e-ISSN: 2289-6589

MODIFIED COMPOUND SMOOTHER IN MEDIAN ALGORITHM OF SPAN SIZE 42

***Nurul Nisa' Khairol Azmi, Mohd Bakri Adam, Norhaslinda Ali**

[1]Faculty of Computer and Mathematical Sciences,
UniversitiTeknologi MARA Negeri Sembilan Branch,
70300 Seremban, Negeri Sembilan, Malaysia.
[2,3]Institute for Mathematical Research,
Universiti Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia.

*Corresponding author's email: nisa@tmsk.uitm.edu.my

Submission date: 30 July 2018 Accepted date: 30 August 2018 Published date:30 March 2019

Abstract

Median smoother of even window size is obtained by averaging the two middle points of arranged sequence using arithmetic mean. Some adjustments were proposed by substituting the algorithm of arithmetic mean using quadratic, geometric, harmonic and contraharmonic mean. The modified running median was appended in 4253HT algorithm. This paper is mainly to assess the performances of proposed adjustment via simulation study. The results show that modified 4253HT using contra harmonic mean performed the best in extracting signal from heavy noise. The extracted signal was then used for forecasting by applying seasonal ARIMA algorithm. Forecasting by smoothed values was found to produce better forecast than forecasting involving actual values which contained heavy noise.

Keywords: compound smoother, running median, non-linear, 4253HT, signal, noise

1.0 INTRODUCTION / BACKGROUND OF THE STUDY

Smoothing is a process of extracting pattern from heavy noise. In time series, the existence of heavy noise blurs pattern and affects the reliability and ease of forecasting process. Velleman (1980) states, it is important to seek for data smoothers that are resistant to noise with occasional "spikes" or long-tailed distribution. Many studies which have been conducted, provide promising evidence that non-linear smoother like running median has a strong ability to reduce heavy noise from a series of data (see Bovik et al., 1983; Hird et al., 2009 and Sargent et al., 2010).

Tukey (1977) introduces a non-linear approach to smoothing, that is compound smoother. Compound smoother is a combination of several algorithms of smoothing, which includes median smoother of various span sizes, weighted moving average, splitting and re-smoothing of the rough. Compound smoother is known as a powerful tool to smooth a data series without excessively disrupting the details of a data series. Vellemen (1981) introduces a compound smoother that involves running median of even and odd span size, Hanning and 'twice'. It is called 4253HT. The 4253HT is only disturbed slightly by long-tailed noise and negligibly by Gaussian white noise.

Sargent et al. (2010) attempted to improve the compound smoother by combining the smoothing algorithm consisting of running median of various span sizes, Hanning and 'twice' to fit the Australian football players' performance. The output of smoother was then used for forecasting using exponential smoothing method. The result found that forecast using smoothed data of compound smoother approach produced better forecast than when using actual data.
Jin and Xiu (2013) proposes a compound smoother that combines moving average, median smoother, maximum smoother and Hanning in order to reconstruct normalized difference vegetation index (NDVI) time series data, called RMMEH. Even though, RMMEH has been found to be better at smoothing the NDVI data based on specific criteria, Jin and Xiu (2013) acknowledge that 4253HT is a good smoother compared to all. However, improvement on the existing compound smoother, 4253HT in comparison has yet been explored.

2.0 METHODOLOGY

2.1 4253HT

4253HT is one of the non-linear smoothing techniques that combines running median, weighted moving average and re-smoothing of the rough. This technique was first introduced by Tukey (1977) and extensively described in different versions by Velleman et al. (1981). Let **X** be a doubly-infinite sequence of real data $\{X_{t-n},...,X_{t-1},X_t,X_{t+1},...,X_{t+n}\}$. A smoother M is defined as an algorithm that works on **X** to generate a new series $M(\mathbf{X}_t)$, smoothed values. The algorithm of 4253HT is as follows:

Step 1: Perform running median of span size two

$$M_1(\mathbf{X}_t) = \text{median}\left(X_{t-2}, X_{t-1}, X_t, X_{t+1}\right) \tag{1}$$

Step 2: Re-center the equation (1)

$$M_2(\mathbf{X}_t) = \text{median}\left(M_1(\mathbf{X}_t), M_1(\mathbf{X}_{t+1})\right) \tag{2}$$

Step 3: Next, equation (2) is smoothed again by applying median smoother with span size three

$$M_3(\mathbf{X}_t) = \text{median}\left(M_2(\mathbf{X}_{t-2}), M_2(\mathbf{X}_{t-1}), M_2(\mathbf{X}_t), M_2(\mathbf{X}_{t+1}), M_2(\mathbf{X}_{t+2})\right) \tag{3}$$

Step 4: Perform running median of span size three

$$M_4(\mathbf{X}_t) = \text{median}\left(M_3(\mathbf{X}_{t-1}), M_3(\mathbf{X}_t), M_3(\mathbf{X}_{t+1})\right) \tag{4}$$

Step 5: Apply weighted moving average or Hanning with coefficients $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$

$$M_5(\mathbf{X}_t) = \frac{1}{4}M_4(\mathbf{X}_{t-1}) + \frac{1}{2}M_4(\mathbf{X}_t) + \frac{1}{4}M_4(\mathbf{X}_{t+1}) \tag{5}$$

Step 6: Re-smooth the rough and add the rough to the smoothed values in (5)

$$M_6(\mathbf{X}_t) = M_5(\mathbf{X}_t) + M_5\left[\mathbf{X}_t - M_5(\mathbf{X}_t)\right] \tag{6}$$

2.2 Modification of 4253HT

The running median of even span size was computed by taking the average of two subsequent points in the middle using arithmetic mean. This value is better than running median of odd span size in the sense that it preserves the significant spike in a data series. The smoothed value produced by running median span size four and re-centered by running median of span size two, was a combination of Equation (1) and (2), which can be expressed as follows:

$$\begin{aligned} M_2(\mathbf{X}_t) &= \frac{1}{4}\left[\text{median}\left(X_{t-2}, X_{t-1}, X_t, X_{t+1}\right) + \text{median}\left(X_{t-1}, X_t, X_{t+1}, X_{t+2}\right)\right] \\ &= \frac{1}{4}\left(X^*_{t-1} + X^*_t + X'_t + X'_{t+1}\right) \end{aligned} \tag{7}$$

where X^* is the ordered observation from window in $X_{t-2}, X_{t-1}, X_t, X_{t+1}$ and X' is the ordered observation from window in $X_{t-1}, X_t, X_{t+1}, X_{t+2}$. Some adjustments were proposed by applying different types of means. The types of means involved were geometric, quadratic, harmonic and contra harmonic. The modifications of running median span size 42 are as follows:

Geometric Mean

$$M_2(\mathbf{X}_t) = \left(X^*_{t-1} \times X^*_t \times X'_t \times X'_{t+1}\right)^{\frac{1}{4}} \tag{8}$$

Quadratic Mean

$$M_2(\mathbf{X}_t) = \left(\frac{X^{*2}_{t-1} + X^{*2}_t + X'^2_t + X'^2_{t+1}}{4}\right)^{\frac{1}{2}} \tag{9}$$

Harmonic Mean

$$M_2(\mathbf{X}_t) = \frac{1}{4}\left(\frac{1}{X^*_{t-1}} + \frac{1}{X^*_t} + \frac{1}{X'_t} + \frac{1}{X'_{t+1}}\right) \tag{10}$$

Contra harmonic Mean

$$M_2(\mathbf{X}_t) = \left(\frac{X^{*2}_{t-1} + X^{*2}_t + X'^2_t + X'^2_{t+1}}{X^*_{t-1} + X^*_t + X'_t + X'_{t+1}}\right) \tag{11}$$

The types of means that produce smaller value than arithmetic mean - geometric and harmonic, are expected to be more resistant to negative impulse or block pulse. On the other hand, quadratic and contra harmonic are expected to be more responsive to positive changes in a data series. Some of the modifications would not work if the observations consist of zero or negative values. Hence, a constant point should be added to a data to ensure the smoothed value can be computed. Modification of 4253HT only involved the running median of span size 42. Upon smoothing rough part, original algorithm was maintained where the middle points for running median of span size four and two were computed via arithmetic mean.

2.3 Simulation Procedure

The evaluation process of smoothing was done through the simulation of signal and noise. The process of simulation was based on procedure from Conradie et al. (2009). Generally, a data can be decomposed into the following components:

$$\text{Data}_t = \text{Signal}_t + \text{Noise}_t = \text{X}_t \tag{12}$$

A signal is a combination of sinusoidal function with linear curve:

$$\text{Signal}_t = \mu_t = \eta t + A \sin B(t - C) \tag{13}$$

with η is the slope of trend,*t* is the index, |*A*| is an amplitude, *B*=$\frac{2\pi}{d}$ where *d* is the period and frequency is $\frac{1}{d}$, and *C* represents the displacement. Hence,

$$\begin{aligned} X_t &= \mu_t + D_t \\ &= \eta t + A\sin B(t-C) + D_t \end{aligned} \quad (14)$$

For sine function, let $\eta = 0.7$, the amplitude |A|=3 and the displacement C=1. The parameter η , |A| and C were chosen according to Conradie et al. (2009). This parameter values will produce a smooth sine curve. Two hundred values from function $\mu_t = \eta t + A\sin B(t-C)$ were simulated for *t* between 0.542 and 19.6416 with the increments of 0.2 at moderate frequency of $\frac{7}{16}$. This frequency mimics a seasonal fluctuation that occurs commonly in real practices. Figure 1 shows a sinusoidal of frequency 0.4375 with linear curve.

The noise, {D_t} was generated as identical and independent random variables from contaminated normal distribution. This is demonstrated as the following:

$$D_t = \begin{cases} \alpha Z_t & \text{if } Y_t = 1, \\ \beta Z_t & \text{if } Y_t = 0 \end{cases} \quad (15)$$

with $\{Y_t\}$i.i.d Bernoulli(*p*) and independent of the $\{Z_t\}$. Thus $P(Y=1)=p$ and $P(Y=0)=1-p$ so that

$$\begin{aligned} P(D_t \le d) &= P(\alpha Z_t \le d \mid Y_t = 1)P(Y_t = 1) + P(\beta Z_t \le d \mid Y_t = 0)P(Y_t = 0) \\ &= p\Phi\left(\frac{d}{\alpha}\right) + (1-p)\Phi\left(\frac{d}{\beta}\right). \end{aligned} \quad (16)$$

with $\{Z_t\}$ i.i.d *N*(0,1). To generate noise with high volatility, let $\alpha = 5.06$ and *p*=0.75, so that Var(X)=(0.75)(5.06)^2 + 0.25 = 23.29. In the simulation of generating highly volatile noise, approximately 75% of the values came from a *N*(0,5.06^2) distribution and approximately 25% was from a *N*(0,1) distribution.

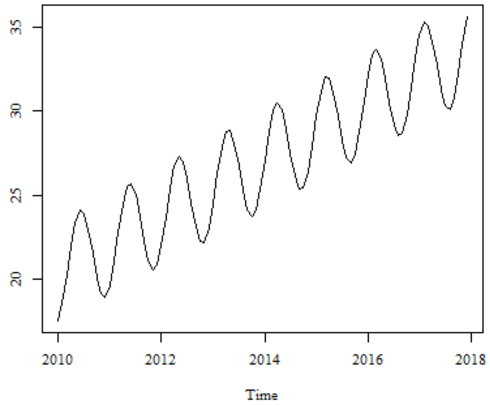

Figure 1 Sinusoidal function of frequency 0.4375 with linear trend

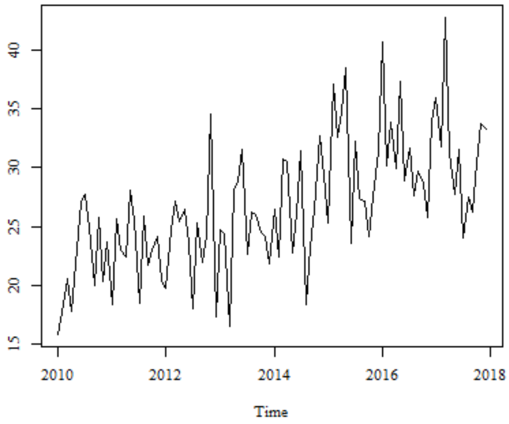

Figure 2 Sinusoidal function of frequency 0.4375 plus linear trend with 75% contaminated normal noise

Figure 2 depicts sinusoidal of frequency 0.4375 plus trend with 75% contaminated normal noise added. It was hard to capture the general trend and existence of seasonal oscillation when 75% contaminated normal noise was added. Two hundred signals plus the generated noise were simulated and applied to the existing and modified 4253HT smoother. The performances of these smoothers were evaluated via estimated integrated mean square error (EIMSE):

$$\text{EIMSE} = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n}\sum_{t=1}^{n}\left(X_{tj} - \mu_j\right)^2 \quad . \tag{17}$$

Low EIMSE indicates the ability for a smoother to recover signal from heavy noise.

2.4 Forecasting

The extracted signal was then further used for forecasting. The method of forecasting applied in this paper was seasonal ARIMA algorithm which takes into account the trend and seasonality at the same time. Seasonal ARIMA can be expressed as follows, according to Box, et al. (2015):

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B)Z_t\,, \qquad Z_t \sim WN\,(0,\sigma^2) \tag{18}$$

where $Y_t = (1-B)^d\left(1-B^s\right)^D X_t$, $\phi(z) = 1-\phi_1 z - \ldots - \phi_p z^p$, $\Phi(z) = 1-\Phi_1 z - \ldots - \Phi_p z^p$, $\theta(z) = 1-\theta_1 z - \ldots - \theta_p z^p$ and $\Theta(z) = 1-\Theta_1 z - \ldots - \Theta_p z^p$.

The data was divided into two parts, namely estimation and evaluation. In the first part of the data, about 72 observations were for the estimation of parameter and expected values; whereas the rest were for the evaluation of the forecast performance. The performance was measured via Mean Square Error (MSE):

$$\text{MSE} = \frac{1}{h}\sum_{t+1}^{t+h}\left(\hat{X}_t - X_t\right)^2 \tag{19}$$

3.0 RESULT AND DISCUSSION

Table 1 shows the performance of smoothers measured via EIMSE. The modified smoother using contra harmonic mean was found to be the best avenue to extract sinusoidal signal of frequency 0.4375 from heavy noise. This was vouched by the low value of EIMSE.

Table 1 Performance of existing and modified smoother measured by EIMSE

Type of modification	EIMSE
Arithmetic	3.991004
Geometric	4.054631
Quadratic	3.983015
Harmonic	4.016506
Contra harmonic	**3.979602**

The extracted signal from smoothing process was subsequently used for forecasting purpose. In this study, seasonal ARIMA algorithm was applied. The performances of forecasting with the inclusion of extracted signal and actual values, were compared. The last 24 observations were used for evaluation to determine whether forecasting using smoothed values is better than using actual values. The MSE for forecasting using smoothed value was 2 2.5226 and forecasting with actual values resulted MSE score of 25.1077. The results indicating forecasting using smoothed values produce better forecast than forecasting using actual values. Figure 3 shows the forecast using actual value while Figure 4 exhibits forecast involving the application of smoothed values from modified 4253HT using contra harmonic mean.

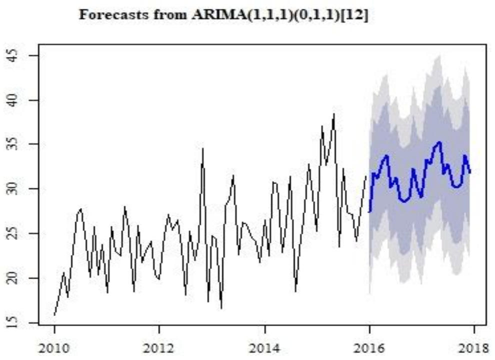

Figure 3 Forecast using actual values

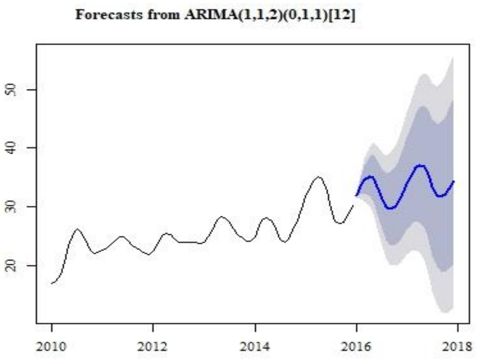

Figure 4 Forecast using smoothed values

4.0 CONCLUSION AND FUTURE WORKS

This study is mainly to assess the performance of modified 4253HT in capturing sinusoidal plus linear trend signal with heavy noise added. Noise with high volatility was added to the signal and the performances were measured by recruiting EIMSE. The results show that modified 4253HT using contra harmonic mean performed the best in extracting signal from heavy noise. The extracted signal was then used for forecasting by applying seasonal ARIMA algorithm. Forecasting involving smoothed values was found to produce better forecast than forecasting with the inclusion of actual values containing heavy noise. For future works, the performance of proposed adjustment to compound smoother will be assessed with the inclusion of different types of signals and noise.

Acknowledgement

This research is partially funded from Putra Grant GP/2018/969400.

References

Bottema, M. J. (1991). Deterministic properties of analog median filters. *IEEE Transactions on Information Theory* 37 (6): 1629-1640.

Bovik, A. C., Huang, T. S. and Munson, D. C. (1983). A generalization of median filtering using linear combinations of order statistics. *IEEE Transactions on Acoustics, Speech and Signal Processing* 31 (6): 1342-1350.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons. New York, USA.

Conradie, W., De Wet, T. and Jankowitz, M. D. (2009). Performance of nonlinear smoothers in signal recovery. *Applied Stochastic Models in Business and Industry* 25 (4): 425-444.

Hird, J. N. and McDermid, G. J. (2009). Noise reduction of NDVI time series: An empirical comparison of selected techniques. *Remote Sensing of Environment* 113 (1): 248-258.

Jin, Z. and Xu, B. 2013. A novel compound smoother RMMEH to reconstruct MODIS NDVI time series. IEEE Geoscience and Remote Sensing Letters 10 (4): 942-946.

Sargent, J. and Bedford, A. (2010). Improving Australian Football League player performance forecasts using optimized nonlinear smoothing. *International Journal of Forecasting* 26 (3): 489-497.

Tukey, J. (1977). *Exploratory Data Analysis*. Addison-Wesley Publishing Company, Readings.

Velleman, P. F. (1980). Definition and comparison of robust nonlinear data smoothing algorithms. *Journal of the American Statistical Association* 75 (371): 609-615.

Velleman, P. F. and Hoaglin, D. C. (1981). *Applications, basics, and computing of exploratory data analysis.* Duxbury Press, Boston, Massachusetts.

Yin, L., Yang, R., Gabbouj, M. and Neuvo, Y. 1996. Weighted median filters : a tutorial. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing* 43 (3): 157-192.