Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 336 (2004) 538–548
PHYSICA A
www.elsevier.com/locate/physa

# Robust methods for stock market data analysis

I. Antoniou[a,b,*], P. Akritas[a,b,c], D.A. Burak[d], V.V. Ivanov[a,b,e], A.V. Kryanev[a,d], G.V. Lukin[d]

[a] *International Solvay Institutes for Physics and Chemistry, CP-231, ULB, Bd. du Triomphe, 1050 Brussels, Belgium*
[b] *Department of Mathematics, Chaos and Innovation Research Unit, Aristoteles University of Thessaloniki, 54124 Thessaloniki, Greece*
[c] *Institut Superieur de Technologie, 6 rue Richard Coudenhove-Kalergi, L-1359 Luxembourg*
[d] *Moscow Engineering and Physical Institute, 115409 Moscow, Russia*
[e] *Laboratory of Information Technologies, Joint Institute for Nuclear Research, 141980 Dubna, Russia*

Received 31 October 2003

**Abstract**

We consider the problem of extraction of trend and chaotic components from irregular stock market time series. The proposed methods also permit to extract a part of chaotic component, the so-called anomalous term, caused by the transient short-time surges with high amplitudes. This provides more accurate determination of the trend component. The methods are based on the M-evaluation with decision functions of Huber and Tukey type. The iterative numerical schemes for determination of trend and chaotic components are briefly presented, resulting in an acceptable solution within a finite number of iterations. The optimal level for extraction of the chaotic component is determined by a new numerical scheme based on the fractal dimension of the chaotic component of the analyzed series. Forecasting from the realized part of the analyzed series and a priori expert information is also discussed.


*PACS:* 89.65.Gh; 05.45.Tp; 02.05.−r

*Keywords:* Stock market; Robust methods; Analysis and forecasting

* Corresponding author. Department of Mathematics, Chaos and Innovation Research Unit, Aristoteles University of Thessaloniki, Thessaloniki 54124, Greece. Fax: +30-2310-997929.
*E-mail address:* iantonio@math.auth.gr (I. Antoniou).

0378-4371/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2003.12.052

## 1. Introduction

The problem of stock market series analysis aiming to identify the structural changes in dynamics of the underlying process and eventually to predict the time series behavior in the future is a very difficult task. The processes underlying stock market data, noise level and time series volatility are "regime shifting", i.e., non-stationary. The commonly accepted viewpoint is that this problem is compatible to inventing a *perpetuum mobile* or solving problems like the *quadrature of the circle* [1].

A first step in stochastic time series analysis is the decomposition of the time series $x(t)$ as a sum of a predictable, deterministic and a chaotic component:

$$x(t) = \tilde{y}_{det}(t) + \tilde{y}_{ch}(t)\,, \tag{1}$$

where $\tilde{y}_{det}(t)$ is the deterministic (or trend) component and $\tilde{y}_{ch}$ is the chaotic component [2–4].

The deterministic component reflects the time-series changes due to the influence of some defined causes which may not be clear enough. However, as a rule their cumulative influence can be predictable during relatively long periods of time. In that case we have the possibility of forecasting.

The chaotic or stochastic part usually concerns a high frequency "*noise*" where the successive elements are practically uncorrelated. This means that this component is not predictable.

Besides, the analyzed series may also involve the anomalous component $y_{an}(t)$ reflecting the structural changes in the process dynamics. In this case (1) becomes

$$x(t) = y_{det}(t) + y_{an}(t) + y_{ch}(t)\,. \tag{2}$$

The terms $y_{det}(t)$, $y_{ch}(t)$ in (2) may significantly differ from the corresponding components in (1).

Unfortunately, traditional methods do not allow the effective extraction of the anomalous term. Moreover, various traditional methods for extraction of chaotic components may lead to the results which are qualitatively different (see, for example Ref. [5]).

Moreover, traditional methods for the determination of the trend component are not stable with respect to short-time surges of a high amplitude. This may result in significant distortion of the trend component itself and not permit to extract correctly the anomalous component $y_{an}(t)$ that very often is of interest.

The aim of this work is the creation of effective numerical schemes for estimation of trend, chaotic and anomalous components for satisfactory forecasting taking also into account a priori expert information. In Section 1 we describe the numerical schemes for extracting of deterministic and chaotic terms using expansions in orthogonal polynomials. The numerical schemes for determination of trend and chaotic terms on the basis of robust linear splines are presented in Section 2. Section 3 is devoted to the optimal determination of the chaotic component based on its fractal dimension. Section 4 presents first results concerning the application of the forecasting scheme which uses a priori expert estimation.

## 2. Extraction of deterministic and chaotic terms using orthogonal polynomials

The extraction of the anomalous component can be achieved by robust methods [6–10]. We consider a robust nonlinear polynomial model of the trend component and corresponding numerical scheme developed on the basis of robust linear splines. For such a model of the trend component with a polynomial degree higher than five there exists a computational instability related to a bad-conditionality of the information matrix [4,11,12] (the expression of the information matrix will be presented below).

In order to overcome this circumstance, we use here a system of orthogonal polynomials, which permits to construct a robust nonlinear polynomial model of the trend component of any rank without inversion of the information matrix.

Let $x(t)$ be an original time-series and let $n$ realized values $x(t_i) = x_i, i = 1, \ldots, n$ be known.

We represent the trend component in the following form:

$$y_{det}(t) = \sum_{k=0}^{m} u_k \Phi_k(t) \,, \tag{3}$$

where $u_k$ are expansion coefficients and $\Phi_k(t)$ are orthogonal polynomials of the degree $k$ conforming to orthogonality conditions

$$\sum_{i=1}^{n} \frac{1}{\sigma_i^2} \Phi_l(t_i)\Phi_p(t_i) = \delta_{lp} \,, \tag{4}$$

where $\delta_{lp}$ is the Kronecker symbol.

The system of orthonormal polynomials can be calculated using, for example, the Forsythe recurrent scheme [11,13].

The expansion coefficients $u_k, k = 0, \ldots, m$ are the solution of the following extremal problem

$$u = \arg\min_{u} \sum_{i=1}^{n} \rho\left(\frac{x(t_i) - y_{det}(t_i)}{\sigma_i}\right) \,, \tag{5}$$

where $\sigma_i^2$ is the dispersion of a random value $x(t_i), i = 1, \ldots, n$, and $\rho(s)$ is the Huber function

$$\rho(s) = \begin{cases} \dfrac{s^2}{2}, & |s| \leqslant K \,, \\ K|s| - \dfrac{K^2}{2}, & |s| > K \,, \end{cases} \tag{6}$$

where $K$ is the Huber parameter [9].

The values of the anomalous component are defined by the following equations:

$$\begin{aligned} y_{an}(t_i) &= 0, & i \in I_0 \,, \\ y_{an}(t_i) &= x(t_i) - y_{det} - K\sigma_i, & i \in I_+ \,, \\ y_{an}(t_i) &= x(t_i) - y_{det} + K\sigma_i, & i \in I_- \,, \end{aligned} \tag{7}$$

where $I_0, I_+, I_-$ are sets of indices defined by relations

$$I_0 = \{i\colon\ |x(t_i) - y_{det}(t_i)| \leqslant K\sigma_i\}\ ,$$

$$I_+ = \{i\colon\ x(t_i) - y_{det}(t_i) > K\sigma_i\}, \quad i = 1,\ldots,n\ ,$$

$$I_- = \{i\colon\ x(t_i) - y_{det}(t_i) < -K\sigma_i\}\ . \tag{8}$$

The extremal problem (5) is reduced to the system of nonlinear equations [7,14]

$$u = (A^{\mathrm{T}}WA)^{-1}A^{\mathrm{T}}Wz\ , \tag{9}$$

where $W = diag(1/\sigma_1^2,\ldots,1/\sigma_n^2)$ is the diagonal matrix, $\sigma_i^2$ is the dispersion of a random value $x_i$ and $A^{\mathrm{T}}WA$ is the information matrix. The elements of the matrix $A$ and the components of the vector $z$ are

$$\begin{aligned}
&A_{ij} = \Phi_j(t_i), && i = 1,\ldots,n, \quad j = 1,\ldots m\ ,\\
&z_i = x_i, && i \in I_0\ ,\\
&z_i = (Au)_i + K\sigma_i, && i \in I_+\ ,\\
&z_i = (Au)_i - K\sigma_i, && i \in I_-\ ,\\
&I_0 = \{i\colon\ |x_i - (Au)_i| \leqslant K\sigma_i\}\ ,\\
&I_+ = \{i\colon\ x_i - (Au)_i > K\sigma_i\}\ ,\\
&I_- = \{i\colon\ x_i - (Au)_i < -K\sigma_i\}\ .
\end{aligned} \tag{10}$$

To find the solution of the system of nonlinear equations (9)–(10), and, consequently, the solution of extremal problem (5), we apply the following iterative scheme

$$u^{(l+1)} = (A^{\mathrm{T}}WA)^{-1}A^{\mathrm{T}}Wz^{(l)}, \tag{11}$$

where the components of the vector $z^{(l)}$ are given by

$$\begin{aligned}
&z_i^{(l)} = x_i, && i \in I_0^{(l)}\ ,\\
&z_i^{(l)} = (Au^{(l)})_i + K\sigma_i, && i \in I_+^{(l)}\ ,\\
&z_i^{(l)} = (Au^{(l)})_i - K\sigma_i, && i \in I_-^{(l)}\ ,\\
&I_0^{(l)} = \{i\colon\ |x_i - (Au^{(l)})_i| \leqslant K\sigma_i\}\ ,\\
&I_+^{(l)} = \{i\colon\ x_i - (Au^{(l)})_i > K\sigma_i\}\ ,\\
&I_-^{(l)} = \{i\colon\ x_i - (Au^{(l)})_i < -K\sigma_i\}\ .
\end{aligned} \tag{12}$$

We have proved that the iterative scheme (11)–(12) is reduced to the robust solution of problem (5) within a finite number of iterations (see Refs. [4,7,14]). Since the system $\Phi_j(t_i)$, $j=1,\ldots,m$ is orthonormal with weights $1/\sigma_i^2$, $i=1,\ldots,n$, the information matrix $A^{\mathrm{T}}WA$ is the identity matrix and the iterative scheme (11)–(12) could be presented in

the following form:

$$u_j^{(l+1)} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \Phi_j(t_i) z_i^{(l)} , \tag{13}$$

where $z_i^{(l)}$ are defined by Eqs. (12).

**Remark.** Similarly one may construct the iterative scheme when the Tukey function is used in Eq. (5) as a robust function

$$\rho(s) = \begin{cases} s^2, & |s| \leqslant R , \\ R^2, & |s| > R . \end{cases}$$

To define the value of the Huber parameter $K$, the table of $K$ dependence on large outliers share [9] is usually used. When analyzing real time series, the large outliers share is not known initially and this does not permit to define properly the Huber parameter. In Ref. [4] we developed a scheme for proper estimation of the Huber parameter without determination of large outliers share.

Upon extraction of trend and anomalous components

$$y_{ch}(t) = x(t) - (y_{det}(t) + y_{an}(t)) \tag{14}$$

the chaotic component $y_{ch}(t)$ must represent (with a certain degree of approximation) a random Brownian process.

## 3. Extraction of deterministic and chaotic terms using robust linear splines

Another effective numerical scheme for determination of the trend component is based on robust linear splines [4].

The robust linear spline $S_\alpha(t)$ is the solution of the extremal problem for the minimum of the function

$$J_\alpha(S_\alpha(t)) = \sum_{i=1}^{n} \rho\left(\frac{x_i - S_i}{\sigma_i}\right) + \alpha \sum_{i=1}^{n-1} (S_{i+1} - S_i)^2 , \tag{15}$$

where $\rho(s)$ is the robust function, $S_i = S_\alpha(t_i)$ are the required values of the robust linear spline, $\alpha > 0$ is the smoothing parameter (similar to the regularization parameter for ill-posed problems [12]).

If in (15) $\rho(s)$ is the Huber function, then in order to get the required vector $S = (S_1, \ldots, S_n)^{\mathrm{T}}$, we use the following algorithm for extraction of all three components based on the robust smoothing linear splines. The system for solving the extremum problem (15) is

$$AS^{(l+1)} = F^{(l)} , \tag{16}$$

where $A$ is the three-diagonal, positively defined symmetrical matrix and

$$S^{(l)} = (S_1^{(l)}, \ldots, S_n^{(l)})^{\mathrm{T}} ,$$

$$F^{(l)} = (y_1^{(l)}, \ldots, y_n^{(l)})^{\mathrm{T}} ,$$

$$y_i^{(l)} = x_i, \qquad |x_i - S_i^{(l)}| \leqslant K\sigma_i ,$$

$$y_i^{(l)} = S_i^{(l)} + K\sigma_i, \qquad x_i - S_i^{(l)} > K\sigma_i ,$$

$$y_i^{(l)} = S_i^{(l)} - K\sigma_i, \qquad x_i - S_i^{(l)} < -K\sigma_i .$$

The iterative scheme (16) is reduced to the robust linear spline that minimizes the objective function (15) within a finite number of iterations.

The definition of the optimal value of the evening parameter is given in Sections 4 and 5.

The robust extraction of time-series components using the Huber function assumes the symmetry of large outliers with respect to the trend component. In case of symmetry failure, instead of the Huber function one can use the Tukey function.

The iterative scheme (16) with the Tukey function $\rho(s)$ can be constructed in the similar way.

## 4. Determination of fractal dimension

In order to choose the optimal level of smoothness determined by a number $m$ of expansion coefficients (3) or by a smoothing parameter $\alpha$, we use the Hausdorff–Besicovitch (HB) fractal dimension [15,16]. The fractal dimension of Brownian process is 1.5 [17]. Therefore, we find the smoothing level for which the fractal dimension of the chaotic component is closely adjacent to 1.5. We assume of course that the chaotic component is close to Brownian motion.

Usually the determination of time series fractal dimension is based on the estimation of the Hurst exponent [18,19]. We use a new algorithm for the direct determination of fractal dimension of time series [20]. This algorithm is based on the asymptotic formula

$$S(\delta) \sim \delta^{2-d_F}, \delta \to +0 , \tag{17}$$

where $d_F$ is the HB fractal dimension, $\delta$ is the time step for a uniform subdivision of the analyzed time interval on $n$ sub-intervals, $S(\delta) = \sum_{i=1}^{n} (\max(x(t)) - \min(x(t)))\delta$ is the area of minimal function graph covering $x(t)$ by rectangles $\{(t_i \leqslant t \leqslant t_{i+1}) \cap (\max(x(t)) \leqslant x(t) \leqslant \min(x(t)))\}, i{=}1, \ldots, n$, and $\max(x(t)), \min(x(t))$ are defined within interval $t \in [t_i, t_{i+1}]$ [20].

From (17) we have

$$\ln S(\delta) = b + a \ln \delta . \tag{18}$$

Thus, plotting the dependence graph $S(\delta)$ in the double logarithmic scale and fitting by the least-squares method the angular coefficient $a$, we find the HB fractal dimension using the formula $d_F = 2 - a$.

After the extraction of chaotic components corresponding to different values of $m$ (when the robust orthonormal polynomials are used) or $\alpha$ (when the robust splines are used) and plotting the dependence of $m$ or $\alpha$ against the fractal dimensions of corresponding chaotic components, can define the optimal values of $m$ or $\alpha$ that correspond to the minimum of the objective function $(d_F - 1.5)^2$.

The numerical experiments have shown that the amount of experimental data necessary for reliable determination of the HB fractal dimension applying the scheme (17)–(18) is several times (sometimes dozens of times) less compared to data necessary for reliable determination of the Hurst exponent. The last circumstance permits one to increase the efficiency of extraction of chaotic components, and, thus, increase the efficiency of time-series forecasting.

## 5. Forecasting scheme based on a priori expert estimations

We developed the forecasting scheme given a priori expert estimates defined in the form of a set of pairs

$$\{x_{exp}(t_i), \sigma_i^2, i = n+1, \ldots, n+L\}\ , \tag{19}$$

where $x_{exp}(t_i)$ is the most probable expert estimate of the analyzed series for the future time $t_i$, $\sigma_i^2$ is the corresponding dispersion and $L$ is the forecasting horizon.

The forecasting scheme involves only part of the analyzed time series

$$x(t_{n-l}), x(t_{n-l+1}), \ldots, x(t_n), \quad x_{exp}(t_{n+1}), \ldots, x_{exp}(t_{n+L}) \tag{20}$$

for the determination of the trend component by robust orthogonal polynomials or robust linear splines.

The horizon $l$ of realized values of the analyzed series is selected according to the chosen rank $m$ (when the system of robust orthonormal polynomials is used) or the smoothing parameter $\alpha$ (when robust linear splines are used) and the forecasting horizon $L$. The values of the smoothing parameters ($m$ or $\alpha$) and $l$ can be estimated through minimization of the residual functional based on the difference between the real values of the analyzed series and the predicted values (at the stage of a preliminary tuning of the forecasting scheme on realized part of the analyzed time series).

If a priori expert estimates are absent, then, by default, one may assume

$$x_{exp}(t_{n+i}) = x(t_n), \quad \sigma_i = \sigma, \ \ i = 1, \ldots, L\ , \tag{21}$$

where $\sigma$ is the dispersion of differences between realized time series values $x(t_{n-l+i}), \ldots, x(t_n)$ from values $x_{det}(t_{n-l}), \ldots, x_{det}(t_n)$ determined by the trend component on the basis of a realized part of the analyzed series. As forecasted values $x_{for}(t_{n+i}), i = 1, \ldots, L$ are used, the values $x_{det}(t_{n+i})$ corresponding to the trend component determined by the scheme described above.

Figs. 1, 4 and 7 show the time series of close prices (solid curve) for some leading corporation stocks (Fig. 1-Merrill Lynch Corp., Fig. 4-Citibank Inc., Fig. 7-Bank of America Corp.) included in S& P500 together with the trend components (dashed curve) determined using robust orthogonal polynomials and robust linear splines.


Fig. 1. Time series of close prices for Merrill Lynch Corp. (solid curve) superimposed by the trend term (dashed curve).

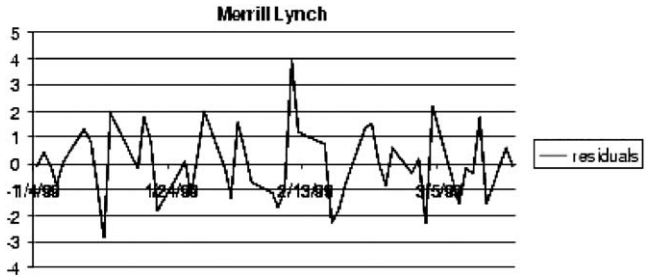


Fig. 2. Chaotic component for time series presented in Fig. 1.

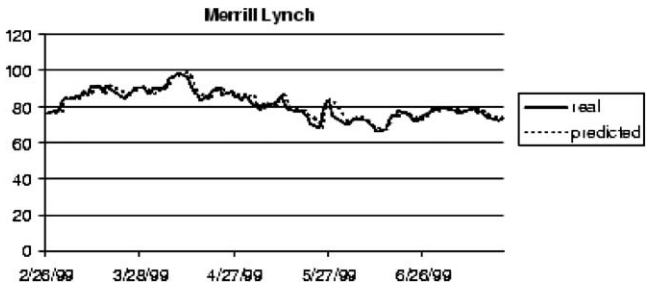


Fig. 3. Time series of close prices for Merrill Lynch Corp. (solid curve) superimposed by the forecasted values (dashed curve).

Figs. 2, 5 and 8 demonstrate the chaotic components of time-series presented in Figs. 1, 4 and 7.

Figs. 3, 6 and 9 show the time series of close prices (solid curve) and the time series of forecasted values.

The preliminary analysis of forecasting results shows that the percentage of coincidence in increase and decrease of predicted and real prices amounts to 65–73%. This bound of the developed approach is obtained in cases of absence of a priori information, when instead of expert estimates $x_{exp}(t_{n+i})$ were used realized values $x(t_n)$.

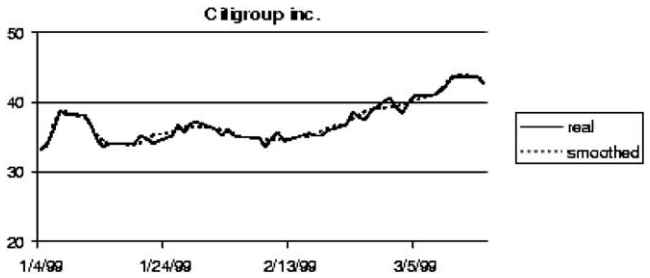


Fig. 4. Time series of close prices for Citigroup inc. (solid curve) superimposed by the trend term (dashed curve).

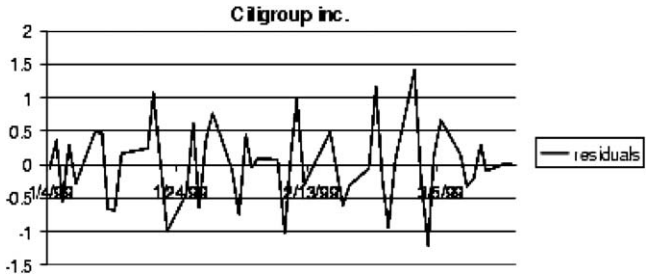


Fig. 5. Chaotic component for time series presented in Fig. 4.


Fig. 6. Time series of close prices for Citigroup inc. (solid curve) superimposed by the forecasted values (dashed curve).

## 6. Conclusion

The developed schemes for estimation of the trend and chaotic components together with the forecasting scheme based on robust orthogonal polynomials and robust orthogonal splines possess computational stability and stability against transient large outliers in realized values of the analyzed series. This approach permits to extract the abnormal large outliers, including those not obviously clear by their time position

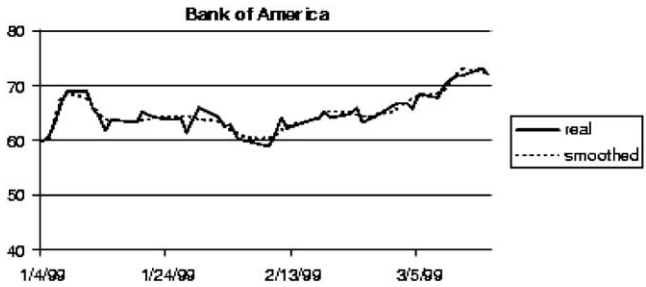


Fig. 7. Time series of close prices for Bank of America Corp. (solid curve) superimposed by the trend term (dashed curve).

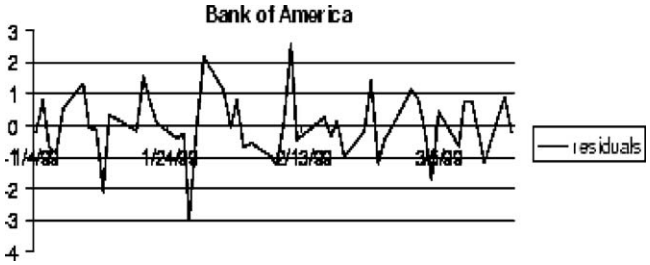


Fig. 8. Chaotic component for time series presented in Fig. 7.


Fig. 9. Time series of close prices for Bank of America Corp. (solid curve) superimposed by the forecasted values (dashed curve).

and by the amplitude, which provides the possibility of detailed analysis of anomalous and critical events in economics. The proposed forecasting scheme allows to take into account additional a priori expert information and, therefore provide more accurate prediction of future events.

The efficiency of our preliminary analysis of the above described approach has been demonstrated. The comparison with other existing methods will be considered in future work.

## Acknowledgements

This work has been partly supported by the Luxembourg Ministry of Culture, High Education and Research under Grant BFR01/068.

## References

[1] T. Hellström, K. Hollström, Predicting the stock market, Center of Mathematical Modeling (CMM), Mälardalen University, P.O. Box 883, S-721 23 Västeras, Sweden, Technical Report Series IMa-Tom-1997-07, August, 1998.
[2] H.D.I. Abarbanel, Analysis of Observed Chaotic Data, Springer, New York, 1995.
[3] S. Haykin, Adaptive Filter Theory, Prentice-Hall, Englewood Cliffs, NJ, 1991.
[4] A.V. Kryanev, G.V. Lukin, Mathematical Methods for Stochastic Data Processing, Nauka Editions, Moscow, 2003, in press.
[5] I. Antoniou, V.V. Ivanov, V. Ivanov Valery, P.V. Zrelov, On the log-normal distribution of stock market data, Physica A, in press.
[6] D.A. Andrews, P.J. Bickel, F.R. Hampel, P.J. Huber, W.H. Rodjers, J.W. Tukey, Robust Estimates of Location: Survey and Advances, Princeton University Press, Princeton, NY, 1972.
[7] V.Ya. Arsenin, A.V. Kryanev, M.V. Tsupko-Sitnikov, Application of Robust Methods for Ill-Posed Problems Solving, USSR Comp. Math. Math. Phys. 29 (5) (1989) 653–661.
[8] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust statistics, The Approach Based on Influence Functions, Wiley, New York, 1985.
[9] P.J. Huber, Robust Statistics, Wiley, New York, 1981.
[10] R.L. Launer, G.N. Wilkinson (Eds.), Robustness in Statistics, Academic Press, New York, 1979.
[11] G.A.F. Seber, Linear Regression Analysis, Wiley, New York, 1977.
[12] A.N. Tikhonov, V.Ya. Arsenin, Solution of Ill-Posed Problems, Winston, Washington, 1997.
[13] G.E. Forsythe, M.A. Malcolm, C.B. Moler, Computer Methods for Mathematical Computations, Prentice-Hall, Englewood Cliffs, NJ, 1977.
[14] V.Ya. Arsenin, A.V. Kryanev, Generalized maximum likelihood method and its application for Ill-posed problems solving, in: A. Tikhonov (Ed.), Ill-Posed Problems in Natural Sciences, VSP-BV, Netherland, 1992, pp. 3–12.
[15] B.B. Mandelbrot, Fractals: Form, Chance and Dimension, W.H. Freeman & Co., San Francisco, 1977.
[16] C.A. Rogers, Hausdorff Measures, Cambridge University Press, Cambridge, 1970.
[17] H.H. Hastings, G. Sugihara, Fractals. A User's Guide for the Natural Sciences, Oxford University Press, Oxford, 1993.
[18] J. Feder, Fractals, Plenum Press, New York and London, 1989.
[19] H.E. Hurst, R.P. Black, Y.M. Simaika, Long-Term Storage, an Experimental Study, Constable, London, 1965.
[20] M.M. Dubovikov, N.V. Starchenko, Variation Index and its Application to the Analysis of Fractal Structures, Almanac Gordon, N1, 2003.