Malaysian Journal of Mathematical Sciences 11(S) February: 1–16 (2017)
Special Issue: Conference on Agriculture Statistics 2015 (CAS 2015)

MALAYSIAN JOURNAL OF MATHEMATICAL SCIENCES

Journal homepage: http://einspem.upm.edu.my/journal

Exploratory Extreme Data Analysis for Farmer Mac Data

Husain, Q. N. *[1,3], Adam, M. B.[1,2], Shitan, M.[1,2], and Fitrianto, A.[1,2]

[1] *Department of Mathematics, Universiti Putra Malaysia, Malaysia*
[2] *Institute for Mathematical Research, Universiti Putra Malaysia, Malaysia*
[3] *Department of Banking and Finance, University of Tikrit, Iraq*

E-mail: qasimalmojamee@gmail.com
**Corresponding author*

ABSTRACT

Maximum extreme data exist in many fields. The exploration of the extreme data using exploratory data analysis approaches is still considered new although it's being introduced four decades ago by Tukey. In this paper, proposition of some modifications of statistical techniques by Tukey focusing on the resistant line and smoothing line with running means methods are introduced. The modified of exploratory extreme data analysis gives comparable findings with the Tukey's method and applied to extreme data set of Farmer Mac.

Keywords: Exploratory Data Analysis, Smoothing, Resistant Line, Running Means.

1. Introduction

Exploratory data analysis (EDA) is a field of statistics introduced by Tukey (1977) for data analysis. In EDA the data appoints an assortment of techniques using graphical visualization. EDA involves looking at the data from many angles to reveal the behavior of the data. The structure of the EDA approaches can identify and interpret trends and relationships among variables.

The resistant line is a technique of EDA uses to describe and analyze data where graphical and numerical summaries are used to uncover interesting structures in Tukey (1977). The resistant line helps us to summarize the dependency of the response factors against other factors in terms of simplest possible description (Velleman and Hoaglin (2004)). The use of the resistant line is equivalent to Tukey's way of resisting outliers (see Shitan and Vazifedan (2011) and Tukey (1977)).

Comparing various smooth ways of the same data to get the best smooth gives us a new easily looks to what we might have seen in the raw data in Tukey (1977). We propose in this paper some modification of statistical techniques of Tukey focusing on the resistant line and smoothing running means methods.

2. Resistant Line

We focus our attention on flexible techniques for plotting the dependent variable y_i on the value of x_i, in terms of simplest possible description represent by straight line in the form:

$$y_i = a + bx_i. \tag{1}$$

Now to summarize any $x - y$ data, we need a and b numerical values which make the lines pass through the data or close to it (Velleman and Hoaglin (2004)).

2.1 Dividing the Batches

We need to sort the data in ascending order by x_i and divide the batches into three portions to obtain the fitting resistant line. The three portions are called the left portion, the middle portion and the right portion. The length of each portion depends on the length n of x_i. If n is a multiple of 3, then the length of each portion will be the same, otherwise it will give either 1 or 2 extra values.

If there is only one extra value it should go to the middle portion. In case of 2 extra values, one goes to left portion and the other goes to the right portion. When there are ties in x_i we have to be careful with dividing the batches. The ties must be located in the same portion (see Shitan and Vazifedan (2011)). The process of dividing the batches into three portions can be done in two main cases according to the value of n. Since the type of data that we deal with is time data, then we focus on the case of n is a multiple of 3 that means no residuals (rough is zero) of dividing the number of batches by 3. In this paper we will focus our attention on the case that n is a multiple of 3.
Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be n data pairs and y_i^* be the initial y_i related to x_i, $i = 1, \ldots, n$ and then the n pairs sorted data in ascending values of x_i be:

$$\left(x_{(1)}, y_1^*\right), \left(x_{(2)}, y_2^*\right), \ldots, \left(x_{(n)}, y_n^*\right)$$

Since n is divided by 3 with no residuals, so we obtain three portions of the same length $\frac{n}{3}$. Let l_p, m_p and r_p be the left, the middle and the right portions respectively, then:

$$\begin{aligned} l_p &= \left(x_{(1)}, y_1^*\right), \left(x_{(2)}, y_2^*\right), \ldots, \left(x_{\left(\frac{n}{3}\right)}, y_{\frac{n}{3}}^*\right) \\ m_p &= \left(x_{\left(\frac{n}{3}+1\right)}, y_{\frac{n}{3}+1}^*\right), \ldots, \left(x_{\left(\frac{2n}{3}\right)}, y_{\frac{2n}{3}}^*\right) \\ r_p &= \left(x_{\left(\frac{2n}{3}+1\right)}, y_{\frac{2n}{3}+1}^*\right), \ldots, \left(x_{(n)}, y_n^*\right). \end{aligned}$$

2.2 Summary Points

The summary points are the median values of each portion and to get these points we follow the following process:
Let x_l, x_m and x_r be the median of the left, the middle and the right portion of the sorted x_i respectively and let y_l, y_m and y_r be the median of the three portions of y_i respectively then we calculate the median as:

$$\begin{aligned} x_l &= \text{median}\left[x_{(1)}, x_{(2)}, \ldots, x_{\left(\frac{n}{3}\right)}\right] \quad , \quad y_l = \text{median}\left[y_1^*, y_2^*, \ldots, y_{\frac{n}{3}}^*\right] \\ x_m &= \text{median}\left[x_{\left(\frac{n}{3}+1\right)}, \ldots, x_{\left(\frac{2n}{3}\right)}\right] \quad , \quad y_m = \text{median}\left[y_{\frac{n}{3}+1}^*, \ldots, y_{\frac{2n}{3}}^*\right] \\ x_r &= \text{median}\left[x_{\left(\frac{2n}{3}+1\right)}, \ldots, x_{(n)}\right] \quad , \quad y_r = \text{median}\left[y_{\frac{2n}{3}+1}^*, \ldots, y_n^*\right] \end{aligned}$$

2.3 Slope and Intercept

The next step is to find the value of the slope b and the intercept a of the resistant line that can be done as follows:
Let a be the intercept of the resistant line and b be the value of the slope, and then the slope for the resistant line is given by:

$$b = \frac{y_r - y_l}{x_r - x_l}, \tag{2}$$

and the intercept is given by:

$$a = \frac{1}{3}\left[(y_l + y_m + y_r) - b\,(x_l + x_m + x_r)\right]. \tag{3}$$

2.4 Straightening Out Plot

Not all data that we deal with gives approximately straight lines. It will be useful to do a re-expression for the data so as to linearize the plot, and calculate the half slope ratio to get a linear relationship between x and y.

2.4.1 Half Slope Ratio

We shall define b_l to be the left slope, and b_r to be the right slope. The two slopes b_l and b_r are given by:

$$b_l = \frac{y_m - y_l}{x_m - x_l}\ ,\ b_r = \frac{y_r - y_m}{x_r - x_m}. \tag{4}$$

Let b_h be the half slope ratio, then

$$b_h = \frac{b_r}{b_l}. \tag{5}$$

If the value of b_h is close to 1, then we get an approximately straight line plot, otherwise there is no linear relationship between x and y then it will be necessary to re-express the data in order to get a linear relationship.

2.5 Re-express the Data

It is important to consider applying certain mathematical transformations to the data since many data analysis programs will have difficulty making sense of the data in its raw form (see Myatt (2007)). Furthermore, re-express means

using simple mathematical functions such as the logarithm and square root that help to simplify behaviors and clarify analysis to straighten out the data. After displaying the data as curves, we can take three representative points on any curves. Suppose we have three points say $p_1(x_1, y_1), p_2(x_2, y_2)$ and $p_3(x_3, y_3)$. A simple way to see whether the three points do not lie on a single straight line is to find the slopes b_1 and b_2 of the straight lines through the first pair p_1, p_2 and the second pair p_2, p_3 respectively; this gives:

$$b_1 = \frac{y_2 - y_1}{x_2 - x_1} \ , \ b_2 = \frac{y_3 - y_2}{x_3 - x_2}.$$

If $b_1 > b_2$, the curve is convex which means the middle point is above the line joining the other two points (Figure 1).

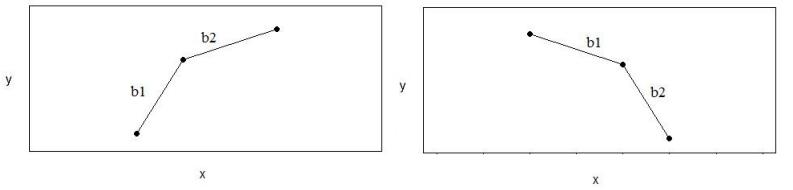

Figure 1: Left: the three points from the data curve shows left convex. Right: the three points from the data curve shows right convex

If $b_1 < b_2$, the curve is concave which means the middle point is below the line joining the other two points, as is shown in Figure 2.

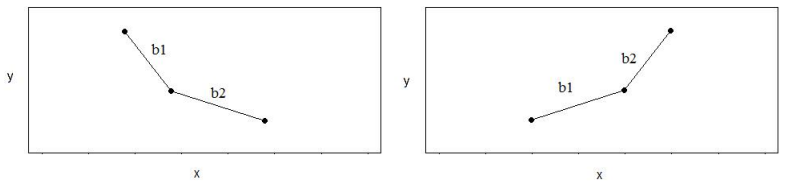

Figure 2: Left: the three points from the data curve shows left concave. Right: the three points from the data curve shows right concave.

If any pair of choices of expression is going to straighten out the early portion of the data curve, these same choices will have to do a reasonably good job in straightening out these three points. We can save a lot of effort by screening

our pairs of expressions on these three points. We will then need to try only the one best expression or perhaps the few best expressions on the whole data. The three summary points $(x_l, y_l), (x_m, y_m)$ and (x_r, y_r) are enough to be chosen for re-expression process. The common ladder ways to re-express y are $\left(y^3, y^2, \log y, \sqrt{y}, -\frac{1}{\sqrt{y}}, -\frac{1}{y}, -\frac{1}{y^2}, -\frac{1}{y^3}\right)$ and the same ways can be used for x. There is a simple rule that can be followed in case of y which is moving on the ladder as the convex appears. Figure 1 shows the upper ladder selections using one of the powers of y later, we check the result to see if b_1 is close enough to b_2 or else we must choose another expression (see Tukey (1977)). The way of choosing the suitable expression can also be guided by a moving point (x, y) in $x - y$ plane if we refer to up by $(+)$ and to down by $(-)$. A brief explanation to the way of choosing a suitable expression can be seen in Figure 3.

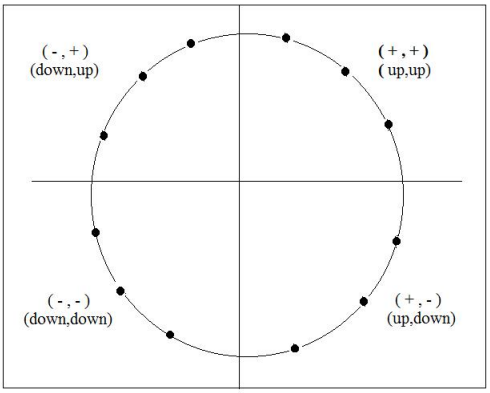

Figure 3: A simple guide of choosing the move for the ladder of powers x and y.

3. Resistant Smoothing

The aim of smoothing is to give a general idea of relatively slow changes of value with little attention paid to the close matching of data values in Myatt (2007)). There are many methods in the literature that used to smooth the curves like smoothing splines, roughness penalty, and many other different methods (see Ramsay and Silverman (2005)). We focus in our paper on running means smoothing method.

3.1 Running Means

The technique of running means to smooth the curves is sometimes used with temporal data. Let $x_i, i = 1, 2, \ldots, n$ be n observations in a data set. We can choose the span of the length optionally. In this technique we classify the process into three types; the symmetric, the right and the left running means. The symmetric mean applies on the odd spans $3, 5, 7, \ldots$, etc., while both the right and left running means use with the even spans $2, 4, 6, \ldots$, etc. In our paper we present symmetric running means. We shall denote the running mean by M such that $2 \leq M \leq n-1$. The spans 3, 5, 7 are used as examples to clarify the symmetric running means with value of n as a multiple of 3.

Symmetric Running Mean of Span 3

The steps start with running mean $3M$ using the form:

$$\begin{aligned} x_1^{3M} &= x_1 \\ x_i^{3M} &= \frac{\sum_{i-1}^{i+1} x_i}{3}, i = 2, 3, \ldots, n-1 \\ x_n^{3M} &= x_n \end{aligned} \tag{6}$$

Symmetric Running Mean of Span 5

The steps of running mean $5M$ are:

$$\begin{aligned} x_1^{5M} &= x_1 \\ x_2^{5M} &= x_2^{3M} \\ x_i^{5M} &= \frac{\sum_{i-2}^{i+2} x_i}{5}, i = 3, \ldots, n-2 \\ x_{n-1}^{5M} &= x_{n-1}^{3M} \\ x_n^{5M} &= x_n^{3M} = x_n \end{aligned} \tag{7}$$

Symmetric Running Mean of Span 7

The steps of running mean $7M$ are:

$$\begin{aligned}
x_1^{7M} &= x_1^{5M} = x_1^{3M} = x_1 \\
x_2^{7M} &= x_2^{5M} = x_2^{3M} \\
x_3^{7M} &= x_3^{5M} \\
x_i^{7M} &= \frac{\sum_{i-3}^{i+3} x_i}{7}, i = 4, \ldots, n-3 \\
x_{n-2}^{7M} &= x_{n-2}^{5M} \\
x_{n-1}^{7M} &= x_{n-1}^{5M} = x_{n-1}^{3M} \\
x_n^{7M} &= x_n^{5M} = x_n^{3M} = x_n
\end{aligned} \tag{8}$$

Generalized Running Means

We can derive a general form to do the process of running means method in two directions. The first is to put the main form to find running mean in general and the second is to get the additional conditions to any chosen span.

Generalized Symmetric Running Means

Let $k \in \mathbb{Z}^+$ be an odd span where $1 < k < n$, then the generalized form for computing the symmetric running mean is:

$$x_i^{kM} = \frac{\sum_{i-\left(\frac{k-1}{2}\right)}^{i+\left(\frac{k-1}{2}\right)} x_i}{k}, \; i = \frac{k+1}{2}, \ldots, n - \frac{k-1}{2} \tag{9}$$

Let the number of additional conditions be n_a and the number of spans be n_s. To calculate n_a we use the form:

$$n_a = n_s - 1.$$

Let $j = 1, 2, \ldots, n$ and $h = 1, 2, \ldots, \frac{m-3}{2}$ such that $2h + 1 \leq m < n$, then the additional conditions can be calculated by the form:

$$x_j^{(2h+3)M} = \begin{cases} x_1 & \text{if } j = 1 \\ x_j^{(2h+1)M} & \text{if } j \neq 1 \\ x_n & \text{if } j = n \end{cases} \tag{10}$$

Take $h = \frac{m-3}{2}$ then:

$$x_j^{\left[2\left(\frac{m-3}{2}\right)+3\right]M} = x_j^{\left[2\left(\frac{m-3}{2}\right)+1\right]M} \Rightarrow x_j^{mM} = x_j^{(m-2)M}$$

A diagram representation shown in Figure 4 that visualizes the process of symmetric running means technique helps understand relationships and the properties among the variant spans using equations (9) and (10).

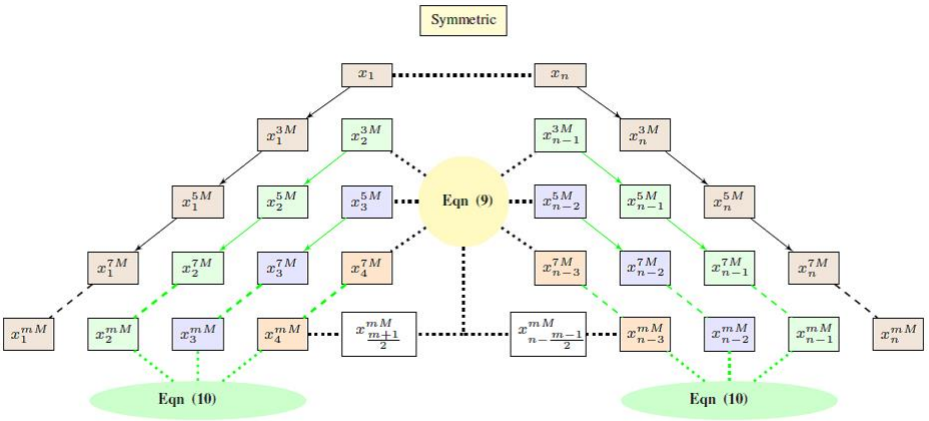

Figure 4: The process of symmetric running means shows the relationships among variant spans.

4. Real data Application

We apply our proposed smoothing techniques using data set of 20 years daily Farmer Mac agriculture data downloaded from the website retrieved on 2/11/16 (`https://www.quandl.com/data/GOOG/NYSE_MAC-Macerich-Co-MAC`).

The data in terms of the maximum indexes of agricultural production for each month for 20 years starting from 1995 to 2014 were organized into 20 rows by 12 columns matrix. Box plots summarizing the extreme data set are displayed in Figure 5.

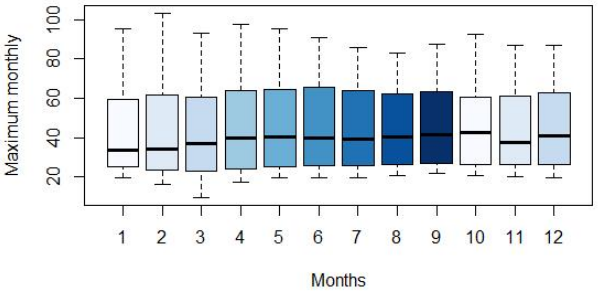

Figure 5: Box plots for 12 maximum monthly values for 20 years represented in this data set.

4.1 Resistant Line smoothing of Farmer Mac

Table 1 shows the initial steps in applying resistant line by dividing the data matrix into three portions l_p, m_p and r_p. The length of x_i which is considered in our study is the number of the months in every year. We have three equal lengths of portion. Each portion has 4 months (columns). The discrete month variable was defined as the set $[1, 2, \ldots, 12]$ to represent January, February,..., December, respectively.

Table 1: The three portions obtained as the first step of resistant line smoothing process.

	l_p				m_p				r_p			
Year	1	2	3	4	5	6	7	8	9	10	11	12
1995	21.50	20.75	20.75	20.50	20.25	19.88	20.62	20.88	21.88	21.00	20.00	20.50
1996	20.12	19.88	20.12	19.75	21.25	21.25	21.25	21.88	22.88	23.12	23.88	26.12
1997	28.12	27.62	29.38	28.25	27.12	28.62	29.25	29.56	29.00	29.25	27.12	29.06
1998	28.88	28.44	30.00	29.62	28.62	29.31	29.38	27.25	28.56	28.00	28.25	27.12
1999	26.69	24.81	24.31	25.56	26.94	27.06	26.62	25.06	24.56	22.62	20.88	20.81
2000	23.94	22.25	20.62	23.19	24.00	22.56	23.94	24.75	21.88	20.75	20.19	19.88
2001	20.94	21.13	21.95	22.77	23.55	24.80	25.01	25.20	25.07	24.20	25.51	26.60
2002	27.57	27.96	30.15	31.48	30.28	31.00	31.00	31.00	31.04	30.75	30.41	31.17
2003	30.90	32.15	33.17	33.60	35.35	47.00	37.50	37.45	38.44	40.55	42.49	44.50
2004	48.55	50.80	53.90	54.30	45.25	48.39	49.88	54.50	55.79	59.75	61.94	64.66
2005	62.15	62.00	59.43	60.30	63.35	67.32	70.22	71.19	66.03	65.73	68.58	68.54
2006	73.57	72.87	75.13	73.66	74.05	71.25	72.75	74.66	77.11	80.35	86.72	87.00
2007	95.53	103.3	93.43	97.69	95.18	90.66	85.61	83.00	87.58	92.66	83.26	81.64
2008	70.54	72.13	71.40	74.91	75.36	72.77	62.15	62.60	67.81	61.51	30.51	21.44
2009	19.85	16.43	9.58	17.53	19.65	21.17	19.67	30.04	35.35	31.06	32.60	37.62
2010	35.95	35.87	41.07	45.61	46.50	43.22	41.45	43.32	45.12	45.00	49.40	48.24
2011	48.94	50.64	49.53	52.82	54.37	53.50	56.20	52.59	48.05	50.00	51.04	51.18
2012	55.29	56.44	57.75	62.24	62.29	59.05	60.15	59.92	61.34	59.67	57.54	58.48
2013	59.99	62.03	64.38	70.05	70.84	63.94	65.93	62.35	59.96	59.76	59.93	60.49
2014	59.02	61.26	62.33	66.08	66.07	67.91	68.77	66.27	66.52	70.50	79.08	84.87

Using the values of summary points from equations (2) and (3), the process will find each value of the slopes b_i and the intercepts a_i for 20 rows of data matrix that display in Table 3. The 20 fitted lines have been computed using a_i and b_i, obtained from equation (1).

Straighten out plot stages result in order to get the linearity relationship between x_i and y_i can be applied using the three summary points $(x_l, y_l), (x_m, y_m)$ and (x_r, y_r) represented in Table 2 and the calculation of both b_l and b_r then b_h using equation (4) and (5) presented in Table 4. After that we compare the values of two slopes for example take $b_l = 1.07500$ and $b_r = 1.14000$ then the value of $b_h = 1.06046$ is closed to 1. Hence the three points are almost on the same line. On the other hand, Figure 3 shows an example of $b_l = 0.18750$ and $b_r = 0.02375$ so $b_l > b_r$ indicating the convex and $b_h = 7.89473$ which is far from 1, in this case an effective solution to deal with this problem is to re-express. We shall either move down the ladder of the powers of x using $\log x, -\frac{1}{x}, \ldots$, etc. or move up the ladder of the powers of y using $y^2, y^3, \ldots$, etc.

Table 2: The summary points as the second step of resistant line smoothing process.

	x_l		x_m		x_r
	2.500		**6.500**		**10.500**
y_l	20.750	y_m	20.435	y_r	20.750
	20.000		21.250		23.500
	28.185		28.935		29.030
	29.250		28.965		28.125
	25.185		26.780		21.750
	22.720		23.970		20.470
	21.540		24.905		25.290
	29.055		31.000		30.895
	32.660		36.960		41.520
	52.350		49.135		60.845
	61.150		68.770		67.285
	73.615		73.400		83.535
	96.610		88.135		85.420
	71.765		67.685		46.010
	16.980		20.420		33.975
	38.510		43.270		46.680
	50.085		53.935		50.520
	57.095		60.035		59.075
	63.205		64.935		59.945
	61.795		67.090		74.790

Table 3: The intercept points and the slopes from resistant lines smoothing process.

a	b
0.000000	20.645000
0.437500	18.739583
0.105625	28.030104
-0.140625	29.694063
-0.429375	27.362604
-0.281250	24.214792
0.468750	20.864792
0.230000	28.821667
1.107500	29.847917
1.061875	47.207812
0.766875	60.750312
1.240000	68.790000
-1.398750	99.146875
-3.219375	82.745937
2.124375	9.983229
1.021250	36.181875
0.054375	51.159896
0.247500	57.126250
-0.407500	65.343750
1.624375	57.333229

Table 4: The left and right slopes of the summary points and the ratio between them

b_l	b_r	b_h
-0.07875	0.07875	-1.00000000
0.31250	0.56250	0.55555556
0.18750	0.02375	7.89473684
-0.07125	-0.21000	0.33928571
0.39875	-1.25750	-0.31709742
0.31250	-0.87500	-0.35714286
0.84125	0.09625	8.74025974
0.48625	-0.02625	-18.52380952
1.07500	1.14000	0.94298246
-0.80375	2.92750	-0.27455167
1.90500	-0.37125	-5.13131313
-0.05375	2.53375	-0.02121362
-2.11875	-0.67875	3.12154696
-1.02000	-5.41875	0.18823529
0.86000	3.38875	0.25378089
1.19000	0.85250	1.39589443
0.96250	-0.85375	-1.12737921
0.73500	-0.24000	-3.06250000
0.43250	-1.24750	-0.34669330
1.32375	1.92500	0.68766234

Figure 6 shows the four faces re-expression that can be done to reach the linearity for each three summary points.

Figure 6a displays the re-expression $-\frac{1}{x}$ and y^{48} to get the linearity and the result is $b_h = 0.99877$ which is considered as a very satisfying result to stop the process.

Figure 6b shows an example of hollow down with $b_l = -1.0200$, $b_r = -5.41875$ and the half ratio $b_h = 0.18823$, but after moving up the ladder of powers for y only using the expression $y^{11.5}$ the half slope ratio $b_h = 1.02935$ which is so closed to 1.

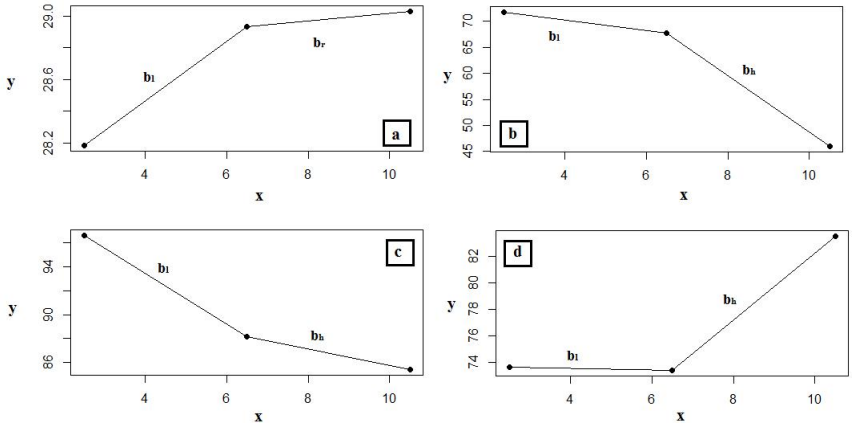

Figure 6: Four variant cases for three summary points represent the convex and the concave examples.

Figure 6c presents the moving down the ladder of the powers for both x and y using the expressions $\log_{10} x$ and $-\frac{1}{y^6}$ that gave $b_h = 1$.

Finally, Figure 6d displays the re-expression of x alone by $\sqrt{x^{5.5}}$ which was enough to get a straight line since $b_h = 0.99713$ meaning that we move down the ladder of the powers of x while keeping y in same power. The changing in shape after the chosen re-expression process can clearly be seen in Figure 7.

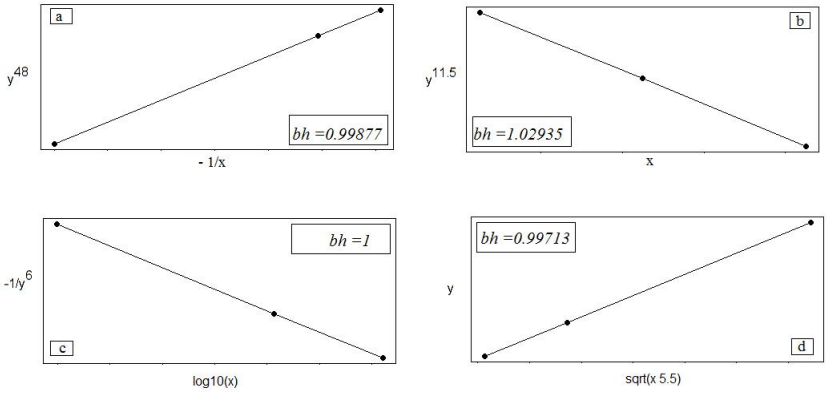

Figure 7: The lines results from re-expression of four examples of summary points using different forms.

5. Running Means

To test the performance of the proposed technique practically, we have applied it to smooth our data curves. The numerical results have been computed using the software R. The process of running means according to the chosen spans with equations (9), (10) and Figure 4 gives the smoothed curves shown in Figure 8 and Figure 9.

The raw data before smoothing is shown in Figure 8a. A slight change in the pattern of raw data appears in Figure 8b by applying the $3M$ technique. In Figure 8c the shape of the data curve become smoother compared to the two previous plots. Finally, it is clear that a better smoothed curve is the one displays in Figure 8d.

The end points $x_1^{7M} = X_1^{5M} = X_1^{3M} = X_1$ and $x_n^{7M} = X_n^{5M} = X_n^{3M} = X_n$ computed by equation (9) and shows in Figure 4 are depicted in the four parts of Figure 8, meaning that the end points of the curves are temporary out of smoothing.

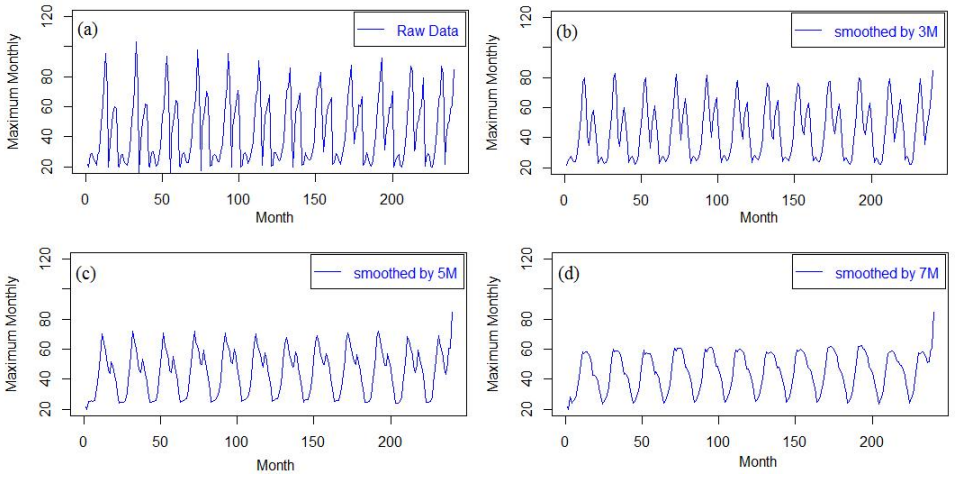

Figure 8: The raw and the smoothed curves result from running mean with variant spans.

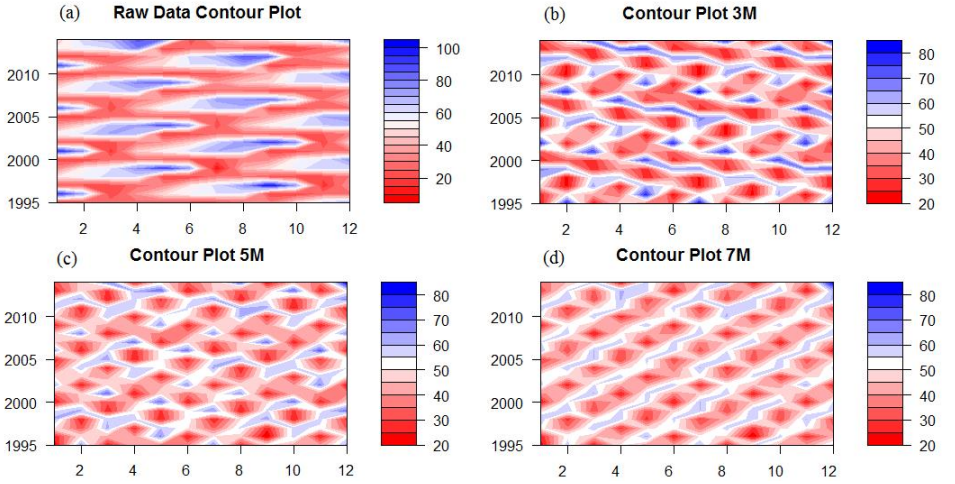

Figure 9: The raw and the smoothed curves result from running mean with variant.

6. Conclusions

Extended approach of Tukey's resistant smoothing and running means approaches as tools of EDA method have been proposed in this paper to have a

better analysis and smoothing curves from the raw data curves. Our proposed approach differs from Tukey's method in the following aspects:

(i) Tukey's technique presented shape of numeric exhibitions while our approach adopts the scatter and contour plots besides presenting general mathematical forms.

(ii) Tukey deals with the calculations manually that makes the process slow with limited kinds of data sets, in contrast the proposed approaches have the technologic preference of using a new software R that can reduce the time and simplify the application parts in spite of the large data observations.

Here, we discuss some limitations of the proposed approach in terms of the following aspects: First, the running means uses mean to find the majority of the results; it cannot deal with the existence of outliers in any data observations. Second, the tails of the raw data are out of smoothed.

Acknowledgments

The authors would like to thank the anonymous referees for their valuable comments that helped improve this paper.

References

Myatt, G. J. (2007). *Making Sense of Data. A Practical Guide to Exploratory Data Analysis and Data Mining.* John Wiley & Sons, Inc., New York, USA.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis Second Edition.* Springer, New York, USA.

Shitan, M. and Vazifedan, T. (2011). *Exploratory Data Analysis for Almost Anyone.* Universiti Putra Malaysia Press, Serdang, Malaysia.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Canada.

Velleman, P. F. and Hoaglin, D. C. (2004). *Application, Basics and Computing of Exploratory Data Analysis.* Duxbury Press, Boston, USA.