Piecewise Linear Approximation of Time Series on the Base of the Weierstrass-Mandelbrot Function

E.A.Pitukhin and L.V.Shchegoleva

Petrozavodsk State University, Petrozavodsk, Russia
eugene@petrsu.ru, schegoleva@petrsu.ru
https://petrsu.ru

Abstract. The paper is devoted to the construction of the algorithm of piecewise linear approximation of time series, the distinctive feature of which is the preservation of sharp "peaks" and data outliers. As a basis, the iterative algorithm of piecewise linear approximation proposed by E. K. Bely in 1994 was taken. This algorithm was used to describe the experimental medical data on respiratory function of the lungs. The algorithm was modified to significantly increase the precision and to reduce the number of iterations. The rule of determining the optimal number of time series splitting by the minimum of the adjusted coefficient of determination was obtained. A new criterion for the algorithm stopping was proposed. As a testing data, the two-factor function of Weierstrass-Mandelbrot was used, which allows to generate data in a wide range of shapes and variability. In numerical experiment, the control parameters of the function were set by random variables with a uniform distribution law. The estimations of probability densities of the output parameters of the algorithm were obtained by the Monte Carlo method. The convergence of the approximation algorithm was studied and the regions of shape and variability parameters, at which the algorithm converges, were revealed.

Keywords: Time series modelling · Algorithm of piecewise linear approximation · Weierstrass-Mandelbrot function.

1 Introduction

The main approaches for time series modelling use smoothing methods [3]. In some cases, on the contrary, it is necessary to keep the "peak" values of time series in the model. For example, when modelling some medical processes: heartbeat, breathing, tremor and others. Then the problem of constructing a time series model in the form of a piecewise linear function arises.

In [1]-[2] the algorithm which allowed to build such functions was offered. The main idea of the algorithm was to construct some partition of the interval of the time series $[t_s, t_f]$ into n parts, where n was specified by researcher. Let

$\{t_0, t_2, \ldots, t_{n-1}\}$ is the initial partition of the interval $[t_s, t_f]$, and $t_0 = t_s$, $t_{n-1} = t_f$. Then, in the loop for each interval $[t_i, t_{i+1}]$, a linear function approximating the values of the time series on the interval is constructed, and then for each pair of intervals $[t_i, t_{i+1}]$ and $[t_{i+1}, t_{i+2}]$, the point of intersection of the constructed lines is calculated. If the abscissa of the intersection point is inside the interval $[t_i, t_{i+2}]$, it replaces the split point t_{i+1} in the partition. The loop stops when all distances between the intersection points and the split points do not exceed the specified value.

The main drawback of this algorithm is the long convergence (about a thousand iterations) [2]. In addition, it is necessary to specify the number of split points and the value of maximum distance to stop the algorithm. These problems were not investigated by the author of [1]-[2].

Therefore, in this paper the following modification of the algorithm is proposed:

1. The method of least squares is used to evaluate values of parameters of the linear functions.
2. The new split point is determined only if it falls within the interval between the point under consideration with the index i and the point with the index $i+2$ ($[t_i + \Delta t, t_{i+2} - \Delta t]$), where Δt is set so that each interval ($[t_i, t_{i+1}]$ and $[t_{i+1}, t_{i+2}]$) contains at least one point of the original time series.
3. The new conditions of loop stopping are based on three metrics of quality of the approximation.

The algorithm for determining the number of partition intervals is also proposed.

2 Algorithm of time series approximation

As a testing data, the two-factor function of Weierstrass-Mandelbrot [4] was used, which allows generating data in a wide range of shapes and variability.

The Weierstrass-Mandelbrot (W-M) function is continuous everywhere but differentiable nowhere:

$$W(t) = \sum_{m=-\infty}^{\infty} \frac{(1 - e^{ib^m t}) \cdot e^{i\phi_m}}{b^{(2-D)m}}. \tag{1}$$

It has two parameters: $b > 1$, $1 < D < 2$, where D – fractal dimension. The time series obtained by the W-M function are non-stationary and have fractal properties [5]. Its sharpened shape makes it suitable for testing the algorithm instead of real time series.

Let $\phi_m = 0$. The real part of W-M function $\Re W(t)$ (the cosine fractal function) is:

$$C(t) = \Re W(t) = \sum_{m=-\infty}^{\infty} \frac{1 - \cos(b^m t)}{b^{(2-D)m}}. \tag{2}$$

Approximation of $\Re W(t)$ till the $2N$-th term of the series is:

$$\overline{C}(t) = \sum_{m=-N}^{N} \frac{1 - \cos\left(b^m t\right)}{b^{(2-D)m}}. \tag{3}$$

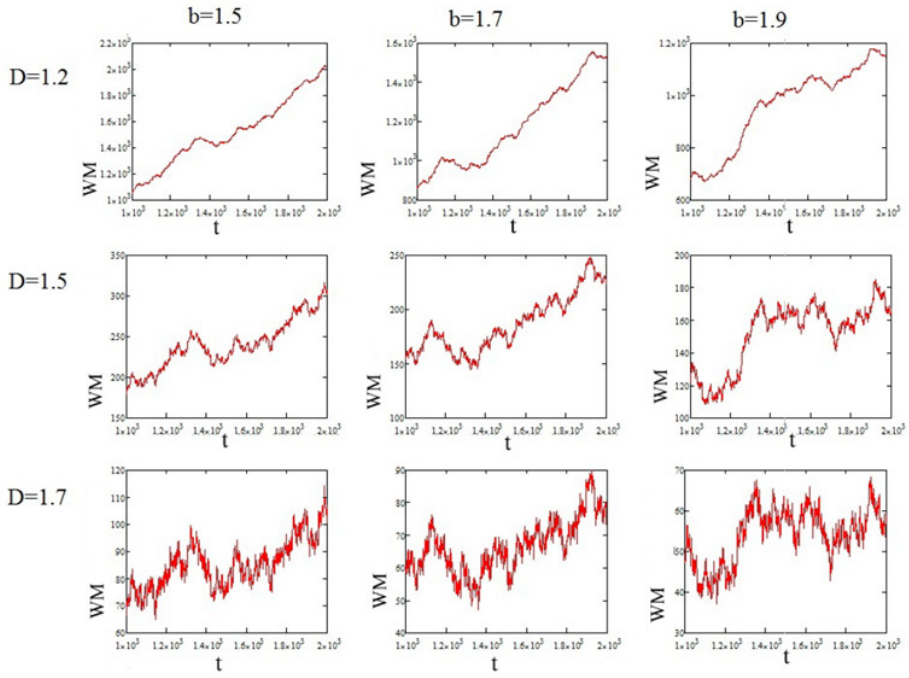

Fig. 1. The Weierstrass-Mandelbrot function with different values of parameters b and D

It can be seen from Fig. 1, how the W-M function changes to different values of the b and D parameters. To simulate real time series, the ranges of parameters b and D were determined as $b \in [1.3, 1.9]$ and $D \in [1.1, 1.8]$ so that the W-M function simulation algorithm converges. A sufficient number of series terms in the simulation of the W-M function equal to 30 ($N = 30$) is also determined.

As a metrics of quality of the approximation, the adjusted coefficient of determination (R^2_{adj}), the mean absolute percentage error (MAPE), root-mean-square error (RMSE), MAPE for centered data, R^2_{adj} for centered data, Akaike information criterion (AIC), and difference between the points (DBP) were used. With the exception of DBP, all other metrics are traditional metrics in statistics. The difference between the points is calculated as follows: $DBP = \frac{1}{n}\sum_{i=1}^{n} |t_i^{k+1} - t_i^k|$, t_i^k is the partition point in the k-th iteration.

Numerical experiments showed some regularities in the change of metric values, which allowed formulating the conditions for stopping of the algorithm. From Tabl. 1, you can see that the DBP statistics become constant, and the MAPE statistics become cyclical with the length of the cycle 2.

Number of iteration	Difference between the points	MAPE	MAPE for centered data	RMSE	R^2_{adj}	R^2_{adj} for centered data	AIC
0	0.70	0.00933	0.19403	18.56172	0.99937	0.99225	2.96023
1	3.00	0.00365	0.15335	13.13462	0.99937	0.99243	2.93975
2	2.95	0.00349	0.15227	16.35711	0.99933	0.99293	2.86395
3	2.35	0.00752	0.14747	13.97332	0.99990	0.99416	2.67669
4	2.05	0.00743	0.14337	12.75732	0.99991	0.99467	2.53523
5	1.40	0.00737	0.14325	12.44206	0.99991	0.99430	2.56026
5	0.35	0.00737	0.14313	12.33232	0.99991	0.99483	2.55545
7	0.75	0.00734	0.14307	12.29559	0.99991	0.99486	2.54842
8	0.40	0.00734	0.14290	12.23998	0.99991	0.99487	2.54796
9	0.40	0.00734	0.14299	12.27526	0.99991	0.99487	2.54676
10	0.30	0.00733	0.14236	12.27616	0.99991	0.99487	2.54634
11	0.30	0.00734	0.14303	12.23003	0.99991	0.99487	2.54715
12	0.30	0.00733	0.14236	12.27616	0.99991	0.99487	2.54634
13	0.30	0.00734	0.14303	12.23003	0.99991	0.99487	2.54715
14	0.30	0.00733	0.14236	12.27616	0.99991	0.99487	2.54634

Table 1. The quality of the approximation

This property is satisfied for different values of the shape (b) and the number of intervals of the partition (n) (see Fig. 2).

As a result, to determine the optimal number of iteration the stopping conditions of the algorithm can be formulated (Condition-K):

1. R^2_{adj} has stabilized or looped;
2. Difference between the points (DBP) has stabilized or looped;
3. MAPE has stabilized or looped.

And the new algorithm of approximation time series is:

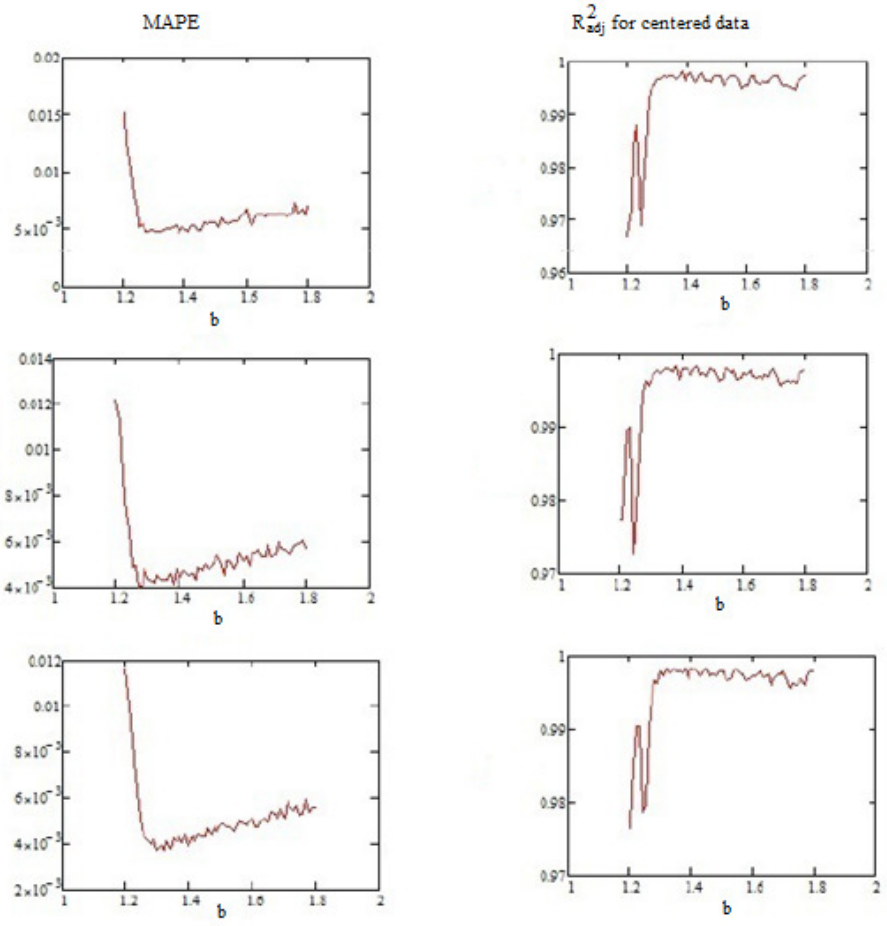

Fig. 2. The approximation metrics for different form values of parameter b and the number of partition intervals n

Algorithm for estimation of number of iterations

Set the time interval $t \in [t_s, t_f]$
Set n – number of intervals
Set $K = 1$ – number of iterations of algorithm
Set the initial partition of the interval:
$t_s = t_0 < t_1 < \ldots < t_{n-1} = t_f$
While Condition-K
  For every $i \in [1, n]$:
    Estimate coefficients of line regressions a_i and b_i for interval $[t_i, t_{i+1}]$ by Ordinary Least Squares method
    Calculate t^*_{i+1} the point of intersection of lines $y = a_i + b_i t$ and $y = a_{i+1} + b_{i+1} t$
    If $t^*_{i+1} \in (t_i + \Delta t, t_{i+2} - \Delta t)$ then $t_{i+1} = t^*_{i+1}$, otherwise $t_{i+1} = t_{i+1}$
  $K = K + 1$

3 Problem of optimal number of partition intervals

The next problem is to determine optimal number of partition intervals n. To solve the problem the second algorithm is proposed. It uses Monte Carlo method.

Algorithm for estimation of number of partition intervals

For every $j \in [1, J]$:
  Get a uniformly distributed value of b_j on the interval $[1.3, 1.9]$
  Get a uniformly distributed value of D_j on the interval $[1.1, 1.8]$
  Simulate W-M function with parameters b_j and D_j
  Set the time interval $t \in [t_s, t_f]$
  Set $n = n_{min}$ – initial number of intervals
  Set $K_j = 30$ – number of iterations of algorithm
  Set the initial partition of the interval: $t_s = t_0 < t_1 < \ldots < t_{n-1} = t_f$
  While Condition-n
    For every $i \in [1, n]$:
      Estimate coefficients of line regressions a_i and b_i for interval $[t_i, t_{i+1}]$ by Ordinary Least Squares method
      Calculate t^*_{i+1} the point of intersection of lines $y = a_i + b_i t$ and $y = a_{i+1} + b_{i+1} t$
      If $t^*_{i+1} \in (t_i + \Delta t, t_{i+2} - \Delta t)$ then $t_{i+1} = t^*_{i+1}$, otherwise $t_{i+1} = t_{i+1}$
    $n = n + 1$
  Calculate MAPE and R^2_{adj} statistics
Estimate average value of n, MAPE, R^2_{adj}

In this algorithm the stopping conditions to determine number of partition intervals (Condition-n) are below (also see Fig. 3):

1. The maximum number of intervals is reached (n_{max});
2. $sign[R^2_{adj}(n) - R^2_{adj}(n-1)] \cdot sign[R^2_{adj}(n-1) - R^2_{adj}(n-2)] = -1$;
3. $R^2_{adj}(n)$ has maximized to critical value;

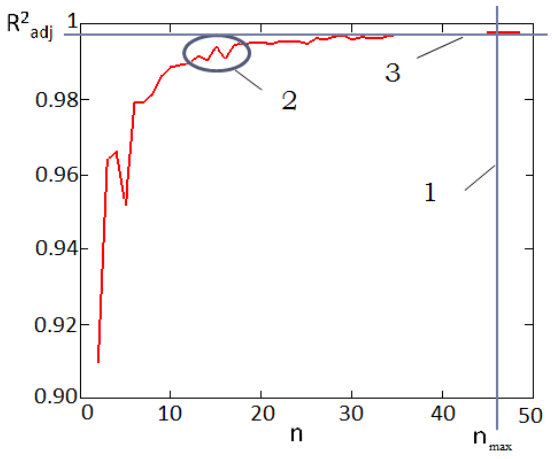

Fig. 3. Visualization of stopping conditions to determine number of partition intervals (Condition-n)

4. MAPE(n) has minimized to critical value.

   In the algorithm J is a number of experiments.

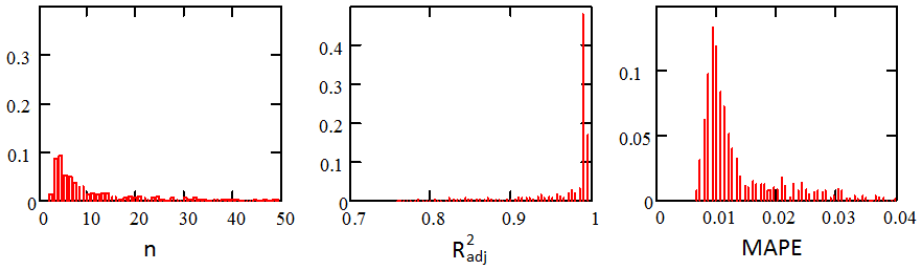

Fig. 4. The output metrics of the algorithm for estimation of number of partition intervals

On the Fig. 4 histograms of the output metrics of the algorithm are presented. The estimations of parameters: $\overline{n} = 25$; $\overline{R^2_{adj}} = 0.968$; $\overline{MAPE} = 0.013$ show that the convergence of the algorithm is quite acceptable.

4 Conclusion

The presented algorithms make it possible to obtain an approximation of the time series that preserves its "peaks". The first algorithm is designed to de-

termine the optimal number of iteration. The second algorithm is designed to estimate number of partition intervals. Both algorithms were tested for the Weierstrass-Mandelbrot function. For each algorithm, the stopping conditions were formulated on the basis of time series approximation quality metrics.

As a result of numerical experiments, estimates of the following parameters were obtained.

1. The range of changes in the input parameters of the Weierstrass-Mandelbrot function was determined: b=[1.3, 1.9], D=[1.1, 1.8].
2. The sufficient number of members of the series in the simulation of the Weierstrass-Mandelbrot function was determined: N=30.
3. The sufficient number of iterations for the convergence of the algorithm was determined: K = 30.

References

1. Bely, E.: About one method of smoothing the experimental curves (in Russian). Works of Petrozavodsk State University. Applied mathematics and computer science **3**, 8–12 (1994)
2. Bely, E.: Signal restoration in radio diagnostic studies (in Russian). Works of Petrozavodsk State University. Applied mathematics and computer science **6**, 51–58 (1997)
3. Box, G.E.P., Jenkins, G.M., Reinsel, G.: Time Series Analysis: Forecasting and Control. Prentice Hall, San Francisco (1994)
4. Edgar, G.: Classics On Fractals. Westview Press, Boulder, CO (2004)
5. Erkuş, S.: Q-periodicity, self-similarity and Weierstrass-Mandelbrot function. Ph.D. thesis, Izmir (2012)