International Journal of Innovative
Computing, Information and Control
Volume **8**, Number **4**, April **2012**

 ISSN 1349-4198
pp. 2705–2711

DATA ZOOMING FOR THE DETECTION OF OUTLIERS AND SUBSEQUENCE DISCORDS

Jamal Ameen[1,2] and Rawshan Basha[3]

[1]Faculty of Advanced Technology
University of Glamorgan
Pontypridd, CF37 1DL, UK
jrmameen@glam.ac.uk

[2]Ministry of Planning
Kurdistan Regional Government, Erbil, Iraq

[3]College of Computer Information Technology
American University in the Emirates
Dubai International Academic City-Dubai, P.O. Box 503000, UAE
rawshan.basha@aue.ae

Received September 2010; revised January 2011

Abstract. *Zooming in and out of images has lately become a norm to provide panoramic as well as detailed views of subjects under consideration taking advantage of advances in computer and information technology including the Global Information Systems. However, the power of zooming as a generic activity has not been yet fully realized. In this paper, we explore this idea for the first time in data mining and use it as a powerful technique to link classical methodologies of outlier detection with their counterparts, subsequence discord identification in time series when large and more detailed data are observed where in the latter case, most of the classical methodologies and techniques have become obsolete and fresh approaches of data analysis are overdue. The idea will be demonstrated using a time series of daily household consumption of electricity featuring high frequency, multiple periodicities observed on a near continuous time periods basis. The zooming approach is shown to be both easy to adopt and powerful in linking and identifying outliers and subsequence discords efficiently saving considerable processing time and opens the way for further research in different fields.*

Keywords: Data zooming, Subsequence discords, Time series processes, Multiple periodicities

1. **Introduction.** The history of time and the role that scientific developments in time series processes have played in the progresses seen around us [1] and the use of this branch of science in various aspects of our daily life have been great [2,3]. The near systematic impact that seasonal variations are making on the process of life on our planet has given the opportunity for modeling techniques to assist decision making. Furthermore, the analysis and modeling of time series processes have played a major role in improving life on our planet for generations.

Just like other allied sciences and technologies, the process of modeling time series has gone through major advances starting from very elementary approaches helping to learn about process behavior [4-6]. In this, it is natural to assume that the amount of information extracted from the observed data would positively be related to their detail, validity and reliability and each of the latter has also gone through changes, again, in line with our ability in observing data. Based on that, information technology and the fast advancements this field has encountered have created great challenges as well as opened

new avenues both in the amount and methods with which information can be extracted from observed data.

As a scientific tool for representing life scenarios from which better understanding, monitoring and control can be achieved, modeling has played its part. Modeling has also gone through different phases as we have changed tactics for observing data at every opportunity that has become available for improvement, starting from a classical highly aggregated data with no time considerations to current near continuous time observations. Some researchers have worked on partitioning the observed time series dataset into subgroups for analysis and comparison to identify changes [7].

The challenges that these changes have created have been so great that they often led to the abandonment of the tools and methodologies in hand and the development of new approaches that can accommodate these advancements. To some extent, this is almost the case now as our ability to observe detailed data has greatly advanced. The introduction of bar codes and the use of scanning devices in market places and the use of Laser Doppler Flux for measuring the effects of positive pressure on cutaneous blood microcirculation, for example, have made most of the previously developed modeling tools redundant.

The birth of data mining has therefore come as a natural progression in an attempt to exploit the power of technology, the very tool that has brought the changes about, to bridge the gap [8,9].

This paper attempts to enrich the process and principles of data mining for a more efficient use of modern technology in line with classical modeling tools and the changes that have taken place in data collection. We revise the smoothing methods and propose an alternative aggregation method that is friendly in dealing with large scale data and helping identify discord subsequences in general. Although the example used in this paper would be a time series dataset with multiple periodicities representing daily recorded data on consumer use of electricity in a region of the United Kingdom over a period of four years, the applicability of the approach extends to other fields of application.

2. **Related Work.** The acknowledgment of fuzziness in observed time series data for dealing with higher frequency and non-stationary data has been documented in [8] where different model performances including fuzzy and stochastic models have been assessed with *accuracy* being the comparative feature. These were also reflected in the development of fully stochastic models that extended the application of ARIMA models [4] and in a series of papers [5,11-14], fuzzy and fully stochastic time series models were defined and implemented in various fields of application in which model performances are expressed through their reliability in probability terms. The use of accuracy as opposed to reliability in model assessment reflects the belief or disbelief of the impossibility of having *true* models as it has been discussed in the literature by the first author of this paper [15] reflects the views of two schools of objectivity and subjectivity.

When data are observed and presented at a disaggregated level, observations that are known as outliers in the terminology of classical statistics, appear as sequences of successive outlying observations as they are currently termed as sub-sequence discords. In all areas of statistics and data analysis, the issues of outlier detection and modeling in the presence of outliers have been a major concern for researchers. This has covered both classical and modern time series analysis [4-6]. However, these techniques have become redundant in dealing with the more general form of outliers that are observed at disaggregated level using current and modern digital data collection tools. In these cases, while an outlying observation is still a possibility, an unusual event lasting for an insignificant period of time is likely to lead to a set of subsequent outlying observations. These are known in the recent literature as subsequence discords [8,9].

The latter studies on subsequence discords have been conducted independently from their historically related sisters (outliers). More recently, these two paradigms were linked benefiting from the nature of the discords together with specific patterns that could be established within the time series data set and classical data mining techniques [16,17].

3. **Smoothing.** It is accustomed that time series observations are recorded over equally distanced time points perhaps for ease of use. In practice this is hardly possible to control. As a result and other factors concerned, random variations usually occur as long as these time differences are not significantly large. However, modelers have dealt with these differences through introducing different techniques. One of the most common of these methods is known as smoothing and different moving average techniques have been introduced for this purpose [18]. For non seasonal time series, the span on which moving averages are applied, is subjective depending on the observed variability of the time series while for seasonal time series processes, it will be determined by the seasonality span (phase) [19]. For example, in an additive seasonal process, a 12 period moving average would make it possible for underlying trends to be extracted from the observed original time series based on the facts that the sum of the additive seasonal deviations from its underlying trend will be zero.

In general, given a time series of length N, $\{x_t\}$; $t = 1, 2, 3, \ldots, N$, the moving average of the above time series with a span of n is usually defined as:

$$y_k = \frac{1}{n} \sum_{i=0}^{n-1} x_{k+i};\ k = 1, 2, 3, \ldots, N - n + 1$$

For example, if $N = 50$ and $n = 3$, the number of observations from the generated three point moving average series would be 48 as two pints are lost one from the start and one from the end.

4. **Time Series Data Zooming.** In addition to the main feature of variability reduction by a factor proportional to the smoothing span of the adopted moving average as a smoothing technique, the number of newly generated data points would reduce by the smoothing window span less one. That is an n point moving average smoothing process of a time series of length N, would result a smoothed time series of with its variance reduced by a factor of n but its length also reduced by $n-1$ points which is disadvantageous when data are scarce. Furthermore, apart from the first and the last $n-1$ points, every point in the original time series equally influence the resulting smoothed time series. In this way, a single outlier point will translate to a level change with a span of n (the length of the smoothing window) for the smoothed time series.

However, most of the current data analysis and modeling challenges are due to the greater availability of more detailed data both in their technical and logistical handling.

Time Series Data Zooming (TSDZ) is therefore a new smoothing technique we introduce here that does not translate outlier points to level changes as it happens with the ordinary smoothing techniques. It also helps linking classical outlier data in time series with their subsequence discord counterpart in observed data using current technological tools. In this, no time series point is processed more than once. In other words, the time series is partitioned into successive equally spaced spans of n, each of which is replaced by its ordinary average. That is each non overlapping window of n points is replaced by its average value to form a new (smoothed) zoomed out time series as expressed below.

$$y_k = \frac{1}{n} \sum_{i=n(k-1)+1}^{nk} x_i;\ k = 1, 2, 3, \ldots, n[N/n]$$

The selection of n is subjective depending on the variability of the original series and the degree of smoothness required in non seasonal time series and would be determined by the seasonality span otherwise to make sure that the smoothed series is free from seasonal effects.

Furthermore, given that the observed x_i s are drawn from a distribution with mean μ and variance σ^2, the y_k zoomed out observation will be expected to have come from the same distribution with the same mean μ and variance $\frac{\sigma^2}{n}$.

5. **Subsequence Discords and Outliers Detection Using TSDZ.** In earlier publications on data mining ([16,17]), we have discussed the origins of subsequence discords in time series as they occur as a result of modern observation tools and their relation to single or multiple outlier points when data are observed at aggregated levels. This idea has been further developed in this paper to introduce Time Series Data Zooming (TSDZ) to zoom in and out of data so that classical data mining methodologies are used to identify outliers which in their being, they represent sub-sequence discords in the original disaggregated (zoomed in) time series data. This approach can be used efficiently to identify discords easily and with more clarity and considerable gain in process time.

As an example, the time series data we use to demonstrate the idea represents daily consumption of electricity in thousands of kilo watts by households in part of the United Kingdom for the period of four years starting from the year 1995 (Figure 1). As the time scale is irrelevant here, we have intended to take the time series process only for what we intend to demonstrate and ignored the time reference.

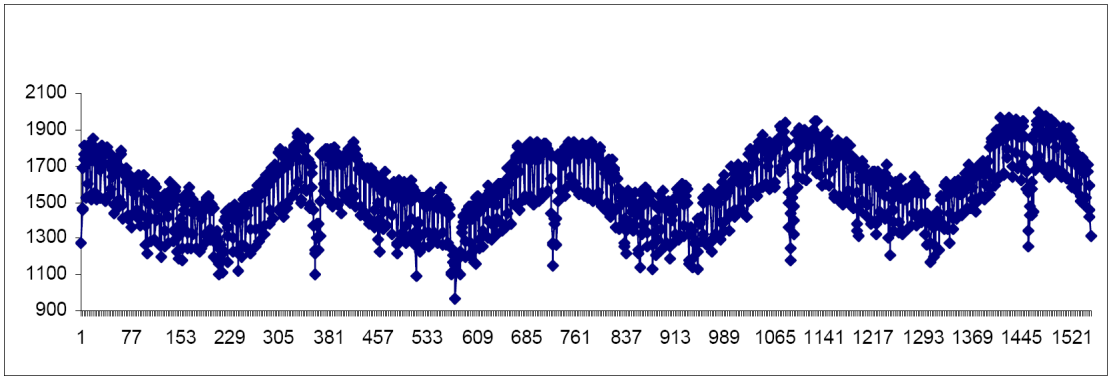


FIGURE 1. Daily electricity consumption

It is evident therefore that a multiple periodicity of seven days per week, four weekly, thirteen readings per year are observed. Such information on the structure of the time series data is of value and can efficiently be used to enhance the search for sub-sequence discords by adopting the zooming technique.

The first stage zooming out of the weeks will generate Figure 2.

The second stage zooming out of every four consecutive weeks will produce thirteen observations per year. This is known as four-weekly time series.

A further zooming out process would produce one observation per year as seen in Figure 4.

In the above hierarchical and systematic zooming process, the nature of the dataset can be investigated for the presence of sub-sequence discords more efficiently. For example, an outlier at an annual level would refer to a sub-sequence discord for part or the entire year. The same principle holds when outliers identified at four-weekly or weekly levels. Furthermore, when investigating for sub-sequence discords using distance measures, it

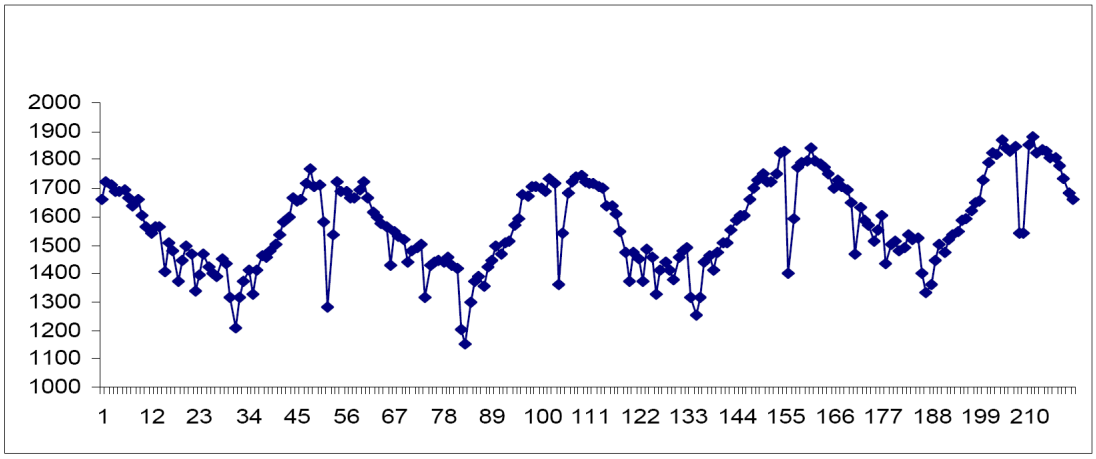


FIGURE 2. Weekly household electricity consumption

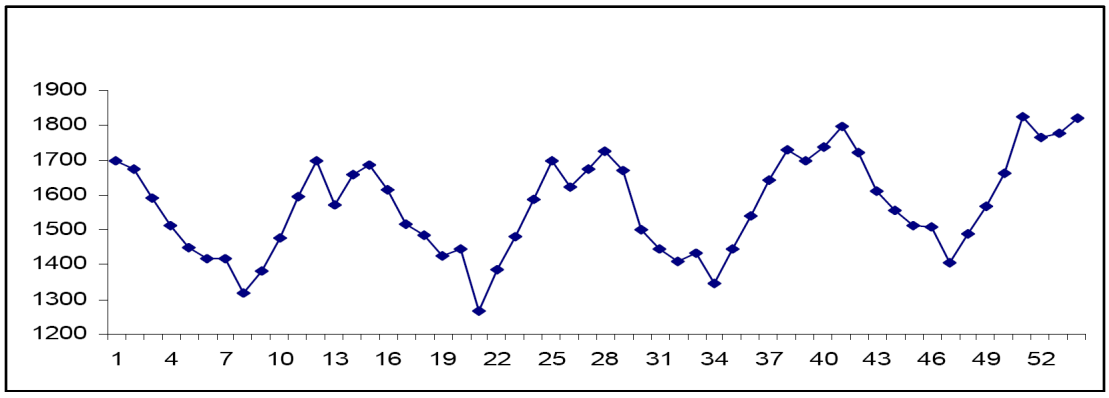


FIGURE 3. Four weekly household electricity consumption

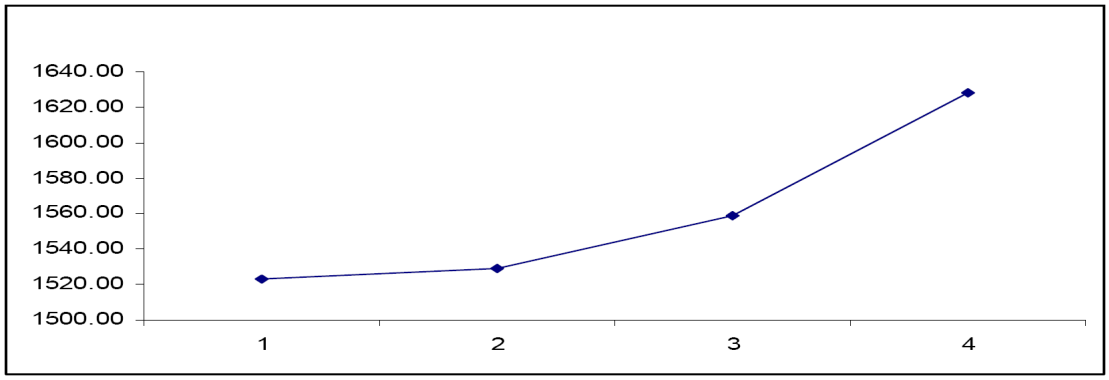


FIGURE 4. Annual electricity consumption

becomes clear that there is often an underlying pattern. It is therefore necessary to state the definition of:

Definition 5.1. *An Outlier (sub-sequence discord): Any observed data point (sub-sequence) that is significantly distanced from the underlying trend (pattern) of the observed data.*

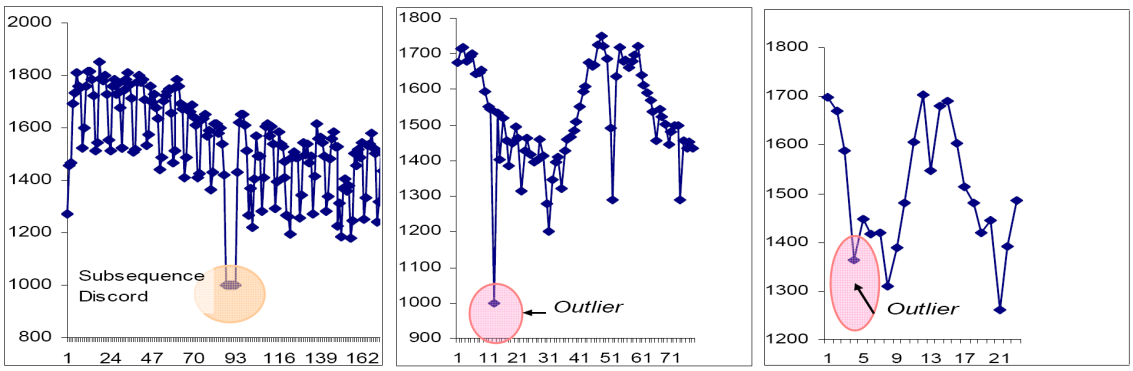
Significance in the above definition is meant to be statistical significance. For example, if a time series is observed under normality assumptions ($Y_t \sim N[\phi_t; V]$), where ϕ_t is the underlying trend, any observation(s) satisfying the criteria $P(|y_t - \varphi_t| > k_\alpha\sqrt{V}) = \alpha$

(with the value of α specified subjectively (greater than 0.2, for example) to reflect the required tracking sensitivity) will be assumed as an outlier. In this, instead of investigating for discords in the original time series, outliers and discords are investigated within the generated zoomed out data after the identification of its underlying trend. In this case, given an observed point $P_t = (p_1, p_2, p_3, \ldots, p_k)_t$ on the time series, its distance from its corresponding point $Q_t = (q_1, q_2, q_3, \ldots, q_k)_t$ on the underlying trend is calculated using the distance formula:

$$d(P,Q)_{k,t} = \sqrt{\sum (p_i - q_i)^2}$$

This will reduce into $d_{CTS}(\widehat{P}, \widehat{Q})_t = \left|\widehat{P} - \widehat{Q}\right|_t$ for univariate time series.

In the daily household electricity consumption time series, a seven point (or multiples of 7) subsequence discord, will show as an outlier(s) in the weekly and four weekly TSDZ data as shown in Figure 5 below.



(a) (b) (c)

FIGURE 5. Subsequence discords and outliers

The time series points from $t = 85$ to $t = 92$ in the original series shown in Figure 5(a) is transformed into the weekly time series point $t = 13$ as shown in Figure 5(b) above and the four-weekly time series point $t = 4$ as in Figure 5(c).

Following the procedures introduced in [12], at $t = 13$, the time series has a mean = 1520 with a variance of 2962 after correcting for the annual trend. This gives a 95% confidence interval of $[1413, 1627]$. The observed point as indicated to be an outlier in Figure 5(b) has a value of 1000 which is clearly outside the calculated 95% confidence interval. Similarly, the time series point at $t = 4$ indicated in Figure 5(c), has a mean of 1520 and a variance of $2962/4 = 740.5$ providing a 95% confidence interval of $[1467, 1573]$ around the mean.

The time series at $t = 4$ in Figure 5(c) is 1364 which is again outside the 95% confidence interval for the mean and hence indicated as an outlier.

6. **Discussion.** The process of zooming in and out of data is expected to be a powerful technique adding to the library of data mining tools and a natural way to link classical to modern data mining tools and techniques. Furthermore, the zooming process makes it easy to establish non-linear trends in the series and account for step changes between zoomed subsequences as it has happened in our example above. The methods used in this paper are only to demonstrate its possible uses and are not to be taken as conclusive.

REFERENCES

[1] J. L. Klein, *Statistical Visions in Time: A Short History of Time Series Analysis*, Cambridge University Press, 1997.

[2] O. Barinova and V. V. Gavrishchaka, Generic regularization of boosting-based optimization for the discovery of regime-independent trading strategies from high-noise time series, *ICIC Express Letters*, vol.4, no.4, pp.1107-1112, 2010.
[3] Y. Liu, J. Zhao and W. Wang, A time series based prediction method for a coke oven gas system in steel industry, *ICIC Express Letters*, vol.4, no.4, pp.1373-1378, 2010.
[4] G. E. P. Box, G. M. Jenkins and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, 3rd Edition, Holden-Day, 1994.
[5] J. R. M. Ameen and P. J. Harrison, Normal discount Bayesian models (with discussion), in *Bayesian Statistics 2*, J. M. Bernardo et al. (eds.), pp.271-298, 1985.
[6] M. West and P. J. Harrison, *Bayesian Forecasting and Dynamic Models*, 2nd Edition, Springer-Verlag, New York, NY, 1997.
[7] Y. Matsumoto and J. Watada, Knowledge acquisition from time series data through rough sets analysis, *International Journal of Innovative Computing, Information and Control*, vol.5, no.12(B), pp.4885-4898, 2009.
[8] E. Keogh, K. Lin and A. Fu, Hot sax: Finding the most unusual time series subsequence: Algorithms and applications, *International Conference on Data Mining*, 2005.
[9] E. Keogh, S. Lonardi and W. Chiu, Finding surprising patterns in a time series database in linear time and space, *The 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, pp.550-556, 2002.
[10] H.-L. Wong, C.-C. Wang and Y.-H. Tu, Optimal selection of multivariate fuzzy time series models to non-stationary series data forecasting, *International Journal of Innovative Computing, Information and Control*, vol.6, no.12, pp.5321-5332, 2010.
[11] J. R. M. Ameen, Non-linear predictor models, *Journal of Forecasting*, vol.2, pp.309-324, 1992.
[12] J. R. M. Ameen, Sequential discount smoothing, *The Statistician*, vol.37, pp.227-237, 1989.
[13] J. R. M. Ameen and P. J. Harrison, Discount weighted estimation, *Journal of Forecasting*, vol.3, pp.285-296, 1985.
[14] J. R. M. Ameen and P. J. Harrison, Discount Bayesian multiprocess models with QUSUM's, in *Time Series Analysis: Theory and Practice*, O. D. Anderson (ed.), 5th Edition, North Holland, Amsterdam, 1984.
[15] J. R. M. Ameen, Comments on model uncertainty, data mining and statistical inference, *J. R. Statist. Soc. A.*, 1995.
[16] R. Basha and J. R. M. Ameen, Unusual sub-sequence identifications in time series with periodicity, *International Journal of Innovative Computing, Information and Control*, vol.3, no.2, pp.471-480, 2007.
[17] J. R. M. Ameen and R. Basha, Higherrarchical data mining for unusual sub-sequence identifications in time series processes, *Proc. of the 2nd International Conference on Innovative Computing, Information and Control*, Kumamoto, Japan, pp.177, 2007.
[18] Sir M. Kendall and J. K. Ord, *Time Series*, 3rd Edition, Hodder Arnold, 1990.
[19] C. Chatfield, *The Analysis of Time Series: An Introduction*, 2nd Edition, Chapman and Hall, London, 1980.