Journal of ELECTRICAL ENGINEERING, VOL. 55, NO. 12/s, 2004, 91–94

MULTIVARIATE THRESHOLD AUTOREGRESSIVE MODELS IN GEODESY

Tomáš Bacigál *

Recently, the research in time series analysis has changed turning 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 a horizontal coordinate system are taken to obtain bivariate time series, which consequently are tested for nonlinearity and modeled using bivariate threshold autoregressive model. The whole procedure, of course, can easily be generalized to more than two-variate series.

K e y w o r d s: Time series, GPS, multivariate TAR models, AIC, Tsay's test

2000 *Mathematics Subject Classification*: 37M10

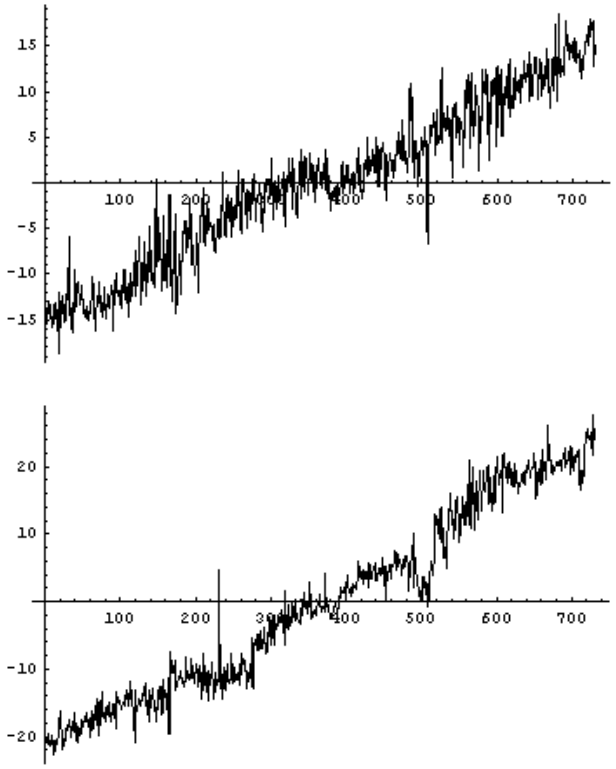

Fig. 1. Two vectors of GPS observations, a) north [mm] and b) east component [mm] of length $n = 730$ days

1 INTRODUCTION

Let us consider time series y of n time-points (Fig. 1). There are several ways to model it. One large family of models, that are strongly suitable for modelling stochastic processes, are those arising from Box-Jenkins methodology such as ARMA etc. [1]. We will be interested in autoregressive (AR) models, defined as

$$y_t = \Phi_0 + \Phi_1\, y_{t-1} + \cdots + \Phi_p\, y_{t-p} + \varepsilon_t\,. \tag{1}$$

This is a linear model and 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 their behaviour by activation of some factor.

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

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

where z is a threshold variable, r is a threshold and their relation delimits constituent regimes of the model. Letter d denotes the 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

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

where $\boldsymbol{y}_t = (y_{1t}\ \ldots\ y_{kt})$, $\boldsymbol{\Phi}_0^{(j)}$ is a constant term for regime j, and y_{kt} denotes k^{th} univariate time series nested in $\boldsymbol{y}_t$.

For $\boldsymbol{y}$ we 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. That is why we have focused on multivariate modelling.

* Department of Mathematics and Descriptive Geometry, Faculty of Civil Engineering, Slovak University of Technology, Radlinského 11, 813 68 Bratislava, Slovak Republic, E-mail: bacigal@math.sk

Research supported by VEGA-grant 1/1033/04.

ISSN 1335-3632

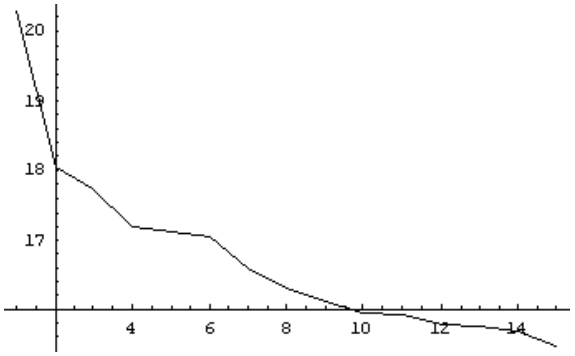

Fig. 2. Determinants of covariance matrices vs. order p

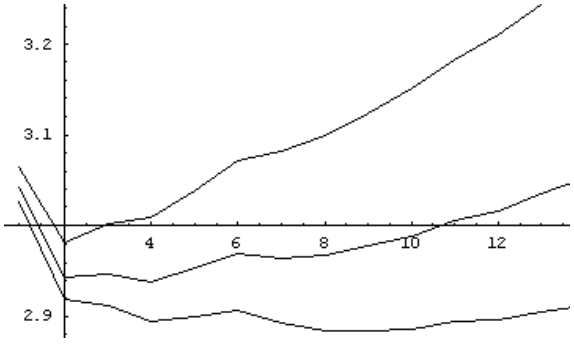

Fig. 3. Information criteria vs. order p

Now, as we have data, the 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 we 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 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.

2 FINDING ORDER OF AUTOREGRESSION

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

a) Using a Levinson-Durbin estimation procedure ($p_{\max}$), especially its outcome — covariance matrices (Fig. 2). Order p is chosen according to plot steepness.

b) Employing three information criteria AIC, BIC, HQIC which are to be minimized by the most appropriate order (Fig. 3).
Order p is chosen as an dominating argument of minimal criteria values.

From the plots in Figs. 2 and 3 $\boxed{p=2}$ seems to be the most adequate.

3 TESTING

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

Following [4], we utilize standard least square regression framework:

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

where $h = \max(p,d)$, $\boldsymbol{X}_t = (1 \quad \boldsymbol{y}_{t-1} \quad \boldsymbol{y}_{t-2}\dots\boldsymbol{y}_{t-p})$ is regressor and $\boldsymbol{\Phi}$ denotes parameter matrix. If H_0 holds, then the 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 $S = \{z_{h+1-d},\dots z_{n-d}\}$ and $t(i)$ is the time index of $z_{(i)}$. Therefore $z_{(i)} = z_{t(i)}$ and the autoregression is

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

It is important to see that the dynamics of the $\boldsymbol{y}_t$ series has not changed (i.e., the independent variable of $\boldsymbol{y}_t$ is $\boldsymbol{X}_t$ 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 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 $\boldsymbol{X}_{t(i)+d}$). Let

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

be the standardized predictive residual of regression (5), where

$$\boldsymbol{V}_m = \left[\sum_{i=1}^{m} \boldsymbol{X}^\top_{t(i)+d}\boldsymbol{X}_{t(i)+d}\right]^{-1}$$

and $\hat{\boldsymbol{\Phi}}_m$ 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

$$\hat{\boldsymbol{\eta}}_{t(l)+d} = \boldsymbol{X}_{t(l)+d}\boldsymbol{\Psi} + \boldsymbol{w}_{t(l)+d}, \quad l = m_0+1,\dots,n-h. \tag{7}$$

where m_0 denotes the starting point of recursive least squares estimation ($m_0 \approxeq 3\sqrt{n}$). The problem of interest is to test the hypothesis H_0: $\boldsymbol{\Psi} = 0$ versus H_1: $\boldsymbol{\Psi} \neq 0$ in (7). Tsay [4] designed a test statistic

$$C(d) = [n-h-m_0-(kp+1)]\times[\ln(\det S_0) - \ln(\det S_1)] \tag{8}$$

where

$$S_0 = \frac{1}{n-h-m_0}\sum_{l=m_0+1}^{n-h} \hat{\boldsymbol{\eta}}^\top_{t(l)+d}\hat{\boldsymbol{\eta}}_{t(l)+d},$$

$$S_1 = \frac{1}{n-h-m_0}\sum_{l=m_0+1}^{n-h} \hat{\boldsymbol{w}}^\top_{t(l)+d}\hat{\boldsymbol{w}}_{t(l)+d},$$

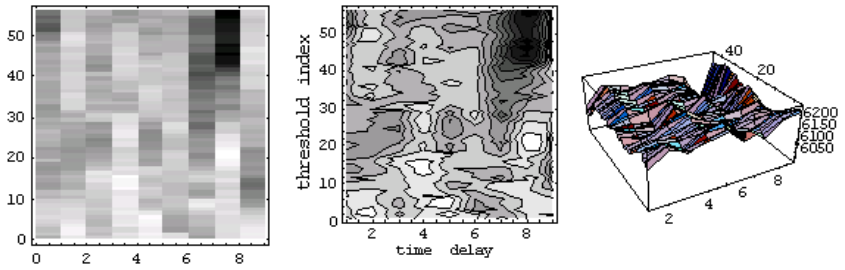

Fig. 4. Density, contour and 3D plot of $S(r,d)$; lower axis represents delay d in days, $r \in \langle -2.6,\ 3.0\rangle$ [mm]

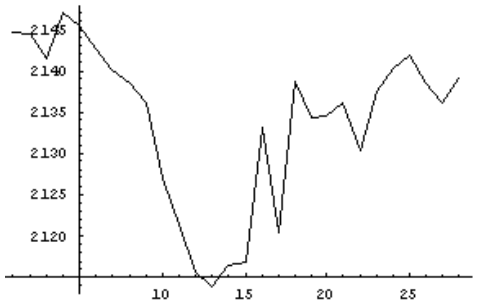

Fig. 6. AIC vs. threshold grid index for $d = 3$

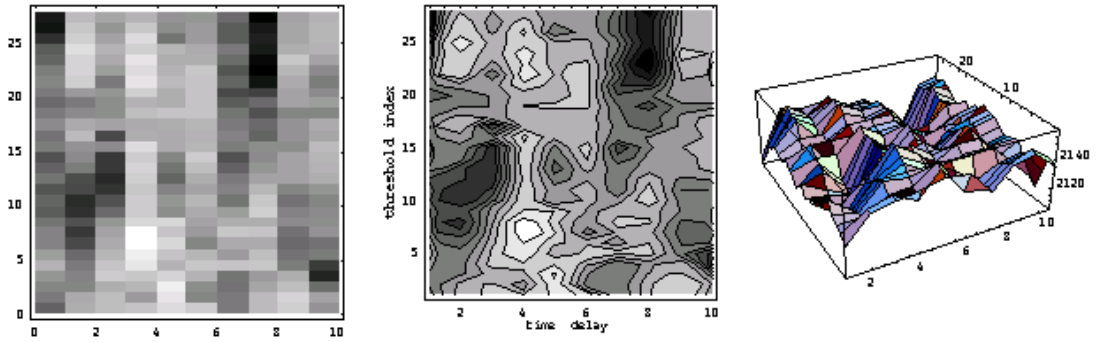

Fig. 5. AIC mapped over grid $r \times d$, $r \in \langle -2.6,\ 3.0\rangle$ [mm], $d \in \{1, 2, \ldots, 10\}$ [day]

Table 1. Results of testing for nonlinearity

p (df)	d	$C(d)$	χ^2		p-value
			$\alpha = 0.05$	$\alpha = 0.01$	
2 (10)	1	29.4	18.3	23.2	0.0010
	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 (18)	1	41.4	28.9	34.8	0.0014
	2	21.1			0.278
	3	30.2			0.035
	4	14.2			0.281
	5	15.6			0.383

Table 2. Results of conditional estimation

p	r [mm]	d [day]	S [mm²]
2	1.89	8	6013.9
	- 0.36	1	6136.5
	- 1.06	1	6137.9
	- 0.35	3	6138.4

and $\hat{\boldsymbol{w}}_t$ is the least square residual of regression (7). Under the null that 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 $C(d) < \chi^2_{df}$, we do not reject the null hypothesis.

N o t e . The test is most powerful, if d is correctly specified.

3 BUILDING UP THE MODEL

First we aim at choosing the best values of 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)

$$\boldsymbol{y}_t = \begin{cases} \boldsymbol{X}_t\boldsymbol{\Phi}_1 + \boldsymbol{\Sigma}_1^{1/2}\,\boldsymbol{a}_t & \text{if } z_{t-d} \le r\,, \\ \boldsymbol{X}_t\boldsymbol{\Phi}_2 + \boldsymbol{\Sigma}_2^{1/2}\,\boldsymbol{a}_t & \text{if } z_{t-d} > r\,, \end{cases} \tag{9}$$

where $\boldsymbol{a}_t = (a_{1t}\ \ldots\ \boldsymbol{a}_{kt}) \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\mathrm{I}})$,
are $(\boldsymbol{\Phi}_i, \boldsymbol{\Sigma}_i, r, d)$. Putting the possible values of r and d into grid $\{1, 2, \ldots d_0\} \times \{r_{\min}, r_{\min} + step, \ldots r_{\max}\}$ model (9) reduces to two separated multivariate linear regressions from which the least squares estimates of $\boldsymbol{\Phi}_i$ and $\boldsymbol{\Sigma}_i$ $(i = 1, 2)$ are readily available:

$$\hat{\boldsymbol{\Phi}}_i(r,d) = \Big(\sum_t^{(i)} \boldsymbol{X}_t^\top \boldsymbol{X}_t\Big)^{-1}\Big(\sum_t^{(i)} \boldsymbol{X}_t^\top \boldsymbol{y}_t\Big), \tag{10}$$

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

where $\sum_t^{(i)}$ denotes summing over observations on regime i, $\hat{\boldsymbol{\Phi}}_i^* = \hat{\boldsymbol{\Phi}}_i(r,d)$, n_i is the number of data points in regime i and k $(k < n_i)$ the dimension of $\boldsymbol{X}_t$. It becomes clear that conditional least squares estimates of r and d should minimize the sum of squares of residuals

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

Table 3. 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

Table 4. Model variables and characteristics

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

Table 5. Parameter and covariance 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

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

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

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

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

$$AIC(p,s,d,r) = \sum_{(j=1)}^{s} [n_j \ln(\det \hat{\mathbf{\Sigma}}_j) + 2k(kp+1)] \quad (13)$$

with

$$\hat{\mathbf{\Sigma}}_j = \frac{1}{n_j} \sum_{t}^{(j)} \hat{\boldsymbol{\varepsilon}}_t^{(j)\top} \hat{\boldsymbol{\varepsilon}}_t^{(j)},$$

where n_j is the number of data points in regime j, $\sum_t^{(j)}$ denotes summing over observations in regime j and $\hat{\boldsymbol{\varepsilon}}_t^{(j)}$ are residuals.

Pretty good agreement between these two methods is easily seen. However, they shall be the subject of further study. Basically, we prefer values confirmed by the majority of demonstrated procedures, rather smaller than higher values, *etc.* But, of course, the choice of a method should depend also on practical expectations, see [2] [3].

4 FINAL RESULTS

Respecting all previous results, the final shape of model has been selected, built up and is shown in Tabs. 4 and 5 and visually compared with original data in Fig. 7. However, decision is not so easy and some comparisons to other methods and confrontation with practical purposes are needed.

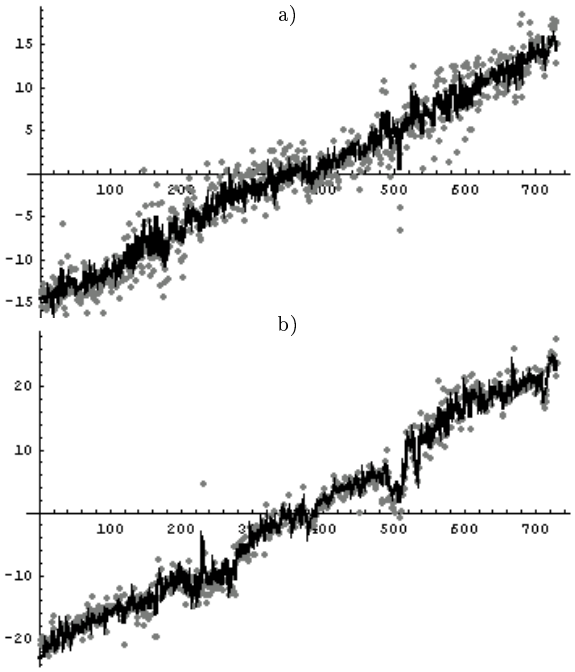

Fig. 7. Visualized fit of the built model. Original data are represented by dotted, model by joined plot of a) north and b) east component of horizontal coordinate system vs. time. [mm vs. days]

Here we have shown one possible way of processing of geodetic data that may be extended to three-or-more-regimes models and models including some exogenous variables. Our major contribution to the application of time series analysis in geodesy is treating the data as set of mutually depending variables effectively describable by multivariate modelling approach rather than by the univariate one.

Acknowledgement

The author gratefully acknowledges many helpful suggestions of Professor Magda Komorníkova.

References

[1] ARLT, J.— ARLTOVÁ, M.: Fiantial Time Series (Finanční časové řady), Grada publishing, Praha, 2003. (in Czech)

[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 Journal of Civil Engineering (2001), 31–37.

[4] TSAY, R.S.: Testing and Modeling Multivariate Threshold Models, Journal of the American Statistical Association **93** (1998), 1188–1202.

Received 3 June 2004

Tomáš Bacigál (Ing) is a PhD student. His PhD-thesis supervisor (in applied mathematics) is Professor Magda Komorníková.