Change-point analysis in environmental time series

A. Manuela Gonçalves¹, Marco Costa², Lara Teixeira¹

¹ Departamento de Matemática e Aplicações, CMAT-Centro de Matemática, Universidade do Minho, Portugal
² Escola Superior de Tecnologia e Gestão de Águeda, Universidade de Aveiro, CMAF-UL, Portugal

E-mail for correspondence: mneves@math.uminho.pt

Abstract: Change-points are present in many environmental time series. 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. In this study, it is proposed an alternative approach for the application of the change-point analysis by taking into account this data structure (seasonality and autocorrelation) based on the Schwarz Information Criterion (SIC). The approach was applied to time series of surface water quality variables measured at eight monitoring sites.

Keywords: Change-point analysis; SIC; Autocorrelation; Seasonality; Mean and variance shift.

1 Introduction

In this study is proposed the application of the Schwarz Information Criterion (SIC) to detect the change-point in mean and variance in time series of water quality variables. The data concerns the River Ave hydrological basin situated in the Northwest of Portugal, where monitoring has become a priority in water quality planning and management in this watershed. The water quality variable analyzed is Dissolved Oxygen (DO), one of the most important variables in assessing surface water quality in a river’s hydrological basin (Costa and Gonçalves, 2011 and Gonçalves and Costa, 2012), measured (in milligrams per liter (mg/l)) monthly from January 1999 to December 2011 in eight monitoring sites: Cantelães (CANT), Taipas (TAI), Ferro (FER), Golães (GOL), Vizela Santo Adrião (VSA), Riba d’Ave (RAV), Santo Tirso (STI) and Ponte Trofa (PTR). In this work, the behavior study of the time series of DO water quality variable is addressed in line with the research of Gonçalves and Costa (2011), Gonçalves and Alpuim (2011), who recently studied trend alterations in environmental variables, including time series of water quality variables. By performing an
2 Change-point analysis

exploratory analysis, we concluded that the DO observed values over time (each time series consists at most of 156 observations) presented changes in mean and/or variance in the series (in particular between 2004 and 2006). As regards the average, it apparently increases or decreases according to the monitoring site, but it reduces the variability of the observations 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 with the weather patterns throughout the year, particularly temperature changes and precipitation intensity.

2 The informational approach

In order to detect changes in time series, the case of a change-point in both the mean and the variance, the aim is to test the following hypothesis

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_n = \mu \land \sigma^2_1 = \sigma^2_2 = \ldots = \sigma^2_n = \sigma^2$$  \hspace{1cm} (1)

versus the alternative hypothesis

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

$$\land$$

$$\sigma^2_I = \sigma^2_1 = \ldots = \sigma^2_k \neq \sigma^2_{k+1} = \ldots = \sigma^2_n = \sigma^2_{II}.$$  \hspace{1cm} (2)

Based on Akaike’s work, in 1978 Schwarz proposed the Schwarz Information Criterion (SIC). The SIC is defined as following

$$\text{SIC}_j = -2 \ln L(\hat{\Theta}_j) + p_j \ln n, \quad j = 1, 2, \ldots, M,$$  \hspace{1cm} (3)

where $n$ is the sample size. This criterion is based on the maximum likelihood function of a given model penalized by the number of parameters that are estimated in the model. Under $H_0$, the SIC is denoted by $\text{SIC}(n)$ and it is obtained as

$$\text{SIC}(n) = -2 \ln L_0(\hat{\mu}, \hat{\sigma}^2) + 2 \ln n, \quad \text{(4)}$$

$$= n \ln 2\pi + n \ln \sum_{i=1}^{n} (X_i - \overline{X})^2 + n + (2 - n) \ln n. \quad \text{(5)}$$

where $L_0(\hat{\mu}, \hat{\sigma}^2)$ is the maximum likelihood function with respect to $H_0$. Under $H_1$, the SIC is denoted by $\text{SIC}(k)$ for fixed $k$, $2 \leq k \leq n - 2$, is obtained as

$$\text{SIC}(k) = -2 \ln L_1(\hat{\mu}_I, \hat{\mu}_{II}, \hat{\sigma}^2_I, \sigma^2_{II}) + 4 \ln n \quad \text{(6)}$$

$$= n \ln 2\pi + k \ln \hat{\sigma}^2_I + (n - k) \ln \sigma^2_{II} + n + 4 \ln n, \quad \text{(7)}$$
where \( L_1(\hat{\mu}_I, \hat{\mu}_{II}, \hat{\sigma}_I^2, \hat{\sigma}_{II}^2) \) is the maximum likelihood function under \( H_1 \). The decision to accept \( H_0 \) or \( H_1 \) is based on the principle of minimum criterion. According to the information criterion principle, we are going to estimate the position of the change-point \( k \) such that \( \text{SIC}(k) \) is the minimal. Then, the estimation of the position of the change-point by \( \hat{k} \) is given by:

\[
\text{SIC}(\hat{k}) = \min_{2 \leq k \leq n-2} \text{SIC}(k)
\]

In order to assess significance, a critical value \( c_\alpha \) can be included in the decision rule for a significance level \( \alpha \), where \( c_\alpha \geq 0 \). The model with a change-point \( \text{SIC}(k) \) is selected if

\[
\min_{2 \leq k \leq n-2} \text{SIC}(k) + c_\alpha < \text{SIC}(n)
\]

otherwise, the model with no change-point \( \text{SIC}(n) \) is more reasonable. The approximate critical values for different series lengths that were obtained through the asymptotic distribution are presented in Chen and Gupta (1999).

3 Change-point detection procedure

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

\[
X_{t}^{(M1)} = \mu + s_t + \epsilon_t, \quad t = 1, \ldots, n,
\]

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

The aim is to detect change-points in both the mean and the variance, i.e., to test the null hypothesis (1) versus the alternative hypothesis (2), through SIC application to the new series \( \{\epsilon_t\}_{t=1,\ldots,n} \), corresponding the SIC(n) to the model (5) and the SIC(k) to the model (7). For a better understanding of the differences between information criterion values of the different models, will be represented SIC(k) values and the SIC(n) – \( c_\alpha \) values for two significance levels, \( \alpha = 0.05 \) and \( \alpha = 0.01 \). If, statistically, a change-point is detected, a second model will be adjusted to the original data,

\[
X_{t}^{(M2)} = \mu_t + s_t + \epsilon_t, \quad t = 1, \ldots, n,
\]

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

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

After the adjustment of the model (10), it follows the binary segmentation process with the second change-points detection, in the two errors
sequences, before and after change-point. However, the data analysis was conservative by taking into account the performed simulation study (not presented in this article) and in agreement with Beaulieu et al. (2012): the presence of autocorrelation in the observations, even weak ones ($\phi \approx 0.3$), tends to originate the detection of false change-points. Thus, in this study when $\text{SIC}(n)$ and $\text{SIC}(k)$ values are very close, even if the change-point is statistically significant, we decided not to consider the existence of a second change-point.

4 Results and discussion

Taking into account the previous studies about this hydrological basin (Gonçalves and Alpuim 2011, Costa and Gonçalves 2011, Gonçalves and Costa 2011) and the inspection of data series, it is reasonable to consider that the series do not present trends (for instance, a linear trend). Moreover, works that compare DO data series (and other water quality variables, Gonçalves and Alpuim 2011) in different water monitoring sites concluded that there is a common pattern in the evolution of these variables considering the same hydrological basin. Thus, it is reasonable to consider the same change-point model for all eight water monitoring sites. The linear model $M_1$ (9) was adjusted to DO data series (original data, without any transformation).

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

The SIC procedure was applied to all series according to the methodology shown above, considering the asymptotic critical values at a 5% significance level. Table 1 summarizes the results of SIC procedures. For all series was detected a change-point significant considering the respective critical value. One should notice that in all series the differences $\text{SIC}(n) - \text{SIC}(k)$ are clearly superior to the approximate critical values at a 5% significance level. Moreover, considering a 1% significance level, only the difference $\text{SIC}(n) -$
FIGURE 1. SIC(\(k\)) values for Cantelães series and adjustment of linear model considering the change-point in Cantelães.

SIC(\(\hat{k}\)) relatively to the Taipas series (TAI) 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 mean and variance in each series, even considering a conservative significance level. For instance, Figure 1 represents SIC(\(k\)) values, \(2 \leq k \leq 154\), for Cantelães series and the values SIC(\(n\)) − \(c_\alpha\) with \(\alpha = 1\%, 5\%\). As the assumptions of normality and independence are not present in some time series, a simulation study was carried out (not presented in this paper) in order to evaluate the methodology’s performance when applied to non-normal data series with or without time correlation.

References


