Proceedings of the International Conference on Computational Modelling and Simulation 2017

17–19 May
Colombo, Sri Lanka

"Computational and Simulation Perspectives: Looking ahead and moving across boundaries"

Faculty of Science,
University of Colombo,
Colombo, Sri Lanka.

Proceedings of the International Conference on Computational Modelling and Sim

Published by *Faculty of Science , University of Colombo*

ISBN : 978-955-703-011-1

Cover : T. Indika Sanjeeva

Printed by : S & S Printers, Jayantha Weerasekara Mawatha, Colombo 10, Sri Lanka

Case study on measuring central tendency of a set of time series based on discrete Haar wavelet

CP Waduge[1*] and NC Ganegoda[2]

[1]Department of Information Technology, Faculty of Computing, General Sir John Kotelawala Defence University, Southern Campus, Sooriyawewa, Sri Lanka

[2] Department of Mathematics, University of Sri Jayewardenepura, Nugegoda, Sri Lanka

*wadugechekha@gmail.com

Abstract

Knowledge discovery in data has been emerged in a rapid phase with the necessity of effective and efficient methods to extract beneficial information from data and simulated results. With this background, this study focuses on investigating two methods used as measures of central tendency for set of time series namely average series or mean series and tamed series. Average series, the most commonly using method of summarizing set of time series may limit the interpretations on variability as some patterns may be vanished because of its solid nature. In this concern, tamed series can be used as an alternative to the average series, a measure of central tendency with a stochastic description. The proposed method of taming is based on Haar wavelet, generated by taking the deviation of two time series as a spectrum called decomposed error spectrum (DES-W) and further manipulations to obtain a measure of central tendency. The functionality of decomposing used in DES-W evades the link between consecutive data points lying in different couples, leading to different spectrums, hence different tamed series, for the same set of input time series. The order of concern of time series in the taming process is another fact courses to varied outcomes, which notifies the sensitivity to the case specification, intimating the applicability to stochastic models.

Keywords: **Average series, Central tendency, Decomposed error spectrum, Haar wavelet, Set of time series, Tamed series.**

I. INTRODUCTION

Knowledge discovery in data has been emerged in a rapid phase due to digital data with the advancement in computer technology. Thus, we always need effective and efficient methods to extract information from data and simulated results [1][4]. With this platform, this study focuses on conducting case studies on prevailing methods used to summarize set of time series data. The most commonly utilized approach of summarizing a set of time series is acquiring the mean series where all the entries of a set of time series are averaged point-wise This method may restrict the interpretations on variability since some patterns may be disappeared once we go through the process of averaging. Thus in 2013, a preliminary way of taking the deviation of two time series as a spectrum and further manipulation to obtain a measure of central tendency called tamed series has been proposed [2][3] as an alternative to the mean series. This enables a localized analysis contrasting to the cases like Fourier analysis [6], which is more with a view of frequency-domain rather than time-domain.

This taming approach has not been investigated with considerable amount of cases and no comparison with the mean series approach has been conducted yet. Thus, several issues raising with tamed series concept, proposed by Ganegoda et. al. has been identified using case studies. Further it has been assessed with the mean series, in order to perceive the advantages and disadvantages, in particular circumstances. In this paper three cases have been highlighted to illustrate these issues.

II. LITERATURE REVIEW

Summarizing data plays a substantial role in extracting information from data. Mean series, where all the entries of a set of time series are averaged point-wise, is the most commonly using method in many aspects like simulation packages designed for mathematical models, to obtain the final output representing outputs of many simulation runs [7], which may restrict the interpretations on variability in certain cases. Incapacitating this limitation of mean series Ganegoda et. al. has proposed a taming method to acquire a representative, a measure of central tendency, to a set of time series via Haar wavelets.

Haar wavelet is the first literature in wavelet transformation, introduced by Alfred Haar in 1909 and further developed to the present platform by Jean Morlet, Alex Grossmann, Y. Mayer, Stephane Mallat and Inqrid Daubechies [5]. This can be manipulated to decompose a time series into two sequences, termed as scaling coefficients (father wavelets) and wavelet coefficients (mother wavelets), allowing to grasp the time series in different levels defined by scaling function. Scaling coefficients represent the average of the two consecutive elements after coupling the elements of the series two by two, concurrently the wavelet coefficients represent slope or gradient arising from those two elements, gained via a

variant of classical Haar Matrix $\frac{1}{2}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}$. With the proposed decomposition for two coupled series $V = ((d_1, d_2), (d_3, d_4), \dots, (d_{2^n-1}, d_{2^n}))$, and $S = ((s_1, s_2), (s_3, s_4), \dots, (s_{2^n-1}, s_{2^n}))$, can obtain $W = ((a_1, b_1), (a_2, b_2), \dots . (a_{2^{n-1}}, d_{2^{n-1}}))$ and $U = ((p_1, q_1), (p_2, q_2), \dots . (p_{2^{n-1}}, q_{2^{n-1}}))$ where, $a_i = \frac{d_{2i-1}+d_{21}}{2}$, $b_i = \frac{d_{2i-1}-d_{21}}{2}$, $p_i = \frac{s_{2i-1}+s_{21}}{2}$ and $q_i = \frac{s_{2i-1}+s_{21}}{2}$.

Based on Haar wavelet decomposition, Decomposed Error Spectrum using Wavelets (DES-W), has been proposed with the purpose of encounter the quality of the error in the sense of increasing and decreasing patterns of the series.

To construct the error spectrum, the decomposed error is defined as,

Average error (scaling error) $x_i = a_i - p_i$

Slope error (wavelet error) $y_i = b_i - q_i$

Then the error spectrum is generated considering the point-wise dominating error type determined as follows:

If $|x_i| > |y_i|$, then average error is the dominant cause of error (denoted by A)

If $|x_i| < |y_i|$, then slope error is the dominant cause of error (denoted by S)

If $|x_i| = |y_i|$, then both average and slope errors are equally responsible (denoted by E)

Further an extension of this process is proposed by performing decomposition to further levels. By incorporating DES-W approach, an error taming technique has been introduced, facilitating a central tendency measure called tamed series for a set of time series. This taming process is carried out by the following algorithm.

If spectrum entry is A (average taming)

$$fwN_1 = (fwS_1 + fwS_2)/2$$
$$mwN_1 = mwS_1$$

If spectrum entry is S (slope taming)

$$fwN_1 = fwS_1$$
$$mwN_1 = (mwS_1 + mwS_2)/2$$

If spectrum entry is E (both average and slope taming)

$$fwN_1 = (fwS_1 + fwS_2)/2$$
$$mwN_1 = (mwS_1 + mwS_2)/2$$

where,

fwS_1 and fwS_2 are series of father wavelets of S_1 and S_2 respectively; mwS_1 and mwS_2 are series of mother wavelets of S_1 and S_2 respectively and fwN_1 and mwN_1 are series of father and mother wavelets respectively of tamed series (say N_1) of S_1 and S_2. After determining fwN_1 and mwN_1, tamed series (N_1) can be obtained by inverse wavelet transforms. This taming process can be extended for more than two series, by taking the tamed series with the next series and so on.

This proposed algorithm has not been validated with different scenarios and the issues raising in generating tamed series has not been discussed. Further no investigation has been carried out to ensure the reliability of outcomes and to compare it with the mean series technique. The next section will portray the concerns identified in this study.

III. CASE STUDY RESULTS

Decomposed error spectrum using wavelets (DES-W), proposed in Ganegoda et. al.'s work is primarily based on coupling consecutive data points. This process precedes to evade the rapport between the consecutive points belongs to different couples. For an instance, the decomposition of the series $\{d_i\}$ will be $V = ((d_1, d_2), (d_3, d_4), \dots, (d_{2^n-1}, d_{2^n}))$ which disrupts the relationship of d_{2i} and d_{2i+1}, $i = 1, 2, \dots$. Since the error spectrum entirely depends on this decomposition, a shift by a single point, provided that an extra data point is available (d_{2^n+1}) , $(V_1 = (d_2, d_3), (d_4, d_5), \dots, (d_{2^n}, d_{2^n+1}))$ performed at the commencement of coupling, produces a different error spectrum. This impact of shifting, in generating the tamed series was investigated using data sets taken from UCI Machine Learning Repository. An example is shown below.

Consider the series

D = {0.9, 1.5, 1.2, 1.6, 1.8, 1.1, 1, 1.9, 2, 2.5, 2.1, 2.5, 3.6, 3.9, 4, 3.6}

And S = {1.3, 1.8, 1.4, 1.2, 1.7, 1.6, 1.4, 1.6, 3, 3.6, 3.2, 3.3, 4, 3.1, 3.4, 3.8}

Then the corresponding error spectrum will be,

Level 1	A	S	S	S	A	A	S	S
Level 2		S		A		A		A
Level 3			A				A	
Level 4				A				

FIG. 1 Error spectrum

By shifting one data point from each series D and S to the left, another two series were formed namely D_1 and S_1

D_1={1.3 , 0.9, 1.5, 1.2, 1.6, 1.8, 1.1, 1, 1.9, 2, 2.5, 2.1, 2.5, 3.6, 3.9, 4}

S_1={1.5, 1.3, 1.8, 1.4, 1.2, 1.7, 1.6, 1.4, 1.6, 3, 3.6, 3.2, 3.3, 4, 3.1, 3.4}

The error spectrum generated considering D_1 and S_1 is as follows:

Level 1	A	A	A	A	S	A	A	A
Level 2		A		A		A		S
Level 3			A				A	
Level 4				A				

FIG. 2 Error spectrum after shifting

The next feature identified in taming, associates with the order of concern of time series in the taming process. The algorithm related to generating the tamed series provides the space to result different outputs, to the same set of inputs, according to the order of consideration of each time series. Following example will illustrate the described issue.

Consider the series P, Q and R.
P={1.5, 1.25, 1.24, 1.47, 1.02, 2.22, 2.1, -0.24}
Q={2.5, 2.75, 2.76, 2.53, 2.98, 1.78, 1.9, 4.24}
R={4, 8, 8.96, 4.39, 6.54,5.5, 8.56, 10}

With the order of taming PQR the resultant N1, where
N1= {3.46, 3.21, 3.44, 3.67, 3.09, 3.09, 3.71, 3.71}
With the order of taming RQP the resultant N2, where
N2={1.33, 5.33, 5.84, 1.27, 3.86, 2.82, 3.71, 5.15}

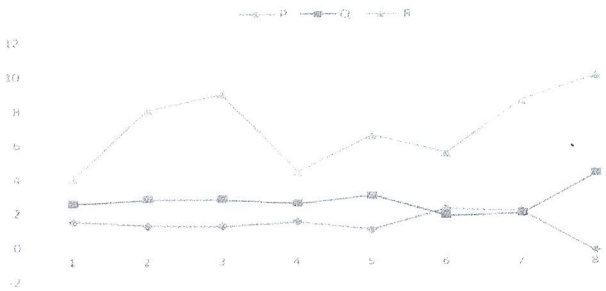

FIG. 3 The series P Q and R

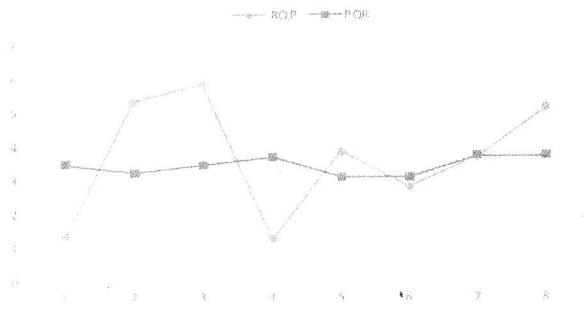

FIG. 4 Comparison of the tamed series generated in different inserting order

These diagrams clearly show the different patterns and the deviation obtained with different taming order.

The next case highlighted in this paper relates with the behaviours of the time series. When using the mean series in certain cases like localized symmetrical behaviours of time series, it may restrict the interpretations, while the taming algorithm generates a series with fluctuating behaviours.
Consider the series Y and Z.
Y= {1.5, 1.25, 1.24, 1.47, 1.02, 2.22, 2.1, -0.24, 8.45, -2.34, 1.04, -0.39, -2.54, 0.44, -4.56, -2.06}
Z= {2.5, 2.75, 2.76, 2.53, 2.98, 1.78, 1.9, 4.24, -4.45, 6.34, 2.96, 4.39, 6.54, 3.56, 8.56, 6.06}

Then the tamed series N1 will be,
N1= {2.125, 1.875, 1.885, 2.115, 1.62, 1.62, 0.93, 0.93, 3.055, 3.055, 2.715, 1.285, 0.51, 3.49, 0.75, 3.25}
The average series corresponding to Y and Z, E, will be
E = {2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2}

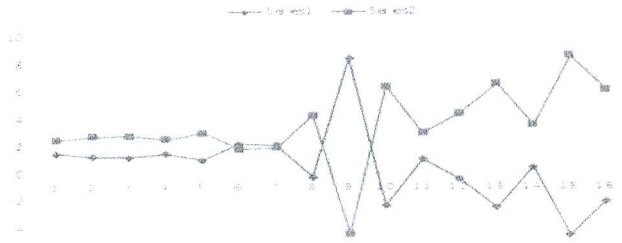

FIG. 5 Series Y and Z

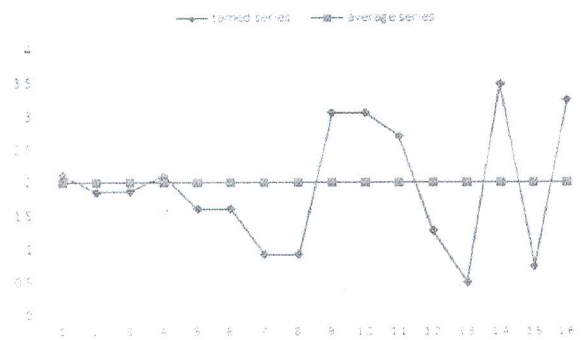

FIG. 6 Tamed series and the average series

Figure 6 depicts the tamed series and the average series, where tamed series is with a flexibility or a stochastic nature.

IV. DISCUSSION

In order to analyse data, both numerical and graphical methods are available such as mean mode median boxplots and graphs. As an alternative to these, Decomposed error spectrum DES-W, and the tamed series have been proposed to analyse a set of time series, where DES-W is with a touch of a graphical interpretation while

tamed series is with the flavour of a numerical representation.

DES-W can be utilized to investigate two time series with a comparison of the behaviour as it gives a spectrum describing the fluctuations and regularities within the series. The decomposing technique applied in DES-W leads to ignore the relationship between consecutive data points lying in different couples resulting different spectrums for the same set of inputs in its first level of decomposing. This effect is suppressed up to some extend in the next level of decomposing, as it is blended couple wise. On the other hand, when dealing with a time series consists of large number of data points, this effect could be ignored.

In generating the tamed series, the considering order of series in the process of taming, results different outcomes for the same set of input series, as it gives the priority to the first comes and so on. Therefore, a noise in the set of time series can highly affect the outcome if we select the particular time series at the beginning of taming. In order to reduce this effect, the DES-W can be used as it gives a graphical idea about the deviations among the series. By giving less priority in the order of concern to the highly-deviated series, the same drawback can be turned into an advantage.

Tamed series concept can be applied in certain cases, where mean series has restrictions in interpreting a set of time series. Also with the sensitivity for order, it will be useful in the stochastic models. Further this concept can be applied locally as it does not make an effect globally like of Fourier analysis.

V. FUTURE WORK

Tamed series is a good approach to be applied in the areas such as signal processing, weather forecast which are vastly susceptible to localized trends. Therefore, more supportive measures would be developed to enhance the credibility of the technique. The effect on the resultant due to different DES-W patterns caused by shifts in the series should be investigate to confirm the trustworthiness of the outcome. In the developed algorithms for taming, a precedence has been given to the order of concern of the time series. Thus, a quantitative measure of dispersion would be developed to identify the order of concern, in order to improve the reliability of the outcome. On the other hand, the necessity identifying the level of taming should be investigated.

Generating tame series is a tedious process compared to calculating the mean series. Hence, a measure consists of both qualitative and quantitative qualities would be developed to determine the localization usage of the tame series approach.

REFERENCES

[1] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery in databases. AI Mag. 17, 37.

[2] Ganegoda, N.C., Kumara, K.K.W.A.S., Tantrigoda, D.A., Boralugoda, S.K., Perera, S.S.N., 2016. An approach for visualizing error and obtaining a measure of central tendency regarding a set of time series using discrete Haar wavelet. J. Wavelet Theory Appl. 10, 1–18.

[3] Ganegoda, N.C., Kumara, K., Tantrigoda, D.A., Boralugoda, S.K., Perera, S.S.N., 2013. A Discrete Haar Wavelet Based Approach for Visualizing Error Regarding a Simulated Time Series. GSTF J. Comput. JoC 3, 56.

[4] Goebel, M., Gruenwald, L., 1999. A survey of data mining and knowledge discovery software tools. ACM SIGKDD Explor. Newsl. 1, 20–33.

[5] Gomez, J. (2013). Wavelet methods for time series analysis.

[6] Merry, R.J.E., Steinbuch, M., 2005. Wavelet theory and applications. Lit. Study Eindh. Univ. Technol. Dep. Mech. Eng. Control Syst. Technol. Group.

[7] Subramanian, S., 2004. Modelling lymphatic filariasis transmission and control. Ph.D. thesis - Erasmus University Rotterdam, Netherlands.