CHANGE POINTS DETECTION AND MODELS FOR PRECIPITATION EVOLUTION. CASE STUDY

ALINA BĂRBULESCU[1] and J. DEGUENON[2]

[1] Technical University of Civil Engineering, Doctoral School Bucharest, Lacul Tei Bvd. 122-24, 020396, Bucharest, Romania, E-mail: alinadumitriu@yahoo.com
[2] UPAC, EPAC, 01 BP 2009 Cotonou, Benin, E-mail: tjudy73@yahoo.fr.

Received December 4, 2013

We discuss the problem of change point detection for the monthly precipitation series collected two meteorological stations in Dobrudja region and we model the series trend before and after the break points.

Key words: change point, decomposition model, trend, seasonality.

1. INTRODUCTION

The study of precipitation evolution is important for a country like Romania where the alimentary security depends on the water supply. It is known that desertification affects the soil fertility, with productivity losses up to 50% in some areas, determining food insecurity, and rising social, economic and political tensions. Downward pluvial trend was noticed in western Africa, from the end of 1960s and China for the period 1954–1976 [1–5]. Land degradation and drought affect many parts of Central and Eastern Europe. A decrease between 5 and 20% of the total annual quantity of precipitation was observed in Romania, but its northern and north-eastern regions, for which an increase between 5 and 10% of the total annual amount of precipitation was predicted. In this context, the water resources study and its management is of high importance to design irrigation systems and for the water supply, in a region where the drought period is between five and six months per year, as Dobrudja [6–7]. Therefore, the objective of the current article is twofold: the determination of change points for precipitation series registered at two meteorological stations and building models for the sub-series. Different statistical tests and have been performed to detect change points in the time series, and subsequently, models have been built for the identified sub-series. Two distinct ways were followed: the series decomposition and a non-parametric approach.

2. INPUT DATA AND METHODOLOGY

Dobrudja is situated in the south-eastern part of Romania, between the Danube River and the Black Sea, between 27°15'05"–29°30'10" eastern longitude and 43°40'04"–45°25'03" northern latitude. The data was collected in the period Jan 1965–Dec 2005, at Mangalia and Sulina, which are the northernmost and southernmost meteorological stations in the region.

The first objective of the study was the change points detection in precipitation series. This problem has a rich history in the literature [8–12]. It appears often in hydrology, being solved by classical tests, segmentation [13–15] and CUSUM procedures [16]. A problem that arises in the change point selection is the choice of the true segmentation, when different methods provide different results. If the method is based on the calculation of means of the neighbours segments, the outliers' presence can affect the segmentation. An outlier is defined as an observation that deviates so much from the others as to arouse suspicion that it was generated by a different mechanism [17] or a value that deviates markedly from other sample's elements in which it occurs [18]. For a non- Gaussian variable, a boxplot may be suggestive in outliers'detection.

After founding the change points, the next step is the determination of models for the sub-series delimited by these points. Classical and modern methods may be used [19-20] and the seasonality implications may be analysed.

Before the modeling process, preliminary analyses have been performed and the results are presented in [21]. The change points detection have been done by the Pettitt test, the segmentation procedures of Hubert [14] and mDP [13]. The results have been compared to those obtained by the frequentist algorithm of Bai and Perron [22], implemented in the R package strucchange [23]. The Bayesian algorithm [24], implemented in the R package bcp [25] completes the study, detecting the posterior mean and probability of a change for each position in a sequence. We shall refer to these algorithms respectively by BH and BP.

Then, a multiplicative model has been built for the series (y_t):

$$y_t = Y_t \cdot S_t \cdot \varepsilon_t, \tag{1}$$

where: Y_t is the trend, S_t – the seasonal component, ε_t – the random component.

Taking into account the seasonality, (1) becomes:

$$y_{ij} = Y_{ij} \cdot S_j^* \cdot \varepsilon_{ij}^*, \tag{2}$$

where: $i \in \overline{1, p}$ is the period number and $j \in \overline{1, m}$ - the sub-period number.

If $t \in \overline{1, n}$, then $t = m(i-1) + j$, for $p \cdot m = n$.

The steps in this method are [26]: (1) the trend determination, by an analytical method or by the moving average method; (2) the seasonal factors calculation by the moving average method and (3) the determination of residual, by:

$$\varepsilon_{ij}^{*} = y_{ij} /(Y_{ij} \cdot S_{j}^{*}) . \tag{3}$$

The method can also be applied by inter - changing the first two steps.

The second approach utilized the wavelets technique (with the hard threshold estimator) to solve the nonparametric regression problem of estimating a function f on the basis of observations y_i at time points, t_i, modelled as

$$y_i = f(t_i) + \varepsilon_i, \tag{4}$$

where ε_i is a noise.

Using an orthonormal wavelets basis for $L^2(R)$ [27], $\{\psi_{j,k}\}_{j,k\in Z}$, any squared integrable function can be represented by:

$$f(x) = \sum_{k=-\infty}^{+\infty} \alpha_{0k} \phi_k(x) + \sum_{j=0}^{\infty} \sum_{k=-\infty}^{\infty} \beta_{jk} \psi_{jk}(x), \tag{5}$$

$$\phi_k(x) = 2^{1/2} \phi(2x-k), \ \psi_{jk}(x) = 2^{j/2} \phi(2^j x - k), \tag{6}$$

$$\alpha_{0k} = \int_{-\infty}^{+\infty} f(x) \phi_{0k}(x) dx, \ \beta_{jk} = \int_{-\infty}^{+\infty} f(x) \psi_{jk}(x) dx. \tag{7}$$

where ϕ is a scaling function, ψ is the mother wavelet.

The steps followed in the nonparametric regression procedure were [5]: the detection of preliminary estimate, the shrinkage and the function reconstruction.

3. RESULTS

3.1. CHANGE POINTS DETECTION

The preliminary analysis leads us to the rejection of the normality hypothesis for the brute series, but its acceptance for both series after a Box-Cox transformation. So, only the Pettitt test, the segmentation and the BP procedures have been performed to test the null hypothesis: H_0: "There is no break in the time series" *vs*. its alternative: H_1: "There is at least a break point in the time series".

All tests were performed at a confidence level of 95%.

For *Mangalia* series, after performing the Pettitt test, H_0 was accepted. The segmentation procedure of Hubert, with Scheffe's test at the significance level of 1%, leaded to the rejection of H_0 and providee the segments: Jan 1965–Aug 2005 (the mean = 35.97), Sept 2005 and Oct 2005–Dec. 2005 (the mean = 47.83). It is clear that the value registered in Sept 2005 (331 mm) is an outlier, being more than six times greater than the maximum average of the other segments.

The mDP algorithm, with the Scheffe's selection criterion, at the significance level of 1%, provided Aug and Oct 2005 as change points. Using the Bayesian Information Criteria (BIC) for the segments selection, the results didn't change. By BP procedure, the change points were also Aug and Oct 2005.

For *Sulina series,* H_0 was rejected after performing the Pettitt test; it was found that Aug 1982 is a break point. By the procedure of Hubert, the segments detected are: Jan 1965–July 1972 (the mean = 27.45), Aug 1972–Sept 1972 (the mean = 102.75) and Oct 1972–Dec 1972 (the mean = 21.341) and Oct 1972 – Dec 2005. The average of the 2nd segment is more than three times higher than those of the neighbouring segments, so we think that the values from Aug 1972 and Sept 1972 are outliers. mDP detected Aug and Sept 1971, July and Sept 1972 as change points (Fig. 1(a)). For comparison reasons, the significance level in Scheffe's test was kept the same as in the procedure of Hubert. Using BIC for the segments selection, the result didn't change. The break points detected by the BP are July 1972 and Sept 1972.

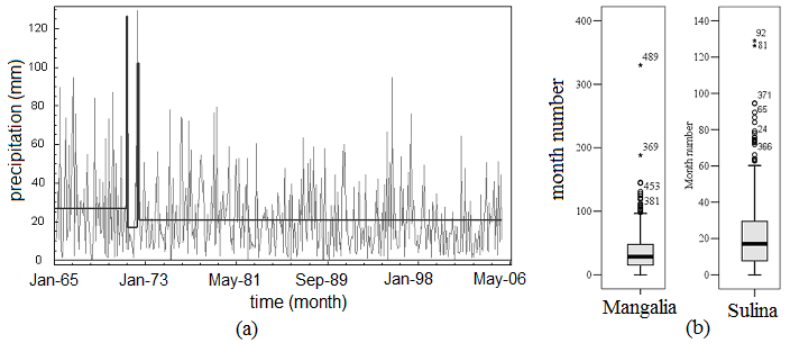

Fig. 1 – (a) Change points obtained by mDP, implemented in segmenter.21 (http://web.itu.edu.tr/~gedikliab/Segmenter/) and (b) the boxplot.

The boxplots analysis (Fig. 1(b)) confirms 330 mm (Sept 2005), respectively 126.3 mm (Sept 1971) and 129 (Aug 1972) as outliers for Mangalia, respectively Sulina. They correspond to the points 489, for Mangalia, 92 and 81, for Sulina. So, Sept 2005, respectively Sept 1971 and Aug 1972 can not be accepted as change points for Mangalia, respectively Sulina series. There is no evidence that Aug 1982 (40.2 mm) is an outlier for Sulina, so it is accepted as a change point. The result of the Pettitt test (Aug 1982) for Sulina, is concordant with that of mDP procedure.

Another approach to decide if the values from Sept 1971, and Aug 1972 are outliers is to remove them from Sulina series and to perform again the tests. Proceeding this way, the break point detected by all methods was Aug 1982.

Summarizing, the results of the break tests are presented in Table 1.

Table 1

Break points of Mangalia and Sulina series

Station	Procedure	Break points			
Mangalia	Pettitt	-			
	Hubert	Aug. 2005	Oct. 2005		
	mDP	Aug. 2005	Oct. 2005		
	BP	Aug. 2005	Oct. 2005		
Sulina	Pettitt	Aug. 1982			
	Hubert			July 1972	Sept 1972
	mDP	Aug. 1971	Sept.1971	July 1972	Sept.1972
	BP			July 1972	Sept.1972

The similarities of the last three tests were expected because all are based on analogous optimization principles. Different break points were detected by the Pettitt test because this is a nonparametric procedure, based on the comparison of the sum of the ranks of the components of each sub-sample in the total sample.

Sulina series, (Z_t), is not Gaussian, but the series (y_t), obtained after the Box-Cox transformation $y_t = (Z_t^{0.34} - 1)/0.34$, is Gaussian. Performing the same tests, as well as the Buishand, Lee & Heghinian tests and BH Bayesian procedure, Aug 1982 was accepted as a break point of (y_t). The BH and mDP procedures confirm the result. Computing the z - scores for (y_t), we decided that the value registered in Aug 1982 is not an outlier, so it can be accepted as a break point.

3.2. MODELS

Since no break point has been detected for *Mangalia* series, Sept 2005 is an outlier and only 3 values are registered after this data, three ways can be followed: (1) to model the entire series, (2) to remove the outlier and to build a model, (3) to remove the outlier and the data after it and to design a model. Since the outlier's presence affects the model quality, the first way was excluded. Models have been built for the last alternatives. Since they do not differ too much, we present only the third alternative (Jan 1965–Aug 2005), which is the best in term of residual.

The equation of the trend has been determined:

$$Y_t = 34.38 + 3.72 \cdot \cos(0.02\,t - 2.33), \tag{8}$$

where t is the (month), numbered from 1 to 488 (Fig. 2(a)).

The seasonality factors (Table 2) vary between 71.5% and 122.8%. The months most affected by the seasonal variations were Sept and Nov and the less affected one was February. A third of months (May, Dec, Aug, July) has the seasonality factors close to one, proving a relative stability in the seasonal variation especially at the beginning of Summer and in December. The residuals have been obtained by the equation (2), taking into account (8) and the seasonality factors; they are plotted in Fig. 2(b). Their values range between 0 and 4.52, with a variance of 0.574. The highest correspond to Aug 1984, Sept 1995, June 1983, Jan 1966, Apr 1997, when precipitation over 100 mm have been registered. This means that the highest rainfall is not very well fitted by the model, even if the residuals are homoskedastic, independent, with a small variance.

Table 2

Seasonality factors (%) for Mangalia series

Month	January	February	March	April	May	June
Factor	83.6	71.9	84.8	92.6	99.4	115.0
Month	July	August	September	October	November	December
Factor	104.1	101.8	122.4	105.8	118.2	100.7

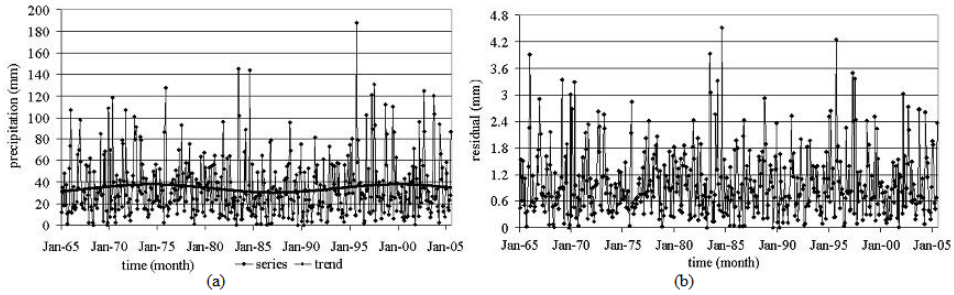

Fig. 2 – (a) Mangalia series and its trend given by (8); (b) residual in the decomposition model.

Nonparametric methods can provide accurate models of data analysis because they make minimal assumptions about the data - generating process. Wavelets provide a spatial frequency resolution, useful in smoothing problems, in particular in density and regression estimation, having excellent statistical properties in data smoothing. They offer a frequency/time representation of data allowing adaptive filtering, reconstruction and smoothing.

The trend of Mangalia series obtained by the wavelets (hard threshold) method is presented in Fig. 3, where the trend determined by (8) is also represented for comparison. The wavelets model captures very well the extreme values and the outliers. For example, the picks near Aug 1981 mark local minimum values, as March 1983 (0.24 mm), June 1983 (122 mm) etc., the highest pick close to April 1998 mark the precipitation of July 1997 (128 mm) etc.

We remark the similarities in the trend given by (8) and those determined by wavelets, if the oscillations around the minimal and maximal values are ignored. The differences between them are due to the characteristics of sin/cos waves and of wavelets family that may be described by the localization property: the sine wave is localized in frequency domain, but not in time domain, while a wavelet is localized both in frequency and time domain.

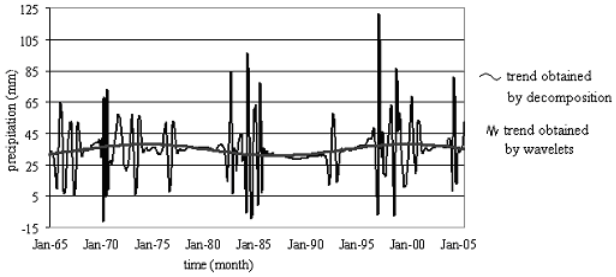

Fig. 3 – The trend of Mangalia series obtained by wavelets method and by equation (8).

The models of *Sulina* series have been built for (y_t), defined by $y_t = (Z_t^{0.34} - 1)/0.34$. The trend in the model (2) is:

$$Y_t = 4.67 + 0.53\cos(0.0077t - 0.36), \tag{9}$$

where t is the time (Fig. 4).

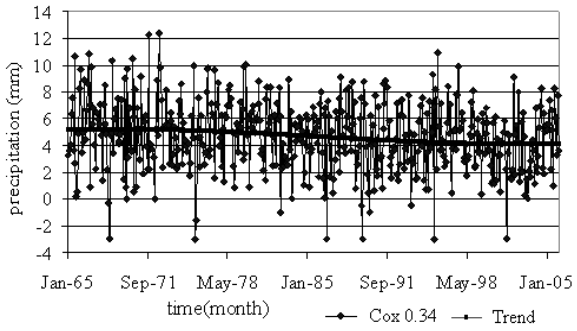

Fig. 4 – Sulina series, after the Box-Cox transformation and its trend given by (9).

From Table 3, 2nd column, we remark that the months most affected by seasonal variations were June and Sept, and the less affected one was March. The residual in the model (2), with the trend (9) and the seasonality factors have a small variance (0.26), is normally distributed and homoskedastic.

The sub-series Jan 1965–Aug 1982 is also Gaussian and homoskedastic, after the same Box-Cox transformation. This sub-series will be called Sulina 1. The trend in the model (2) (Fig. 5(a)) has been determined by the moving average

method. The residuals (Fig. 5(b)) have a variance of 0.31 and are not Gaussian, independent or homoskedastic. July and Aug were the month most affected by seasonal variations. The residuals (Fig. 5(b)) have a variance of 0.31 and are not Gaussian, independent or homoskedastic. July and Aug were the month most affected by seasonal variations.

Table 3

The seasonality factors (%) of Sulina series

Month	Jan 1965-Dec 2005	Jan1965-Aug 1982	Sept 1982 -Dec2005
1⁰	2⁰	3⁰	4⁰
January	87.25	95.5	117.5
February	89.58	96.6	103.2
March	86.79	80.9	114.1
April	90.82	91.3	97.6
May	103.08	109.6	84.0
June	124.47	106.8	87.4
July	98.37	117.1	87.3
August	105.92	113.1	90.8
September	110.48	110.2	97.6
October	91.00	69.4	144.9
November	108.08	99.3	87.9
December	104.12	110.1	87.7

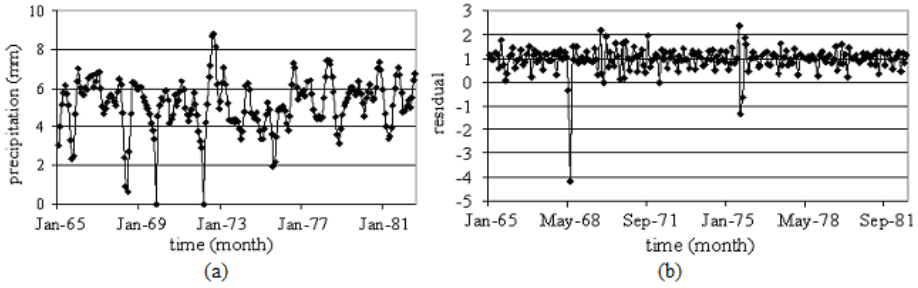

Fig. 5 – (a) The trend in the model (2) for Sulina 1; (b) The residual in model (2) for Sulina 1.

After the same Box-Cox transformation for the sub-series Sept 1982–Dec 2005, the trend of the resulted series, Sulina 2, has been detected by the moving average method (Fig. 6(a)); the corresponding residual (Fig. 6(b)) is Gaussian, homoskedastic and uncorrelated, with a variance of 0.32. In this case, the most affected by the seasonal variations were Oct and Jan.

The wavelets models for Sulina, Sulina 1 and 2 are given in Fig. 7. The model of Sulina is periodic. Some periodicity is also noticed for the trend of Sulina 2. The highest peaks in the models appear where the extreme values were registered. We remark that the extreme values captured by the model for Sulina can be found between the extreme values captured by the models for Sulina 1 and 2.

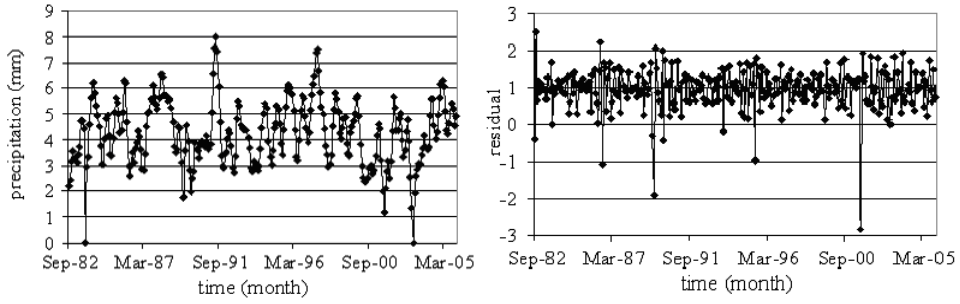

Fig. 6 – The trend and the residual in the multiplicative model for Sulina 2.

The residual standard deviations in the last models were respectively 2.54, 3.75, 2.80, proving that the wavelets method is an interesting alternative to the previous one. Its main advantage is that it is highly adaptive to irregular signals as well as smooth ones. Also, the wavelets decomposition of time series into different scales provides an interpretation of the series structure and extracts the significant information about its former behaviour, using a small number of coefficients, and giving a stable prediction of the series evolution.

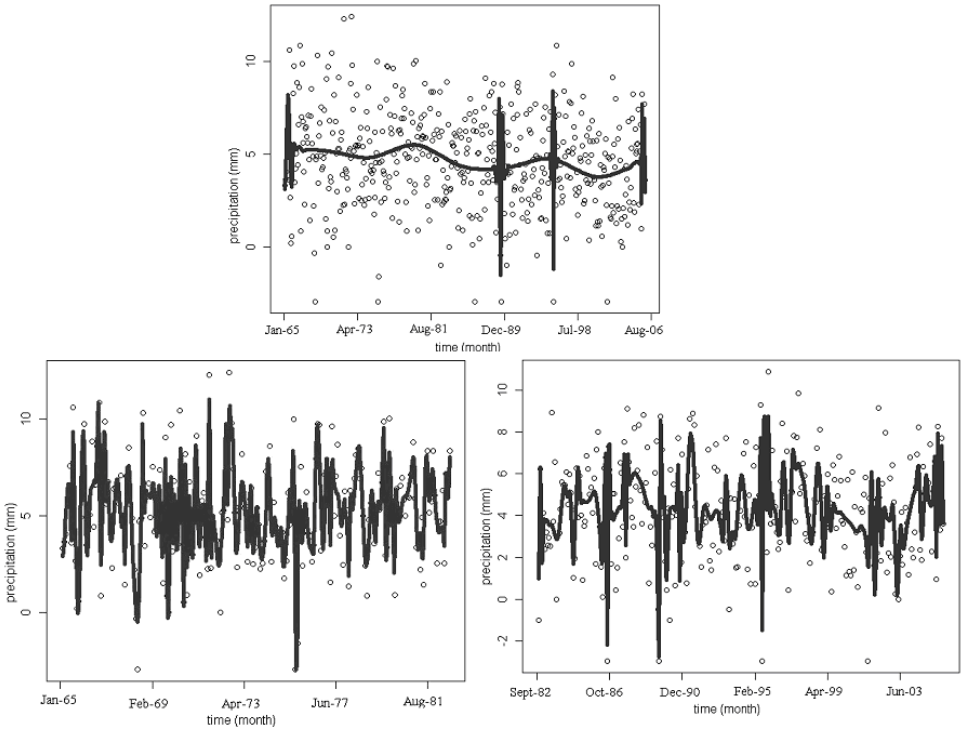

Fig. 7 – Wavelets models for Sulina, Sulina 1 and Sulina 2.

4. CONCLUSIONS

Two types of inter-related problems have been studied: the change point detection and the models determination for two precipitation series. Methods based on dynamic programming and Bayesian procedures have been used for change point detection. The first group of methods estimates the change point moment and the second one gives the probability of a change point in each location of a sequence. The true change point can be selected performing them together.

It has been proved that the presence of very high outlying values can influence the change point detection. Between the non - bayesian procedures, the best performances have been registered by the mDP algorithm, which has the advantage to calculate the optimal segmentation order corresponding to the change points and the possibility of choice between Scheffe's test and BIC for the points' selection. This allow the comparison to other algorithms that use different selection criteria (as Scheffe's one for Hubert procedure and BIC for BH).

The equation of the trend proposed in the first approach for both series was of cosine type. The analysis of the charts of the models determined by non-parametrical methods confirms the existence of a cyclical component in the evolution of precipitation series. To describe the series trend, the first method is suitable, but if one wants to locally look at its variability in time, the non – parametrical method provides more details. Another advantage of the second approach is that it captures the periodical variations without the series decomposition. If a good model can not be found by the first approach, the second one could give good results. Therefore, to analyze the series variations, one can start with the non-parametrical method, which could help to determine the form of an analytical function describing the global trend. Even if the wavelets model is a model of trend detection, that permits to describe the series evolution before and after the change point, it will be interesting to study the relation between the break points existence and the wiggles in the wavelets model.

REFERENCES

1. J. P. Bricquet *et al.,* PHI - V **6**, 83–95 (1997).
2. D. Y. Gong, S.W. Wang, Clim Res **16**, 51–59 (2000).
3. G. Mahe, J.C. Olivry, Sécheresse **6**(1), 109–117 (1995).
4. W. H. Qian, Y. Zhu, Clim Change **50**, 419–444 (2001).
5. C. Schafer, L. Wasserman, 2008, astrostatistics.psu.edu/samsi06/tutorials/tut2larryl_all.pdf
6. E. M. Carstea *et al.*, Rom. Rep. Phys., **65**, (3), 1092–1104 (2013).
7. R. Dumitrache *et al.*, Rom Rep Phys **63**(1), 208–219 (2011).
8. B. Brodsky, B. Darkhovsky, *Nonparametric Methods in Change - Point Problems*, Springer - Verlag, New York. 1993.
9. B. P. Carlin, A. E. Gelfand, A.F.M. Smith, Appl stat **41**, 389–405 (1992).
10. J. Chen, A. Gupta, *Parametric Statistical Change Point Analysis*, Birkhauser Verlag, 2000.

11. H. Chernoff, S. Zacks, Ann. Math. Statist. **35**, 999–1018 (1964).
12. B. Ray, R. Tsay, J Time Ser Anal **23**, 687–705 (2002).
13. A. Gedikli et al., Stoch Env Res Risk A **24**(5), 547–557 (2010).
14. P. Hubert, J. P. Carbonnel, A. Chaouche, J Hydrol **110**, 349–367 (1989).
15. A. Kehagias, E. Nidelkou, V. Petridis, Stoch Environ Res Risk A **20**(1-2), 77–94 (2006).
16. P. Liu *et al.*, Hydrolog Sci J **55**(4), 540–554 (2010).
17. D. Hawkins, *Identification of Outliers*, Chapman and Hall, London – New York, 1980.
18. Barnett, T. Lewis, *Outliers in Statistical Data*, John Wiley & Sons, New York, 1994.
19. V. Kumar, S. K. Jain, Y. Singh, Hydrolog Sci J **55**(4), 484–496 (2010).
20. M. Slimani, C. Cudennec, H. Feki, Hydrolog Sci J **52**(6), 1088–1102 (2007).
21. A. Bărbulescu, C. Șerban (Gherghina), C. Maftei, WSEAS T Math **10**(9), 791–800 (2010).
22. J. Bai, P. Perron, J Appl Econom **18**, 1–22 (2003).
23. A. Zeileis *et al.*, strucchange: A Package for Testing, monitoring and dating structural changes in (linear) regression models. R package version 1.4-0, 2010.
24. D. Barry, J. A. Hartigan, J Am Stat Assoc **35**(3), 309–319 (1993).
25. C. Erdman, J. W. Emerson. bcp: A Package for Performing a Bayesian Analysis of Change Point Problems. R package version 1.8.4, 2007.
26. A. Bărbulescu, *Time series with applications*, Junimea, Iasi, 2002 (in Romanian)
27. I. Daubechies, *Ten lectures in wavelets*, SIAM, Philadelphia PA, 1992.