Outlier detection using some methods of mathematical statistic in meteorological time-series

Michal Elias and Jan Dousa

Citation: AIP Conference Proceedings **1738**, 480059 (2016); doi: 10.1063/1.4952295

View online: http://dx.doi.org/10.1063/1.4952295

View Table of Contents: http://scitation.aip.org/content/aip/proceeding/aipcp/1738?ver=pdfcov

Published by the AIP Publishing

Outlier Detection Using Some Methods of Mathematical Statistic in Meteorological Time-Series

Michal Elias[1,2,a)] and Jan Dousa[1,b)]

[1] *Research Institute of Geodesy, Topography and Cartography, Department of Geodesy and Geodynamics, Ústecka 11, Zdiby, Czech Republic*
[2] *Czech Technical University in Prague, Faculty of Civil Engineering, Department of Mathematics, Thákurova 7, 166 29 Prague 6, Czech Republic*

[a)] Corresponding author: michal.elias@pecny.cz
[b)] jan.dousa@pecny.cz

Abstract. In many applications of Global Navigate Satellite Systems the meteorological time-series play a very important role, especially when representing source of input data for other calculations such as corrections for very precise positioning. We are interested in those corrections which are related to the troposphere delay modelling. Time-series might contain some non-homogeneities, depending on the type of the data source. In this paper the outlier detection is discussed. For investigation we used method based on the autoregressive model and the results of its application were compared with the regression model.

INTRODUCTION

Our research deals with an effective modelling of tropospheric delay corrections in relation with the GPS applications. The satellite geodesy includes several systems and methods (GPS, DORIS, VLBI, etc.) and all of them have at least one thing in common. The techniques are based on propagation of electromagnetic signal through gas medium called the atmosphere [5].

The troposphere represents the lowest layer of the atmosphere and the delay can be simply illustrated as the difference of the signal that should have been propagated in vacuum and signal which is propagated in the analyzed medium.

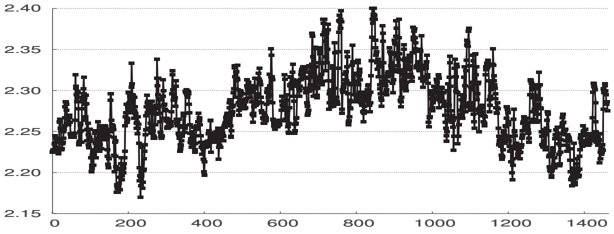

FIGURE 1. Input data source for outlier investigation

Figure 1 shows the typical state of the tropospheric delay corrections during one year. The X-axis includes indexes of the time (so-called time span). The shown data were recorded every six hours and the time span includes more than 1460 observations. The Y-axis of the graph gives the values of the tropospheric delay corrections which represent studying parameter discussed in this paper.

We focus on the situation when the so-called outliers are explored. Generally, the outlier candidate is defined as an extreme observation that differs from the other observations of the statistical sample. The majority of the methods related to the outliers detection are based on criterion of relative distance between the suspected point and the mean

International Conference of Numerical Analysis and Applied Mathematics 2015 (ICNAAM 2015)
AIP Conf. Proc. 1738, 480059-1–480059-4; doi: 10.1063/1.4952295
Published by AIP Publishing. 978-0-7354-1392-4/$30.00

value of the statistical sample. The situation with time series data seems to be more complex. It may occur that the "outliers" in time series are not necessarily "outliers". According to the paper [3], sometimes the outlier is not a globally extreme point and may be hidden in a marginal view of the data. For the purpose of this paper, we used two methods: the first approach is based on the hat matrix technique [3]. The second method is based on the regression model and the rule of $X \times \sigma$ application [4].

LINEAR AUTOREGRESSIVE MODEL AND OUTLIER DETECTION

Let the autoregressive sequence of order k, hereafter denoted by AR(k) [4], be defined as

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_k X_{t-k} + \varepsilon_t \tag{1}$$

where ε_t's are i.i.d. and $\varepsilon_t \sim \left(0, \sigma^2\right)$. Let us assume

$$\mathbf{z}_t^T = (X_{t-1,}, X_{t-2}, \cdots, X_{t-k})$$

and

$$\mathbf{\Phi}^T = (\varphi_1, \varphi_2, \cdots, \varphi_k)\,.$$

Then the equation (1) can be rewritten into the form

$$X_t = \mathbf{z}_t^T \mathbf{\Phi} + \varepsilon_t. \tag{2}$$

Let $\mathbf{X}$ be the vector of observations $X_1, X_2, \cdots, X_n$. We get

$$\mathbf{X} = \mathbf{\Gamma}\mathbf{\Phi} + \varepsilon \tag{3}$$

where ε is the vector of random errors and

$$\mathbf{\Gamma} = \begin{bmatrix} \mathbf{z}_1^T \\ \mathbf{z}_2^T \\ \vdots \\ \mathbf{z}_n^T \end{bmatrix}$$

is the so-called design matrix.

The goal is to estimate vector $\hat{\mathbf{\Phi}}$ of $\mathbf{\Phi}$ which can be simply computed using conventional Least-Square procedure. The Least-Square fitted residuals $\hat{\varepsilon}$ are given as

$$\hat{\varepsilon} = \mathbf{X} - \hat{\mathbf{X}}, \tag{4}$$

where $\hat{\mathbf{X}}$ are the fitted values. Note that $\hat{\mathbf{X}} = \mathbf{\Gamma}\hat{\mathbf{\Phi}} = \mathbf{\Gamma}\left(\mathbf{\Gamma}^{\mathbf{T}}\mathbf{\Gamma}\right)^{-\mathbf{1}}\mathbf{\Gamma}^{\mathbf{T}}\mathbf{X} = \mathbf{H}\mathbf{X}$ and $\mathbf{H} = [h_{ij}]$ is known as the hat matrix in the context of outliers. The Least-Square estimate of $\hat{\sigma}^2$ is given by

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n}. \tag{5}$$

By stationarity of $\{X_t\}$, it holds that $n \to \infty$, $n^{-1}\left(\mathbf{\Gamma}^{\mathbf{T}}\mathbf{\Gamma}\right) \to \mathbf{\Sigma}$ where $\mathbf{\Sigma}$ is Toeplitz matrix [3]. Let us define $d_t = \mathbf{z}_t^T \mathbf{\Sigma}^{-1} \mathbf{z}_t$ for $t = 1, 2, \cdots, n$. Then d_t is the Mahalanobis distance between z_t and the zero vector. Note also that

$$nh_t = \mathbf{z}_t^T \left(\frac{\mathbf{\Gamma}^T\mathbf{\Gamma}}{n}\right)^{-1} \mathbf{z}_t \to d_t \quad \text{as} \quad n \to \infty, \tag{6}$$

where h_t denotes the t-the diagonal element of the hat matrix. This means that we can interpret nh_t as the Mahalanobis distance between the vector $\mathbf{z}_t$ and the mean values vector (zero vector).

For outliers detection, now let us have the realization of size n, $X_1, X_2, \cdots, X_n$, from autoregressive AR(k) process. From the state space point of view, at time t, it is the relative position of $X_t, X_{t-1}, \cdots, X_{t-k+1}$ in the k dimensional space that we are interested in, not just the X_t itself. Therefore, it is reasonable to refer to the state vector, i.e. "remote" state vectors. Geometrically speaking, we look for remote points in the k-dimensional space spanned by the columns of $\mathbf{\Gamma}$

[3]. For outlier $\mathbf{z}_t$ detection we can use distance d_t. Under the hypothesis that the autoregressive process is Gaussian and there is no outlier, it holds that for every $t = k+1, \cdots, n \;\; d_t \sim \chi^2_k$. Therefore, by equation (6), nh_t leads itself as a useful measure for outlier detection within the linear Gaussian context (this result is not valid for the conventional regression analysis,[3]). If nh_t is larger than a critical value of χ^2_k, we can conclude that $\mathbf{z}_t = (X_{t-1}, X_{t-2}, \cdots X_{t-k})$ is the outlier vector.

If we suppose that the value X_{t-1} is large, then the vectors $\mathbf{z}_t, \mathbf{z}_{t+1}, \cdots, \mathbf{z}_{t+k-1}$ are influenced by this parameter and the numbers $h_t, h_{t+1}, \cdots, h_{t+k-1}$ are large as well. Therefore, if h_{t-1} is small and h_t is large, we may identify the value X_{t-1} as the outlier in the observation vector.

ORDER ESTIMATION OF *AR(k)* MODEL

The described procedure was applied to detect some possible outliers in the meteorological time-series. For this purpose the autoregressive model of the order k is considered. We tested three different approaches:

1. Order estimation based on a partial autocorrelation function ρ_{kk}. The estimates r_{kk} of the partial autocorrelation function ρ_{kk} are given in the recurrent form

$$r_{11} = r_1, \tag{7}$$

$$r_{kk} = \frac{r_k - \sum_{j=1}^{k-1} r_{k-1,j} r_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-1,j} r_j} \quad \text{for} \quad k > 1 \tag{8}$$

and where

$$r_{k,j} = r_{k-1,j} - r_{kk} r_{k-1,k-j} \quad \text{for} \quad j = 1, 2, \cdots, k-1.$$

For an autoregressive process of order k, the partial autocorrelation function ρ_{kk} will be non-zero for k less than or equal to k and zero for k greater than k. Quenouille suggested an approximation $\sigma(r_{kk})$ for r_{kk} [1].

2. Order estimation based on the white noise variance estimation given by the equation (5).
3. Order estimation based on the informative criterion methods [2]. We used, for example, the AIC criterion developed by Akaike which is based on information theory and has a form

$$AIC(k) = \log \hat{\sigma}^2_k + \frac{2k}{n}. \tag{9}$$

We also used BIC criterion. This criterion can be given as

$$BIC(k) = \log \hat{\sigma}^2_k + \frac{k \log n}{n}. \tag{10}$$

and the last criterion based on a principle of iterated logarithm is known as HQ criterion and can be given in a form

$$HQ(k) = \log \hat{\sigma}^2_k + ck \frac{\log(\log n)}{n}, \tag{11}$$

where c is constant (usually $c > 1$).

NUMERICAL RESULTS DISCUSSION

We can demonstrate the outliers detection in meteorological time-series using method described in the theoretical section. It was necessary to determine the order of autoregressive model. We decided to use three approaches as were presented in this paper. The statistical hypothesis testing partial autocorrelation r_{kk} shows that the order k should have been $k = 2, 3, 4$ or 11. However, Figure 2(a) shows other coefficients periodically repeated. The white noise variance estimations plot (Figure 2(b)) shows that the estimated values vary around the same value between the orders 3 to 10. Then we can see a small decrease from order 11 and the same effect we can see on the order 27. Higher orders we do not cover in our analysis. The most objective method how to choose the order of AR model is to use the informative criterion methods. Based on the results showed in Figure 3(a), the minimal value of BIC criterion is given for order $k = 3$. Both the AIC criterion and HQ determine that the minimal value for the order equals $k = 27$.

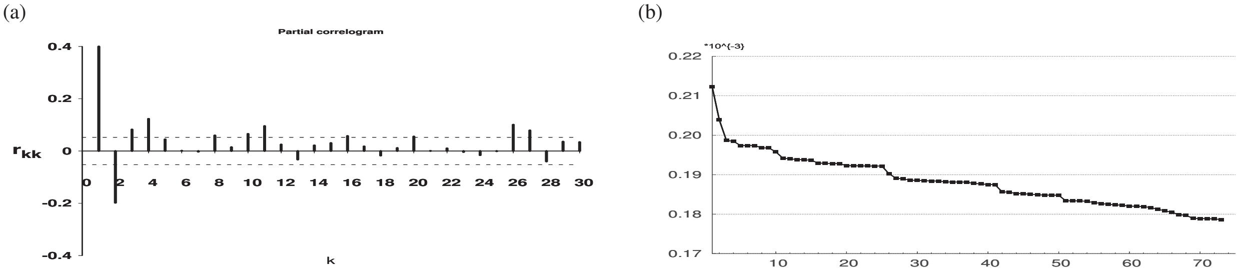

FIGURE 2. On the left plot we see partial autoregressive model. Figure on the right shows the result of white noise variance estimation for the study parameter.

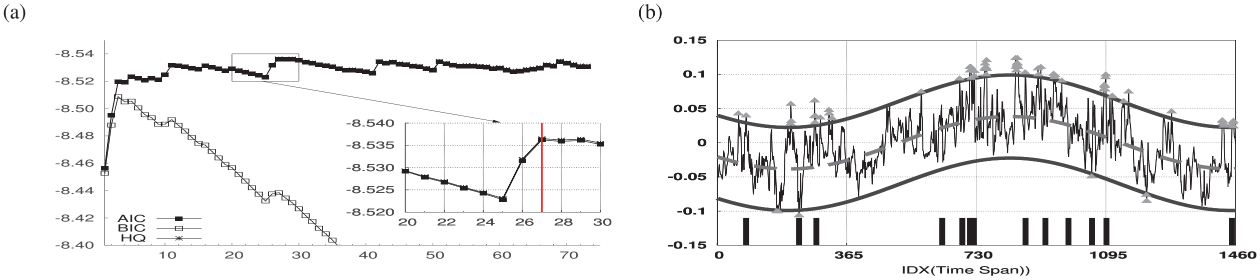

FIGURE 3. The left plot shows the progress of informative criterion methods. On the right there are displayed the results of outlier identification. The plot comprises fitted regression model and the critical borders as given in $2 \cdot \sigma$ rule. The triangles represent the outlier candidates when regression model was accepted. Vertical lines point to those observables which were detected as outliers using the procedure explained in theoretical part.

For the meteorological time-series introduced in the first part of the paper we decided to use model AR(27). We identified the outliers using the procedure described in theoretical part (showed in Figure 3(b) as bold vertical lines): $X_{82}, X_{231}, X_{280}, X_{634}, X_{691}, X_{712}, X_{723}, X_{869}, X_{925}, X_{990}, X_{1056}, X_{1097}, X_{1450}$.

The alternative method based on regression model fitting was demonstrated as well [4]. We approximated input data by approximation model

$$f_t = -0.0325\sin(2\pi t/1249.2) - 0.0203\cos(2\pi t/1249.2),$$

when Least-Square procedure was applied. Based on the obtained results we proved that this chosen method might act as a full-valued replacement for the regression model.

ACKNOWLEDGEMENTS

This work was supported by the project COST-CZ (LD14102)

REFERENCES

1. G.E.P. Box, G.M. Jenkins, G.C. Reinsel: Time Series Analysis. Forecasting and Control. Wiley, 2008.
2. Ch. Chen, R.A. Davis, P.J. Brockwell, Z.D. Bai: Order Determination for Autoregressive Processes Using Resampling Methods. In: Statistica Sinica 3, pp. 487-500, 1993.
3. M.C. Hau, H. Tong: A Practical Method for Outlier Detection in Autoregressive Time Series Modelling. In: Stochastic Hydrol. Hydraul. 3, pp. 241-260, 1989.
4. M. Mudelsee: Climate Time Series Analysis. Classical Statistical and Bootstrap Methods. Springer, 2014.
5. G. Seeber: Satellite Geodesy: foundations, methods and applications. Walter de Gruyter, 1993.