JEL Classification: C14, C60, G14

PREPROCESSING OF INPUT DATA FOR THE STOCK MARKET MONITORING SYSTEM

Oleksandr V. PISKUN
Candidate of Technical Sciences, Associate Professor of the Department of Higher Mathematics and Informational Technologies, Cherkasy Institute of Banking of the University of Banking of the National Bank of Ukraine (Kyiv)
E-mail: piskun@ukr.net

Summary. In the article the nonparametric smoothing methods of time series are considered. Optimum preprocessing methods of data series for system of monitoring of the stock markets are determined.

Key worlds: *stock market, monitoring, nonparametric smoothing methods.*

Research Problem Formulation. Modern financial markets are characterized by considerable complexity of the processes occurring in them. There is a globalization of international markets; currency, interest rate, rate of securities, staple prices volatility are increasing, and as a result, financial markets have become more volatile, complex and risky. An instrument of monitoring system of a stock market is needed in order to effectively control its state. One of the newest methods for studying time series is recurrence quantification analysis (RQA). Previous study showed the ability of measure laminarity (LAM) of RQA to reveal different periods of financial market functioning and to analyze crisis events on them [1, 2].

Financial market is highly susceptible to turbulence, and therefore the dynamics of the corresponding recurrence measures will contain the stochastic component. When building the monitoring system it is necessary to provide smoothing of LAM in order to automate the detection of critical points of transitions between periods of market functioning.

Recent Research and Publications Analysis. Methods of processing and time series analysis are studied in the works of famous Ukrainian and foreign scholars and experts, including: John Yule, M. Kendall, A. Sewart, W. Hartley, J. Pollard, S. Ayvazian, G. Kildishev and others. However, the choice of strategies and methods for pre-processing and analysis of time series depends on the researcher's ultimate goal.

Purpose of this paper is to review, analyse and determine the optimal method of pre-processing (smoothing) data series for monitoring system of the stock markets.

Justification of Scientific Results. For long series, it is usually impossible to specify an appropriate parametric curve to smooth the series for its entire length. In this case, one uses a variety of nonparametric methods for the analysis of time series, such as moving average smoothing, frequency filtering, etc. [3].

Let us analyse the most common methods of nonparametric smoothing of time series.

Moving Average. While smoothing with this method the actual values of a dynamic series are replaced with mean values, that characterize the midpoint moving period [4].

Simple smoothing is based on the creation of a new series of simple arithmetic mean, calculated for time periods with the length k:

$$\bar{y}(t) = \frac{1}{k} \sum_{t=i}^{i+k-1} y(t),$$

where k – is the length of the smoothing period. It depends on the nature of the time series, as well as the purpose of smoothing and is selected by a researcher,

i – is a serial number of a mean,

n – is the length of the series.

Average weighted smoothing is about determining the weighted averages for different points of the data series. The basis of the method is the idea of local polynomial approximation trend of low degree.

The values of trend at point *tj* are being approximated on the levels of a series with time interval $[t_j - k, t_j + k]$ by a polynomial of a given order *l*:

$$\bar{y}(t) = \sum_{i=0}^{l} a_i t_j^i,$$

parameters of which are calculated by the method of least squares using equations of the following type:

$$a_0 \sum_{i=-k}^{k} t_j^i + a_1 \sum_{i=-k}^{k} t_j^{i+1} + \cdots + a_l \sum_{i=-k}^{k} t_j^{i+k} = \sum_{i=-k}^{k} y_i t_j^i.$$

Solving the received equations to a_i, we obtain a sequence of weights that depend only on the width of the interval (2*k*+1) and the order of the polynomial *l*. The calculation of the evaluation value of the trend at point *t* is equivalent to creating a weighted sum of the series values in the interval $[t_j - k, t_j + k]$.

The value of weights for different *k* and *l* are defined and presented in the tables [4, 5]. For polynomials of zero and first order weights a_i are equal, that leads this method to a simple smoothing.

According to the presented structure of the local smoothing, first and last *k* elements of the input series are left unsmoothed. For smoothing the last segments of the series appropriate formulas were developed [4].

The Method of Local Regression (generalized representation of the Methods of Weighted Moving Average). Smoothed value $\bar{y}_j$, corresponding to the middle segment of the local smoothing [*tj* - *k*, *tj* + *k*] is calculated using the vector

$$\hat{A} = (K^T K)^{-1} K^T Y_{t_j - k}^{t_j + k},$$

where *K* is matrix of size (2*k* + 1) (*l* + 1), the elements of which are calculated according to the formula

$$k_{ij} = (i - k - 1)^{j-1}, i = 1,2,\ldots,2k-1;\ \ j = 1,2,\ldots,l+1,$$

The value a_0 is taken in the meaning of the trend value in the point t_j. Other components of the vector of polynomial coefficients are used in calculating the extreme points of the trend, as the presented in the above form algorithm does not allow to estimate the trend of the first *k* and last *k* points of time series. For the extreme values calculation, the vectors of coefficients of the polynomial are used. They are calculated for the points numbered *k* + 1 and *n* - *k* by the formulas:

$$\bar{y}_j = \sum_{i}^{l} t_{i+1}^{(k+1)} (j - k - 1)^i, j \le k,$$

$$\bar{y}_j = \sum_{i}^{l} t_{i+1}^{(n-k)} (j + k - n)^i, j > n - k.$$

In addition, for smoothing the first and the last segments one can apply algorithms such as recursive least squares method.

The Method of Locally Weighted Regression (LOWESS or LOESS) implies that each interval of a smoothed signal is found as value of a function of locally weighted regression, which is estimated in the corresponding neighborhood [6]. Local weights are calculated by the formula

$$w_i = \left(1 - \left|\frac{t_k - t_i}{d(t_k)}\right|^3\right)^3,$$

where t_k – is a point, in which the regression function is calculated,

t_i – is the point for which weight function is calculated,

d(t) – is a half of the interval width on which regression is calculated.

The most commonly used is linear (Lowess) or quadratic (Loess) regression function; much less one uses polynomial functions of higher order.

Parameter of the above methods of smoothing, which determines the quality and characteristics of the smoothed signal is width of the window, i.e. the number of signal intervals in a corresponding neighborhood, which are used for smoothed values calculation. In general, the larger the width of the window is, the more high-frequency components and features of the signal are eliminated as a result of smoothing.

Smoothing Splines. Let the interval [*a, b*] be divided into (*k* +1) by the points $a_1 < \ldots < a_k$. Spline or piecewise polynomial function of degree m with k connections at the points $a_1, \ldots, a_k$ is a function *S(t)*, which in each interval $(a_j, a_{j+1}), j = 0, \ldots, k$ is described by algebraic polynomial $P_j(t)$ of the degree *m*. Coefficients of polynomials $P_j(t)$ are consistent with each other so that the conditions of continuity of *S(t)* and its (*m*-1) derivatives at the nodes connections were seen [7].

The best way of approaching splines is approximation of equidistant nodes; the most commonly used are the third-degree splines (cubic splines) [8].

For a cubic spline function *S(t)* on each interval $[t_{i-1}, t_i]$ is a polynomial of the third degree $S_i(t)$, coefficients of which should be defined:

$$S_i = a_i + b_i(t - t_i) + c_i(t - t_i)^2 + d_i(t - t_i)^3$$

Terms of continuity of all derivatives up to the second order are written as

$$S_i(t_{i-1}) = S_i(t_{i-1})$$
$$S_i'(t_{i-1}) = S_{i-1}'(t_{i-1})$$
$$S_i''(t_{i-1}) = S_{i-1}''(t_{i-1})$$

Smoothing cubic spline is defined as a spline that minimizes the following functional, which also depends on a parameter p:

$$I(s,p) = p\sum_{k=1}^{n} w_k\,(y_k - s(t_k))^2 + (1-p)\int_{t_1}^{t_n}\left(\frac{d^2s(t)}{dt^2}\right)^2 dt,$$

where (t_k, y_k), k=1,2,...,n – is the approaching data;

w_k – is weights of data;

p – is smoothing parameter that varies from 0 to 1, which determines the curvature of the spline.

Weights w_k are usually chosen as the generalized quadrature formula weight trapezoids. In the case of equidistant points with step h:

$$w_1 = w_n = \frac{h}{2}, \quad w_k = h, \quad k = 2,3,\dots,n-1.$$

If you set the value of the smoothing parameter p close to zero, the smoothing spline looks like a straight line that approximates the data by the method of least squares as the second term becomes the main in the functional that is minimizing

$$(1-p)\int_{t_1}^{t_n}\left(\frac{d^2s(t)}{dt^2}\right)^2 dt,$$

it is the second term that is just responsible for the smoothing; its minimization corresponds to the construction of the spline with the smallest value of the second derivative (zero for first-order polynomial). Conversely, if the value of the smoothing parameter is close to 1, the first term becomes main in the functional that is minimizing

$$p\sum_{k=1}^{n} w_k\,(y_k - s(t_k))^2,$$

It is the first term that is responsible for spline passing through the given points. When p = 1 smoothing spline converges to a regular cubic spline.

We have analyzed the methods of local approximation. There is another approach, which can be called as non-local smoothing methods. In this case, the adjusted series is determined by all the output values of the series. If the value is changed or new terms are added, the components are also changing. One of these methods is the wavelet transformation.

Wavelets are special functions in the form of short waves (bursts) with zero integral value and are localized along the axis of the independent variable, capable of displacement along this axis and scaling (stretching/compression). Any of the commonly used types of wavelets generates a complete orthogonal system of functions. In the case of wavelet analysis (decomposition) of a process (signal) due to changes in scale, wavelets are able to detect differences in the characteristics of the process at different scales, and with the shear one can analyze the process features at various points throughout the tested interval. It is due to the properties of the completeness of the system, one can make restoration (reconstruction or synthesis) of a process by reverse wavelet transform [9].

Wavelet transformation of one-dimensional signal which is financial time series, is its expansion in the system of the basic functions

$$\psi_{ab}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right),$$

is constructed from the parent (original) wavelet $\psi(t)$, which has certain properties, due to the time shift operations (b) and the time scale change (a). For given values of the parameters a and b function $\psi_{ab}(t)$ and the wavelet is generated by a parent wavelet $\psi(t)$.

Basic function $\psi(t)$ must be localized in time and frequency domain and possess such properties as zero average value, limitations and automodelity (the latter means that the scale transformation does not change the number of oscillations). The choice of $\psi(t)$ is determined by the purpose of the study. Each function $\psi(t)$ has its own characteristics in the time and frequency domain, so using different functions can better identify the properties of the process.

Wavelet-transform signal is divided into:

- continuous (continuous wavelet transformation, CWT) – conversion parameters (a, b) take any real value;

- discrete (discrete wavelet transformation, DWT) – conversion parameters (a, b) take discrete values. However, large-scale transformations and basic wavelet shifts are provided with the integer exponents of 2.

Since the set of coefficients obtained by using continuous wavelet transformation is redundant, during

the study of time series discrete wavelet transform is being used [10].

There are two functions for orthonormal wavelets: scaling function $\varphi(t)$ and wavelet $\varphi(t)$. In the implementation of wavelet transform signal parental wavelet $\varphi(t)$ reveals the detailed (high-frequency component) of the signal being analyzed, while scaling function $\varphi(t)$ reveals a smoothed (low-frequency) component.

Thus, the signal can be represented as a set of successive approximations of rough (approximate) $Am(t)$ and refined (detailing) $Dm(t)$

$$y(t) = A_m(t) + \sum_j^m D_j(t),$$

components with further refinement iteration method. Each step corresponds to a refinement of the scale $a = 2^m$ (ie level m) analysis (decomposition) and synthesis (reconstruction) of a signal.

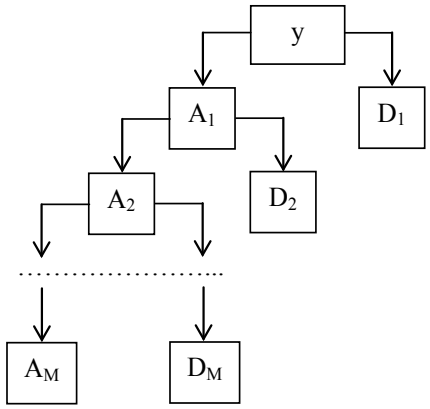
In the first step the signal $y(t)$ can be decomposed into two components:

$$y(t) = A_1(t) + D_1(t) = \sum_k a_{1k}\varphi_{1k}(t) + \sum_k d_{1k}\psi_{1k}(t).$$

The process of decomposition can be extended by $A_1(t)$, then – $A_2(t)$, etc. The signal $y(t)$ at the level of decomposition of m is represented by a set of coefficients a_{mk} and d_{mk} [10]:

$$y(t) = \sum_k a_{mk}\varphi_{mk}(t) + \sum_k d_{mk}\psi_{mk}(t) + \sum_k d_{m-1,k}\psi_{m-1,k}(t) + \cdots + \sum_k d_{1k}\psi_{1k}(t).$$

Wavelet decomposition of the signal $y(t)$, conducted up to the level of M, can be represented graphically as a tree (Fig. 1): decomposition of the signal – top – down and reconstruction – bottom – up.



$m = 0\ y = \{y_i\}$

$m = 1\ y = A_1 + D_1$

$m = 2\ y = A_2 + D_2 + D_1$

$m = M\ y = A_M + D_M + D_{M-1} + \ldots + D_1$

Fig. 1. Decompositional Tree

While using this algorithm at each level, the low-frequency component, which is formed in the previous step, is decomposed into high-frequency and low-frequency component of the next level.

We apply the methods to smooth the laminarity and based on the results we select the best method for preprocessing the data series for a system of monitoring of the stock markets.

In order to receive the plausible model the general level of smoothing is important. When monitoring for the analysis of the dynamics of the market, the most important points are those located in the right of the series – the most recent observations. They determine the main dynamic trends in the indicator in the present and the future.

Let us consider DJI for the period from 09.08.1999 to 11.07.2011 with the length of 3000 values (http://finance.yahoo.com/) and the appropriate measure of LAM (Fig. 2). According to Fig. 1 we can define the period of normal functioning, market crisis and a period of relaxation. Visual analysis of the measure LAM enables us to recognize different periods of functioning. When the market is functioning normally, the trajectory of a measure has horizontal trend in some range. Decreasing LAM indicates the crisis, and increasing – relaxation and renewal.

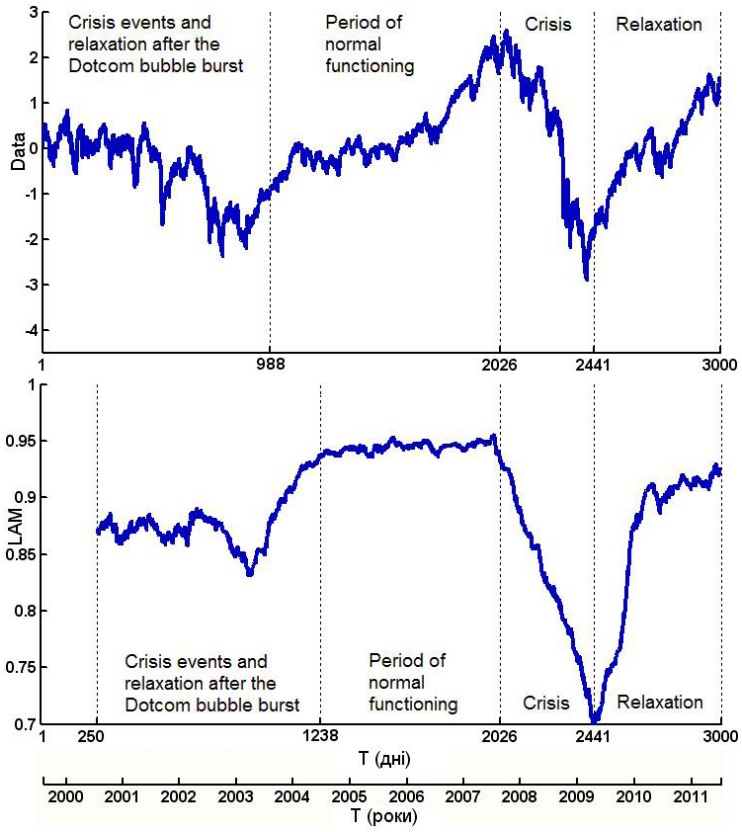


Fig. 2. DJI index and corresponding measure LAM

To study methods of smoothing it is proposed to simulate the process of monitoring in a real-time. To do this, let us choose the length of the series, which will serve as the initial window, 1500 values. Let's take part of the row (from 1 to 1500 points) and calculate its LAM with the window of 250 points. Then take the next part of the series (from 2 to 1501 points), calculate LAM and so on until the end of the series. The result is a set of LAM rows, which corresponds to the daily monitoring in real time. Further, we smooth LAM series in the neighborhood of points where the market changes its regime of functioning (point 1 (2026 in Fig. 2) – during normal functioning of the market/crisis, point 2 (2441 in Fig. 1) – crisis/relaxation) with different methods and compare the results.

For the method of local approximation calculations were performed by using the Curve Fitting Toolbox for MATLAB R2011b and MS Excel. The results showed that the methods of weighted moving average of the first order (WMA), the first-order local regression (LR), Lowess and smoothing splines provide sufficient smoothing. Satisfactory smoothing of the edging intervals of the time series showed only LR (Fig. 3).

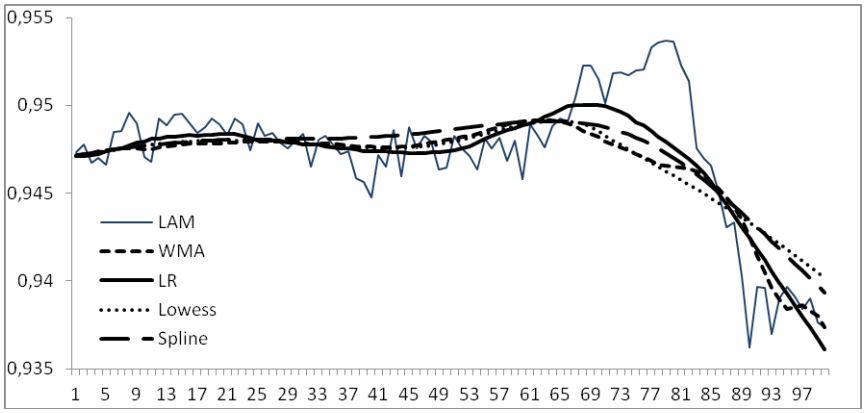


Fig. 3. Methods comparison in the point 1 neighborhood

Smoothing series by wavelet-transformation (orthogonal wavelets were used) was conducted using Wavelet Toolbox for MATLAB R2011b. The best results on the criterion of minimum delays in a number of significant smoothing series was shown by Daubechies wavelet of the 6th order, of the 5th level.

Thus, we have identified two methods that meet the requirements of the LAM smoothing. Let us compare them with each other. For this we take LAM series with the step 5 in the points 1 and 2 neighborhoods, smooth them by means of those methods and consider last 50 points (Fig. 4, 5).

Series 1_12 ends 15 points before the point 1, series 1_15 goes to the point 1, series 1_16, 1_18 end 5 and 15 points after the point 1 accordingly.

Series 2_12 has poin2 as the last point, series 1_15, 1_16, 1_18 ended 15, 20 and 30 points after the point 2 accordingly.

Comparison of the results (Fig. 4, 5) showed that both methods provide adequate level of smoothing series, and satisfactory smoothing of the edging intervals.

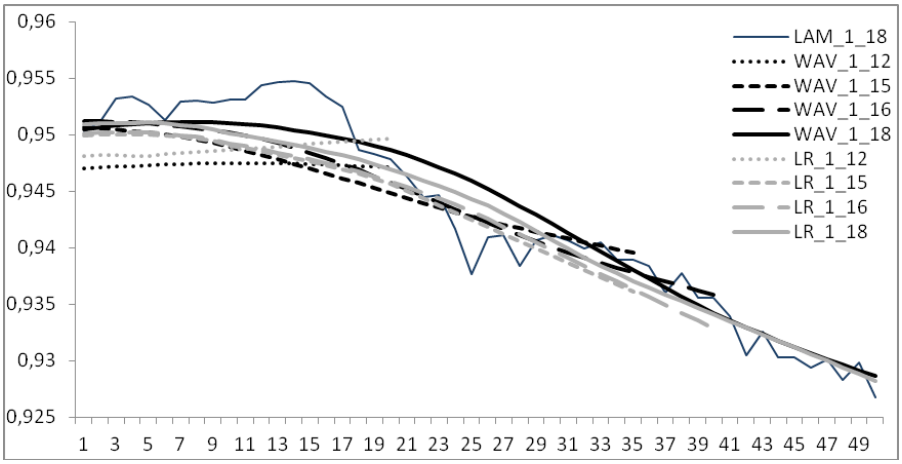


Fig. 4. LAM smoothing in point 1 neighborhood

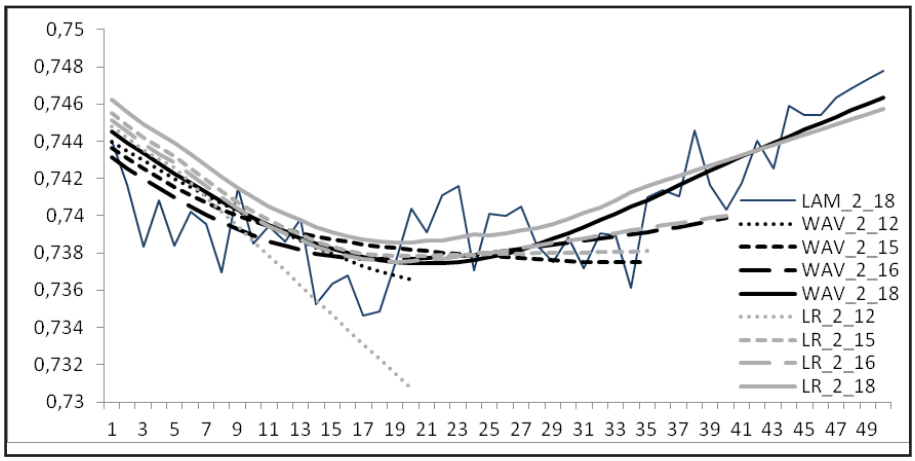


Fig. 5. LAM smoothing in point 2 neighborhood

Conclusion. The analysis of nonparametric smoothing methods revealed two methods (first order local regression and wavelet-transformation) that meet all the requirements as preprocessing tools of data series to automate the detection of transition points between periods of stock market functioning. The choice of a method depends on the practical implementation of the monitoring system of the stock markets.

References

1. Piskun O. V. (2011) Osoblyvosti zastosuvannia rekurentnykh diahram i rekurentnoho kilkisnoho analizu dlia doslidzhennia finansovykh chasovykh riadiv [Features of Application of Recurrent Charts and Recurrent Quantitative Analysis for the Study of Financial Time]. Finansovyy prostir, 3, 111–118.

2. Piskun O. Recurrence Quantification Analysis of Financial Market Crises and Crashes. Retrieved from http://arxiv.org/pdf/1107.5420.pdf

3. Hardle W. (1989). Applied nonparametric regression. Wolfgang Hardle. Cambridge: Cambridge University Press.

4. Kendall M. G., J. Keith Ord (1989). Time Series Third Edition. Oxford: Oxford University Press.

5. Pollard J. H. (1977). A Handbook of Numerical and Statistical Techniques. Cambridge: Cambridge University Press.

6. Cleveland W. S., Devlin S. J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. Jornal of American Statistical Association. Vol. 83, 403, 829–836.

7. Arcangeli R. The Multidimensional Minimizing Splines – Theory and Applications / Remi Arcangeli, Maria Cruz Lopez de Silanes, Juan Jose Torrens. – Boston: Kluwer Academic Publishers, 2004. – 280 p.

8. Pollock D. S. G. (1999) A Handbook of Time-Series Analysis, Signal Processing and Dynamics. London: ACADEMIC PRESS.

9. Daubechies I. (1992) Ten Lectures on Wavelets. Philadelphia: SIAM.

10. Crowley P. M. (2007) A Guide to Wavelets for Economists. Journal of Economic Surveys. Vol. 21, 2, 207–267.