomplete datasets potentially pose another problem to our approach as missing data corrupts the accuracy of the DFT calculations. Since this is the first step in all our algorithms, any errors caused in this step can be catastrophic. To mitigate the effect of missing data on our algorithms, we consider the time series with missing data to be a time series with unevenly sampled data and represent each sample by a 2-tuple *(time instant, actual value)*. This enables us to use the algorithm for DFT calculations with unevenly sampled data as proposed by Press and Rybicki [15]. Figure 5 shows a time series with 20% missing data overlayed with a pulse train at the detected periods. From the figure, we can see that the proposed algorithm is robust to missing-data. A detailed study of the effect of missing data is presented in next section.

Summary

In this section, we have proposed algorithms for periodicity detection that enable us to handle non-stationary datasets containing possibly simultaneous periodic processes. Hence, given a particular dataset, our analysis starts by first applying a sliding window to the signal. Then, within each window, we estimate the dominant period by using the Fourier decomposition followed by an autocorrelation analysis. These estimates are then smoothed across time. In those windows that yield a non-zero period estimate, we reexamine the data for other potential periods. An iterative algorithm

that first detects a plausible period and then subtracts its harmonic components performs this reexamination. Finally, we have demonstrated how our algorithms are able to handle missing-data by performing Fourier analysis with unevenly sampled data.

4. RESULTS

In this section we present the results of our proposed algorithm on various datasets. These datasets originate from very different domains like astronomy, meteorology, medicine, mathematics, automotive, video surveillance and geography. All the datasets unless otherwise noted are from The UCR Time Series Data Mining Archive [5].

4.1 Dataset Description and Detected Periodicities

Figure 11 shows the results on 14 publicly available datasets. The plot shows the actual series and the vertical lines show the periodic segment. The gap between two successive vertical lines specifies the detected period. The x axis represents the time (or index) and the y axis represents the actual value of the series.

Motor Current- This dataset classifies the state of motor in 3 classes: *Healthy, Broken Connector and Broken Bar*. Figures 11(a) and 11(b) show the results on one broken connector and one broken bar. We have already shown the results for the healthy case in Section 3, Figure 1(c). The size of each series is 1×1500.

Eamonn noGun- This dataset is a two-dimensional time series of x and y co-ordinates. Figure 11(c) shows the detected periodicity for only the x co-ordinates. Figure 11(d) shows the results for the y-coordinates. The size of data is 2×8853.

Great Lakes- This data monitors the water level of 5 lakes *Erie, Huron, Ontario, StClair, Superior*. Figures 11(e) and 11(f) shows the results for *Ontario* and *StClair* respectively. The size of the dataset is 5×984.

Pseudo Periodic- This is a synthetic dataset where the signal appears highly periodic, but never exactly repeats itself. This can be thought of as the presence of random noise in the data. The datasets contains 10 pseudo periodic datasets generated from 10 different simulation runs. We show the results on first 4 signals shown in Figure 11(g-j). It is evident that even when the signal is not exactly repeating, we are able to find the right periodicity. The total size of dataset is $10 \times 100,000$. For the figures we have reduced the dataset size by sampling every 50^{th} point.

A- A.dat dataset was used in the Sante Fe competition[1] . A.dat is a univariate time record of a single observed quantity, measured in a physics laboratory experiment. The size of A.dat is 1×1000. A.cont provides approximately $10,000$ points beyond the end of the competition data set. Please note that since the dataset comes from same experiments, the inherent periodicity should be same. This is precisely what we found. A period of 9 was found in both the datasets as shown in Figure 11(k-l). *Also note that it may appear that A.cont has more periods. This is because the time range shown*$(1-200$ *is double the one shown for A.dat* $(1-100)$. This was done intentionally to show that the signal's amplitude and mean do not effect our periodicity detection algorithms.

Daily Temperatures - Melbourne, Australia- This dataset contains the average minimum and maximum temperature of Melbourne, Australia[2] for a period of 10 years resulting in 3650 values. In both series, we found the period to be exactly 365 days, i.e. 1 year. Figure 11(m) and 11(n) show the result for minimum and maximum daily temperature respectively.

Flutter: Non Stationary Dataset- This data originates from industry. No other description about the data is available. The size of the data is 1024×2. The second attribute is the attribute of interest and is used for analysis. This data exhibits non-stationarity. We used our algorithms for segmentations and periodicity detection algorithm. We used a window size of 10. Figure 6(a) shows the original data. The vertical lines show the result of the segmentation algorithm; we found 4 different segments. Figure 6(b) shows the detected period in the second segment $[200-320]$. Similarly, figure 6(c) shows the result on the next segment $[340-500]$. *Please note that periodicity was not detected on a priori segmented signals. The segmentation is a by-product of our algorithm when dealing with non-stationary signals.*

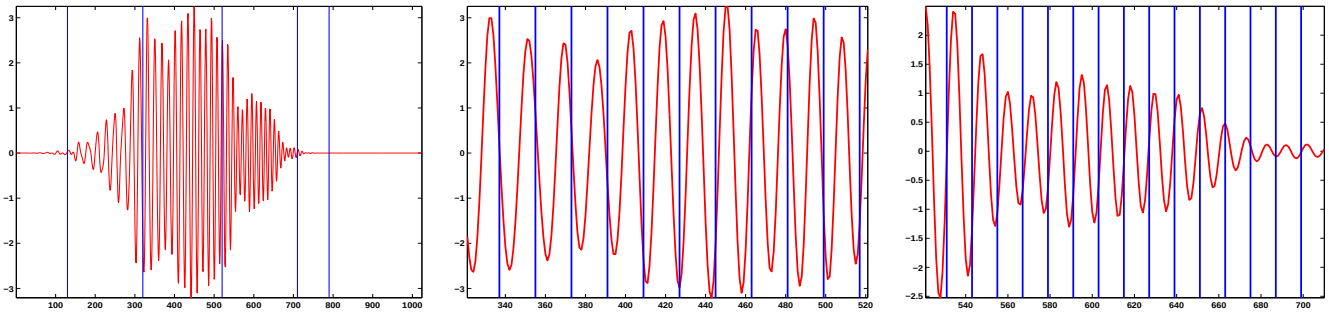
Sunspots: Multiple Interlaced Periodicities- The dataset provided monthly mean sunspot numbers for 240 years. The size of the datasets is 1×2880. The periodicity in sunspot series dataset is known to be 11 years which is very close to actual observed period of 11 years and 27 days. We successfully detected 132 months (11 years) as the primary period. However, when searching for the presence of other possible interlaced periods, we found two additional periods at 120 months (10 years) and 540 months (45 years).

Figure 8(a) shows the original signal and the primary period of 132 months. The next plot shows the signal after the frequencies corresponding to the primary period are removed. The detected period is 120 months. Figure 8(c) shows the final period (540 months), detected after filtering the frequency components associated with the secondary period. Kanjilal et al. [9] also report three periodicities at 11, 10 and 12 years (in order) on this dataset. We did not find the period at 12 years. *The authors would like to point out that period of* 45 *years is not simply concatenation of* 4 *periods of* 11 *years. The last periodicity was detected from the dataset after we have already removed the frequency component corresponding to the first and the second periods. Therefore, we are not grouping multiple periods of* 11 *years to find this larger period.*

ECG Dataset: Discovering Hidden Relationships- Detection of periodicities can reveal hidden relationships between seemingly uncorrelated variables. For example, the top panel in Figure 9 shows how the three variables of heart rate (blue), chest volume (green) and blood oxygen con-

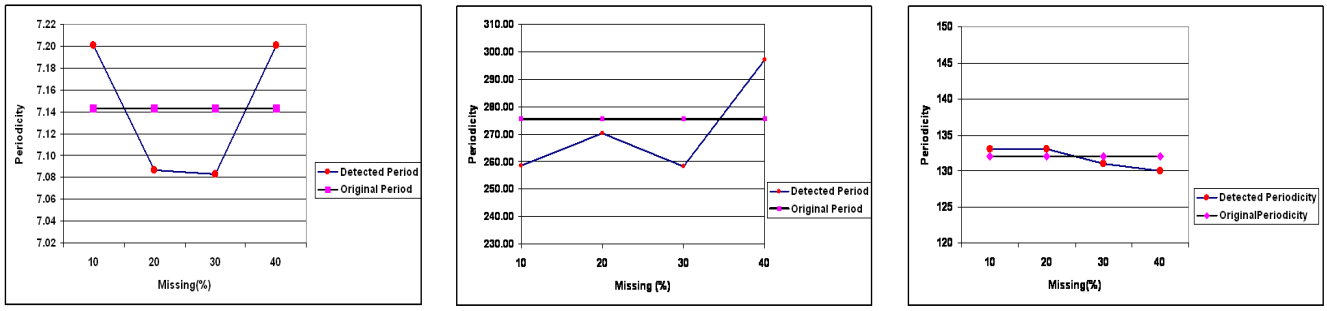
[1]http://www-psych.stanford.edu/~andreas/Time-Series/SantaFe.html

[2]http://www-personal.buseco.monash.edu.au/~hyndman/TSDL/



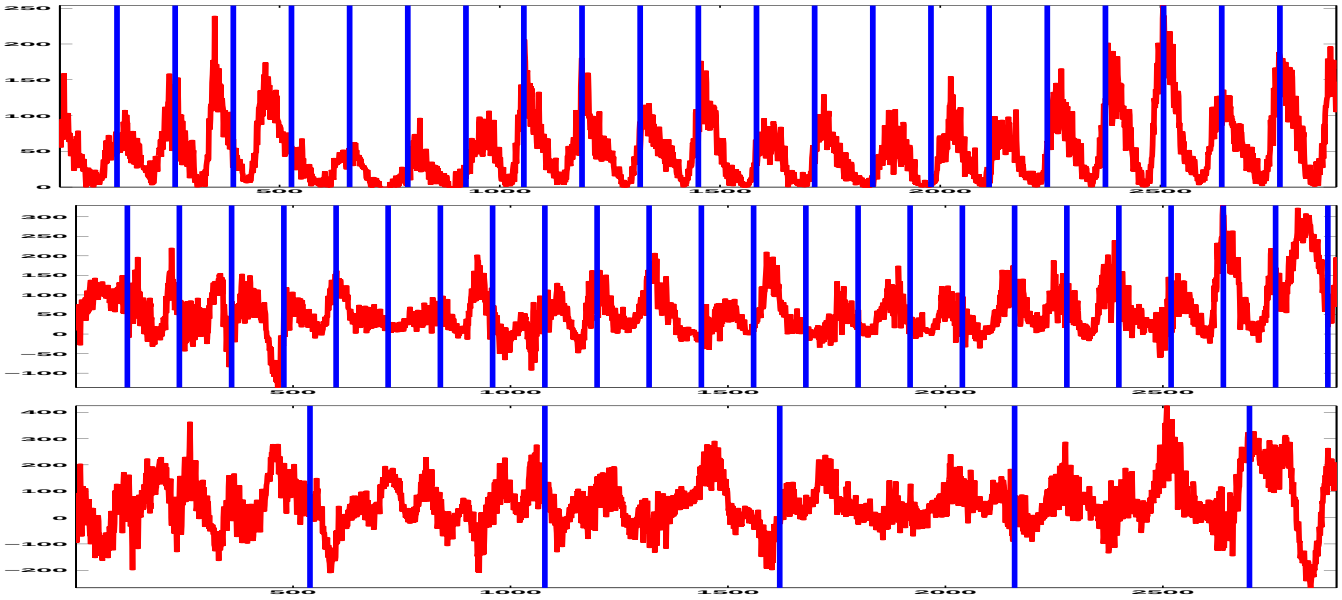
(a)Original Dataset with segmentation (b) second segment (c) Third segment

Figure 6: Non Stationary dataset "Flutter" containing different periodicities detected in different parts.



(a) Sine Wave (b) Motor Current (c)Sunspot

Figure 7: Effect of Missing Values on Periodicity Detection



(a) Primary Period at 132 months (b) Period at 120 months (c) Period at 540 months

Figure 8: Multiple Interlaced Periodicities found in the Sunspot dataset

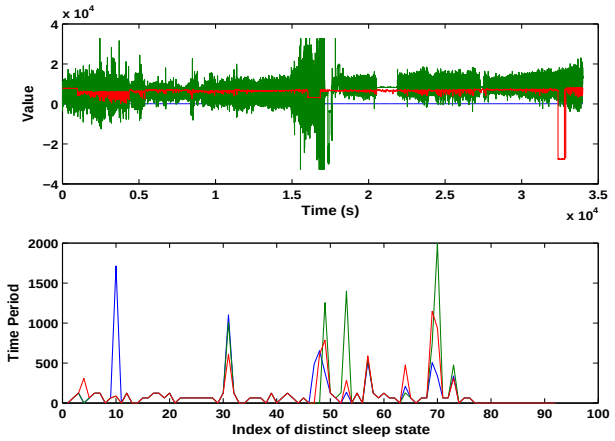


Figure 9: Periodicity detection for multivariate ECG dataset

centration (red) vary in the Sante Fe ECG dataset[3]. As the figure shows, the sample values of these variables are uncorrelated although from physiological studies we know that they must be correlated. Indeed, this correlation may be observed in the variation of the detected periodicities of the three variables. The correlation coefficients between the trajectories of the detected periods of heart rate and chest volume is 0.39, between heart rate and blood oxygen concentration is 0.52 and between chest volume and blood oxygen concentration is 0.78. The corresponding values computed directly on the samples are -0.07, 0.67 and -0.02. Hence, only the the correlation between the heart rate and the blood oxygen level can be observed directly. Therefore, we believe that periodicity can be a useful derived attribute that can reveal how seemingly uncorrelated variables are related.

4.2 Periodicity Detection in Incomplete Dataset

In this section we show the results highlighting the robustness of our algorithm when handling incomplete datasets. We choose three different datasets i)a synthetic sine series generated at 70 Hz, ii) the healthy motor current data (see Section 3) and iii) the Sunspot numbers. We found the primary period in each of the dataset. Next, we randomly removed 10% of the data and detected the period again. This experiment is run systematically by varying the percentage of missing data from 10% to 40%. Figure 7 show the detected periods on the incomplete data. For comparison, we also show the periodicity with complete data (shown by a horizontal line). Even when 40% of the data is missing, the detected periodicity is very close to the true one. The maximum deviation as percentage of true periodicity is 0.84%, 7.84% and 1.52% for the sine wave, the motor current and the SunSpot numbers respectively. Note that the motor current data is sampled at a very high sampling rate of 33300 Hz. Hence, even a very small error in the estimate of the period in terms of samples translates into a large error in the period estimate in terms of seconds.

4.3 Periodicity Detection in Presence of Noise

We now discuss the performance of our algorithm in the presence of noise. We generated a series of same length as the original signal by sampling from a uniform distribution $U \sim [-1, 1]$. We then scaled the noise such that $\sum(signal)^2 = \sum(noise)^2$. Finally, we multiplied noise with β, where β specifies the signal to noise ratio and added the resulting scaled noise to the signal. We systematically varied β in the range $[0, 1]$ in steps of 0.1. At $\beta = 0$, no noise is added and at $\beta = 1$, the contribution of noise and signal are the same. We ran the experiments on the same three datasets used in studying impact of missing value. For all three datasets we found the same periodicity till $\beta = 0.9$. At $\beta = 1$, the quality of results start degrading and the correct periodicity is not discovered. Thus, even when the noise strength is as high as 90% (at $\beta = 0.9$) of original signal, we are able to identify the correct periodicity information. We attribute this nice property of our algorithm to the simultaneous use of information from both time-frequency autocorrelation domains.

4.4 Avoiding False Positives

In this section, we evaluate the quality of results for the case where no period is present in the input signal. Instead of using a signal which is comprised of noise, we have decided to use a one dimensional logistic map [10, 18]. The function is recursively defined as $x(k) = x(k-1) + r * (1 - x(k-1))$. The function generates a strictly periodic signal for values of $r \in [3.0, 3.57]$. However, when $r > 3.57$, the signal is chaotic. We generated two signals with $r = 3.5687$ and $r = 3.92$. We then concatenate these two signals to form one large signal. Our periodicity detection results are shown in Figure 10. The top graphs shows the segmentation of the signal. We correctly detect two different signal periods. The periodicity in the first part was detected as 32, which matches the theoretical value of the periodicity of this signal at $r = 3.5687$. However, in the next segment, no period was detected. Figure 10(b) shows the resultant periods on the first segment.

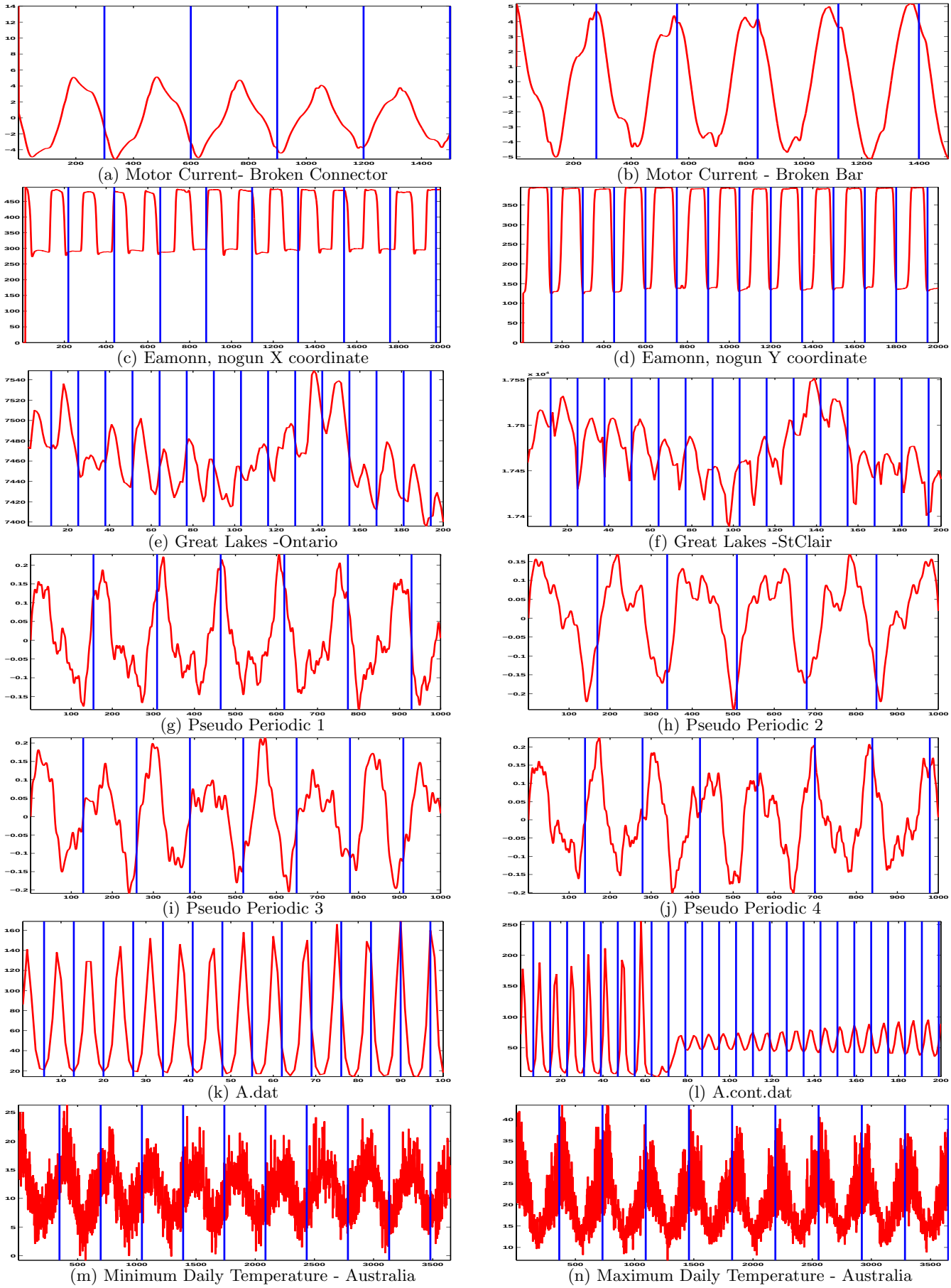
5. DISCUSSION AND CONCLUSIONS

In this article, we have presented robust periodicity detection algorithms for cyclic time series datasets. The algorithms combine information from the time-frequency domain and the autocorrelation space to find meaningful periods. The algorithms are extended to handle non-stationary datasets, multiple interlaced periodicities and incomplete datasets. We have also showed that the design of our algorithm makes it highly robust to the noise. We empirically evaluated the quality of results of our algorithms on a large number of time series datasets. We also showed the proposed algorithm can discover hidden relationships among attributes in multidimensional time series.

If the true period of a signal is T, in a few cases our algorithm discover $2 * T$ as the true period. In such cases, we manually divide the periodicity by the factory of 2 after a visual inspection of the signal.

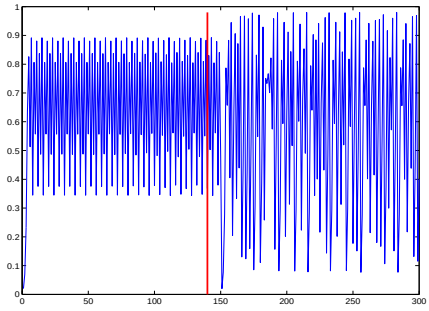
For the non stationary case, we have used a sliding window over the length of the signal. Except for the change points, the signal is assumed to be stationary within each window. Deciding the optimal window size is highly dependent on the signal. For a slowly (fast) changing signal, the

[3]http://www.physionet.org/physiobank/database/santa-fe/

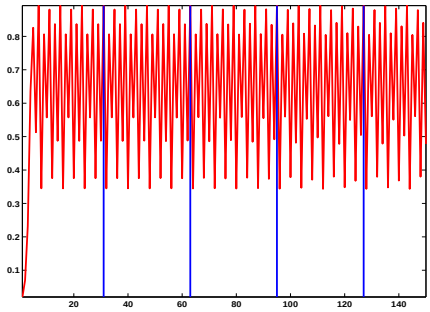


(a) Motor Current- Broken Connector
(b) Motor Current - Broken Bar
(c) Eamonn, nogun X coordinate
(d) Eamonn, nogun Y coordinate
(e) Great Lakes -Ontario
(f) Great Lakes -StClair
(g) Pseudo Periodic 1
(h) Pseudo Periodic 2
(i) Pseudo Periodic 3
(j) Pseudo Periodic 4
(k) A.dat
(l) A.cont.dat
(m) Minimum Daily Temperature - Australia
(n) Maximum Daily Temperature - Australia

Figure 11: Experimental Evaluation on several publicly available datasets



(a) Segmented Signal



(b) Periodicity showed for the first segement only

Figure 10: Avoiding false positive analysis on the Logisitic Map data

window length should be high (low). In all our experiments, we have fixed the length of the window at 128.

The two issues discussed above can be handled if some domain knowledge about data generating process is available. For example in speech analysis, the size of the sliding window is usually fixed at 10 ms and any period greater than 2 ms is not allowed [7].

Future work will attempt to detect joint periods in multi-dimensional datasets and also analyze data from even wider range of domains than the ones reported here. A potentially interesting application involves the problem of segmenting human motion trajectories [4]. We believe that periodicity could be a important cue that characterizes individual human motions.

6. ACKNOWLEDGMENTS

This work is funded by the following NSF grants NGS-0326386, ACI- 0234273 and NSF Career Award IIS-0347662.. The authors would like to thank Dr. Eamonn Keogh for providing the time series datasets.

7. REFERENCES

[1] B. D. O. Anderson and J. B. Moore. *Optimal filtering*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1979.

[2] C. Berberidis, W. G. Aref, M. J. Atallah, I. P. Vlahavas, and A. K. Elmagarmid. Multiple and partial periodicity mining in time series databases. In *ECAI*, pages 370–374, 2002.

[3] R. Cutler and L. S. Davis. obust periodic motion and motion symmetry detection. In *CVPR*, pages 2615–2622, 2000.

[4] J. Davis and S. Taylor. Analysis and recognition of walking movements. In *International Conference on Pattern Recognition*, 2002.

[5] Eamonn Keogh. The UCR Time Series Data Mining Archive. Website. `www.cs.ucr.edu/~eamonn/TSDMA/index.html`. Riverside CA, University of California.

[6] J. Han, G. Dong, and Y. Yin. Efficient mining of partial periodic patterns in time series database. In *ICDE*, pages 106–115, 1999.

[7] X. Huang, A. Acero, and H. Hon. *Spoken Language Processing*. Prentice Hall PTR, Upper Saddle River, NJ, 2001.

[8] L. Jackson. *Digital filters and signal processing* . Kluwer Academic Publishers, Boston, MA, 1986.

[9] P. P. Kanjilal, J. Bhattacharya, and G. Saha. Robust method for periodicity detection and characterization of irregular cyclical series in terms of embedded periodic components. *Phys. Rev. E*, 59(4):4013–4025, 1999.

[10] R. M. May. Simple mathematical models with very complicated dynamics. *Nature*, 261:459, 1976.

[11] R. Meddis and M. J. Hewitt. Virtual pitch and phase sensitivity of a computer model of the auditory periphery I: pitch identification. In *J. Acoust. Soc. Am.,Vol 89, no. 6 2866-2882*, 1991.

[12] A. V. Oppenheim, R. W. Schafer, and J. R. Buck. *Discrete-time signal processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, 2nd edition, 1999.

[13] B. Özden, S. Ramaswamy, and A. Silberschatz. Cyclic association rules. In *ICDE*, pages 412–421, 1998.

[14] R. K. Pearson, H. Lähdesmäki, H. Huttunen, and O. Yli-Harja. Detecting periodicity in nonideal datasets. In *SDM*, 2003.

[15] W. H. Press and G. B. Rybicki. Fast Algorithms for Spectral Analysis of Unevenly Sampled Data. In *The Astrophysics Journal,Vol 338, 277-280*, 1989.

[16] H. Stark and J. W. Woods. *Probability and random processes with applications to signal processing*. Prentice-Hall, Upper Saddle River, NJ, 3rd edition, 2002.

[17] P. Stoica and R. L. Moses. *Introduction to spectral analysis*. Prentice-Hall, Upper Saddle River, NJ, 1997.

[18] S. H. Strogatz. *Nonlinear dynamics and chaos : with applications to physics, biology, chemistry, and engineering*. Westview Press, Cambridge, MA, 2000.

[19] M. Vlachos, P. S. Yu, and V. Castelli. On periodicity detection and structural periodic similarity. In *SDM*, 2005.

[20] D. L. Wang and G. J. Brown. Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Trans. on Neural Networks*, 10(3):684–697, 1999.

[21] J. Yang, W. Wang, and P. S. Yu. Mining asynchronous periodic patterns in time series data. *IEEE Trans. Knowl. Data Eng.*, 15(3):613–628, 2003.

[22] W. Yang and G. Lee. Efficient partial multiple periodic patterns mining without redundant rules. In *COMPSAC*, pages 430–435, 2004.