International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.9, No.10, (2016), pp.353-364
http://dx.doi.org/10.14257/ijsip.2016.9.10.34

A Rapid Steady-State Data Extracting Method Fusing Outliers Detection Based on Automatic Piecewise Curve Fitting

Hanming Fang[1*], Aishe Shui[2], Fuxing Zong[3] and Yuheng Wu[4]

Department of Logistics Information & Logistics Engineering, LEU, Chongqing 401311, China
fang_hanming@163.com

Abstract

Aiming at the problem of extracting steady-state data automatically and rapidly from the process data which contains outliers, a rapid steady-state data extracting method fusing outliers detection based on automatic piecewise curve fitting is proposed. Firstly, the method carries out outliers detection according to local deviation and replaces it with grey theory. Then the noise is minimized through sliding mean filter and the quasi-steady-state data is extracted rapidly by involved rules. Lastly, the quasi-steady-state data is further judged by automatic piecewise curve fitting. The simulation test shows that the proposed method can not only eliminate the effect of outliers and noise, but also extract the steady-state data conforming to human's experience quickly and efficiently.

Keywords: *steady-state detection, data extraction, outliers, curve fitting*

1. Introduction

Steady-state detection and steady-state data extraction are widely used in system modeling, process optimization control and process fault diagnosis. With the development of science and technology, all kinds of instruments are used in industrial production and produce a lot of process data. How to mine the steady-state data which play an important role in guiding chemical production from the process data is worth researching.

The existing steady-state detection methods can be summarized as three categories: the method based on the mechanism, the method based on statistical analysis and the method based on trend extraction [1]. The method based on the mechanism is rarely used in general because of the need to know the running mechanism of process and its poor universality [1-2]. The method of statistical analysis represented by the method of combination of statistical test, confidence and R test can carry the general steady-state detection. But this method can only detect the stability of a period and is influenced easily by its parameters [3-4]. Most studies have focused on the trend extraction method at present.

Li *et. al.,* [3] ruled out the fault factors' influence on the steady-state detection by improved filtering. But this method is high to the requirement of filter function. Literature [4] carried out the steady-state detection by the method of combination of statistical test. It is easy to make the first type mistake. Gao *et. al.,* [5] utilized the Gaussian filtering to de-noise, then R test is adopted to detect the steady state. But R test is sensitive to three filter coefficients, and more difficult to practice. The references on selecting fitting functions and determining piecewise intervals for curve fitting method were given in the literature [6]. Kelly *et. al.,* [7] put forward the method of Hypothesis testing primarily, then used test for steady-state detection. Chen *et. al.,* [8] adopted the method of histogram to eliminate outliers on the basis of literature [7], but these two methods omitted outliers easily when the changes were larger between data before and after. Lv *et. al.,* [9] got a

[*] Corresponding Author

polynomial fitting curve for overall data by data weighting and gave a method for calculating the steady-state index. But this method did not fully consider the random error or the gross error. The adaptive smoothing algorithm and threshold value method were applied to eliminate the random error and gross error [10]. But this method was easy to appear a situation that the denominator is zero.

Therefore, the paper proposes an extracting method of steady-state data based on automatic piecewise curve fitting containing outliers detection. This method is able to apart the steady-state data from process data which include outliers. It mainly contains two steps: outliers detection and substation, quasi-steady-state data screening and steady-state extraction. The details of this method are described as follows.

2. Outliers Detection and Substation

2.1. Outliers Detection Based on the Improved Local Deviation

Outliers will affect the effect of steady-state detection directly. Outliers detection has attracted wide attention [11]. Focusing on the problem of outliers detection which change rather greatly during the process, an improved local deviation algorithm is put forward to solve this problem on the basis of literature [12-[13]. Specific steps are depicted as below.

Process data $X = \{x(1), x(2), .., x(n)\}$, use the sliding window technology, mobile step is 1, the width of sliding window is $2m+1$,take the data centered $x(i)$ ($\{x(i-m), ..., x(i), ..., x(i+m)\}$) as examples.

(1) Calculate the absolute value of longitudinal distance between point $x(i)$ and other $2m$ points.

$$d_{ij} = |x(i) - x(j)| \tag{1}$$

Where d_{ij} is the absolute value of longitudinal distance between point $x(i)$ and point $x(j)$.

(2) Calculate the average longitudinal distance (D_i) between point $x(i)$ to other $2m$ points.

$$D_i = \frac{\sum\limits_{j=i-m(j\neq i)}^{i+m} d_{ij}}{2m} \tag{2}$$

(3) Calculate the degree of deviation (Dev_i) for point $x(i)$.

$$Dev_i = \sqrt{\frac{\sum\limits_{j=i-m(j\neq i)}^{i+m} \left|D_j - d_{ij}\right|^2}{2m}} \tag{3}$$

(4) Calculate the degree of deviation (Dev) for other points among the data sequence according to the above steps, take the maximum deviation of the other point as $Dev_{\max}$, and calculate the Dev_i - $Dev_{\max}$ ratio.

$$Dev_{\max} = \max\{x(i-m), ..., x(i-1), x(i+1), ..., x(i+m)\} \tag{4}$$

$$k = \frac{Dev_i}{Dev_{\max}} \tag{5}$$

Where k is the relative degree of deviating from the other data within a window for point $x(i)$. The point $x(i)$ is thought to be outliers when the k value is greater than 1.5. The threshold value can also be adjusted on the basis of the running time of process.

2.2. Outliers Substitution Based on Grey Prediction Theory

There exist three ways concerning the prediction for points which true value is uncertain: grey system theory, probability statistics and fuzzy mathematics. The method of probability statistics not only need to know the typical distribution of data, but also is in view of the large sample data. Fuzzy mathematics is a kind of forecasting method with experience, whereas, membership functions need to be constructed for it [14]. The grey system theory aims at the objective law of data. At the same time, it can forecast and estimate well for small sample data through grey sequence operator [15].

2.2.1. GM(1,1) Model Building: Extract the n data before a certain outlier, such as $X'=\{x'(1),x'(2),..,x'(n)\}$, let $z'(k)=mx'(k)+(1-m)x'(k-1)$. Where $x'(n)=\sum_{k=1}^{n}x(z)$, $k=1,2,\ldots,n$; $\alpha\in[0,1]$. Determine whether X' satisfies the following conditions.

$$\rho(k)=\frac{X(k)}{X'(k-1)}<0.5 \tag{6}$$

$$\delta(k)=\frac{X'(k)}{X'(k-1)}\in[1,1.5] \tag{7}$$

$$\sigma(k)=\frac{X(k)}{X(k-1)}\in\left(e^{-\frac{2}{n+1}},e^{\frac{2}{n+1}}\right) \tag{8}$$

Make further adjustment when X' does not satisfy these conditions. Otherwise, establish the differential coefficient equation.

$$\frac{dX'}{dt}+aX'=b \tag{9}$$

Where a is development and b is control grey number.

2.2.2. Model Solution:

$$\hat{a}=(B^TB)^{-1}B^Ty_N \tag{10}$$

$$\hat{x}'(k+1)=\left[x(1)-\frac{b}{a}\right]e^{-ak}+\frac{b}{a} \tag{11}$$

where $\hat{a}=\begin{bmatrix}a\\b\end{bmatrix}$, $y_N=[x(2),x(3),...,x(n)]^T$, $B=\begin{bmatrix}m*x'(1)+(1-m)*x'(2) & 1\\ m*x'(2)+(1-m)*x'(3) & 1\\ \vdots & \vdots\\ m*x'(n-1)+(1-m)*x'(n) & 1\end{bmatrix}$.

Determine the value of m when the sum of squares error between estimate value and true value is minimum.

$$SS=\sum_{k=1}^{n}(\varepsilon(k))^2 \tag{12}$$

Use the equation (13) to obtain the predictions.

$$\hat{x}(k+1) = \hat{x}'(k+1) - \hat{x}'(k) \tag{13}$$

3. Quasi-Steady-State Data Screening and Steady-State Data Extraction

3.1. The Quasi-Steady-State Data Screening Rules

(1) To eliminate the disturbance caused by noise to the fitting curve , the moving average filtering method is used for the data which have completed outliers detection and substation.

(2) Calculate the difference $\Delta x_{\max}$ between the maximum and minimum values and the standard deviation σ by the steady-state operating data determined as a matter of human experience.

(3) Calculate the formula $e(i) = |x(i) - x(i-1)|$, $(i = 2,3,...,n)$.

(4) Extract the data automatically when there are continuous ten points meet the condition that $e(i)$ is less than $\Delta x_{\max}$. Read the process data $x(0)$ in the history database when calculating the value $e(1)$.

3.2. The Steady-State Extraction Rules

3.2.1. Automatic Piecewise Curve Fitting:

(1) Set the minimum amount q of fitting points, and carry out the least squares polynomial fitting for the first piece of the first batch($i=1$， $c=1$) extracted quasi-steady-state data.

$$\hat{x}(t)^c{}_i = a^c{}_{i0} + a^c{}_{i1}t + a^c{}_{i2}t^2 + \cdots + a^c{}_{im}t^m \tag{14}$$

Where $\hat{x}(t)^c{}_i$ is the fitting value of the c piece among the i batch, m is the highest number of fitting function, q is the amount of initial matching points.

(2) Calculate the absolute error value $y_{it}{}^c$ (the t time of the c piece among the i batch) between the fitting value and the actual value.

$$y_{it}{}^c = \left|\hat{x}(t)^c{}_i - x(t)^c{}_i\right| \tag{15}$$

(3) Let $t_{i1}{}^c - t_{i0}{}^c = q-1$ when there is point which its $y_{it}{}^c$ value is greater than y_g . Where $t_{i1}{}^c$ and $t_{i0}{}^c$ stand for the end and start time of the c piece fitting function respectively. Let $q = q+1$ when there are no point which its $y_{it}{}^c$ value is greater than y_g . Continue the three-steps before until such points are found. The begin time of the next piece fitting curve is $t_{i0}{}^{c+1}$, where $t_{i0}{}^{c+1} = t_{i1}{}^c - p$, p is the quantity of overlapping data among the two pieces fitting curve before and after.

(4) Let $c = c+1$, repeat the above three steps to fit the i batch of data. As is shown in Figure 1:

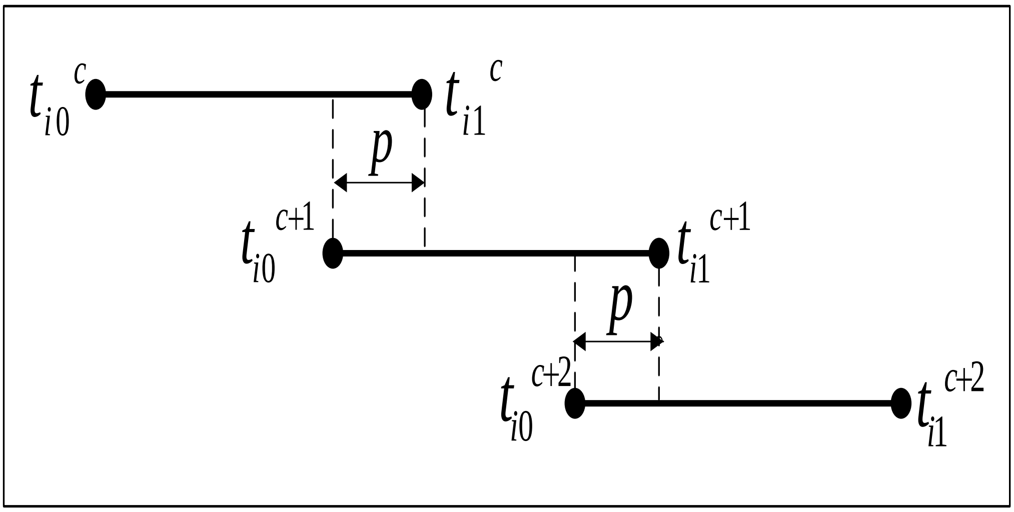

Figure 1. The Schematic Chart of Automatic Piecewise

(5) Let $i = i + 1$, repeat the above four steps to complete fitting all extracted data.

3.2.2. Fitting Curve Derivation:

(1) Weighting processing is needed for the overlapping curves to make the first order derivation continuous throughout the whole curve.

$$\hat{x}(t) = \begin{cases} \hat{x}(t)^{c}{}_{i}, & t_{i0}{}^{c} \le t \le t_{i0}{}^{c+1} \\ w_1\hat{x}(t)^{c}{}_{i} + w_2\hat{x}(t)^{c+1}{}_{i}, & t_{i0}{}^{c+1} \le t \le t_{i0}{}^{c+2} \end{cases} \tag{16}$$

Where $w_1 = \frac{1}{2} + \frac{1}{2}\sin(\frac{k-1}{p-1}\pi + \frac{\pi}{2})$, $w_2 = \frac{1}{2} - \frac{1}{2}\sin(\frac{k-1}{p-1}\pi + \frac{\pi}{2})$, $1 \le k \le p$.

(2) Calculate $s(t) = \left|\hat{x}'(t)\right|$.

(3)The piece of data is steady-state data when there are continuous ten points meeting the condition that $s(t)$ is less than σ.

4. Simulation Experiment

The function (x=1.2×exp(-0.213×t/6)+5.4×exp(-0.17×t/6)×sin(1.23×t/6)+3) is used to produce 200 data ranging from t=1 to t=200 in the simulation experiment. In order to verify the accuracy and reliability of the method the paper proposed, join the Gaussian white noise that the Signal to Noise Ratio is 40, and change some points as follows: x(11)=9.5 , x(43)=6 , x(120)=3.9 , x(161)=4.1 , x(179)=3.2 . The raw data distribution of simulation experiment is shown in Figure 2. The data distribution of the simulation function added noise and outliers is shown in Figure 3.

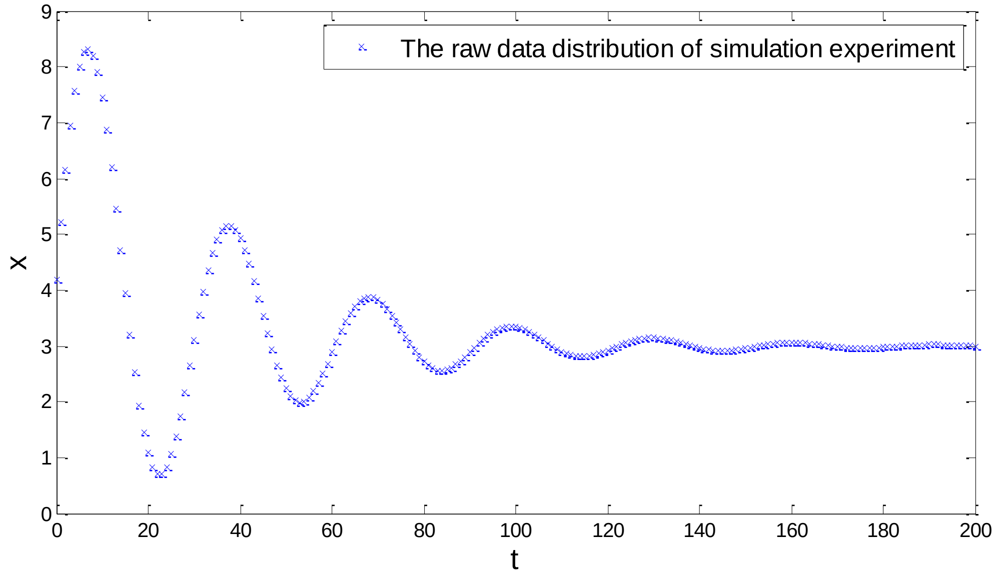

Figure 2. The Raw Data Distribution of Simulation Experiment

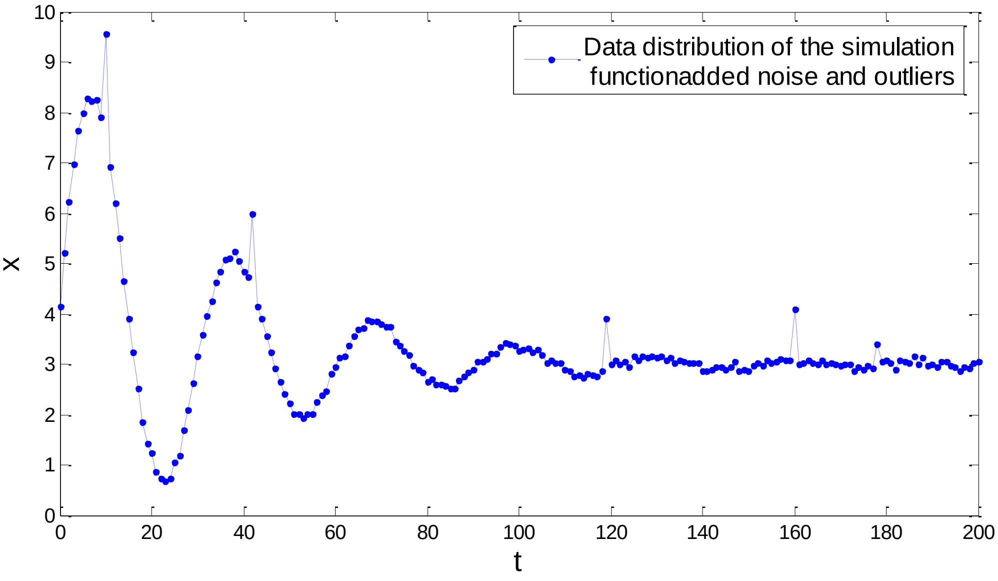

Figure 3. Data Distribution of the Simulation Function Added Noise and Outliers

(1) Use the sliding window technology, mobile step is 1, the width of sliding window is 5. Calculate the k values from $t = 3$ to $t = 198$. Read other process data in the history database when calculating other k value. The curve of k value is shown in Figure 4. In Figure 4, it is easily to find that these points $(11, 9.55)$, $(43, 6)$, $(120, 3.9)$, $(161, 4.1)$, $(179, 3.4)$ are outliers.

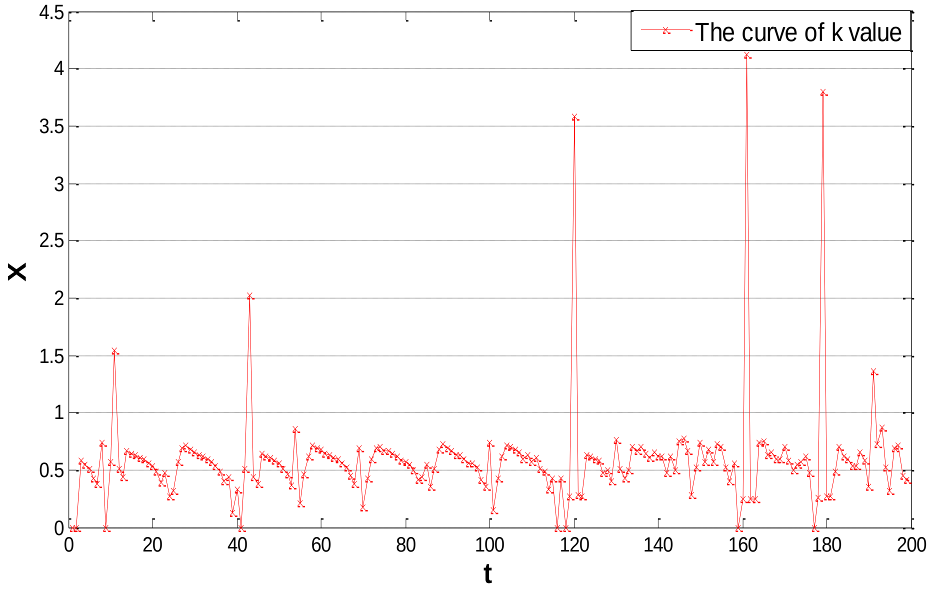

Figure 4. The Curve of k Value

(2) The simulation curve changes so complex that the value of n should not be too large, otherwise the accuracy of prediction will be affected. Where take $n=5$. Use the data sequence $X=\{x(6),x(7),x(8),x(9),x(10)\}$ to predict the point $\mathrm{x}(11)$. Select different m values to conduct experiments. The error sum of squares curve is obtained under different values of m , and is shown in Figure 5.

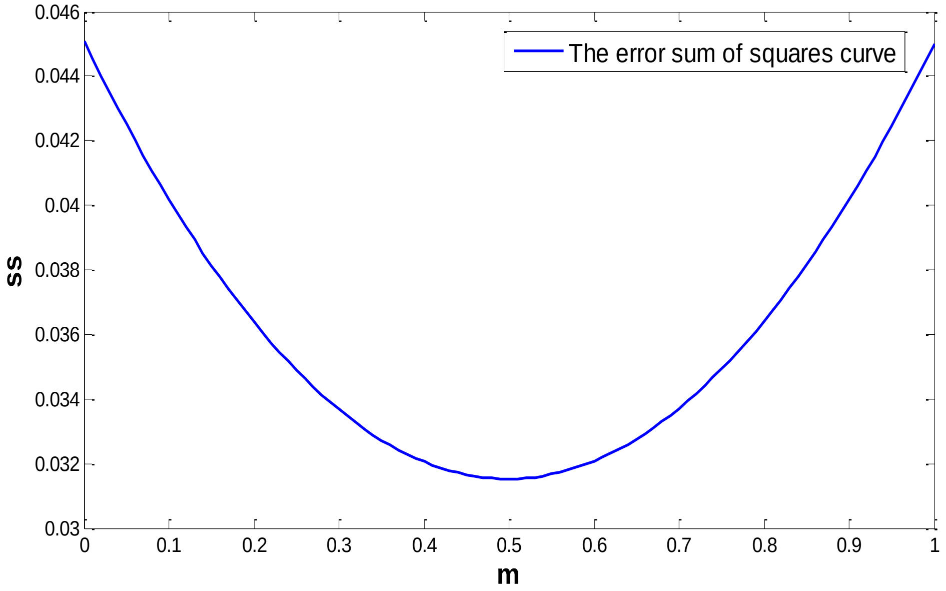

Figure 5. The Error Sum of Squares Curve

In Figure 5, the error sum of squares curve is minimum when the value of m is 0.5. Other outliers will be replaced by using the similar steps. The results are shown in the following Table 1.

Table 1. The Comparison of Two Methods for Outliers Forecasting

	X(11)	X(43)	X(120)	X(161)	X(179)
Ture value	7.45	4.47	2.89	3.06	2.97
Means method	7.39	4.40	2.91	3.09	2.94
GM(1,1)	7.61	4.51	2.90	3.06	2.98
Relative error of means method/%	0.86	1.55	0.59	0.91	0.90
Relative error of GM(1,1)/%	2.10	0.82	0.07	0.02	0.43

In Table 1, we can see that the sum of relative error using GM (1, 1) method is smaller than means method. The method of GM (1, 1) forecasts the outliers according to the development law of data.It can predict well in the presence of noise.

(3) Replace outliers with predictive values, then use the moving average filter which template is 1×5 to deal with the data. The result is shown in Figure 6.

(4) Calculate the difference $\Delta x_{\max}$ ($\Delta x_{\max}=0.16$)between the maximum and minimum values and the standard deviation σ ($\sigma=0.046$) by the steady-state operating data determined according to human experience. Calculate the value of $e(i)$, as is shown in Figure 7.

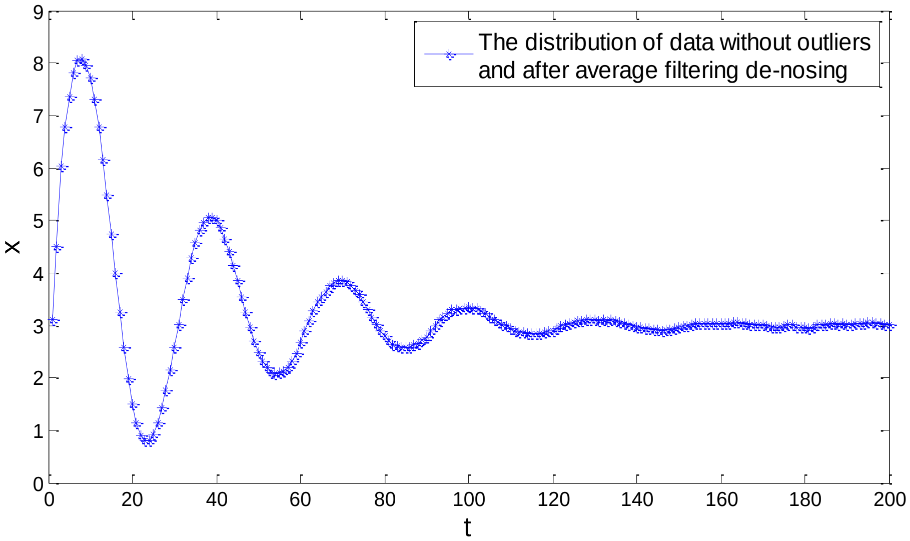

Figure 6. The Distribution of Data Without Outliers and After Average Filtering De-Nosing

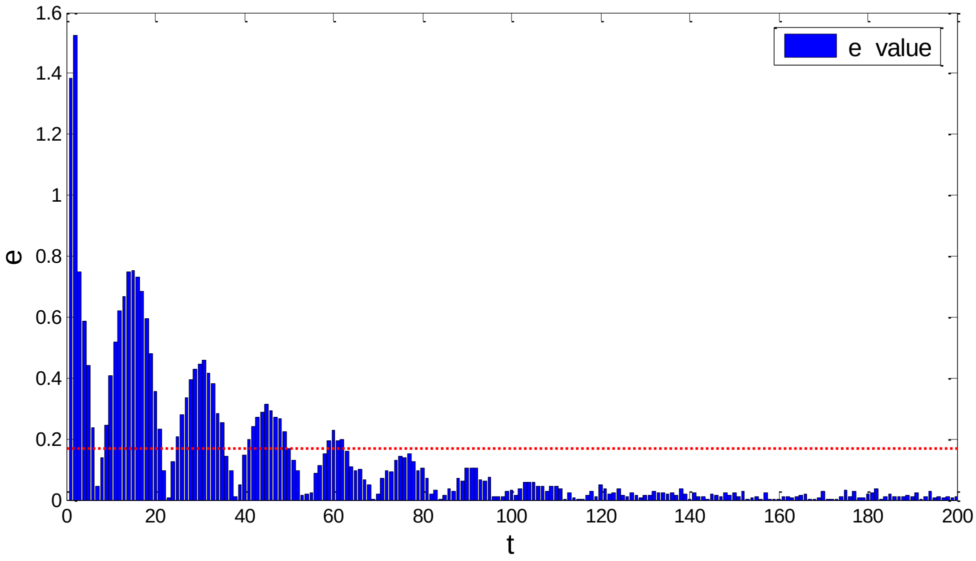

Figure 7. The Curve of e Value

In Figure 7, we can see the value of e is smaller than $\Delta x_{\max}$ after t ($t=63$). There are no continuous ten points meeting the condition that $e(i)$ is less than $\Delta x_{\max}$ in other time. Extract the data after t ($t=63$) as quasi-steady-state data.

(5) Considering the requirements of fitting accuracy and robustness , let $i=1, q=10, m=3, p=5, y_g=0.05$. Carry out automatic piecewise curve fitting, the result is shown in Equation (17), and the fitting curve is shown in Figure 8.

$$x(t)=\begin{cases} 3.69-0.4859\times((t-72.5)/5.916)-0.3061\times((t-72.5)/5.916)^2+0.1102\times((t-72.5)/5.916)^3 & (63\le t\le 77) \\ (0.5+0.5\times sin((t-78)\times\pi/4+\pi/2))\times x(t)_1+(0.5-0.5\times sin((t-78)\times\pi/4+\pi/2))\times x(t)_3 & (78\le t\le 82) \\ 2.894+0.6014\times((t-91)/7.9376)+0.08118\times((t-91)/7.9376)^2-0.2078\times((t-91)/7.9376)^3 & (83\le t\le 99) \\ (0.5+0.5\times sin((t-99)\times\pi/4+\pi/2))\times x(t)_3+(0.5-0.5\times sin((t-99)\times\pi/4+\pi/2))\times x(t)_5 & (100\le t\le 104) \\ 2.885-0.1512\times((t-112.5)/7.649)+0.1378\times((t-112.5)/7.649)^2+0.02282\times((t-112.5)/7.649)^3 & (105\le t\le 120) \\ (0.5+0.5\times sin((t-120)\times\pi/4+\pi/2))\times x(t)_5+(0.5-0.5\times sin((t-120)\times\pi/4+\pi/2))\times x(t)_7 & (121\le t\le 125) \\ 3.015-0.1637\times((t-138)/10.25)-0.00268\times((t-138)/10.25)^2+0.07388\times((t-138)/10.25)^3 & (126\le t\le 150) \\ (0.5+0.5\times sin((t-150)\times\pi/4+\pi/2))\times x(t)_7+(0.5-0.5\times sin((t-150)\times\pi/4+\pi/2))\times x(t)_9 & (151\le t\le 155) \\ 3.035-0.033392\times((t-163.5)/7.649)-0.01914\times((t-163.5)/7.649)^2+0.01275\times((t-163.5)/7.649)^3 & (156\le t\le 171) \\ (0.5+0.5\times sin((t-171)\times\pi/4+\pi/2))\times x(t)_9+(0.5-0.5\times sin((t-171)\times\pi/4+\pi/2))\times x(t)_{11} & (172\le t\le 176) \\ 3.011+(0.04595\times((t-186)/8.515))-(0.002324\times((t-186)/8.515)^2)-(0.01238\times((t-186)/8.515)^3) & (177\le t\le 200) \end{cases} \tag{17}$$

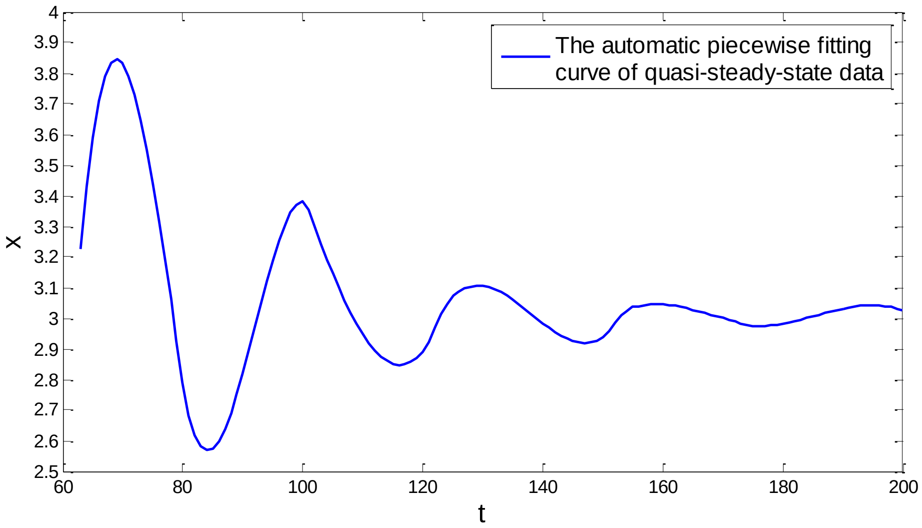

Figure 8. The Automatic Piecewise Fitting Curve of Quasi-Steady-State Data

(6) Calculate the value of $s(t)$ and compare the value of $s(t)$ with σ. The data after t ($t = 105$) are steady-state data according to the rules mentioned above, and are shown in Figure 9.

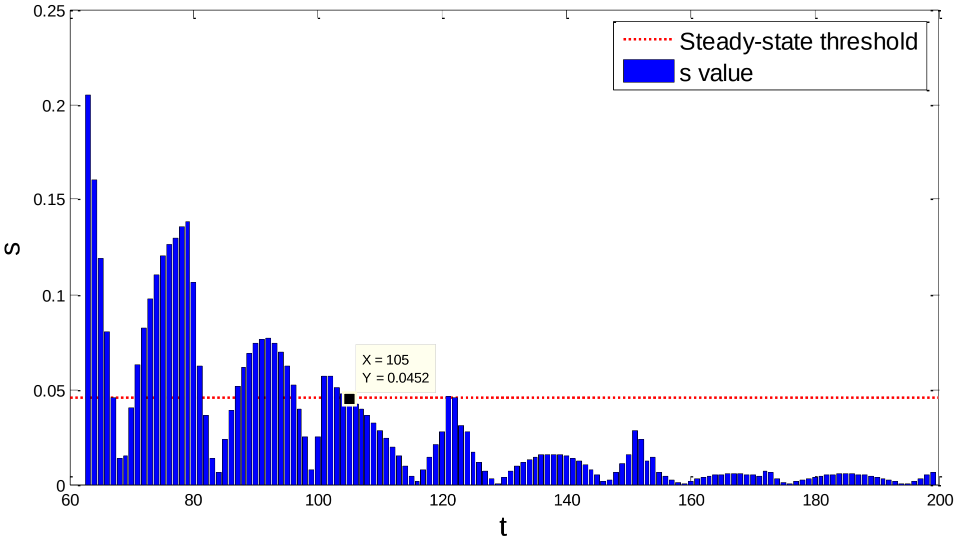

Figure 9. The Curve of s Value

5. Conclusion

The steady-state extraction rules and automatic piecewise curve fitting algorithm are put forward on the basis of introducing outliers detection and substation based on the improved local deviation and grey prediction theory. The flow of extracting method of steady-state data based on automatic piecewise curve fitting containing outliers detection is given. The simulation experiment shows that the method not only can eliminate the influence of outliers for steady-state detection, but also extract the steady-state data in line with human experience effectively and conveniently.

References

[1] J. Liu, M. Gao, Y. Lv and T. Yang, “Overview on the steady-state detection methods of process operating data”, Chinese Journal of Scientific Instrument, vol. 34, **(2013)**, pp. 1739-1748.

[2] L. E. Torfason and G. A. Wood, “Mechanism modeling”, US. Patent 4,062,130, **(1977)**.

[3] C.f. Li, B. Z. Chen, X. R. He, T. Qiu and S. Y. Hu, “Improved filtering method for steady state tests with measurements containing gross errors”, Journal of Tsinghua University, vol. 44, **(2004)**, pp. 1160-1162.

[4] S. Narasimhan, R. S. H. Mah, A. C. Tamhane, J. W. Woodward and J. C. Hale, “A composite statistical test for detecting changes of steady states”, Aiche Journal, vol. 32, **(1986)**, pp. 1409-1418.

[5] M. Gao, J. Liu, R. Wang, H. Zhang and X. Zhang, “Steady-state detection of power plant history data based on adaptive gauss filter”, Journal of Chinese Society of Power Engineering, vol. 9, **(2014)**, pp.708-713.

[6] K. Pahnchawatt, R. Lipikorn and S. Keeratipibul, Listeria monocytogenes Colony Counting from Microscopic Images Using Nonlinear Piecewise Least-Square Curve-Fitting Filter. International Conference on Information Science & Applications (ICISA), IEEE, **(2013)**, June 24-26; Suwon.

[7] J. D. Kelly and J. D. Hedengren, “A steady-state detection (ssd) algorithm to detect non-stationary drifts in processes”, Journal of Process Control, vol. 23, **(2013)**, pp. 326-331.

[8] L. J. Chen, Y. H. Zhang and R. R. Hai, “A new steady-state detection method for fusing outliers”, Control & Instruments in Chemical Industry, vol. 5, **(2013)**, pp. 582-586.

[9] Y. Lv, J. Liu, W. Zhao and T. Yang, “Steady-state detecting method based on piecewise curve fitting”, Chinese Journal of Scientific Instrument, vol. 33, no. 1, **(2012)**, pp. 194-200.

[10] W. C. Chen and F. Liu, “An improved steady state identification method based on polynomial filtering”, Control Engineering of China, vol. 2, **(2012)**, pp. 195-202.

[11] Y. Liu, H. Zhang, F. Han and J. Tan, “An efficient rfid data cleaning method based on wavelet density estimation”, Journal of Digital Information Management,vol. 13, **(2015)**, pp. 10-14.

[12] K. Zhang, M. Hutter and H. Jin, “A New Local Distance-Based Outlier Detection Approach for Scattered Real-World Data”. Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining; Bangkok, Thailand, **(2009)**, April 27-30.

[13] S. Zhou and W. Xu, “Deviation-based local outlier detection algorithm”, Chinese Journal of Scientific Instrument, vol. 35, **(2014)**, pp. 2293-2298.

[14] W. Y. Qian, Y. G. Dang and S. F. Liu, “Study on Grey Prediction Model gGM(1,1) and Its Application Based on Amplitude Compression Transformation”, Proceedings of 2013 IEEE International Conference on Grey systems and Intelligent Services (GSIS), Macao, **(2013)** November 15-17.

[15] H. L. Yang, J. X. Liu and B. Zheng, “Improvement and application of grey prediction gm(1,1) model”, Mathematics in Practice & Theory, vol. 23, **(2011)**, pp. 39-46.