УДК 53.087:004.021

Detection of Regularity Violations of Cyclic Processes in a Temperature Monitoring System Using Patterns Form

Hussein Sh. Hussein
Alexey G. Yakunin*
Information Technologies Faculty
Altai State Technical University
Lenin av., 46, Barnaul, 656038
Russia

Received 29.01.2015, received in revised form 04.02.2015, accepted 13.03.2015

Periodicity mining is used for predicting trends in time series data. Discovering the rate at which the time series is periodic has always been an obstacle for fully automated periodicity mining. In this paper, a method for detecting the weather temperature series periodicity is proposed. The proposed method, based on DFT, effectively discovered the series periodicity and determined the periodic patterns and their repetition frequencies. Then, the series has been divided into equal time slots based on the pattern repetition frequency. A reference series has been constructed as repetitions for a template pattern, which was constructed from the patterns averages of the original temperature series. The reference series is very useful in temperature series analysis, as the patterns deviations, the future patterns predictions, and the anomalies detections. Experimental results show that the proposed method accurately discovers periodicity rates and periodic patterns.

Keywords: patterns, cyclic, periodicity, air temperature, Z-scores, DFT.

Introduction

Periodicity mining is a method that helps in predicting the behavior of time series data [1] Many data mining researches concerned to identifying and extracting different types of patterns in massive data. These patterns include temporal patterns [2], sequential patterns [3–7], and partial periodic patterns [8–12]. The drawbacks of these techniques are that, some of them require the user to specify a period for the time series periodicity, and others use a trial-and-error scheme, which is obviously not efficient. Time series analysis has an important role in weather measurements changes analysis. It has been carried out for various weather parameters such as air temperature [13, 14] and rainfall data analysis [15], The periodicity analysis of temperature time series is very important to study the effects of climate change. It is difficult to know if there is any cyclic behavior on temperatures by looking at the time-domain signal (as shown in Fig. 1). However, it becomes evident in its frequency-domain representation. In this work, a new technique will be developed to detect the periodicity in temperature data. Also, a reference series will be constructed to analysis the deviation of measured data.

Notes

- The data for experimental test is extracted from the weather monitoring system database in Altai State Technical University.

*yakunin@agtu.secna.ru

- Matlab will be used for all simulation and experimental analysis.

1. Pattern detection procedure

The following steps summarize the algorithm for detecting similarities and patterns:

- **Smoothing the measured data**

  To remove the noise and distortion, the measured temperature will be smoothed using any of well-known smoothing method. Moving average, the most common smoothing algorithm, will be applied here.

- **Measured Data normalization**

  To concentrate analysis on the series fluctuations, the measured temperature samples "*f(t)*" will be normalized with Z-scores [16], the most commonly used method, which can be carried out using the following equation:

$$z(t) = \frac{f(t) - f'}{\delta}, \tag{1}$$

  where f' *and* δ *are the mean and the standard deviation of f(t) respectively.*

- **Analyzing Cyclic Behavior of the Temperature**

  To analysis the periodicity of the normalized measured data, It will be transformed into frequency-domain representation using Discrete Fourier Transform "DFT", DFT can be calculated with the following equation:

$$\mathrm{Z}_k = \sum_{n=0}^{N-1} z_n . e^{\frac{-j2\pi kn}{N}}, \tag{2}$$

  where N *is the normalized data samples.*

  The sinusoid's frequency is k cycles per N samples, where $1 \leqslant k \leqslant N$.

  Each Z_k is a complex number that encodes both amplitude and phase of a sinusoidal component of function z_n.

  The periodicity frequencies of the temperature series can be obtained from the observing of the significant peaks in the frequency-domain series Z.

2. Reference series construction

The reference series R(t), which will be constructed from the patterns averages of the original data, will be used to analysis the devotion of the input data.
The following sequence will describe how the reference series can be prepared.

- ***Template pattern formation***

The template pattern will be constructed from the normalized temperature series according to the procedure: the normalized temperature series z(t) will be divided into equal time slots "S", the period of each slot "T" is the inverse of the periodicity frequency. Each point $\mathrm{P}(\mathrm{t}_i)$ in the

template pattern will be equal the average of the analogous points in every time slot, and can be calculated according to the following equation:

$$P(t_i) = \frac{1}{n_s} \sum_{j=1}^{n_s} S_j(t_i), \tag{3}$$

where $i = 1, 2, \ldots, m$ (number of samples in every time slot) and n_s "the total number of time slots " $= N/T$.

$P(t)$ will be repeated n_s times to get a reference series has the same number of patterns that for the normalized data.

To increase the similarity between the reference series and the normalized series, each pattern in the reference series will be shifted by a factor C. The adjustment factor C is not constant, but linearly changes according to the averages of the present time slot, the next time slot and the previous one. One of the suggestions to calculate the adjustment C as follows:

$$C_k(t_i) = \begin{cases} \frac{\frac{m}{2} - i}{m} \overline{S}_{k-1} + \frac{\frac{m}{2} + i}{m} \overline{S}_k, & i = 1, 2, \ldots, \frac{m}{2}, \\ \frac{1.5 * m - i}{m} \overline{S}_k + \frac{i - \frac{m}{2}}{m} \overline{S}_{k+1}, & i = \frac{m}{2} + 1, \ldots, m, \end{cases} \tag{4}$$

where $k = 1, 2, \ldots, n_s$ and $\overline{S}_k$ is the average of the time slot k.

Every time slot of the reference series $R(t)$ can be calculated with the following equation:

$$R_k = P(t_i) + C_k(t_i). \tag{5}$$

To measure the similarity between the normalized series and the reference series, the cross-correlation [17–19] will be carried out between them.

The following section will summarize the experimental results for proposed algorithm.

3. Experimental results

The proposed method has been applied to a randomly selected sample of temperature measured every 30 second along one month (April 2014) (exactly 25 days, with total 72,000 sample) using the high resolution DS18S20 sensors, which is a part of a full academic weather monitoring project. "*More details about the project can be found on the website* ***abc.altstu.ru***".

To remove the distortions, the measured data has been smoothed using the moving average method with half-hour (or 60 sample) moving window.

The measured data has been normalized with z-scores as shown in Fig. 1. To detect the series periodicity, it has been transformed to frequency domain with DFT.

The result for frequency transformation is shown in Fig. 2, which indicates that the DFT of the measured data has two spectral lines that are clearly larger than any other frequency component, one of them near 1, which means the series almost has one cycle/week, the other peak at 7 means the series repeated daily with lower temperatures during the night and higher temperatures during the day.

The daily periodicity of the measured data has been tested according to the proposed procedure.

The normalized data has been divided into one day time slots and the template pattern has been constructed using equation (3). Also the adjustment factor $C(t)$ has been calculated using

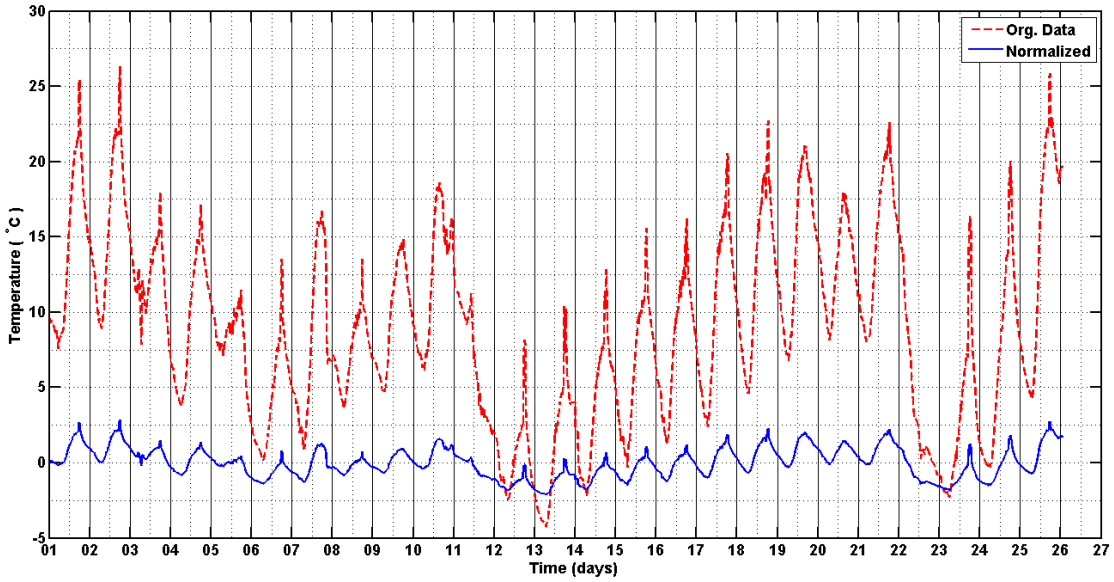

Fig. 1. Measured Temperature and its normalization

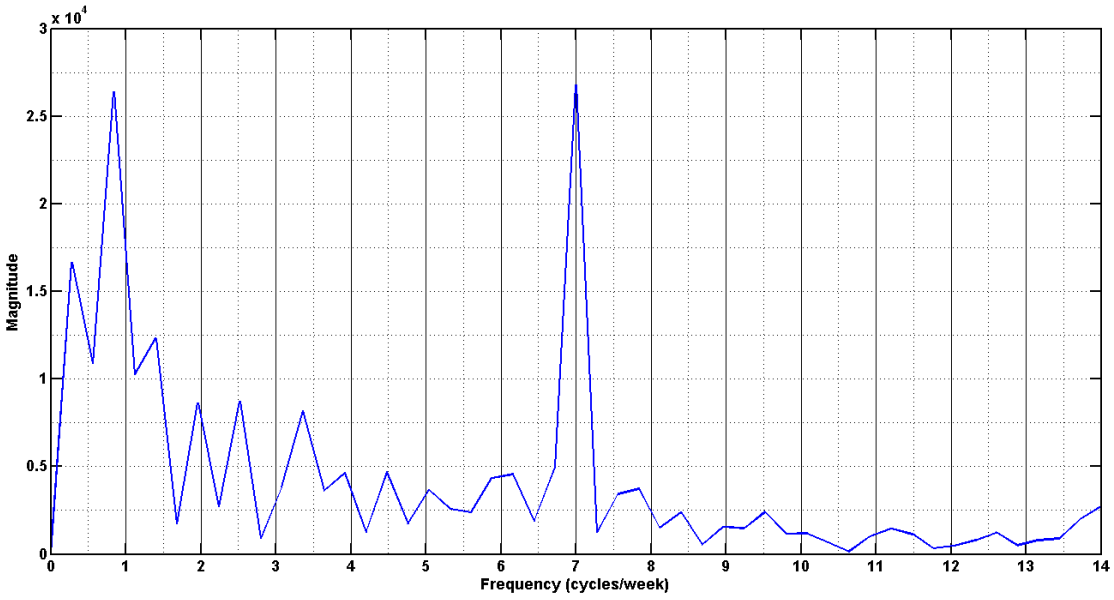

Fig. 2. DFT of the normalized temperature

equation (4) as well as the reference series $R(t)$, the calculation results are shown in Fig. 3. The cross-correlation between the reference and the normalized series has been carried out to measure the similarity between them.

Fig. 4 shows that the cross-correlation between the reference series and normalized series has a significant correlation factor equals 0.93.

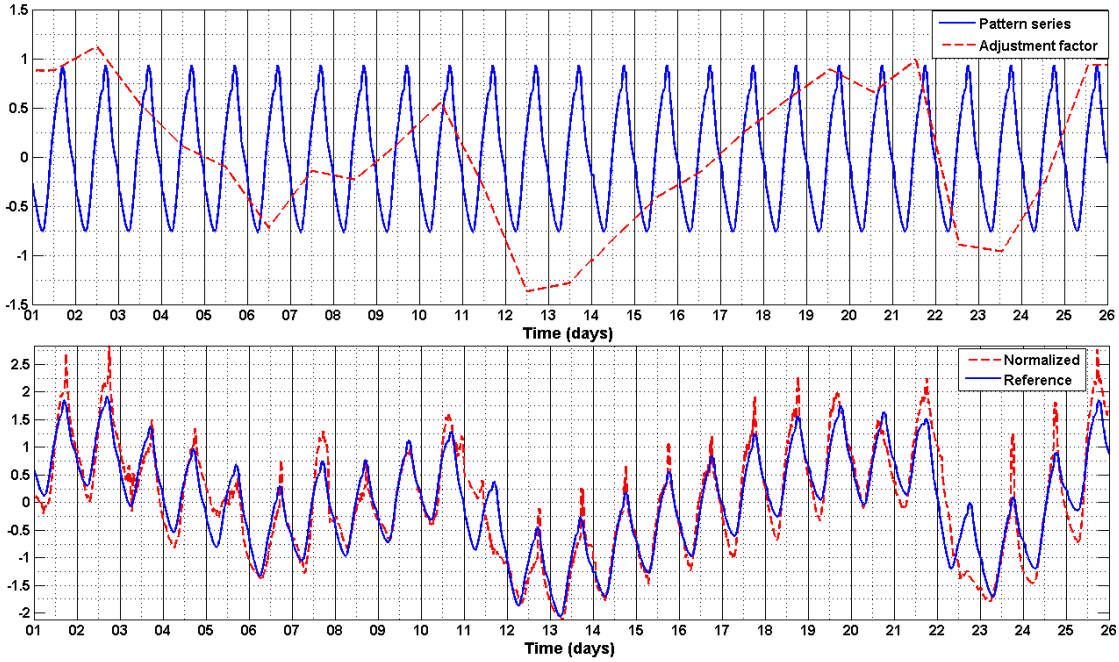

Fig. 3. The pattern repetition series, the adjustment factor, the normalized temperature and the reference series

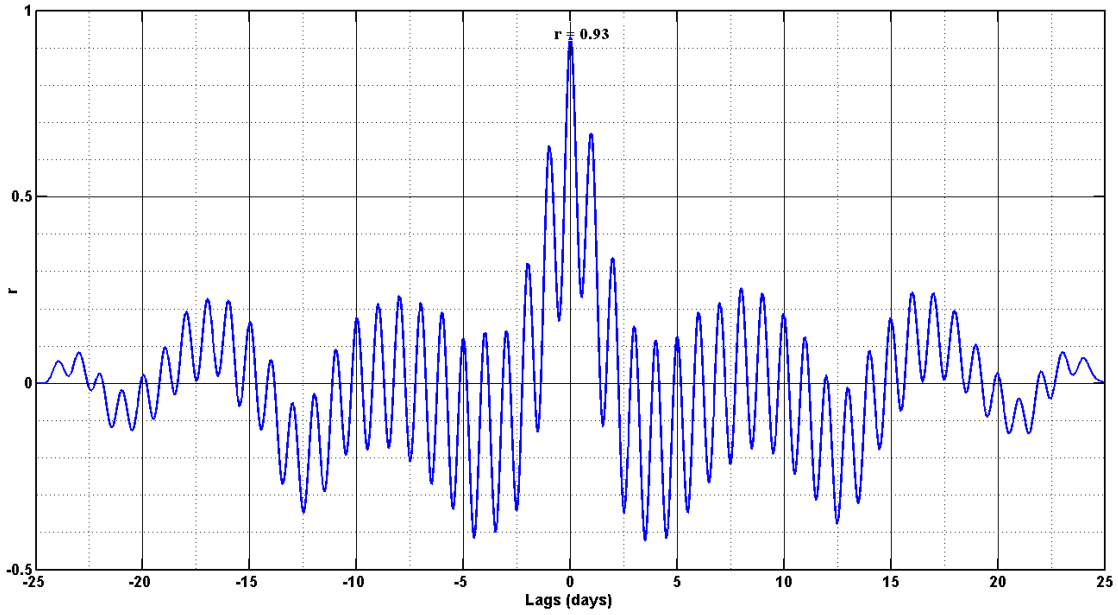

Fig. 4. The cross-correlation between the normalized and the reference series

The residual between the two series is shown in Fig. 5, which indicates that the residual is bounded within a threshold 0.5 (except small deviations), which reflects strong similarities between the reference signal and the original one.

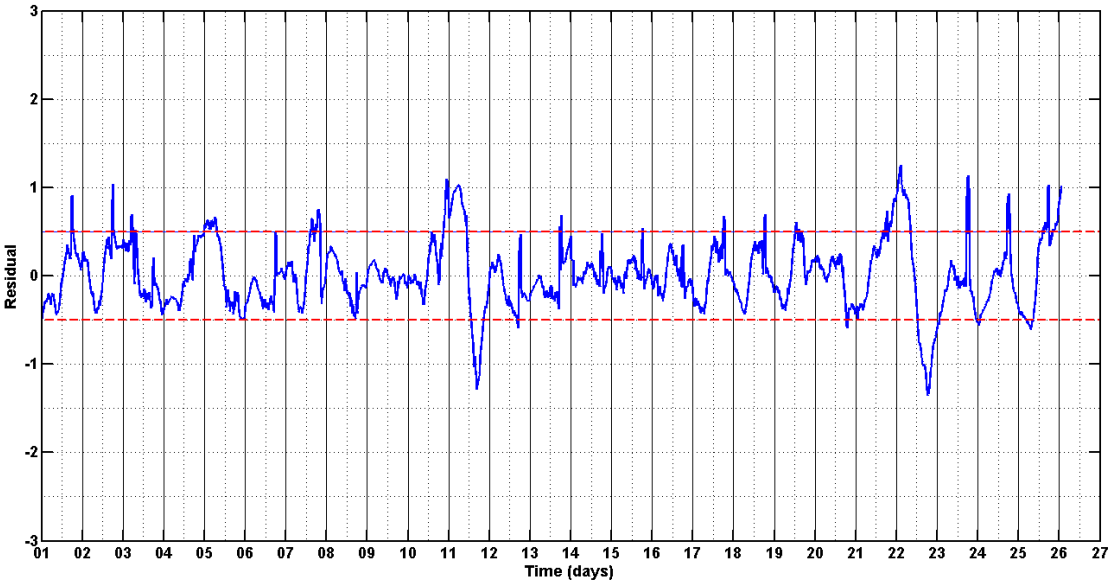

Fig. 5. The Residual between normalized and reference series

The following are examples for using the reference signal

- **Measurement of temperature stability**

  A strong relationship between the two series indicates more stable temperature measurements.

- **The anomalies and outliers detection**

  For examples, in Fig. 5, the residuals that lay outside the threshold lines indicate anomalies. Some of them, with narrow peak above the threshold (like those in days 1, 2, 24 and 25 in Fig. 5), arise from heating of the temperature sensor as a direct effect of sunlight. Other peaks with longer duration (like in days 11 and 22) arise from the abnormal changes in the temperature.

- **Temperature predictions**

  The repetition of the reference signal patterns can be used to predict the future temperatures. The adjustment factor for the predicted temperature can be calculated from the same time in previous years.

Conclusion

In this paper, an efficient method for detecting the weather temperature series periodicity has been proposed based on DFT.

A reference series, which was constructed from the patterns averages of the original temperature series, was very useful to analysis temperature series, For examples, the pattern deviations, the future patterns predictions, and the anomalies detections.

The proposed method can be generalized to involve, not only weather parameter series, but also stock prices, power consumption in factories, and other similar series.

References

[1] S.Makridakis, Time series prediction: Forecasting the future and understanding the past, *International Journal of Forecasting*, (1994), no. 10, 463–466.

[2] C.Bettini, X.S.Wang, S.Jajodia, J.L.Lin, Discovering frequent event patterns with multiple granularities in time sequences, *IEEE Trans. Knowl. Data Eng.*, **10**(1998), 222–237.

[3] R.Agrawal, R.Srikant, Mining sequential patterns, Proc. Elev. Int. Conf. Data Eng., 1995, 3–14.

[4] J.Han, J.Pei, Y.Yin, R.Mao, Mining sequential patterns by pattern-growth: The prefixspan approach, *IEEE Trans. Knowl. Data Eng.*, **16**(2004), 1424–1440.

[5] C.Raпssi, T.Calders, P.Poncelet, Mining conjunctive sequential patterns, *in Data Mining and Knowledge Discovery*, **17**(2008), 77–93.

[6] J.Yin, Z.Zheng, L.Cao, USpan: An efficient algorithm for mining high utility sequential patterns, In 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012, 660–668.

[7] T.P.Exarchos, C.Papaloukas, C.Lampros, D.I.Fotiadis, Mining sequential patterns for protein fold recognition, *J. Biomed. Inform.*, **41** (2008), 165–179.

[8] S.M.S.Ma, J.L.Hellerstein, Mining partially periodic event patterns with unknown periods, Proc. 17th Int. Conf. Data Eng., 2001.

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

[10] D.Chudova, A.Ihler, K.K.Lin, B.Andersen, P.Smyth, Bayesian detection of non-sinusoidal periodic patterns in circadian expression data, *Bioinformatics*, **25**(2009), 3114–3120.

[11] J.Assfalg, et al., Periodic Pattern Analysis in Time Series Databases, Proceedings of the 14th International Conference on Database Systems for Advanced Applications, 2009, 354–368.

[12] F.Rasheed, M.Alshalalfa, R.Alhajj, Efficient periodicity mining in time series databases using suffix trees, *IEEE Trans. Knowl. Data Eng.*, **23**(2011), 79–94.

[13] C.Casty, et al., Temperature and precipitation variability in the European Alps since 1500, *Int. J. Climatol.*, **25**(2005), 1855–1880.

[14] M.Brandt, et al., Environmental change in time series – An interdisciplinary study in the Sahel of Mali and Senegal, *J. Arid Environ.*, **105**(2014), 52–63.

[15] A.Astel, J.Mazerski, Z.Polkowska, J.Namiesnik, Application of PCA and time series analysis in studies of precipitation in Tricity (Poland), *Adv. Environ. Res.*, **8**(2004), 337–349.

[16] R.E.Shiffler, Maximum Z Scores and Outliers, *Am. Stat.*, **42**(1988), 79–80.

[17] R.O.Duda, P.E.Hart, Pattern Classification and Scene Analysis, New York: Wiley, 1973.

[18] J.C.Yoo, T.H.Han, Fast normalized cross-correlation, *Circuits, Syst. Signal Process.*, **28**(2009), 819–843.

[19] D.M.Tsai, C.T.Lin, Fast normalized cross correlation for defect detection, *Pattern Recognit. Lett.*, **24**(2003), 2625–2631.

Обнаружение нарушений закономерностей протекания циклических процессов в системах температурного мониторинга с применением паттернов формы

Хуссейн Ш. Хуссейн
Алексей Г. Якунин

Выявление периодичности широко используется для предсказания трендов при исследовании временных рядов. Нахождение периода следования всегда представляло определенную проблему в полностью автоматических системах анализа. В данной статье предлагается метод для выделения периодических циклов на графиках изменения температуры окружающей среды. Основанный на дискретном преобразовании Фурье, он эффективно выделяет периодические участки и частоту их повторения. В предлагаемом методе весь временной ряд разбивается на ряд эквидистантных временных интервалов. Опорный ряд восстанавливается в виде серии повторяющихся паттернов, форма которых определяется путем усреднения форм интервалов описывающего изменение температуры оригинального ряда. Такой ряд очень полезен для анализа оригинальной серии наблюдений, такого как обнаружение отклонений на отдельных интервалах, экстраполяция результатов, выявление аномалий в поведении температурного графика. Экспериментальные результаты показывают, что предложенный метод точно находит длительность периода и форму циклически повторяющихся фрагментов ряда.

Ключевые слова: шаблоны, циклические, периодичности, температура воздуха, Z-scores, DFT.