Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v9n2p267

Change-point detection in environmental time series based on the informational approach
By Costa, Gonçalves, Teixeira

Published: 14 October 2016

Electronic Journal of Applied Statistical Analysis
Vol. 09, Issue 02, 2016, 267-296
DOI: 10.1285/i20705948v9n2p267

Change-point detection in environmental time series based on the informational approach

Marco Costa*[a], A. Manuela Gonçalves[b], and Lara Teixeira[c]

[a]*Escola Superior de Tecnologia e Gestão de Águeda & Centro de Investigação e Desenvolvimento em Matemática e Aplicações, Universidade de Aveiro, Apartado 473, 3754 - 909 Águeda, Portugal*
[b]*Departamento de Matemática e Aplicações & Centro de Matemática, Universidade do Minho, Campus de Azurém - 4800-058 Guimarães, Portugal*
[c]*Faculdade de Ciências da Universidade do Porto, Rua do Campo Alegre 1021/1055, 4169-007 Porto, Portugal*

Published: 14 October 2016

In this study, the Schwarz Information Criterion (SIC) is applied in order to detect change-points in the time series of surface water quality variables. The application of change-point analysis allowed detecting change-points in both the mean and the variance in series under study. Time variations in environmental data are complex and they can hinder the identification of the so-called change-points when traditional models are applied to this type of problems. The assumptions of normality and uncorrelation are not present in some time series, and so, a simulation study is carried out in order to evaluate the methodology's performance when applied to non-normal data and/or with time correlation.

keywords: change-point detection, water quality data, Schwarz Information Criterion, mean and variance shift, simulation study.

1. Introduction

Statistical methodologies are applied in many practical contexts in order to identify changes in a sequence of chronologically ordered data. Usually the change-point analysis presents two goals. The first is to detect if there is any change in the sequence of

*Corresponding author: marco@ua.pt.

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

observed random variables. The second is to estimate the number of changes and their corresponding locations (Chen and Gupta, 2012). The problem of detecting and analyzing change-points is associated with different behavioral changes in the time series that may occur: for instance, changes in the mean, in the variance, both in the mean and in the variance, and also a change-point in regression models coefficients. A description of the various types of change-points can be found in Chen and Gupta (2012) and in Beaulieu et al. (2012).

Several methods with different approaches have been developed to tackle the problem of change-point analysis. In a non-parametric approach, Hájek (1962) used tests ranks for changes in a regression model. The change-point problems through a Bayesian-type approach were studied by Chernoff and Zacks (1964). Likelihood ratio statistics for testing for changes in the mean have been discussed by Hawkins (1977), and later by Worsley (1979) (with known and unknown variance). An informational approach model, the so-called Schwarz Information Criterion (SIC), was developed by Schwarz (1978) to detect the change-point in means and variances in a sequence of normal random variables. The statistic test and its approximate distribution for the multiple changes in the mean vector for a sequence of normal random vectors was derived by Srivastava and Worsley (1986). The corresponding problem of changes in the regression model has also been studied (Krishnaiah and Miao, 1988). Most methods are mainly based on both the normal model and the temporal uncorrelation, and don't take into account other possible time series properties such as seasonality and non-stationarity. There has been significant progress in multiple change-point detection through penalties more advanced than the BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion) variants. The work Caussinus and Mestre (2004) proposes a simple BIC-like multiple change-point penalty that is based on the total number of change-points. The problem of modeling a class of nonstationary time series using piecewise autoregressive (AR) processes where the minimum description length principle is applied to compare various segmented AR fits to the data is analyzed in Davis et al. (2006). These ideas are taken further in environmetrics applications in Li and Lund (2012) and in Lu et al. (2010).

Environmental data analysis has been gaining a great relevance due to the increasing human activity exerted on nature, and so the use of differentiated methodologies for the assessment of the impact and changes that have been occurring is pertinent and essential for the management of the various problems resulting from these sustainability issues. Within this context, it is proposed a work plan that aims at contributing to the construction of more fruitful answers for the problems identified in the implementation of environmental quality management systems. Change-point detection in this type of systems has been frequently reported as extremely relevant in the processes of decision-making by the competent entities.

In the environmental area, the change-point technique analysis has been widely used, particularly in the context of the problems associated with an exhaustive exploration of nature and its consequences. For instance, Lund and Reeves (2002) studied the annual average temperature of Chula Vista, California, and monthly carbon dioxide concentrations reported at Mauna Loa, Hawaii, and Jarušková (2010) studied the monthly temperature average in Stockholm. Regarding air pollution, Barratt et al. (2007) stud-

ied changes in ambient mean air pollution levels following the introduction of a traffic management scheme at Marylebone Road, Central London, and Jarušková (1996) analyzed air pressure time series at Swiss meteorological stations. Changes in a long-term annual dataset by measuring maximum precipitation in South Taiwan were studied by Chu et al. (2012).

In Economics and Finance areas, several change-point studies can also be found, such as Inclán and Tiao (1994), who analyzed data series of International Business Machines (IBM) stock prices, and Hsu (1977) who investigated the potential impacts of the Watergate events on the United States stock market.

This study focus on the informational approach, a general model technique selection that can be adapted to a diverse set of situations, namely to detect the change-point in both the mean and the variance of water quality variable.

The data concerns the Ave River basin situated in Northwest Portugal, where monitoring has become a priority in water quality planning and management in this watershed. The economic base of the Ave Valley is strongly linked to the industry, since water is a key factor in industrial location, but this industrialization has led to poor water quality since the mid-1970s. The variable that will be analyzed is DO, one of the most important variables for assessing surface water quality (Costa and Gonçalves, 2011; Gonçalves and Costa, 2011), measured monthly from January 1999 to December 2011, in eight monitoring sites.

In this work, the study of the time series of Dissolved Oxygen water quality variable is addressed in line with the research of Gonçalves and Costa (2011) and Gonçalves and Alpuim (2011), who recently studied trend alterations in environmental variables, including time series of water quality variables. Nevertheless, the application of change-point techniques can be more fruitful because the instant of a change in a time series may be not known and its estimation can be important.

2. Study area and data set description

The data was collected from the Portuguese National Information System for Water Resources (SNIRH) that was created by the Institute of Water (INAG) and is related to the Ave River basin. The Ave River basin is located in Northwest Portugal (Figure 1). The hydrological basin covers an area of 1400 km^2, of which about 247 Km^2 and 340 km^2 correspond, respectively, to the areas of the basins of its two main adjacent streams, the Este River and the Vizela River.

The Ave River develops in the general east-west direction and runs about 101 km from its source in Serra da Cabreira to its mouth in Vila do Conde, creating a wide and complex basin. In the Ave River basin, water courses present, in general, serious disturbances (both physico-chemical and biological), except near the springs, resulting in poor water quality, which in turn has obvious repercussions on aquatic communities. This situation is mainly due to the strong pressure exerted by urban households that are scattered throughout the basin. The Ave River basin region has an economy highly dependent on the industry, and water has played an important role in this valley's

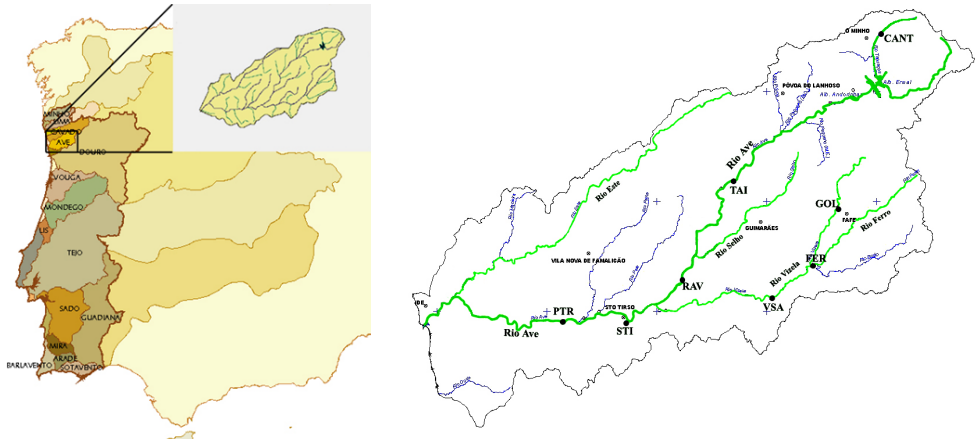

Figure 1: Spatial representation of the Ave River hydrological basin.

industry (mainly textiles and garment). A main reason for the extreme pollution of these waters is that the construction of infrastructure in order to control and avoid pollution has not accompanied the industrial development.

In this study, eight water monitoring sites are considered. These eight monitoring sites result from the restructuring of the water quality monitoring network in 1998 (Table 1). Its spatial representation is shown in Figure 1.

Table 1: Water quality monitoring sites.

Water stream	Monitoring site	Designation
Ave River	Taipas	TAI
	Cantelães	CANT
	Riba d'Ave	RAV
	Santo Tirso	STI
	Ponte Trofa	PTR
Ferro River	Ferro	FER
Vizela River	Golães	GOL
	Vizela (Santo Adrião)	VSA

The variable analyzed is DO, measured in milligrams per liter (mg/l), which is one of the most important indicator variables in determining the pollution degree in a water course. Organic matter oxidation, photosynthesis and respiration are transformation processes that significantly affect this variable. The larger the amount of dissolved oxygen, the better the water quality. The dataset used concerns the period from January

1999 to December 2011.

Table 2 summarizes descriptive statistics, as well as the number of missing values for the monthly measurements of the DO water quality variable at the 8 monitoring sites during the above-mentioned period. The monitoring sites Riba d'Ave, Santo Tirso and Ponte Trofa present mean values slightly lower when compared with the remaining five monitoring sites, thus reflecting poor water quality. These sites present both the lowest mean values and the largest DO dispersion.

Table 2: Descriptive statistics and missing values number of DO in 8 monitoring sites.

Monitoring site	Range	Average	Standard deviation	Skewness	Missing values
CANT	$7.40 - 12.80$	9.76	1.03	0.20	6
TAI	$6.60 - 11.72$	9.34	1.11	-0.04	5
RAV	$1.80 - 11.70$	8.50	1.70	-0.73	1
STI	$1.67 - 12.00$	8.28	2.04	-0.87	2
PTR	$2.40 - 11.70$	8.06	1.85	-0.73	2
FER	$7.30 - 11.70$	9.54	1.06	0.01	4
GOL	$7.00 - 11.70$	9.46	1.06	0.02	5
VSA	$7.20 - 12.40$	9.57	1.11	0.22	5

In an exploratory analysis of the observed values of DO it is indicated the possibility of changes in the mean and/or variance (in particular between 2004 and 2006). As regards the average, it apparently increases or decreases according to the monitoring site, but the observations variability reduces in all monitoring sites, more evidently on some of them. Another important feature is the indication of a seasonal component. This is due to the seasonal relationship between DO concentration and the weather patterns throughout the year, particularly temperature changes and precipitation intensity.

3. Methods

As said before, the change-point statistical inference has two goals: the first is to detect if there is any change in the sequence of observed random variables; the second is to estimate the number of changes and their corresponding locations. The detection of multiple change-points can be performed through the binary segmentation procedure. The binary segmentation procedure and its consistency were presented in Vostrikova (1981).

For instance, the binary segmentation procedure was used in Chen (1998) to search the existence of various change-points in the Boston Stock Exchange monthly sales volume,

and has the advantage of detecting simultaneously the number of change-points and their location.

3.1. General problem formulation

Consider a sequence of independent random variables $X_1, X_2, \ldots, X_n$ with probability distribution functions $F_1, F_2, \ldots, F_n$, respectively. Then, in general, the change-point problem is to test the following null hypothesis,

$$H_0 : F_1 = F_2 = \ldots = F_n \tag{1}$$

versus the alternative hypothesis

$$H_1 : F_1 = \ldots = F_{k_1} \neq F_{k_1+1} = \ldots = F_{k_2} \neq F_{k_2+1} = \ldots = F_{k_q} \neq F_{k_q+1} \ldots = F_n, \tag{2}$$

where $1 < k_1 < k_2 < \ldots < k_q < n$, q is the unknown number of change-points and $k_1, k_2, \ldots, k_q$ are the respective unknown positions that have to be estimated.

If the distributions $F_1, F_2, \ldots, F_n$ belong to a common parametric family $F(\theta)$, where $\theta \in \mathbb{R}^p$, then the change-point problem is to test the null hypothesis about the population parameters θ_i, $i = 1, \ldots, n$:

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_n = \theta \quad \text{(unknown)} \tag{3}$$

versus the alternative hypothesis

$$H_1 : \theta_1 = \ldots = \theta_{k_1} \neq \theta_{k_1+1} = \ldots = \theta_{k_2} \neq \ldots \neq \theta_{k_{q-1}+1} = \ldots = \theta_{k_q} \neq \theta_{k_q+1} = \ldots = \theta_n, \tag{4}$$

where q and $k_1, k_2, \ldots, k_q$ have to be estimated.

These hypotheses together reveal the aspects of change-point inference: determining if any change-point exists in the process and estimating the number and position(s) of change-point(s) Chen and Gupta (2001). Note also that the hypothesis can be adapted where there is a single change-point or multiple change-points in the sequence of observations. Next sections are dedicated to the detection of one change-point in a time series which may be applied in a binary segmentation procedure.

3.2. Change-point in both the mean and the variance

Change-point in both the mean and the variance sometimes occurs. This problem has not been widely discussed and only the most recent studies present examples. For instance, Chen and Gupta (1999) used the informational approach and Hawkins and Zamba (2005) used the likelihood ratio statistics.

In case of a change-point both in the mean and in the variance, we want to test the following hypothesis

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_n = \mu \quad \wedge \quad \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2 \tag{5}$$

versus the alternative hypothesis

$$H_1 : \mu_I = \ldots = \mu_1 = \mu_k \neq \mu_{k+1} = \ldots = \mu_n = \mu_{II}$$

$$\wedge \tag{6}$$

$$\sigma_I^2 = \sigma_1^2 = \ldots = \sigma_k^2 \neq \sigma_{k+1}^2 = \ldots = \sigma_n^2 = \sigma_{II}^2.$$

For instance, assuming the normality, the model representing a shift in both the mean and the variance can be expressed as

$$X_t = \begin{cases} \mu_I + \epsilon_t^I, & \epsilon_t^I \sim N(0, \sigma_I^2), \quad t = 1, \ldots, k \\ \mu_{II} + \epsilon_t^{II}, & \epsilon_t^{II} \sim N(0, \sigma_{II}^2), \quad t = k+1, \ldots, n, \end{cases} \tag{7}$$

where μ_I and σ_I^2 are the mean and the variance before the unknown change-point and μ_{II} and σ_{II}^2 are the mean and the variance after the unknown change-point.

3.3. The informational approach

The Akaike Information Criterion (AIC) was introduced by Akaike (1973) for model selection in Statistics. The general formulation of the AIC to select the "best" model among M models can be expressed by

$$\text{AIC}_j = -2 \ln L(\hat{\Theta}_j) + 2p_j, \quad j = 1, 2, \ldots, M, \tag{8}$$

where $L(\hat{\Theta}_j)$ is the maximum likelihood for model j, as a measure of model evaluation, $\hat{\Theta}_j$ is a estimate for Θ_j, set of parameter values for model j, and p_j is the number of parameters that are estimated in model j. A model that minimizes the AIC (Minimum AIC estimate, MAICE) is considered to be the most appropriate model. However, the MAICE is not an asymptotically consistent estimator of model order (Schwarz, 1978). This criterion has profoundly influenced the developments in statistical analysis, particularly in time series, outliers analysis (Kitagawa, 1979), robustness, regression analysis, multivariate analysis (Bozdogan et al., 1994). Based on Akaike's work, many authors have further introduced various information criteria (Bozdogan, 1987; Rao and Wu, 1989).

One of the AIC modifications is the Schwarz Information Criterion (SIC), proposed by Schwarz (1978). The SIC is defined as following

$$\text{SIC}_j = -2 \ln L(\hat{\Theta}_j) + p_j \ln n, \quad j = 1, 2, \ldots, M, \tag{9}$$

where n is the sample size. This criterion is based on the maximum likelihood of a given model penalized by the number of parameters that are estimated in the model. Also, the model that minimizes the SIC is considered to be the most appropriate model, representing the best compromise between parsimony (few parameters) and good fit (small residuals).

Apparently, the difference between AIC and SIC is in the penalty term: instead of $2p$, it is $p \ln n$. However, SIC gives an asymptotically consistent estimate of the order of the true model and makes use of the sample information (Chen and Gupta, 2012).

In short, the informational approach (in this case by using the Schwarz Information Criterion (SIC)) intends to identify the change-point through the identification of the model that minimizes the SIC, that is, the model considered to be the most appropriate in comparison with the model with no change-point. In this setting we considered two models: one corresponding to the null hypothesis (5) and another to the alternative hypothesis (6).

Under H_0 (5) and assuming the normality and the independence, the maximum likelihood estimators for μ and σ^2 are $\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ respectively. Then, denoting SIC under null hypothesis (5) by SIC(n), we have

$$\mathrm{SIC}(n) = -2\ln\, L_0(\hat{\mu}, \hat{\sigma}^2) + 2\ln n, \tag{10}$$

and the maximum of the likelihood function is given by

$$L_0(\hat{\mu}, \hat{\sigma}^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left[-\frac{(X_i - \hat{\mu})^2}{2\hat{\sigma}^2}\right]. \tag{11}$$

The number two in the second term of the equation (10) represents the number of parameters to estimate: the mean and the variance. Combining (10) through (11), we thus have

$$\mathrm{SIC}(n) = -2\sum_{i=1}^{n}\left\{\ln\left[\frac{1}{\sqrt{\frac{2\pi}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}} \exp\left(-\frac{(X_i - \overline{X})^2}{\frac{2}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}\right)\right]\right\} + 2\ln n \tag{12}$$

and then, after some simple computations, we obtain

$$\mathrm{SIC}(n) = n\ln 2\pi + n\ln\sum_{i=1}^{n}(X_i - \overline{X})^2 + n + (2-n)\ln n. \tag{13}$$

Under the alternative hypothesis (6) it is necessary to estimate four parameters: two means and two variances, before and after the change-point. Under alternative hypothesis the SIC, denoted by SIC(k), is hence obtained as

$$\mathrm{SIC}(k) = -2\ln\, L_1(\hat{\mu}_I, \hat{\mu}_{II}, \hat{\sigma}_I^2, \hat{\sigma}_{II}^2) + 4\ln n. \tag{14}$$

The maximum likelihood function is given by

$$L_1(\hat{\mu}_I, \hat{\mu}_{II}, \hat{\sigma}_I^2, \hat{\sigma}_{II}^2) = \prod_{i=1}^{k}\left\{\frac{1}{\sqrt{2\pi\hat{\sigma}_I^2}} \exp\left[-\frac{(X_i - \hat{\mu}_I)^2}{2\hat{\sigma}_I^2}\right]\right\} \prod_{i=k+1}^{n}\left\{\frac{1}{\sqrt{2\pi\hat{\sigma}_{II}^2}} \exp\left[-\frac{(X_i - \hat{\mu}_{II})^2}{2\hat{\sigma}_{II}^2}\right]\right\}. \tag{15}$$

After some algebraic simplifications, holds

$$\mathrm{SIC}(k) = n\ln 2\pi + k\ln\hat{\sigma}_I^2 + (n-k)\ln\hat{\sigma}_{II}^2 + n + 4\ln n, \tag{16}$$

where $\overline{X}_I = \frac{1}{k}\sum_{i=1}^{k} X_i$, $\overline{X}_{II} = \frac{1}{n-k}\sum_{i=k+1}^{n} X_i$ and $\hat{\sigma}_I^2 = \frac{1}{k}\sum_{i=1}^{k}(X_i - \overline{X}_I)^2$, $\hat{\sigma}_{II}^2 = \frac{1}{(n-k)}\sum_{i=k+1}^{n}(X_i - \overline{X}_{II})^2$. The SIC$(k)$ serves as the test statistics for the model selection.

3.4. Model Selection

According to the information criterion principle, we are going to estimate the position of change-point k such that SIC(k) is the minimal. Notice that in order to obtain the maximum likelihood estimators, we can only detect changes that are located between the second and $(n-2)$ positions. Then, the estimation of the change-point position by $\hat{k}$ is given by

$$\mathrm{SIC}(\hat{k}) = \min_{2\leq k\leq n-2} \mathrm{SIC}(k). \tag{17}$$

Chen and Gupta (1997) presented a theorem and its proof which states that the estimator of k according to (17) is consistent to the true change-point. Furthermore, they presented some properties of the test statistics SIC(k), in particular the characteristic function, the mean and the variance.

The model with a change-point k is selected if $\mathrm{SIC}(k) < \mathrm{SIC}(n)$, otherwise, the model with no change-point SIC(n) is more reasonable. The information criteria, such as SIC, provides a remarkable way for exploratory data analysis with no need to resort to either the distribution or the significance level α. However, when SIC(k) and SIC(n) are very close, we can question if the small difference between SIC(k) and SIC(n) might be caused by data fluctuation, and therefore there may be no change at all. In order to investigate the significance of the change-point (Chen and Gupta, 1997) introduced the significance level α and its associated critical value c_α, where $c_\alpha \geq 0$.

Thus, the null hypothesis should be rejected if

$$\min_{2\leq k\leq n-2} \mathrm{SIC}(k) + c_\alpha < \mathrm{SIC}(n) \tag{18}$$

where c_α and α are computed such that

$$1-\alpha = P\Big[\mathrm{SIC}(n) < \min_{2\leq k\leq n-2} \mathrm{SIC}(k) + c_\alpha | H_0\Big]. \tag{19}$$

The computation of the critical values c_α needs the null distribution of $\min_{2\leq k\leq n-2} \mathrm{SIC}(k)$. Such a distribution, however, still remains unknown so far. Under the null hypothesis, the asymptotic critical values can be computed through the approximate formula

$$c_\alpha \approx -2\ln n + \left\{ -\frac{1}{a(\ln n)} \ln\ln \left[1-\alpha+\exp\Big(-2\exp\big[b(\ln n)\big]\Big)\right]^{-1/2} + \frac{b(\ln n)}{a(\ln n)} \right\}^2, \tag{20}$$

where $a(\ln n) = (2\ln\ln n)^{1/2}$ e $b(\ln n) = 2\ln\ln n + \ln\ln\ln n$. For different significance values c_α and different sample sizes n, Chen and Gupta (1999) computed the c_α values, and the approximate c_α values are tabulated in that work.

4. Change-point detection procedure: an application to a real dataset

The change-point methods are often applied to detect changes in environmental time series, which are complex and hinder the process of change-point detection. The detection of changes in time series includes the behavior knowledge of the studied variable

over time. A common feature of environmental series is the significant time dependence on the observations (autocorrelation), particularly at the monthly scale or at a smaller time scale. The presence of an autocorrelation creates patterns in time series that can be easily confused with change-points, especially if the magnitude of the change-point is small (Jarušková, 1997). In the presence of dependence, the risk of false detection tends to increase and the power of detection diminishes (Beaulieu et al., 2012). However, the effect of serial correlation on the distributions of change-point statistics was studied in some scenarios. Inferences about the change-point in regression models with AR(1) error structure were considered in El-Shaarawi and Esterby (1982). When the variables are not independent but form an ARMA sequence, Antoch et al. (1997) showed that the asymptotic critical values in the CUSUM approach have to be multiplied by $\sqrt{2\pi f(0)/\gamma}$, where γ is the variance and $f(\cdot)$ denotes the special density of the corresponding ARMA process. A method for undocumented change-point detection for series with autocorrelated and periodic features based on a regression F-type statistics was developed by Lund et al. (2007). The detection of changes of one variable observed in a set of locations but analyzed as an independent series was performed in Jarušková and Rencov (2008) and in Alpuim and El-Shaarawi (2009). The change-point detection method of Antoch et al. (1997) was incorporated by Gonçalves and Alpuim (2011) in a state-space modeling in order to identify possible changes of water quality variables.

The monthly environmental series are often skewed. The data transformations must be done very carefully because they can eliminate important data behaviors, which may lead to changes that prevent the change-point detection or the acceptance of existing change-points that do not exist (false change-point detection). Statisticians analyze the hydrometeorological data often transforming the observations, usually by using the logarithmic transformation (Jarušková, 1997). Sometimes the interpretation of a change in parameters causes problems as the mean and the variance of log-normal distribution are functions of both parameters. An example of monthly averages of water discharges of a small creek in the Erzgebirge Mountains was presented in Jarušková (1997) in order to show that tests detect a change in the mean but not in the variance of the transformed data, concluding that the scale characteristic changed but the shape characteristic of the original series remained the same. Thus, variations in these time series can be easily misinterpreted and result in identifying apparent changes, even though there are none. This is a challenging problem in change-point detection, as most techniques were developed for independent observations. In the same paper, the author advises to deal, when possible, with the annual averages instead of monthly averages, because the problem of skewness is usually not so serious (averaging reduces data skewness).

In this work, the main goal is to study the DO time series and to perform the change-point analysis by using the SIC procedure. It will be carried out a change-point analysis for the eight DO time series, at each monitoring site, in order to understand whether the changes suggested by the exploratory analysis performed – relatively to change-points in both mean and variance – are statistically significant or only due to inherent data variation (associated with random hydrological phenomena). As the datasets under study present a seasonal behavior, a special attention was required by addressing this characteristic through the use of linear models. This approach arises as a strategy to

tackle the problem of the presence of a seasonal component in time series, especially in environmental data analysis. Generally, the hydrometeorological data consists of monthly observations, and they usually have a seasonal character (for instance, this can be explained by natural processes such as the seasons). DO observations in the eight monitoring sites are monthly measurements that evidence a seasonal component. So, this impact should be first minimized in the application of the change-point detection procedure.

A simple approach consists of subtracting the overall January average from January's data, the overall February average from the February's data, and so on. This approach is more suitable for series with no evident trend. In this work it is adopted the method followed by Gonçalves and Alpuim (2011): the seasonal component, s_t, takes twelve different values, $\lambda_i, i = 1, \ldots, 12$, associated with each month and expressing the positive or negative deviation from the data due to that month's effect. This effect is usually described with the help of twelve dummy variables, which must add up to zero, by considering the linear model with an intercept term. The seasonal component is represented by a linear combination of eleven explanatory variables, $s_{t,i}$, defined as

$$s_{t,i} = \begin{cases} 1, & \text{if date } t \text{ corresponds to month } i \\ -1, & \text{if date } t \text{ corresponds to month } 12 \\ 0, & \text{otherwise} \end{cases}.$$

The seasonal component for December can be calculated from the other month components through the formula $\hat{\lambda}_{12} = -\sum_{i=1}^{11} \hat{\lambda}_i$. The choice of December to be written as a linear combination of the other months is arbitrary and any month can be used for that purpose. Finally, it is applied a multiple regression model providing estimators with optimal properties.

Time series can present statistical properties such as a constant mean and seasonality, whose parameters can be estimated at the same time. Thus, the adjusted model is

$$X_t^{(M1)} = \mu + s_t + \epsilon_t, \quad t = 1, \ldots, n, \tag{21}$$

where μ is the global series mean, s_t is the seasonal component and ϵ_t is a white noise with $\mathrm{E}(\epsilon_t^2) = \sigma^2$. The change-point detection considers the errors series $\hat{\epsilon}_t = X_t^{(M1)} - \hat{\mu} - \hat{s}_t, t = 1, \ldots, n$.

The aim is to detect change-points in both the mean and the variance, i.e., to test the null hypothesis (5) *versus* the alternative hypothesis (6), through the application of the Schwarz Information Criterion (SIC) to the new series $\{\hat{\epsilon}_t\}_{t=1,\ldots,n}$, corresponding the SIC($n$) to the model (13) and the SIC(k) to the model (16). For a better understanding of the differences between the information criterion values of the different models represented by SIC(k) values and the SIC(n) $- c_\alpha$ we considered values for two significance levels, $\alpha = 5\%$ and $\alpha = 1\%$, and they are represented in the graphics by horizontal reference lines.

If, statistically, a change-point is detected, a second model will be adjusted to the original data,

$$X_t^{(M2)} = \mu_t + s_t + \epsilon_t, \quad t = 1, \ldots, n, \tag{22}$$

where s_t is the seasonal component for $t = 1, \ldots, n$,

$$\mu_t = \begin{cases} \mu_I & \text{if} \quad t \leq k \\ \mu_{II} & \text{if} \quad t > k \end{cases} \quad \text{and} \quad \epsilon_t = \begin{cases} N(0, \sigma_I^2) & \text{if} \quad t \leq k \\ N(0, \sigma_{II}^2) & \text{if} \quad t > k \end{cases} .$$

After the adjustment of the model (22) for each series, it is applied the binary segmentation process with the second change-point detection, in the two errors sequences, before and after change-point. The verification of the assumption of data normality is performed through the histograms and the Shapiro Wilk test. In order to investigate if residuals follow a white noise process, the autocorrelation function (ACF) and partial autocorrelation function (PACF) are estimated in each residual series. In the change-point procedure it is considered a significant level of 5%.

5. Results

Taking into account previous studies about this hydrological basin, namely, Gonçalves and Alpuim (2011), Costa and Gonçalves (2011) and Gonçalves and Costa (2011) in this study it is not considered a trend component in the modeling procedure. However, the data exploratory analysis suggests the incorporation of change in both the mean and the variance at unknown times in the modeling process.

The linear model M1 (21) was adjusted to the DO data series (original data, without any data transformation). Table 3 presents model parameters estimates of model M1. The SIC procedure was applied to all series according to the methodology presented in subsection 3.4, considering the asymptotic critical values at a significance level of 5%. Table 4 summarizes the results of SIC procedures. For all series it was detected a significant change-point considering the respective critical value.

Notice that in all series the differences $\text{SIC}(n) - \text{SIC}(\hat{k})$ are clearly superior to the approximate critical values at a 5% level. Moreover, considering a significance level of 1%, only the difference $\text{SIC}(n) - \text{SIC}(\hat{k})$ related to the Taipas (TAI) series is lower than the approximate critical value of $c_{1\%}$ (for instance, $c_{1\%} \approx 15.079$ when $n = 150$). Thus, change-point procedures are assertive about the existence of a change-point in both the mean and the variance in each series, even considering a conservative significance level. For instance, Figure 2 presents $\text{SIC}(k)$ values, $2 \leq k \leq 154$, for the Cantelães series and the values $\text{SIC}(n) - c_\alpha$ with $\alpha = 1\%, 5\%$.

Once the change-point $\hat{k}_i$ is detected in the series i ($i = 1, 2, ..., 8$), model M2 (Eq. 22) was adjusted in order to estimate the vector of parameters $\Theta = (\mu_I, \mu_{II}, \sigma_I^2, \sigma_{II}^2, s_1, ..., s_{12})$. In order to investigate the verification of assumptions of the change-point procedure, Table 5 presents statistical results of the series of residuals assumptions of uncorrelation and normality. In seven series of residuals time correlation is statistically significant with a AR(1) process behavior suggested by PACF and ACF functions; however, the normality is rejected. The Ponte Trofa (PTR) series presents different statistical properties because the correlation is not significant but the normality is accepted.

Thus, the initial assumptions of the SIC procedures are not satisfied, and so, it was designed a simulation study in order to evaluate its performance in the presence of serial

Table 3: Parameters estimates of model M1 (no change-point) for the eight series and the coefficients of determination.

	CANT	TAI	RAV	STI	PTR	FER	GOL	VSA
μ	9.77	9.34	8.51	8.28	8.05	9.53	9.46	9.57
σ^2	0.55	0.46	1.05	1.65	1.27	0.58	0.55	0.59
s_1	0.77	1.23	1.62	1.89	1.86	0.96	1.04	0.95
s_2	0.93	1.05	1.44	1.72	1.64	1.00	0.83	1.08
s_3	0.77	0.69	0.91	1.09	0.93	0.60	0.73	0.60
s_4	0.29	0.36	0.54	0.67	0.75	0.42	0.30	0.37
s_5	−0.07	−0.03	0.28	0.36	0.45	0.08	−0.15	0.04
s_6	−0.71	−0.80	−0.70	−1.37	−1.02	−0.74	−0.69	−0.80
s_7	−0.95	−1.19	−2.39	−2.73	−1.79	−0.81	−0.82	−1.10
s_8	−0.94	−1.46	−1.43	−1.71	−2.02	−1.18	−1.26	−1.24
s_9	−0.94	−0.82	−1.84	−1.91	−2.06	−1.04	−0.81	−0.98
s_{10}	−0.31	−0.33	−0.92	−1.05	−1.26	−0.24	−0.44	−0.09
s_{11}	0.43	0.50	0.64	1.14	0.63	0.31	0.56	0.33
s_{12}	0.73	0.80	1.85	1.90	1.89	0.64	0.71	0.84
R^2	0.48	0.62	0.64	0.60	0.63	0.49	0.51	0.52

correlation and/or non-normality. For instance, the application of the SIC framework in the context of serial correlation (usually presented in monthly environmental data) needed to be adapted, namely through a simulation study, in order to compute the corrected significant levels corresponding to the usual level of 5%. This study is presented in annex. In this context, approximations of the significance levels c^* were obtained using (20) when there is a serial correlation of $\phi = 0.3$ with Gaussian data and uncorrelated data with a zero mean Exponential distribution (a very asymmetrical distribution) in order to get a real significance level of 5%. The significance levels obtained were 2.4% and 0.4%, respectively. Note that in the case of exponential distribution it is necessary to compute the critical value associated to a very lower significance level in order to obtain a real significance level of 5%. Table 6 presents the corrected critical values for each series, thus confirming the statistical significance of the change-points according to Table 4.

After, it was performed the binary segmentation procedure for testing the existence of any change-points in each subseries. Four new change-points were detected, namely in CANT, TAI, RAV and VSA, considering the critical values associated to 5% and assuming the assumptions of the SIC procedure. However, if the corrected critical values

Table 4: Results of change-point procedures (n_i-number of observations in site i, $\hat{k} = \text{argmin}_{2\leq k\leq 154}\text{SIC}(k)$).

Site	n	SIC(n)	$\hat{k}$	SIC($\hat{k}$)	$c_{5\%}$	change-point
CANT	150	345.67	73	287.25	6.802	Jan/2005
TAI	151	321.79	70	307.96	6.791	Oct/2004
RAV	155	456.53	89	436.03	6.746	May/2006
STI	154	523.33	89	493.54	6.757	May/2006
PTR	154	482.48	83	443.22	6.757	Nov/2005
FER	152	356.58	70	341.12	6.780	Oct/2004
GOL	151	348.35	77	312.58	6.791	May/2005
VSA	151	358.44	74	321.06	6.791	Feb/2005

Table 5: Estimates of significant autoregressive coefficients and p-values of Shapiro Wilk tests for the normality for the residuals series.

	CANT	TAI	RAV	STI	PTR	FER	GOL	VSA
$\hat{\phi}$	0.23*	0.29*	0.27*	0.25*	0.01	0.20*	0.18*	0.22*
SW p-value	0.66	0.52	0.20	0.95	0.03*	0.25	0.34	0.24

* significant at a 5% level

are computed considering correlation, once the correlation is present in these series, only a second change-point (June/2005) remains statistically significant, namely in RAV, as it is presented in Table 7. Thus, model M2 was adjusted to RAV considering two change-points. The global results of the models adjustment are presented in Table 8 and their fit to original data is presented in Figure 3 and Figure 4. Note that models present a good fitting because they have coefficients of determination superior to 0.55. Relatively to the final model adjusted to the RAV series, the residuals present a correlation of $\phi = 0.268$ and the normality was accepted (SW p-value =.996).

Except for RAV, in the remaining monitoring sites there was a decrease of variance, and with regard to the mean, there is a first group (composed by Cantelães, Taipas, Ferro, Golães and Vizela Santo Adrião) which presents DO values that on average are higher in the first subseries in comparison with the observations of the second subseries. The second group (composed by the monitoring sites Santo Tirso and Ponte Trofa) presents lower average values before change-point, which increase, in average, after this. In the

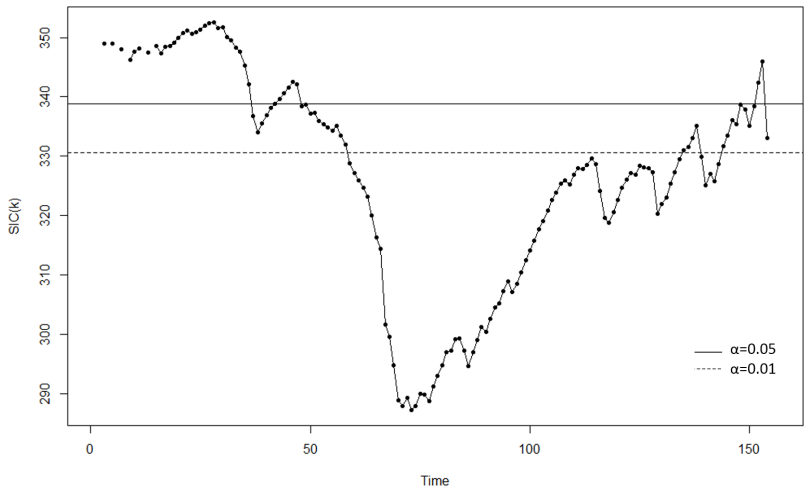

Figure 2: SIC(k) values for the Cantelães series.

Table 6: Approximated critical values c^* for each series considering serial correlation and normality for all locations, with the exception of PTR whose critical value was obtained considering uncorrelated data with Exponential distribution.

	CAN	TAI	RAV	STI	PTR	FER	GOL	VSA
n	150	151	155	154	154	152	151	151
c^*	10.390	10.378	10.328	10.340	24.810	10.365	10.378	10.378

first group the change-points positions occur in the end of 2004 and in the beginning of 2005, and in the second group they occur a little later, in the end of 2005 and in the beginning of 2006. This analysis suggests that are two distinct monitoring sites groups: a group that presents, along the time observed, a water quality improvement relatively to the DO mean concentration, whereas the other group presents water quality degradation. The identification of these two groups corroborates the results obtained in the same hydrological basin in Gonçalves and Alpuim (2011).

The RAV series was the only site where there were identified two change-points (in March 2006 and June 2005). In a more recent period, after March 2006, DO concentration has a mean level greater than before. However, in the period between the two change-points the analysis indicates a bad period with respect to DO. Despite attempts by the authorities, it was not possible to identify an objective reason for this.

In the appendix it is presented a simulation study that was conducted in order to assess

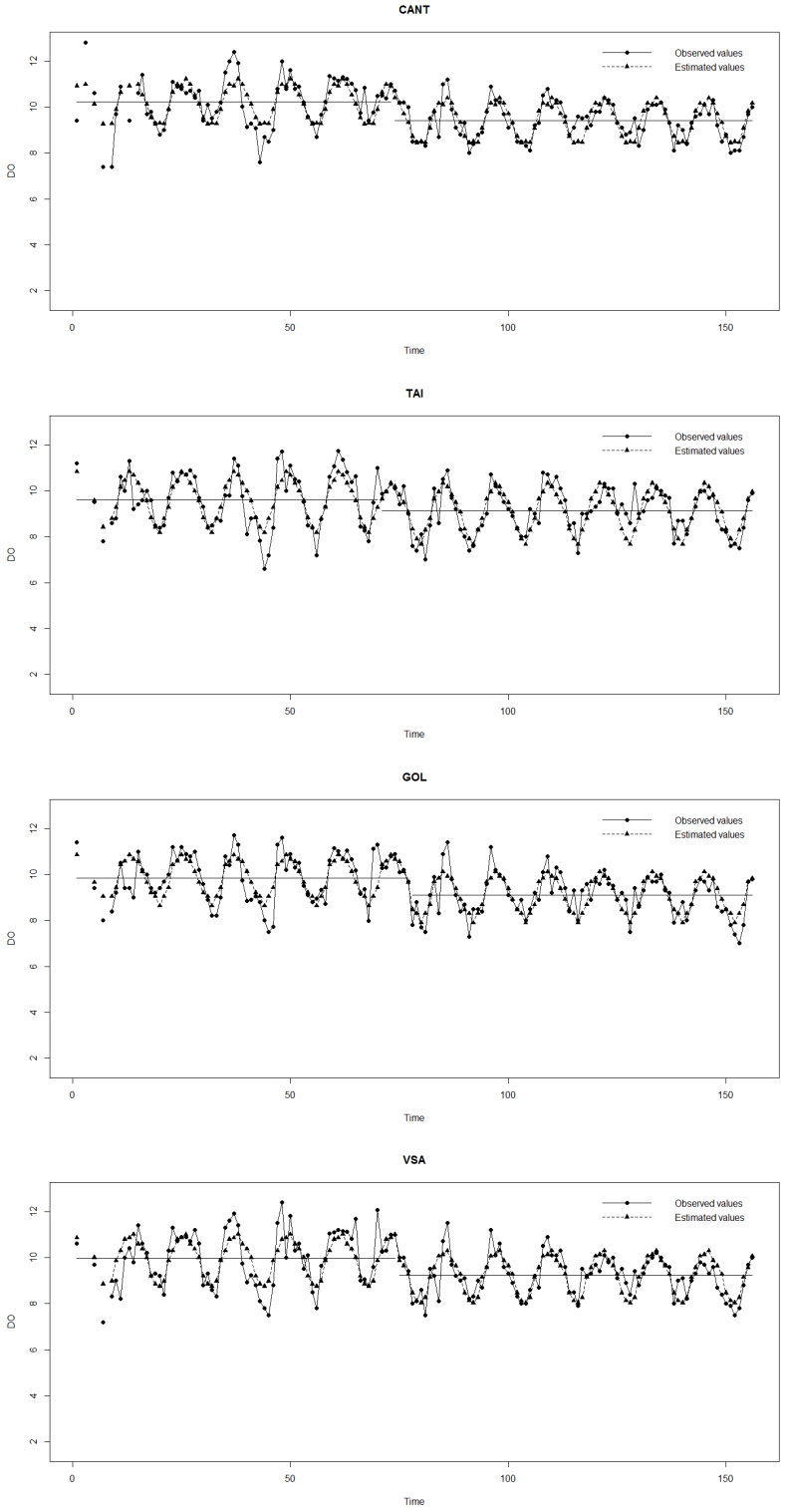

Figure 3: Adjustment of the final models.

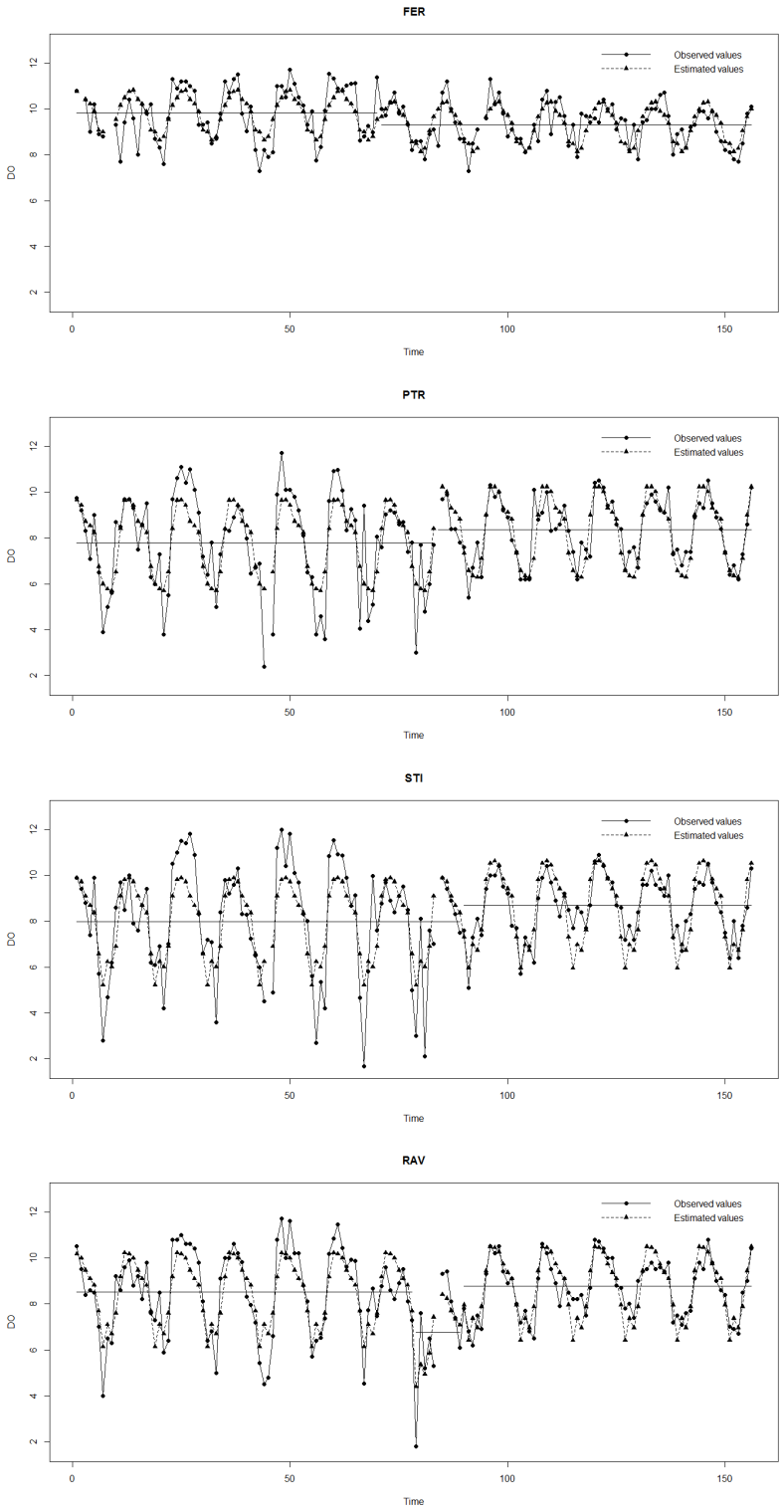

Figure 4: Adjustment of the final models.

Table 7: Results of the second step of the binary segmentation procedure.

	CANT	TAI	RAV	VSA
n	67	86	88	82
α^*	0.026	0.026	0.026	0.026
c^*	11.51	11.03	10.99	11.12
SIC(n)	161.36	159.72	288.33	144.00
SIC(k)	151.01	149.09	273.32	135.57
SIC(n) $- c^*$ >SIC(k)	no	no	yes	no
$\widehat{k}$			78	
change-point			June/2005	

the autocorrelation and non-normality impacts on the change-point detection when it is adopted the SIC approach.

6. Conclusions

We detected change-points in both the mean and the variance in the eight time series observed in the monitoring site of the Ave River hydrological basin. Whereas in seven water monitoring sites was detected one change-point, in the RAV site series were identified two change-points statistically significant according to the binary segmentation procedure.

The residual analysis of the adjusted models showed that some of the assumptions of the applied methodology are not fully verified in some time series, namely independence and normality of errors. Thus, the simulation study developed allowed a better assessment of the impact of non-verification of these assumptions in the change-point detection process.

The main conclusion of this study is that in the presence of positive autocorrelation, even if weak, the methodology tends to detect false change-points, i.e., the real significance is greater than what is considered for purposes of determining the critical point. For example, for samples of size 150 (sample size similar to the DO time series) the empirical significance obtained is approximately 14%, considering a critical point associated to a significance of 5%. The simulation study presented in the appendix revealed that when the errors are not Gaussian distributed and have a very asymmetric distribution (as the Exponential distribution) the SIC procedure's performance is jeopardized. However, the non-Gaussian errors occur only in one site but even in this case the change-point is statistically highly significant.

The analysis of the time series allowed to verify that in every monitoring sites there was a variance decrease. In five monitoring sites (Cantelães, Taipas, Ferro, Golães

and Vizela Santo Adrião) there was a decrease in average, which translates into water quality deterioration, considering only DO concentration. In the sites of Santo Tirso and Ponte Trofa there was a water quality improvement. However, these two monitoring sites continue to present the smallest DO mean concentrations, i.e., they present lower water quality. These results are in agreement with the results presented by Costa and Gonçalves (2011). The RAV series presents an alternating behavior due to the three periods identified.

It was not possible, in spite of diligences made to the official authorities, to identify factors or specific actions that might be at the origin of the detected change-points. However, a consistent result of performed analysis was the decrease of the DO concentration variability, which might be associated with the improvement of measuring instruments.

The main conclusions of this study are both the lowest variability of data and a statistically significant change in the DO concentration mean after the period of May 2004 – October 2006 in all monitoring sites. However, this change in the DO concentration average in a group of sites was for better (higher levels of DO) and in the other group it was verified a deterioration of water quality concerning the DO concentration.

Acknowledgements

Marco Costa was partially supported by Portuguese funds through the CIDMA - Center for Research and Development in Mathematics and Applications, and the Portuguese Foundation for Science and Technology ("FCT– Fundação para a Ciência e a Tecnologia"), within project UID/MAT/04106/2013. A. Manuela Gonçalves was suported by the Research Centre of Mathematics of the University of Minho with the Portuguese Funds from the *"Fundação para a Ciência e a Tecnologia"*, through the Project PEstOE/MAT/UI0013/2014.

References

Alpuim, T. and El-Shaarawi, A. (2009). Modeling monthly temperature data in Lisbon and Prague. *Environmetrics*, 20(7):835–852.

Antoch, J., *Hu*šková, M. and Prášková, Z. (1997). Effect of dependence on statistics for determination of change. *Journal of Statistical Planning and Inference*, 60:291–310.

Akaike H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In *Proceedings of the 2nd International Symposium of Information Theory*, pages 267–281. Akademic Kiado.

Barratt, B., Atkinson, R., Anderson, H.R., Beevers, S., Kelly, F., Mudway, I. and Wilkinson, P. (2007). Investigation into the use of the CUSUM technique in identifying changes in mean air pollution levels following introduction of a traffic management scheme. *Atmospheric Environment*, 41(8):1784–1791.

Beaulieu, C., Chen, J. and Sarmiento, J.L. (2012). Change-point analysis as a tool to

detect abrupt climate variations. *Philosophical Transactions of the Royal Society A*, 370:1228–1249.

Bozdogan, H. (1987). Model selection and Akaike's Information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370.

Bozdogan, H., Sclove, S.L. and Gupta, A.K. (1994). AIC-Replacements for some multivariate tests of homogeneity with applications in multisample clustering and variable selection. In *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, volume 2, pages 199–232. Kluwer Academic.

Caussinus, H. and Mestre, O. (2004). Detection and correction of artificial shifts in climate series. *Journal of the Royal Statistical Society: Series C*, 53(3):405–425.

Chen, J. (1998). Testing for a change point in linear regression models. *Communications in Statistics - Theory and Methods*, 27(10):2481–2493.

Chen, J. and Gupta, A.K. (1997). Testing and Locating variance Changepoints with Application to Stock Prices. *Journal of the American Statistical Association*, 92(438):739–747.

Chen, J. and Gupta, A.K. (1999). Change point analysis of a Gaussian model. *Statistical Papers*, 40(3):323–333.

Chen, J. and Gupta, A.K. (2001). On change point detection and estimation. *Communications in Statistics - Simulation and Computation*, 30(3):665–697.

Chen, J. and Gupta, A.K. (2012). *Parametric Statistical Change Point analysis*. Birkhauser.

Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subject to changes in time. *Annals of Mathematical Statistics*, 35:999–1018.

Chu, H.J., Pan, T.Y. and Liou, J.J. (2012). Change-point detection of long-duration extreme precipitation and the effect on hydrologic design: a case study of south Taiwan. *Stochastic Environmental Research and Risk Assessment*, 26(8):1123–1130.

Costa, M. and Gonçalves, AM. (2011). Clustering and forecasting of dissolved oxygen concentration on a river basin. *Stochastic Environmental Research and Risk Assessment*, 25(2):151–163.

Davis, R.A., Lee, T.C.M. and Rodriguez-Yam, G.A. (2006). Structural break estimation for non- stationary time series models. *Journal of the American Statistical Association*, 101(473):223–239.

El-Shaarawi, A.H. and Esterby, S.R. (1982). Inference About the Point of Change in A Regression Model With A Stationary Error Process. In *Proceedings of an International Conference Held at Canada Centre for Inland Waters, Time Series Methods in Hydrosciences*, pages 55–67.

Gonçalves, A.M. and Alpuim, T. (2011). Water quality monitoring using cluster analysis and linear models. *Environmetrics*, 22(8):933–945.

Gonçalves, A.M. and Costa, M. (2011). Application of Change-Point Detection to a Structural Component of Water Quality Variables. In *Proceedings of the Interna-*

tional Conference on Numerical Analysis and Applied Mathematics, volume 1389, pages 1565–1568. AIP Conference Proceedings.

Gonçalves, A.M. and Costa, M. (2013). Predicting seasonal and hydro-meteorological impact in environmental variables modelling via Kalman filtering. *Stochastic Environmental Research and Risk Assessment*, 27(5):1021–1038.

Hájek, J. (1962). Asymptotically most powerful rank order tests. *Annals of Mathematical Statistics*, 33:1124–1147.

Hawkins, D.M. (1977). Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association*, 72:180–186.

Hawkins, D.M. and Zamba, K.D. (2005). Statistical process control for shifts in mean or variance using a changepoint formulation. *Technometrics*, 47(2):164–173.

Hsu, D.A. (1977). Tests for Variance Shift at an Unknown Time Point. *Journal of the Royal Statistical Society: Series C*, 26(3):279–284.

Inclán, C. and Tiao, G.C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, 89(427):913–923.

Jarušková, D. (1996). Change-Point Detection in Meteorological Measurement. *Monthly Weather Review*, 124:1535–1543.

Jarušková, D. (1997). Some problems with application of change-point detection methods to environmental data. *Environmetrics*, 8:469–483.

Jarušková, D. and Rencov, M. (2008). Analysis of annual maximal and minimal temperatures for some European cities by change point methods. *Environmetrics*, 19:221–233.

Jarušková, D. (2010). Asymptotic behavior of a test statistic for detection of change in mean of vectors. *Journal of Statistical Planning and Inference*, 140:616–625.

Krishnaiah, P.R. and BQ Miao, B.Q. (1988). *Review about Estimation of Change Points.* Handbook of Statistics, volume 7, pages 375–402, Elsevier.

Kitagawa, G. (1979). On the use of AIC for the detection of outliers. *Technometrics*, 21(2):193–199.

Li, S. and Lund, R.B. (2012). Multiple Changepoint Detection via Genetic Algorithms. *Journal of Climate*, 25:674–686.

Lu, Q., Lund, R.B. and Lee, T.C.M. (2010). An MDL Approach to the Climate Segmentation Problem. *Annals of Applied Statistics*, 4(1):299–319.

Lund, R. and Reeves, J. (2002). Detection of Undocumented Changepoints: A Revision of the Two-Phase Regression Model. *Journal of Climate*, 15:2547–2554.

Lund, R. Wang, X.L., Lu, Q.Q.,Reeves, J., Gallagher, C. and Feng, Y. (2007). Changepoint Detection in Periodic and Autocorrelated Time Series. *Journal of Climate*, 20:5178–5190.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna. http//www.R-project.org

Rao, C.R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika*, 76(2):369–374.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Srivastava, M.S. and Worsley, K.J. (1986). Likelihood ratio test for a change in the multivariate normal mean. *Journal of the American Statistical Association*, 81(393):199–204.

Vostrikova, L.J. (1981). Detecting 'disorder' in multidimensional random processes. *Soviet Mathematics Doklady*, 24:55–59.

Worsley, K.J. (1979). On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association*, 74(366a):365–367.

Appendix

A. Simulation study

A.1. Simulation study design

A simulation study was conducted to investigate the performance of the Schwarz Information Criterion procedure in many scenarios, namely when the methodology for the change-point problem described in the previous section is applied when the errors are not normally distributed or when the uncorrelation assumption of errors is violated. From the practical point of view, it is relevant to investigate the methodology's performance by assessing if it is suitable to extend conclusions, even when assumptions are not verified. The conclusions drawn are only valid for the analyzed scenarios once they are established to encompass the behaviors of the studied time series, namely the change-point in the mean and in the variance, the presence of correlation or the lack of normality in the errors distribution. In order to develop the simulation study it was used the freeware statistical software R (R Development Core Team 2011).

Two scenarios are considered. In the first, time series are generated without change-points. In the second, it is set a change-point, both in the mean and in the variance.

When no change-point is simulated, time series are generated according to the model

$$X_t = \mu + \epsilon_t, t = 1, \ldots, n, \tag{23}$$

where μ is the mean, ϵ_t is a white noise error, and n the sample size.

In the second scenario, when a change-point is set, this is induced in the midpoint $t = \frac{n}{2}$ instant. This option is due to the fact that the change-points found in the study of real data (Dissolved Oxygen time series) occurred at central instants in the time series. In this case, the time series are generated according to the model

$$X_t = \begin{cases} \mu_I + \epsilon_t^I, t = 1, \ldots, k \\ \mu_{II} + \epsilon_t^{II}, t = k+1, \ldots, n, \end{cases} \tag{24}$$

where μ_I and μ_{II} are the means before and after change-point, and ϵ_t^I and ϵ_t^{II} are the errors with mean 0 and variances σ_I^2 and σ_{II}^2, respectively.

A comparative study was performed in order to estimate the error of Type I, which is estimated by the empirical significant level calculated by the ratio of the null hypothesis rejected, where the generated series doesn't present a change-point. A comparative study was also be conducted in order to obtain the empirical power of the statistical test, which is estimated by the ratio of the null hypothesis rejected when one change-point was induced in the generated series. Also, it is important to evaluate when the change-point is detected in an appropriate way, i.e., close to the true change-point instant k. In each of the above scenarios were considered different stochastic errors structures (uncorrelated and correlated errors) and the Normal and Exponential distributions. The Exponential distribution is considered due to its strong skewness.

When errors are considered with a correlation structure, this is assumed to be characterized by a first-order autocorrelation process AR(1), i.e., it follows the structure $\epsilon_t = \phi\epsilon_{t-1} + a_t$, with $\mid \phi \mid < 1$, wherein a_t is a white noise with null mean. In this study, $\phi = 0.3$ represents the correlation that was detected in some series in the previous section.

Normality of errors was assumed, since it is one of the assumptions of the adopted methodology and serves as a reference to compare with the series generated from exponential errors. In the exponential case, the errors are obtained by $\epsilon_t = Y_t - \frac{1}{\lambda}$, where $Y_t \sim \text{Exp}(\lambda)$ and $\text{E}(Y_t) = \frac{1}{\lambda}$. We considered samples with size $n = 50$ (small samples), a sample size approximate to the time series studied $n = 150$, and yet high dimensional sample size with $n = 500$.

For each sample size, it is considered a vector of parameters that characterizes the simulated model. When a series does not have a change-point the parameter vector is $\Theta = \{\mu, \sigma^2, \phi\}$, and when there is a induced change-point the parameter vector is $\Theta = \{\mu_I, \mu_{II}, \sigma_I^2, \sigma_{II}^2, \phi\}$. In the series without change-point, the mean considered is $\mu = 0$, without loss of generality. Different shifts are considered when a change-point is generated. In the first subseries it is taken $\mu_I = 0$; after the change-point it is considered $\mu_{II} = 0.2$, $\mu_{II} = 0.5$ or $\mu_{II} = 0.8$. Note that these values were considered according to the practical results obtained in the series studied in the previous section.

The possible variance values are established taking into account the empirical results of the time series. Thus, the series without change-point assumed variances errors $\sigma^2 = 0.5$, $\sigma^2 = 1$ and $\sigma^2 = 1.5$. When a change-point is induced, the following values were considered (0.6, 0.3) and (2, 0.6) for the pair (σ_I^2,σ_{II}^2). In cases where errors follow an AR(1) process, the white noise a_t is simulated with a variance $\sigma_a^2 = (1 - \phi^2)\sigma_\epsilon^2$. Furthermore, when errors have Exponential distribution, these are obtained by considering $\lambda = \sqrt{1/\sigma_\epsilon^2}$, when there is no correlation and $\lambda = \sqrt{1/\sigma_a^2}$, otherwise.

The simulation study was designed by generating 2000 samples per each scenario, without change-point and with change-point, and for each parameter combination Θ, by considering Normal and Exponential random errors distributions. The SIC procedure was performed considering the critical value of the test with a significance level of 5%.

A.2. Simulation results

Tables 9 and 10 present the empirical significances when under H_0. As expected, when the errors are independent and normally distributed ($\phi = 0$), the empirical significance is very close to the considered significance of 5%, even for small samples ($n = 50$).

When the methodology is applied to correlated observations ($\phi = 0.3$) the empirical significance is greater than the adopted significance (5%), being even double or triple. Thus, these results are in accordance with those reported in Beaulieu et al. (2012). It is emphasized, however, that the correlation's impact becomes stronger for larger sample sizes. Bibliographical research carried out did not found any reference to this fact. The results show that the magnitude of the observations variance has no major impact in the SIC test performance, since the empirical significance obtained is similar for different σ_ϵ^2 values.

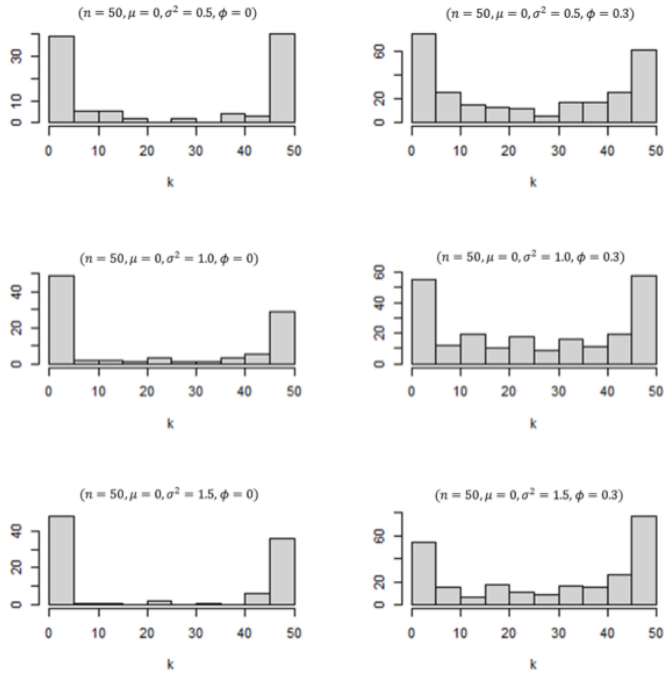

Figure 5: Histograms of false change-points detected by considering samples with $n = 50$ and errors normally distributed.

In order to examine the false change-points detected when series with no change-points were generated, their histograms were drawn. For instance, Figure 5 and Figure 6 show that false change-points correspond to instants close to the beginning or the end of the time series, predominantly where the errors are Gaussian. In fact, when errors are exponential, the false change-point detection is more uniform over the series time interval. However, for larger samples, the results are closer to those obtained for the Gaussian errors.

Tables 11 and 12 present the empirical power where errors are normally and exponentially distributed, respectively, under the alternative hypothesis H_1. Results show that the empirical power is greater when there is a dependency structure in the observations. As expected, when the differences $\mu_{II} - \mu_I$ are higher, the empirical power is greater for both normal and exponential errors.

It should be noted that the impact of the changes in variances does not affect the overall relative pattern of the methodology's performance. However, the results show that when the mean difference is small, the empirical power is higher when the variances difference is superior. When the mean difference is 0.8 and it is associated with a greater change in the magnitude of variances, the empirical power tends to decrease. The applied methodology shows a good performance for large samples ($n = 500$) once it has empirical powers near 100% in almost all scenarios.

In order to allow a more thorough analysis of the proposed methodology's performance

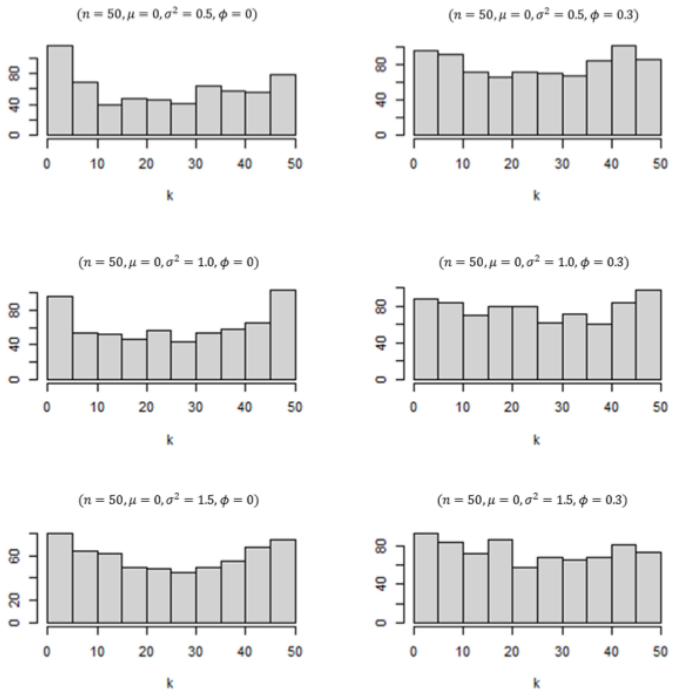

Figure 6: Histograms of false change-points detected by considering samples with $n = 50$ and errors exponentially distributed.

as well the accuracy of the location of the detected change-points, the percentages of the detected change-points close to the true change-point ($k = \frac{n}{2}$) were computed. For this purpose, upper and lower limits were established, among which a change-point was detected with a reasonable accuracy ($19 - 31$, $63 - 87$ and $226 - 274$ for samples of size $n = 50$, $n = 150$ and $n = 500$, respectively).

For each case it was computed the ratio of statistical significant change-point within the established limits under H_1 and the ratio of the change-point within the limits over the number of all change-point detected. The results are presented in Table 13 for the normal errors and in Table 14 for the exponential errors.

The results highlight the tendency that when the observations are correlated, the adopted methodology provided best performance (in change-point detection), even considering the change-points located within the established limits. By comparing the results for the series presenting random errors that follow a normal distribution with errors that follow an exponential distribution, we could conclude that, in the latter case, the performance is inferior compared to the observations from a normal distribution.

Globally, we can say that the difference between the performance in the scenarios of independence or autocorrelation presence is attenuated when we analyzed the percentages of detected change-points located within the limits, mainly for smaller samples ($n = 50$).

Table 8: Parameters estimates of final model M2 considering the change-points in each series (after the binary segmentation procedure) and the coefficient of determination.

Parameter	CANT	TAI	RAV	STI	PTR	FER	GOL	VSA
μ_I	10.22	9.62	8.51	7.97	7.78	9.81	9.85	9.96
μ_{II}	$\downarrow$9.41	$\downarrow$9.12	$\downarrow$6.79	$\uparrow$8.69	$\uparrow$8.37	$\downarrow$9.31	$\downarrow$9.11	$\downarrow$9.24
μ_{III}			$\uparrow$8.78					
σ^2_I	0.58	0.49	1.00	2.21	1.70	0.70	0.51	0.65
σ^2_{II}	$\downarrow$0.24	$\downarrow$0.34	$\uparrow$2.22	$\downarrow$0.64	$\downarrow$0.60	$\downarrow$0.37	$\downarrow$0.34	$\downarrow$0.31
σ^2_{III}			$\downarrow$0.45					
s_1	0.70	1.21	1.66	1.92	1.87	0.95	1.00	0.89
s_2	1.00	1.06	1.48	1.76	1.65	1.01	0.81	1.05
s_3	0.76	0.70	0.95	1.13	0.94	0.59	0.71	0.63
s_4	0.31	0.37	0.58	0.71	0.75	0.41	0.28	0.40
s_5	-0.09	-0.05	0.31	0.40	0.46	0.07	-0.20	0.03
s_6	-0.69	-0.79	-0.82	-1.39	-1.02	-0.75	-0.64	-0.77
s_7	-0.96	-1.21	-2.37	-2.74	-1.79	-0.83	-0.81	-1.11
s_8	-0.92	-1.46	-1.41	-1.72	-2.01	-1.17	-1.22	-1.21
s_9	-0.95	-0.83	-1.82	-1.96	-2.08	-1.03	-0.80	-0.98
s_{10}	-0.32	-0.35	-0.91	-1.06	-1.26	-0.27	-0.43	-0.10
s_{11}	0.41	0.52	0.66	1.12	0.63	0.34	0.57	0.32
s_{12}	0.75	0.83	1.70	1.83	1.86	0.68	0.73	0.85
R^2	0.64	0.67	0.72	0.63	0.65	0.55	0.63	0.63

Table 9: Empirical significance for 2000 replicates simulated from a Normal distribution.

		$n = 50$		$n = 150$		$n = 500$	
μ	σ^2	$\phi = 0$	$\phi = 0.3$	$\phi = 0$	$\phi = 0.3$	$\phi = 0$	$\phi = 0.3$
0	0.5	0.050	0.133	0.052	0.138	0.048	0.161
	1	0.048	0.114	0.047	0.146	0.039	0.163
	1.5	0.048	0.123	0.042	0.141	0.044	0.168

Table 10: Empirical significance for 2000 replicates simulated from a Exponential distribution.

		$n = 50$		$n = 150$		$n = 500$	
μ	σ^2	$\phi = 0$	$\phi = 0.3$	$\phi = 0$	$\phi = 0.3$	$\phi = 0$	$\phi = 0.3$
0	0.5	0.307	0.403	0.443	0.544	0.571	0.687
	1	0.316	0.388	0.442	0.555	0.569	0.677
	1.5	0.298	0.375	0.431	0.532	0.585	0.685

Table 11: Empirical power from 2000 samples generated from a Normal distribution.

				$n = 50$		$n = 150$		$n = 500$	
μ_I	μ_{II}	σ_I^2	σ_{II}^2	$\phi = 0$	$\phi = 0.3$	$\phi = 0$	$\phi = 0.3$	$\phi = 0$	$\phi = 0.3$
0	0.2	0.6	0.3	0.130	0.280	0.475	0.710	0.995	0, 999
		2	0.6	0.310	0.443	0.939	0.953	1.000	1.000
0	0.5	0.6	0.3	0.356	0.716	0.962	0.996	1.000	1.000
		2	0.6	0.409	0.621	0.985	0.995	1.000	1.000
0	0.8	0.6	0.3	0.773	0.973	1.000	1.000	1.000	1.000
		2	0.6	0.617	0.866	0.998	1.000	1.000	1.000

Table 12: Empirical power from 2000 replicates simulated from a Exponential distribution.

				$n = 50$		$n = 150$		$n = 500$	
μ_I	μ_{II}	σ_I^2	σ_{II}^2	$\phi = 0$	$\phi = 0.3$	$\phi = 0$	$\phi = 0.3$	$\phi = 0$	$\phi = 0.3$
0	0.2	0.6	0.3	0.402	0.521	0.717	0.889	0.998	1.000
		2	0.6	0.538	0.608	0.911	0.958	1.000	1.000
0	0.5	0.6	0.3	0.601	0.864	0.992	1.000	1.000	1.000
		2	0.6	0.595	0.786	0.994	0.999	1.000	1.000
0	0.8	0.6	0.3	0.919	0.995	1.000	1.000	1.000	1.000
		2	0.6	0.781	0.938	1.000	1.000	1.000	1.000

Table 13: Ratios of statistical significant change-points within the established limits from all samples under H_1. Between parenthesis are shown the ratios of the change-points within the limits from all change-points detected. Errors with Normal distribution.

				$n = 50$		$n = 150$		$n = 500$	
μ_I	μ_{II}	σ_I^2	σ_{II}^2	$\phi = 0$	$\phi = 0.3$	$\phi = 0$	$\phi = 0.3$	$\phi = 0$	$\phi = 0.3$
0	0.2	0.6	0.3	0.055	0.127	0.336	0.426	0.889	0.875
				(0.421)	(0.451)	(0.706)	(0.600)	(0.893)	(0.875)
		2	0.6	0.214	0.273	0.825	0.773	0.984	0.975
				(0.690)	(0.615)	(0.879)	(0.811)	(0.984)	(0.975)
0	0.5	0.6	0.3	0.252	0.534	0.839	0.894	0.990	0.986
				(0.708)	(0.745)	(0.872)	(0.897)	(0.990)	(0.986)
		2	0.6	0.301	0.450	0.896	0.875	0.994	0.990
				(0.736)	(0.725)	(0.910)	(0.879)	(0.994)	(0.990)
0	0.8	0.6	0.3	0.689	0.886	0.980	0.987	1.000	1.000
				(0.891)	(0.910)	(0.980)	(0.987)	(1.000)	(1.000)
		2	0.6	0.512	0.713	0.959	0.939	0.997	0.996
				(0.828)	(0.824)	(0.960)	(0.939)	(0.997)	(0.996)

Table 14: Ratios of statistical significant change-points within the established limits from all samples under H_1. Between parenthesis are shown the ratios of the change-points within the limits from all change-points detected. Errors with Exponential distribution.

				$n = 50$		$n = 150$		$n = 500$	
μ_I	μ_{II}	σ_I^2	σ_{II}^2	$\phi = 0$	$\phi = 0.3$	$\phi = 0$	$\phi = 0.3$	$\phi = 0$	$\phi = 0.3$
0	0.2	0.6	0.3	0.136	0.208	0.295	0.452	0.749	0.843
				(0.339)	(0.399)	(0.412)	(0.509)	(0.750)	(0.843)
		2	0.6	0.247	0.313	0.572	0.639	0.872	0.920
				(0.458)	(0.514)	(0.628)	(0.667)	(0.872)	(0.920)
0	0.5	0.6	0.3	0.365	0.620	0.825	0.909	0.983	0.997
				(0.607)	(0.717)	(0.832)	(0.909)	(0.983)	(0.997)
		2	0.6	0.348	0.531	0.803	0.867	0.945	0.976
				(0.585)	(0.675)	(0.908)	(0.868)	(0.945)	(0.976)
0	0.8	0.6	0.3	0.777	0.909	0.963	0.983	0.998	0.999
				(0.846)	(0.914)	(0.963)	(0.983)	(0.998)	(0.999)
		2	0.6	0.589	0.765	0.920	0.945	0.972	0.996
				(0.754)	(0.816)	(0.920)	(0.945)	(0.972)	(0.996)