Journal of
Environmental & Analytical Toxicology

Sharma et al., J Environ Anal Toxicol 2015, 5:5
DOI: 10.4172/2161-0525.1000304

Research Article Open Access

Automating the Smoothing of Time Series Data

Shilpy Sharma*, David A Swayne and Charlie Obimbo
School of Computer Science, University of Guelph, Canada

Abstract

Modelling requires comparison of model outputs to measurements, for calibration and verification. A key aspect data smoothing is to "filter out" noise. Often, data must be adjusted to a model's time step (e.g. hourly to daily). For noisy data, LOWESS/LOESS (Locally Weighted Scatterplot Smoothing) is a popular piecewise regression technique. It produces a "smoothed" time series. LOWESS/LOESS is often used to visually assess the relationship between variables. The selection of LOWESS tuning parameters is usually performed on a visual trial and error basis. We investigate the so-called robust AIC (Akaike Information Criteria) for automatic selection of smoothing. Robust Pearson correlation coefficient and mean-squared error are employed to determine the polynomial degree of piecewise regression. The exclusion of outliers is attempted using a Hampel outlier identifier. We illustrate, assuming noisy linear data, how our proposed methods work for auto-tuning both the smoothing parameter and the degree of polynomial for LOWESS/LOESS.

Keywords: LOWESS; LOESS; Data smoothing

Introduction

LOWESS is a powerful non parametric technique for fitting a smoothed line for a given data set either through univariate or multivariate smoothing [1]. It implements a regression on a collection of points in a moving range, and weighted according to distance, around abscissa values in order to calculate ordinal values. "LOWESS" and "LOESS" are acronym for "Locally Weighted Scatterplot Smooth" because, for data smoothing, locally weighted regression is used. Furthermore, a robust weight function can be used to compensate for undue influence of extreme points. Differentiation of a regression model depends on the way it is used: a linear polynomial is used for LOWESS whereas a quadratic polynomial is used for LOESS [2]. From the literature review, most authors consider LOWESS/LOESS same, but they are different. LOWESS is derived from term "locally weighted scatterplot smoothing" whereas according to Potts LOESS stands for "Locally Estimated Scatterplot Smoothing" [3].

LO(W)ESS is one of the most widely used method for data smoothing and trend estimation. Its graphical representation helps to visualize the overall trends in a time series and identify times of changes. As an example, in 2012 presidential elections [4], LOWESS fit is used to predict the presidential candidate. Moreover, Niblett [5] used LOWESS fit for the evolution of legal rules. There are many more other examples where LOWESS fit has been used for trend estimation. This is the easiest way to communicate trends especially for the non-technical people.

There are numerous techniques for data smoothing: splines, beziers, kernel and polynomial regression. Local regression provides some attractive features, according to Cleveland [6]. LOESS fit is extremely informative when the data set is very large [7]. Moreover, it is used to solve the problems of precision, noise filtering, and outliers and is known to adapt well to bias problems, as opposed to these other methods. Also, LO(W)ESS is computationally efficient [8]. Our experiments with polynomial regression and Bezier curves show that if we increase the degree of polynomial, it increases undulations in the curve fit being attempted, and we cannot estimate the actual picture. Moreover, increasing degree of polynomial often creates the problem of data over-smoothing. If outliers are present in the dataset, robust LOWESS/LOESS (rLOWESS, rLOESS) procedure is used to overcome the problem of distorted values. The presence of outliers can be detected using a Hampel identifier. It is considered to be the most efficient and robust method for identification of outliers [9]. A data point is considered as an outlier, based on Hample Identifier (HI), if it goes beyond $\pm 3\sigma$ where the variance estimate σ is calculated by subtracting median absolute deviation and median. It signals the index of the outlier as the final outcome.

The selection of smoothing parameter, α, is often entirely based on a "repeated trial" basis. Some researchers argue that it should lie between 0.2 and 0.8 while others consider 0.5 as an ideal parameter value [10]. There is no specific technique for selection of the exact value of α. Selection of α may lead to "over-smoothing" or "under-smoothing" of data, does not necessarily provide good information for LO(W)ESS fit. Figure 1 shows a sample LO(W)ESS fit using different smoothing parameters. The LO(W)ESS fit which follow the almost all the data points is called "under smoothing" or "over-fitting" whereas if does not follow the data and produce a smooth line is called "lack of fit" or "under-smoothing".

For calculating smoothed values from the robust method, there are two extra steps; firstly, it requires the calculations of residuals from LOWESS/LOESS and then robust weight function is applied using the bisquare function [11]. Once the regression function values are calculated with flexible weights and polynomial degree, LOESS fit is complete [11]. Robust AIC (Akaike Information Criteria) is used for the selection of the best fitted smoothing parameter, α, and akaike weights are used to evaluate the best selected model [12]. Moreover, we can use the robust Pearson correlation coefficient for the selection of degree of polynomial, λ.

We have noted from the literature that there are nine different methods being used for selection of α in non-parametric regression, but, according to Aydin [13], Improved Akaike Information Criteria (AICc) and Generalized Cross Validation (GCV) are the best methods

***Corresponding author:** Shilpy Sharma, School of Computer Science, University of Guelph, Canada, Tel: 15198244120; E-mail: shilpy@uoguelph.ca

Received May 14, 2015; **Accepted** June 20, 2015; **Published** June 24, 2015

Citation: Sharma S, Swayne DA, Obimbo C (2015) Automating the Smoothing of Time Series Data. J Environ Anal Toxicol 5: 304. doi:10.4172/2161-0525.1000304

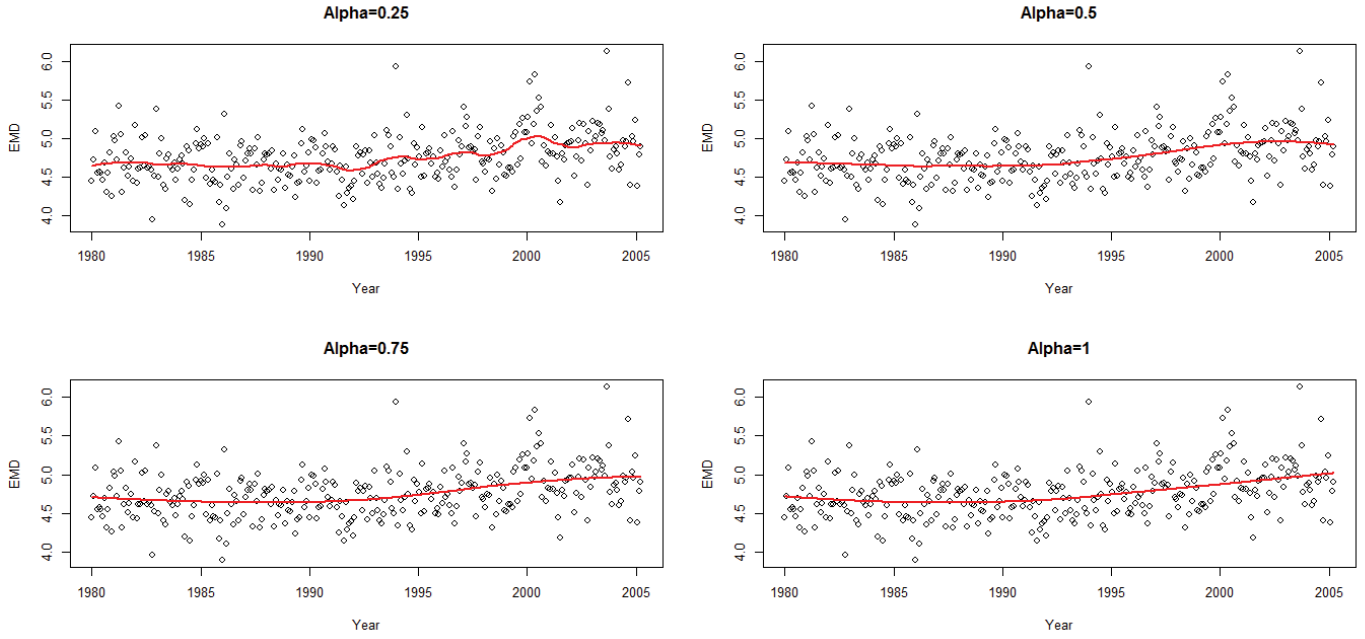

Figure 1: LOWESS with different values of α.

based on small and large datasets [14,15]. The best model is selected is the one with the smallest AIC score. Tharmaratnam [16] proposed the robust version of AIC which produces promising results for model selection in the presence of outliers compared to non-robust AIC. We will be using AIC M-estimator, robust AIC, for α. The best selected model can be estimated using Akaike weights [12].

In this paper, we are using synthetic as well as real data [17], to illustrate the usefulness of our proposed method.

Methods

LOWESS analysis for noisy synthetic linear data

We have generated noisy linear data using (1).

$$Y = X + e \tag{1}$$

An artificial dataset with 50 data points is generated for the experiment. Normally distributed noise, e, with zero mean and 0.01 variance has been added to produce outliers in the linear data, whereas X is generated uniformly.

LOWESS/LOESS analysis for Rock Creek River

The data set used for our research analysis is from Heidelberg University [17]. It is collected from Rock Creek River for 10 different parameters and we will analyze "Suspended Solid" data collected hourly but intermittently between 1982-2013.

Calculating the mean of stratified samples does not necessarily provide an accurate picture of the data. Our data readings vary from hourly to daily data and therefore, weights should be considered to calculate average. We will reduce the values using Flow-Weighted Mean Concentration (FWMC) for accuracy and consistency which should give a clear picture of actual loadings.

The flow-weighted average is considered to be an accurate method for use in calculating average for stratified samples in which the readings have different time intervals, varying from hourly to daily readings [18,19]. These different weights should be considered for calculating average. Calculating FWMC does not have effect of missing data [19]. Equation (2) shows the formula to calculate flow-weighted average is [18,19].

$$FWMC = \frac{\sum_1^n (c_i * t_i * q_i)}{\sum_1^n (t_i * q_i)} \tag{2}$$

where q_i=flow in the i^{th} sample, c_i=concentration of the i^{th} sample, t_i=time window for the i^{th} sample

Proposed Methodology

The steps for appropriate selection of smoothing parameters for LO(W)ESS are shown in Figure 2. The experimental steps for selection of α and λ are as follows:

Analyze the presence of outliers in the data

In order to detect the presence of outliers, a Hample Identifier (HI) is employed.

Differentiate LO(W)ESS from rLO(W)ESS

If outliers are present in the data set rLO(W)ESS is used otherwise, LO(W)ESS is used for analysis.

Examine the presence of a monotonic relationship

This step is extremely important to identify which degree of polynomial is used, linear or quadratic. If X and Y show a monotonic relationship, then λ=1 otherwise, λ=2 [11]. The strength of correlation can be measured based on values give in Table 1. For automatic selection of λ, weighted Pearson correlation, r_w, or Mean Squared Error (MSE) can be used [20,21]. The best fit polynomial is considered to be the one for which MSE is less or having high r_w value. The r_w is more suitable for testing correlation compared to Pearson correlation [20]. The Pearson Correlation coefficient is calculated for the express purpose of identifying whether the data is monotonic.

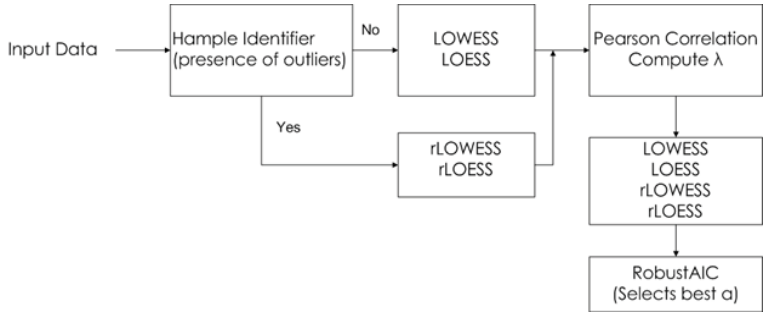

Figure 2: Steps for automatic smoothing parameter selection: LO(W)ESS.

Relationship	Very strong	Strong	Moderate	Weak	Very weak
Value	0.80-1.0	0.60-0.79	0.40-0.59	0.20-0.39	0.00-0.19

Table 1: Correlation coefficient.

Equation (3) shows the formula to calculate MSE and Equation (4) for r_w.

$$MSE = \frac{1\sum_1^n \left(\hat{Y}_i - Y_i\right)^2}{n} \tag{3}$$

where Y is observed values, $\hat{Y}$ is the predicated values and n is the number of values.

$$r_w = \frac{\sum_{i=1}^n w_i \left(x_i - \bar{x}_w\right)\left(y_i - \bar{y}_w\right)}{\sqrt{\sum_{i=1}^n w_i \left(x_i - \bar{x}_w\right)^2 \sum_{i=1}^n w_i \left(y_i - \bar{y}_w\right)^2}} \tag{4}$$

where x, y are the set of variables, w_i is the weight, $\bar{x}_w$, $\bar{y}_w$ are the weighted mean of x and y variables.

Differentiating LOWESS, LOESS, rLOWESS and rLOESS

The selection of best LO(W)ESS model can be performed as follows:

a) Monotonic relationship without outliers: LOWESS

b) Monotonic relationship with outliers: rLOWESS

c) No monotonic relationship without outlier: LOESS

d) No monotonic relationship with outlier: rLOESS

Select the best smoothing parameter using RobustAIC

Tharmaratnam [16] proposed the robust version of AIC, which produces promising results for model selection in the presence of outliers compared to non-robust AIC. RobustAIC score is computed for all the values of α between 0.1 and 1. The model with lowest AIC score is considered as the best value of α based on the dataset. Section below on Computational Steps for LOWESS/LOESS provides the details of computing LO(W)ESS/rLO(W)ESS fit based on the smoothing parameter. The best selected model can be estimated using Akaike weights [12]. The algorithm for calculating Akaike weights is as follows:

a. Calculate AIC for all the models and identify the best model, AIC_{min}

b. Calculate the difference between AIC of every model and AIC_{min}

$$\Delta_i\left(AIC\right) = AIC_i - AIC_{min} \tag{5}$$

c. Compute Akaike weights for each model and normalized relative likelihoods

$$w_i = \frac{exp\left[-0.5 * \Delta_i\left(AIC\right)\right]}{\sum_{i=1}^n exp\left[-0.5 * \Delta_i\left(AIC\right)\right]} \tag{6}$$

Computational steps for LOWESS/LOESS

In the literature, the selection of smoothing parameter, α, is often entirely based on trial and basis. Some researchers argue that it should be between 0.2 and 0.8 while other considers 0.5 as an ideal starting point [10]. There is no specific technique for selection of the exact value of α. Random selection of α may lead to over-smoothing or under-smoothing of data, which in turn does not provide good information for LOWESS fit. The LOWESS/LOESS fit which follow the almost all the data-point is called "under-smoothing" or "over-fitting" whereas if does not follow the data and produce a smooth line is called "lack of fit" or "under-smoothing". The step by step calculation of LOWESS/LOESS and rLOWESS/rLOESS are as follows [1,10,22].

A: Compute tricube weights, equation 7, using scaled distance. These weights are calculated for set of numbers in local neighbourhood.

$$w(x) = \begin{cases} \left(1 - |x|^3\right)^3, & if\ |x| < 1; \\ 0 & otherwise \end{cases} \tag{7}$$

B: Run weighted least square regression for those set of numbers.

C: If outliers are present in data; calculate residuals, median of residuals and robust weight, equation 8, using robust weight function.

$$\delta_i = \begin{cases} \left(1 - \left(\frac{e_i}{6s}\right)^2\right)^2, & if\ |e_i| < 6s; \\ 0 & otherwise \end{cases} \tag{8}$$

D. Run weighted least regression using robust weights.

E. Repeat step 3 and 4 until convergence criteria is met.

Scatterplot smoothing

For demonstration of the LOWESS procedure, we will first examine a synthetic dataset and finally real life data, the suspended solids parameter, from Rock Creek River.

Experiment using synthetic data: Figure 3 shows the scatterplot for our synthetic data. We will examine our dataset with the proposed methods, given in section on proposed methodology.

Step 1 Presence of outliers: The Hample Identifier (HI) detected two outliers at index 21 and 46.

Step 2 LO(W)ESS vs. rLO(W)ESS: Since two outliers are detected

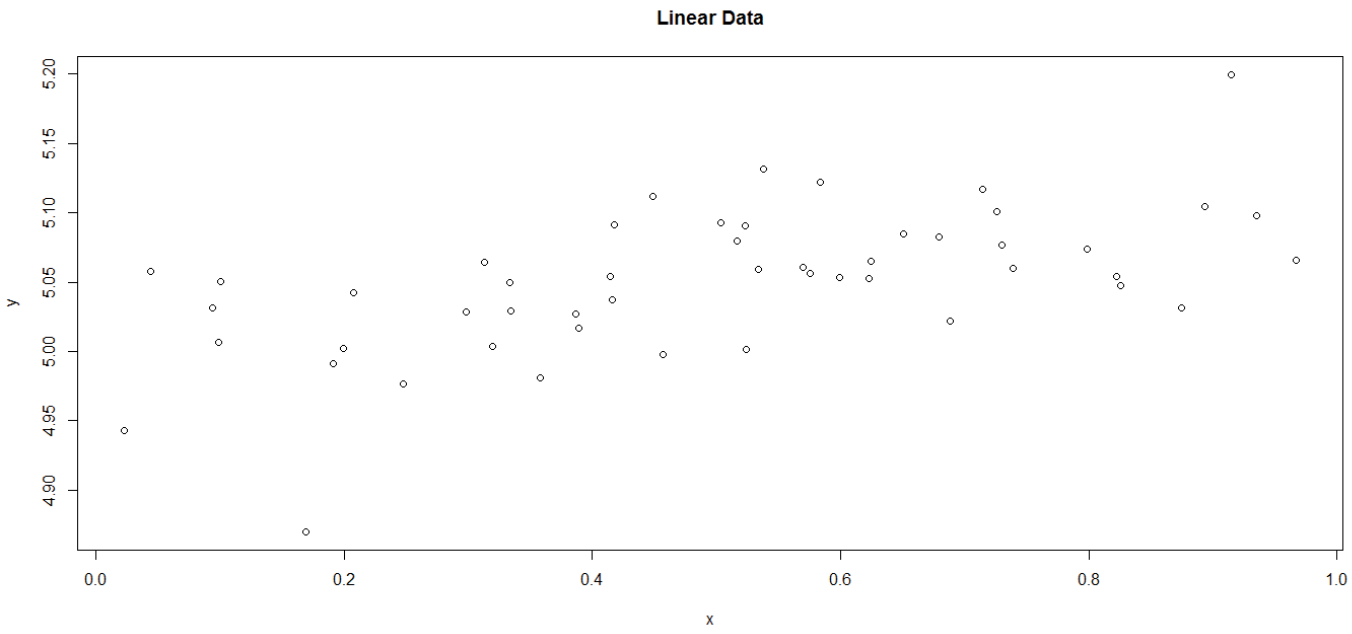

Figure 3: Scatterplot for linear data.

in the dataset, rLO(W)ESS is employed. The computational procedure for LO(W)ESS and rLO(W)ESS is given in section on Proposed Methodology.

Step 3 Identify monotonic relationships: The selection of appropriate α and λ is extremely important. Based on experimental results for linear data, the value of r_w=0:896 whereas r=0.582. This indicates that there is strong relationship; therefore, λ=1 should be used for analysis. Similarly, the calculation value for MSE for λ=1 is 0.00184 whereas for λ=2 is 0.00188. Again, MSE confirms that λ=1 is the best for analysis.

Step 4 Differentiating LOWESS, LOESS, rLOWESS and rLOESS: Since a monotonic relationship exists and presence of outliers is detected in Step 1, rLOWESS is chosen for the analysis.

Step 5 RobustAIC for parameter selection: Table 2 shows the RobustAIC score calculated from the equations given in section computational steps for LOWESS/LOESS robustAIC selects the best value for α based on the data. Based on Tharmaratnam robust AIC [16], Table 2 illustrates that α=0.3 is the smallest AIC score. Table 3 shows the best selected model using akaike weights, calculated from step 5 in section on proposed methodology.

Figure 4 shows the rLOWESS fit based on α=0.3, black line is the original data and blue is the rLOWESS fit. In this situation, the original data to which the noise was "added" is not recovered. This is an artificial dataset. The criteria do not recover the noise-free data, but rather they smooth the dataset according to the noisy data as presented, some of which began as an added noise, and which is retained in the smoothed line.

Experiment on real data:

Step 1 Presence of outliers: Based on data set, HI detected 28 outliers in the dataset

Step 2 LO(W)ESS vs. rLO(W)ESS: Since outliers are detected in the dataset, rLO(W)ESS is chosen.

Smoothing parameter	Robust AIC score
0.1	-323.061
0.2	-340.792
0.3	-347.394
I 0.4	-325.645
0.5	-320.392
0.6	-311.830
0.7	-309.811
0.8	-317.420
0.9	-306.575
1.0	-310.225

Table 2: Robust AIC score for different values of *α*.

Robust AIC	δAIC	w_i (AIC)
-323.061	24.333	5.01685E-06
-340.792	6.602	0.035536013
-347.394	0	0.964439054
-325.645	21.749	1.82616E-05
-320.392	27.002	1.32088E-06
-311.830	35.564	1.82663E-08
-309.811	37.583	6.65626E-09
-317.420	29.974	2.98884E-07
-306.575	40.819	1.3199E-09
-310.225	37.169	8.18708E-09

Table 3: Best selected model using Akaike weights.

Step 3 Identify monotonic relationships: Based on the analysis, the value of r_w=0.0191 shows a very weak relationship; rLOESS is indicated.

Step 4 Differentiating LOWESS, LOESS, rLOWESS and rLOESS: Since monotonic relationship does not exist and presence of outliers is detected in Step 1, rLOESS is chosen for the analysis.

Step 5 RobustAIC for parameter selection: RobustAIC selects the best value for α based on the data, as described in section on experiment

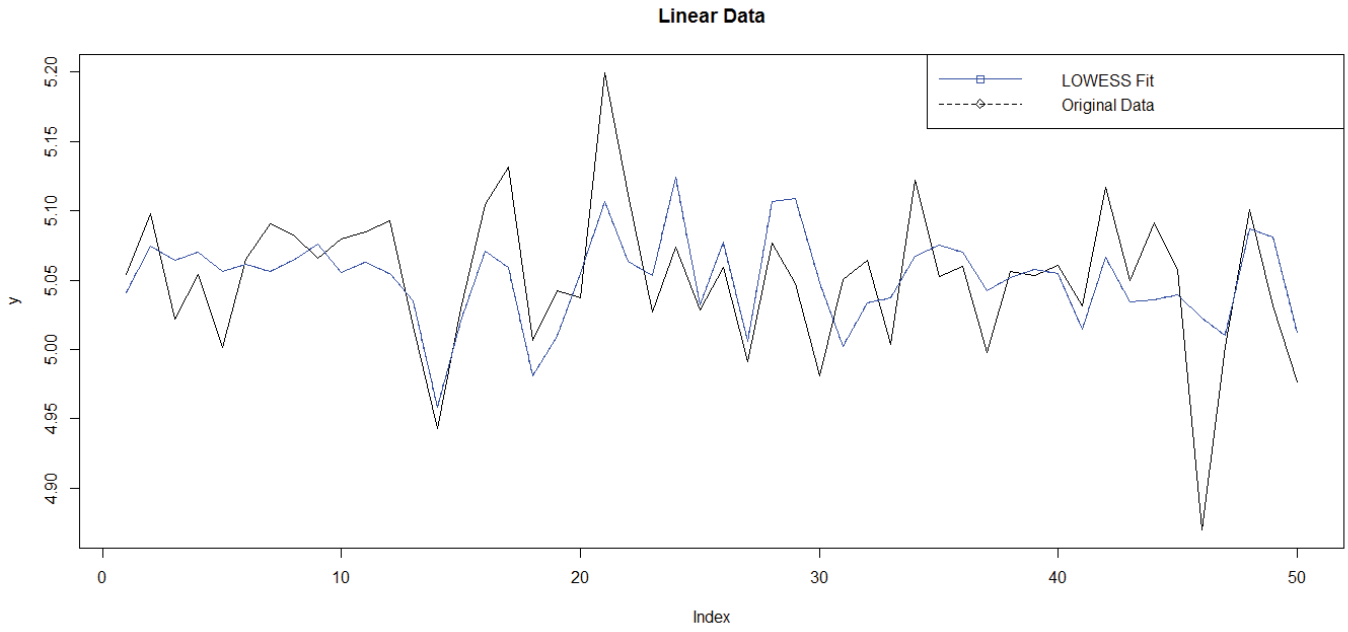

Table 4: Robust AIC score for different values of α.

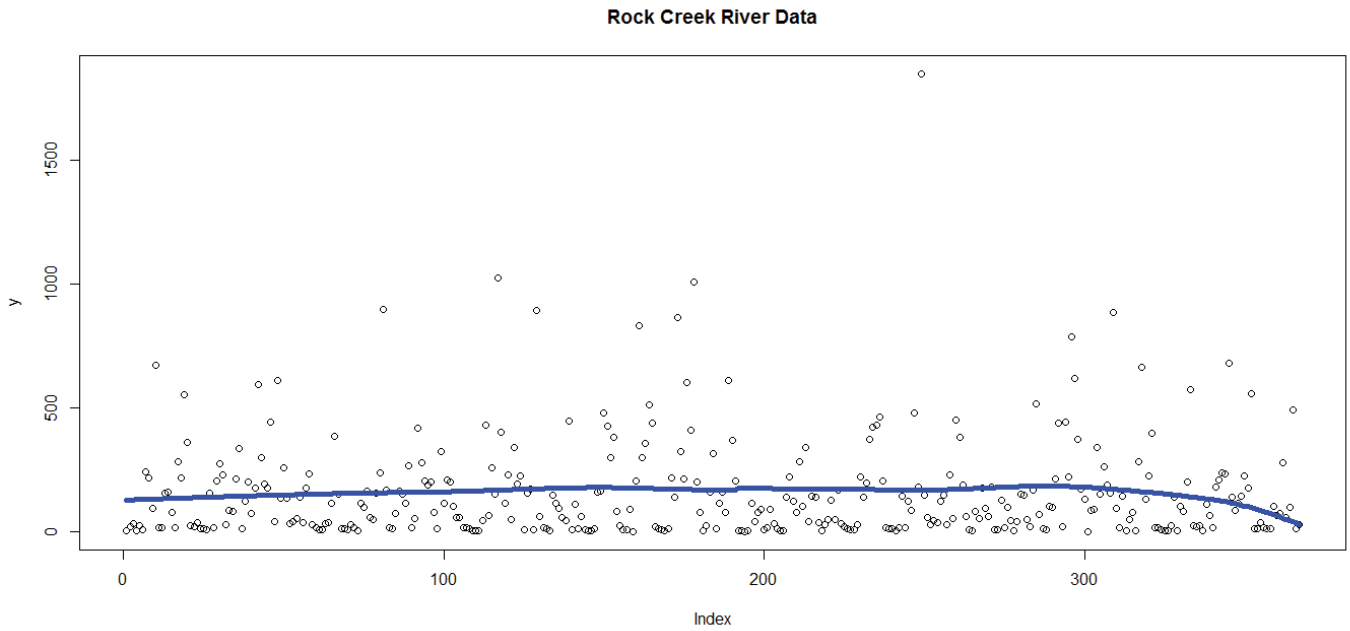

Figure 5: rLOWESS fit with α=0.8.

using synthetic data based on tharmaratnam robust AIC [16], Table 4 shows that α=0.8 is the smallest AIC score. Table 5 shows the best selected model using akaike weights (Figure 5).

Conclusion

LO(W)ESS is widely used in different application areas such as for normalization and accessing non-linear relationships between variables and considered as one of the important member of non-parametric regression in statistical circle. It is unfortunate that despite its wide application area, the important parameters are selected on trial and error basis. Over-smoothing and under-smoothing is neither acceptable nor desirable in such situations. Over-smoothing divulges trend but ignores local variations whereas under-smoothing results in too many local variations.

An automatic approach for selection of smoothing parameters for LO(W)ESS fit has been proposed and tested. The degree of polynomial and presence of outliers is used to select the type of LO(W)ESS. Also, the best value of smoothing parameter is chosen based on the least value of AIC values. AIC with mm-estimator is employed for the

α	Robust AIC score
0.1	3190.12
0.2	2801.927
0.3	2475.304
0.4	2496.164
0.5	2479.504
0.6	2452.204
0.7	2193.253
0.8	1930.11
0.9	1963.701
1.0	2097.753

Table 4: Robust AIC score for different values of *α*.

Robust AIC	δAI C	wi(AI C)
3190.12	1260.01	2.47E-274
2801.927	871.817	4.87E-190
2475.304	545.194	4.10E-119
2496.164	566.054	1.21E-123
2479.504	549.394	5.02E-120
2452.204	522.094	4.25E-114
2193.253	263.143	7.23E-58
1930.11	0	1.00E+00
1963.701	33.591	5.08E-08
2097.753	167.643	3.95E-37

Table 5: Best selected model using Akaike weights.

selection of best model for smoothing parameters that works well in the presence of outliers. Also, weighted MSE is used for the estimation of best degree of polynomial. The accuracy of the aforementioned methodology has been tested and demonstrated using experimental results.

In the first experiment, an artificial data set has been generated to test if the proposed method works as per expectations. It is known in advance that the data is linear with the presence of outliers in it. A real data set from Rock Creek River is examined using the proposed method. Our proposed method is able to automate the smoothing parameter and degree of polynomial. At the same time, it eliminates the problem of over-smoothing and under-smoothing of data. The approach is flexible and easy to implement in a variety of situations.

References

1. Cleveland WS, Devlin SJ (1988) Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting. J Am Stat Assoc 83: 596-610.
2. MathWorks (2015) Lowess Smoothing.
3. (2003) Nutrient scientific technical exchange partnership and support.
4. Rothschild D (2012) As gingrich's fate rises, so does obama's zot.
5. Niblett A, Posner RA, Shleifer A (2010) The evolution of a legal rule. Journal of Legal Studies 39: 325-358.
6. Cleveland WS, Loader CL (1996) Smoothing by Local Regression: Principles and Methods. Statistical Theory and Computational Aspects of Smoothing 10-49.
7. Faraway J (1997) Regression analysis for a functional response. Technometrics 39: 254-261.
8. Hen I, Sakov A, Kafkafi N, Golani I, Benjamini Y (2004) The dynamics of spatial behavior: how can robust smoothing techniques help? J Neurosci Methods 133: 161-172.
9. Liu H, Shah S, Jiang W (2004) On-line outlier detection and data cleaning. Comput Chem Eng 28: 1635-1647.
10. Cleveland WS (1979) Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc 74: 829-836.
11. Jacoby WG (2000) Loess: A Nonparametric, Graphical tool for depicting relationships between variables. Electoral Studies 19: 577-613.
12. Wagenmakers EJ, Farrell S (2004) AIC model selection using Akaike weights. Psychon Bull Rev 11: 192-196.
13. Aydin D, Memmedli M, Omay RE (2013) Smoothing parameter selection for nonparametric regression using smoothing spline. European Journal of Pure and applied mathematics 6: 222-238.
14. Garcia D (2010) Robust smoothing of gridded data in one and higher dimensions with missing values. Comput Stat Data Anal 54: 1167-1178.
15. Francisco-.Fernandez M, Opsomer JD (2005) Smoothing parameter selection method for nonparametric regression with spatially correlated errors. Can J Statistics, 33:279-295.
16. Tharmaratnam K, Claesken G (2013) A comparison of robust versions of the AIC based on M, S and MM-estimators. Journal of Theoretical and Applied Statistics 47: 216-235.
17. Baker D (2013) Data and analysis template files.
18. Baker D (2005) Time-weighted and flow-weighted mean concentrations.
19. Cooke S, Ahmed S, Alpine NM (2005) Introductory Guide to Surface Water Quality Monitoring in Agriculture. Conservation and Development Branch, Alberta Agriculture, Food and Rural Development.
20. Kalogirou S (2013) Testing geographically weighted multicollinearity diagnostics.
21. Stronger D, Stone P (2008) Polynomial regression with automated degree: A function approximator for autonomous agents. Int J Artif Intell Tools 17: 159-174.
22. Kimmel RK, Booth DE, Booth SE (2010) The analysis of outlying data points by robust locally weighted scatter plot smooth: a model for the identification of problem banks. International Journal Operational Research 7: 1-15.