Multivariate Threshold Autoregressive Models in Geodesy

Ing. Tomáš Bacigál

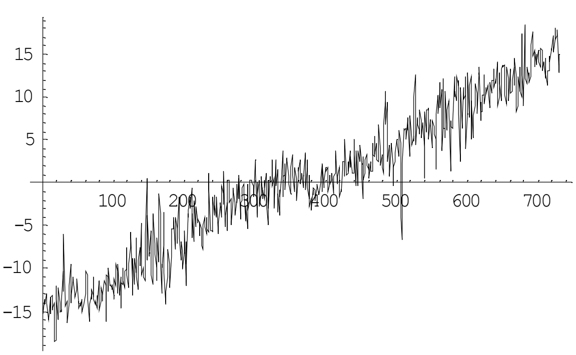
Department of Mathematics and Descriptive Geometry, Faculty of Civil Engineering
Slovak University of Technology, Bratislava
bacigal@math.sk

Abstract

In recent years, the situation in time series analysis has changed turning its concern from linear to nonlinear modeling. In this article we are trying to show how a special case of such a large family of models (as threshold autoregressive ones are) may be applied within processing of continual GPS observations. Two components (north and east) of point position in horizontal coordinate system are taken to obtain bivariate time series, which consequently are tested for nonlinearity and modeled using bivariate threshold autoregressive model. Whole procedure, of course, can be easily generalized to more than two-variate series.

Introduction

May we have time series $\boldsymbol{y}$ of n time-points,



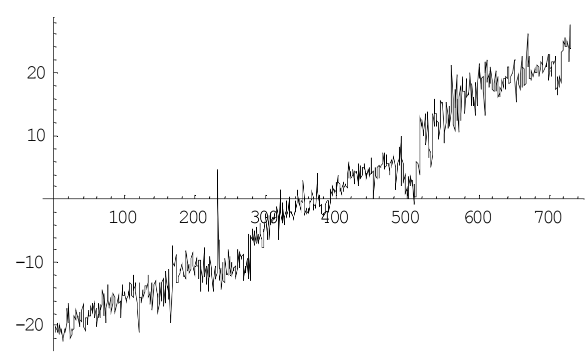


Fig.1: Two vectors of GPS observations, with length n=730 days

there are several ways to model it. One large family of models, that are strongly suitable for modeling stochastic processes, are those arising from Box-Jenkins methodology such as ARMA etc. We will be interested in autoregressive (AR) models, defined

$$y_t = \Phi_0 + \Phi_1 y_{t-1} + \ldots + \Phi_p y_{t-p} + \varepsilon_t . \qquad (1)$$

This is linear model and as such, it may fit only linear dependencies. But what if we know our time series are nonlinear (excluding common trend and seasonality) but piecewise linear, changing it's behaviour by activation of some factor.

We get threshold autoregressive model (TAR), e.g.

$$y_t = \begin{cases} \Phi_1^{(1)} y_{t-1} + \ldots + \Phi_p^{(1)} y_{t-p} + \varepsilon_t^{(1)} & \text{if } z_{t-d} \le r \\ \Phi_1^{(2)} y_{t-1} + \ldots + \Phi_p^{(2)} y_{t-p} + \varepsilon_t^{(2)} & \text{if } z_{t-d} > r \end{cases}, \tag{2}$$

where z is a threshold variable, r is a threshold and their relation delimites constituent regimes of the model. Letter d denotes time lag (delay). Because there is often a need to process more than a single vector of measurements at once (sometimes given with some explanatory time series), we will speak about multivariate TAR model

$$\mathbf{y}_t = \mathbf{\Phi}_0^{(j)} + \sum_{i=1}^{p} \mathbf{\Phi}_i^{(j)} \mathbf{y}_{t-i} + \boldsymbol{\varepsilon}_t^{(j)} \quad \text{if } r_{j\text{-}1} < z_{t-d} \le r_j \,, \tag{3}$$

where $\mathbf{y}_t = (y_{1t} \quad ,\ldots \quad y_{kt})$,
$\Phi_0^{(j)}$ is constant term for regime j, and
y_{kt} denotes k[th] univariate time series nested in $\mathbf{y}_t$.

As for **y** I put to use GPS observations at permanent station Pecny which are given as point coordinates in horizontal coordinate system (*n, e, v* – north, east and vertical component). Usually the components have been processed separately. However, this means a risk of some information loss, as they are obviously somehow correlated. Thats'why I've focused on multivariate modeling.

Now, as we have data, type of model and assume that the threshold variable z is known, but the delay d, the order p of AR model and threshold r are not (for simplicity I restrict the case to 2 regimes).

The goal is threefold:

1. to find proper order p of AR model.
2. to make sure, that time series are not linear using a test developed by prof. Tsay.
3. to choose the best delay and threshold values, and consequently to build up the final shape of multivariate model.

1. Finding order of autoregression

For now, we handle the data as being linear and follow two ways:

a) Using a Levinson-Durbin estimation procedure (p_{max}=15) and specially its outcome – covariance matrices

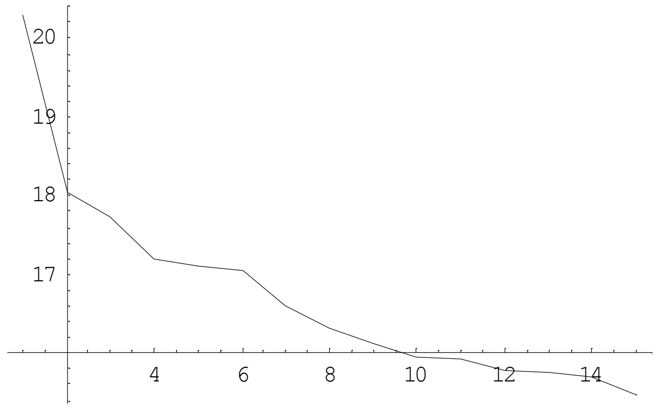


Fig.2: Determinants of covariance matrices vs. order *p*.

Order *p* is chosen subjectively according to plot steepness.

b) Employing three information criteria AIC, BIC, HQIC which the most appropriate order minimizes.

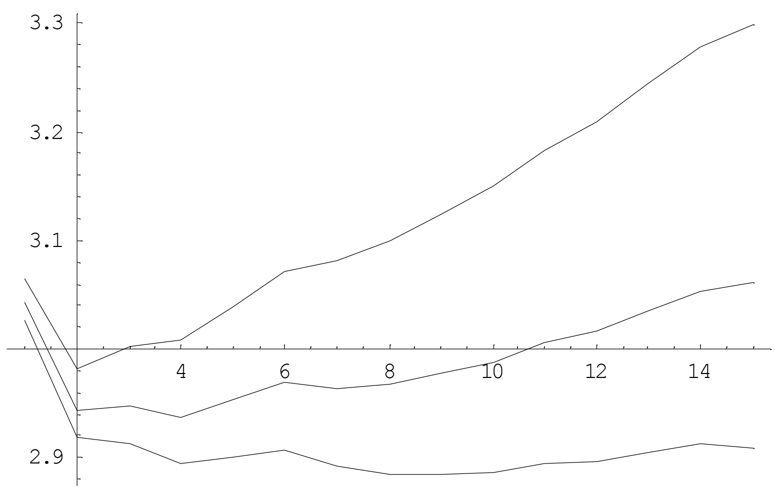


Fig.2: Information criteria vs. order *p*.

Order *p* is chosen as an dominating argument of minimal criteria values.

$p = 2$

2. Testing

Null hypothesis H_0: $\mathbf{y}_t$ is linear.
Alternative hypothesis H_1: $\mathbf{y}_t$ follows a threshold model

We utilize standard least square regression framework:

$$\mathbf{y}_t = \mathbf{X}_t\mathbf{\Phi} + \boldsymbol{\varepsilon}_t \qquad t = h+1, \ldots n \quad , \tag{4}$$

where $h = \max(p, d)$,
$\mathbf{X}_t = \begin{pmatrix} 1 & \mathbf{y}_{t-1} & \mathbf{y}_{t-2} & \ldots & \mathbf{y}_{t-p} \end{pmatrix}$ is regressor,
$\mathbf{\Phi}$ denotes parameter matrix.

If H_0 holds, then least square estimates are useful, otherwise the estimates are biased under H_1.

Now let the ordering of the threshold variable z be rearranged increasingly so that $z_{(i)}$ is the smallest element of S=$\{z_{h+1-d}, \ldots z_{n-d}\}$ and $t(i)$ is the time index of $z_{(i)}$. Therefore $z_{(i)} = z_{t(i)}$ and autoregression is

$$\mathbf{y}_{t(i)+d} = \mathbf{X}_{t(i)+d}\mathbf{\Phi} + \boldsymbol{\varepsilon}_{t(i)+d}, \qquad i = 1, \ldots n-h. \tag{5}$$

It is important to see that the dynamics of the $\mathbf{y}_t$ series has not changed (that is the independent variable of $\mathbf{y}_t$ is $\mathbf{X}_t$ for all t). What has changed is the ordering by which the data enter the regression setup. This means an effective transformation of threshold model into a changepoint problem.

To detect model change consider the idea:

If $\mathbf{y}_t$ is linear, then recursive least squares estimates of the arranged regression is consistent so that the predictive residuals approach white noise (consequently, predictive residuals are uncorrelated with the regressor $\mathbf{X}_{t(i)+d}$).

Let

$$\hat{\boldsymbol{\eta}}_{t(m+1)+d} = \frac{\mathbf{y}_{t(m+1)+d} - \mathbf{X}_{t(m+1)+d}\hat{\mathbf{\Phi}}_m}{[1 + \mathbf{X}_{t(m+1)+d}\mathbf{V}_m\mathbf{X}^{\mathrm{T}}_{t(m+1)+d}]^{1/2}} \tag{6}$$

be the standardized predictive residual of regression (5) where

$$\mathbf{V}_m = [\sum_{i=1}^{m} \mathbf{X}^{\mathrm{T}}_{t(i)+d}\mathbf{X}_{t(i)+d}]^{-1} \tag{7}$$

and $\hat{\mathbf{\Phi}}_m$ is the estimate of arranged regression (5) using data points associated with the m smallest values of z_{t-d}.

Next there comes a regression

$$\hat{\boldsymbol{\eta}}_{t(l)+d} = \mathbf{X}_{t(l)+d}\boldsymbol{\Psi} + \mathbf{w}_{t(l)+d} \qquad l = m_0 + 1, \ldots n - h \tag{8}$$

where m_0 denotes the starting point of recursive least squares estimation ($m_0 \cong 3\sqrt{n}$). The problem of interest is to test the hypothesis H_0: $\boldsymbol{\Psi} = \mathbf{0}$ versus H_1: $\boldsymbol{\Psi} \neq \mathbf{0}$ in (8). Tsay(1998) designed a test statistic

$$C(d) = [n - h - m_0 - (kp + 1)] \times \{\ln[\det(S_0)] - \ln[\det(S_1)]\} \tag{9}$$

where

$$S_0 = \frac{1}{n-h-m_0} \sum_{l=m_0+1}^{n-h} \hat{\boldsymbol{\eta}}^{\mathrm{T}}_{t(l)+d} \hat{\boldsymbol{\eta}}_{t(l)+d}, \qquad S_1 = \frac{1}{n-h-m_0} \sum_{l=m_0+1}^{n-h} \hat{\mathbf{w}}^{\mathrm{T}}_{t(l)+d} \hat{\mathbf{w}}_{t(l)+d}$$

and $\hat{\mathbf{w}}_t^{\mathrm{T}}$ is the least square residual of regression (8). Under the null that $\mathbf{y}_t$ is linear (and some regularity conditions), $C(d)$ is asymptotically a χ^2 random variable with $k(pk+1)$ degrees of freedom.
If $C(d) < \chi^2_{\mathrm{df}}$, we do not refuse the null hypothesis.

Test results:

p	d	$C(d)$	χ^2		p-value	Degr. of freedom
			$\alpha = 0.05$	$\alpha = 0.01$		
2	1	29.4	18.3	23.2	0.0010	10
	2	15.1			0.128	
	3	23.2			0.010	
	4	8.4			0.406	
	5	11.9			0.290	
	6	15.8			0.104	
	7	25.3			0.005	
	8	21.9			0.034	
	9	13.2			0.213	
	10	18.9			0.041	
4	1	41.4	28.9	34.8	0.0014	18
	2	21.1			0.278	
	3	30.2			0.035	
	4	14.2			0.281	
	5	15.6			0.383	

Note. The test is most powerful when d is correctly specified.

3. Building up the model

First we aim at choosing the best values delay and threshold.

a) One way is to apply conditional least squares estimation.

Assume that p and s (number of regimes) are known, then parameters of model (for now a bit simplified)

$$\mathbf{y}_t = \begin{cases} \mathbf{X}_t\mathbf{\Phi}_1 + \mathbf{\Sigma}_1^{1/2}\mathbf{a}_t & \text{if } z_{t-d} \le r \\ \mathbf{X}_t\mathbf{\Phi}_2 + \mathbf{\Sigma}_2^{1/2}\mathbf{a}_t & \text{if } z_{t-d} > r \end{cases} \tag{10}$$

where $\mathbf{a}_t = (a_{1t} \quad \dots \quad a_{kt}) \sim N(0,\mathbf{I})$,

are ($\mathbf{\Phi}_i$, $\mathbf{\Sigma}_i$, r, d). Putting the possible values of r and d into grid $\{1, 2, \dots d_0\} \times \{r_{\min}, r_{\mathrm{rmin}} + step, \dots r_{\max}\}$ model (10) reduces to 2 separated multivariate linear regressions from which the least squares estimates of Φ_i and Σ_i (i=1,2) are readily available:

$$\hat{\mathbf{\Phi}}_i(r,d) = \left(\sum_t^{(i)} \mathbf{X}_t^{\mathrm{T}}\mathbf{X}_t\right)^{-1}\left(\sum_t^{(i)} \mathbf{X}_t^{\mathrm{T}}\mathbf{y}_t\right) \quad ,$$

$$\hat{\mathbf{\Sigma}}_i(r,d) = \frac{\sum_t^{(i)} (\mathbf{y}_t - \mathbf{X}_t\hat{\mathbf{\Phi}}_i^*)^{\mathrm{T}}(\mathbf{y}_t - \mathbf{X}_t\hat{\mathbf{\Phi}}_i^*)}{n_i - k} \tag{11}$$

where $\sum_t^{(i)}$ denotes summing over observations in regime i,
$\hat{\mathbf{\Phi}}_i^* = \hat{\mathbf{\Phi}}_i(r,d)$,
n_i is number of data points in regime i and
k the dimmension of $\mathbf{X}_t$ ($k < n_i$).

It becomes clear that conditional least squares estimates of r and d should minimize the sum of squares ofresiduals

$$(\hat{r},\hat{d}) = \arg\min_{r,d} S(r,d) \tag{12}$$

where

$$S(r,d) = (n_1 - k)\mathrm{Tr}[\mathbf{\Sigma}_1(r,d)] + (n_2 - k)\mathrm{Tr}[\mathbf{\Sigma}_2(r,d)]\,. \tag{13}$$

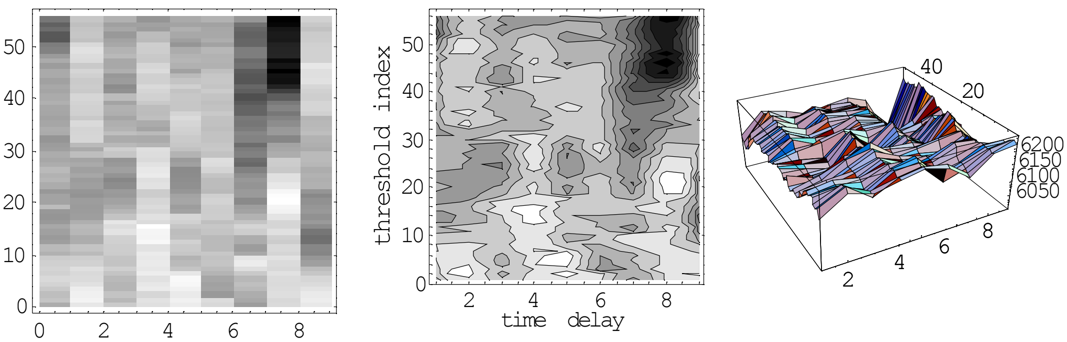


Fig.3: Density, contour and 3D plot of $S(r,d)$;
x-axis represents time delay, y-axis grid index of threshold value

Results of conditional estimation:

p	r [mm]	d [day]	S [mm^2]
2	1.89	8	6013.9
	- 0.36	1	6136.5
	- 1.06	1	6137.9
	- 0.35	3	6138.4

b) Besides this, we may apply Akaike information criterion AIC to the same grid $r \times d$.

In fact, it comes along with and supplement the least squares estimation procedure and, of course, there are other parameters defining the multivariate threshold model that could be selected by the criterion

$$AIC(p,s,d,r) = \sum_{j=1}^{s} [n_j \ln(\det(\hat{\mathbf{\Sigma}}_j)) + 2k(kp+1)] \tag{14}$$

with $\hat{\mathbf{\Sigma}}_j = \frac{1}{n_j} \sum_t^{(j)} \hat{\boldsymbol{\varepsilon}}_t^{(j)\mathrm{T}} \hat{\boldsymbol{\varepsilon}}_t^{(j)}$,

where n_j is the number of data points in regime j,

$\sum_t^{(j)}$ denotes summing over observations in regime j,

$\hat{\boldsymbol{\varepsilon}}_t^{(j)}$ are residuals.

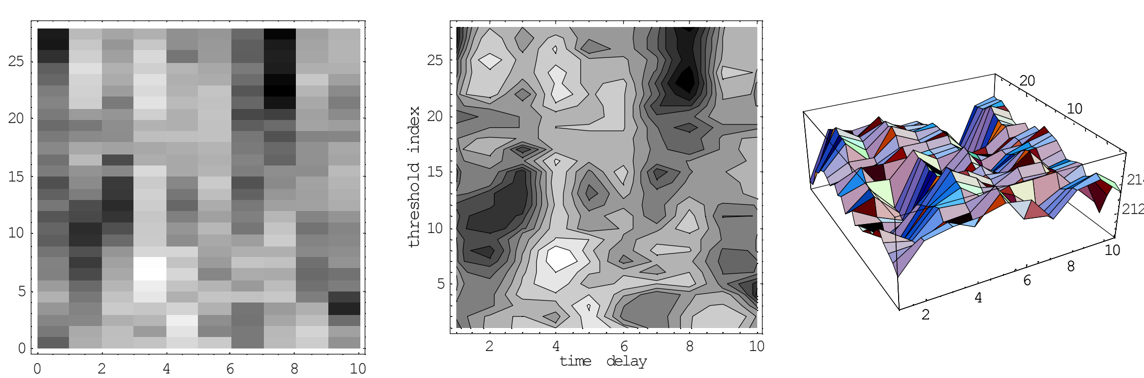


Fig.5: AIC maped over grid $r \times d$, r∈⟨-2.6, 3.0⟩, d = {1,2...10}

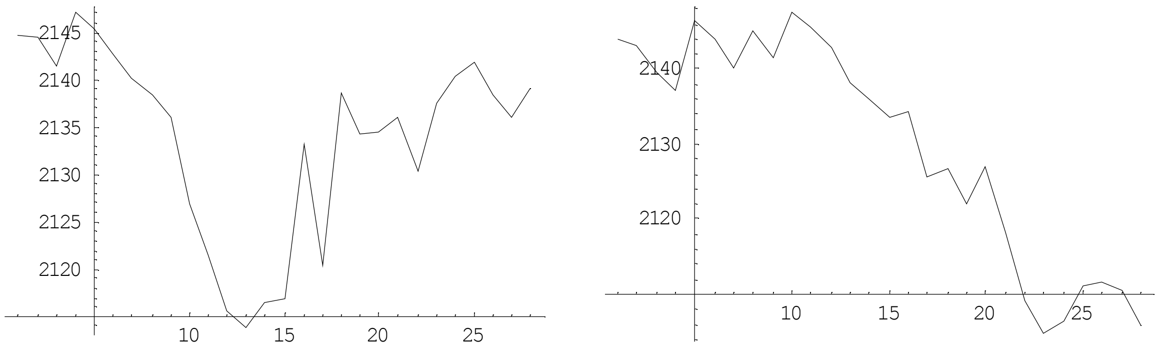


Fig.5: AIC vs. threshold grid index for a) d = 3, b) d = 8

Results of AIC model selection:

p	*r* [mm]	*d* [day]	*AIC*
2	1.91	8	2100
	- 0.30	3	2110
	0.25	1	2120
	- 0.35	1	2121

There's easily seen pretty good agreement among the methods, however still partional and shall be a subject to further study. Basically, I prefer those values confirmed by the majority of demonstrated procedures, rather smaller than higher values... but of course, it should depend on practical expectations at most.

Final results

Model variables and characteristics:

$p = 2$	$d = 1$ day	$r = -0.35$ mm	$s = 2$ regimes	$z_t = y_{1t}$

Parameter matrices:

Φ_1	
- 0.010237	- 0.274496
0.412028	0.0171475
0.005351	0.359622
- 0.017014	- 0.027737
0.053311	0.417337

Φ_2	
0.152080	0.166899
0.226559	0.033515
- 0.108756	0.492913
0.185507	0.041789
0.001387	0.236399

Covariance matrices:

Σ_1 [mm^2]	
4.736	- 0.287
- 0.287	3.194

Σ_2 [mm^2]	
4.399	- 0.898
- 0.898	4.692

Visualization:

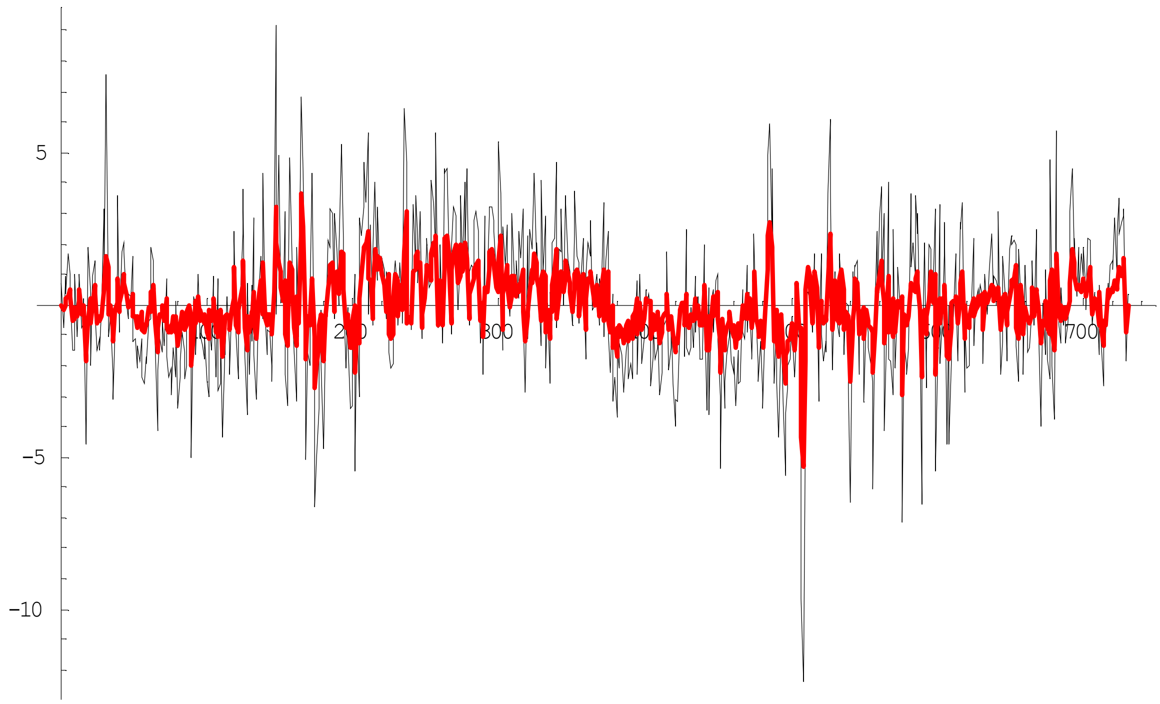


Fig.6 Combined plots of original time series (removed linear trend) and its model: n_{Pecny}

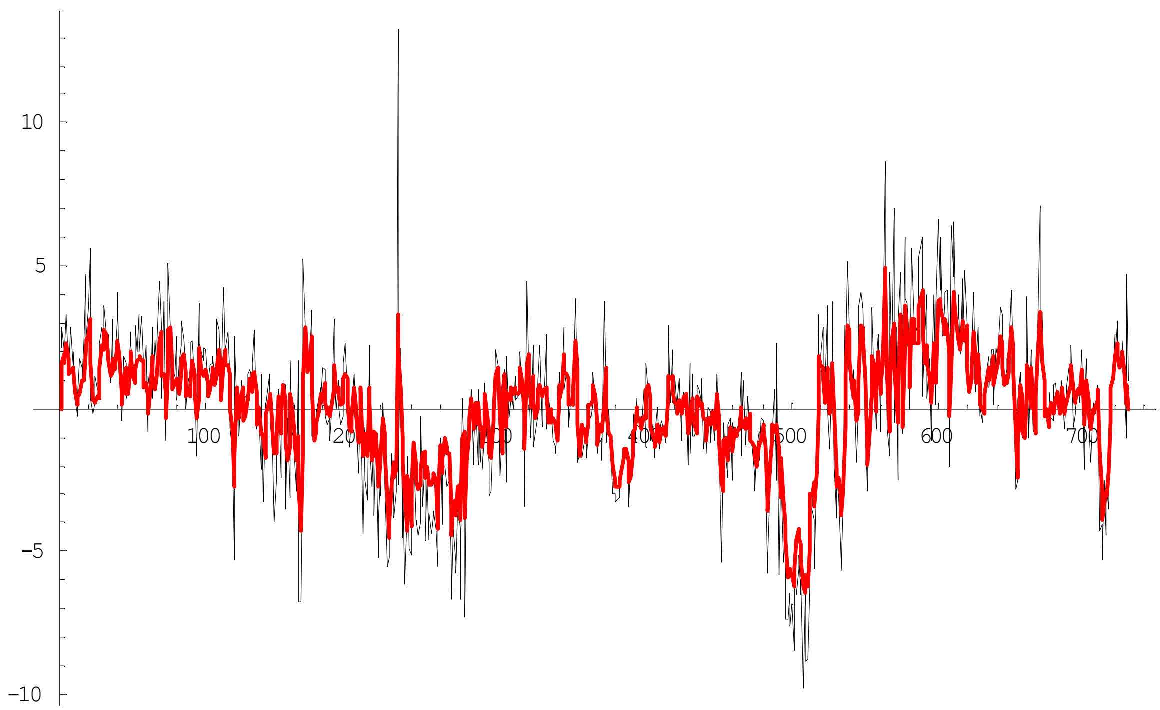


Fig.7 Combined plots of original time series (removed linear trend) and its model: e_{Pecny}

References

[1] Arlt, J., Altová, M.: *Finanční časové řady,* Grada publishing, Praha, 2003.

[2] Bruyninx, C., Kenyeres, A., Takacs, B.: *EPN data and product Analysis for improved velocity estimation: First results*, International Association of Geodesy Symposia, Vol.125, 2001.

[3] Hefty, J.: *The permanent Modra-Piesok GPS station and its long-term and short-term stability*, Slovak Yournal of Civil Engineering, 31-37, 2001.

[4] Tsay, R.S.: *Testing and modeling multivariate threshold models*, Journal of the American Statistical Association, Vol. 93, 1188-1202, 1998.