METMA
VI INTERNATIONAL WORKSHOP ON SPATIO-TEMPORAL MODELLING

Information Criterion on the Change-point Detection in Water Quality Data

A. Manuela Gonçalves[1,*], Marco Costa[2] and Lara Teixeira[3]

[1] *Department of Mathematics and Applications; CMAT-Center of Mathematics, University of Minho, Portugal; mneves@math.uminho.pt*
[2] *Higher School of Technology and Management of Águeda-University of Aveiro; CMAF-UL, Portugal; marco@ua.pt*
[3] *Department of Mathematics and Applications, University of Minho, Portugal; lara-teixeira@hotmail.com*
* *Corresponding author*

Abstract. *The number of studies using change-point methods to detect shifts has been increasing. The Schwarz Information Criterion SIC is used in this study to search for the change-point in time series of water quality variables. The data set is related to the River Ave basin in Northwest Portugal and the variables were observed in some water quality monitoring sites.*

Keywords. *Water quality; Change-point detection; Schwarz Information Criterion; Mean and variance shift.*

1 Introduction

Environmental issues have gained a great importance. The impact of human activity on nature has been increasing and it may cause a change in nature. In this study we focus on water quality, namely on a data set related to the River Ave basin in Northwest Portugal, that consists mainly of monthly measures of physical-chemical and microbiological variables in a network of water quality monitoring sites. To study the changes in nature, the problem of statistical change-point has been an attractive topic in statistical analysis for decades. In this study we used a methodology to determine the exact time and nature of the shift in the data set based on Schwarz Information Criterion. This informational approach can be useful to discriminate among several change-point models with different types of changes (shift in the mean, shift in variance, shift in the parameters of a linear regression model, etc.) and it is generally used for model selection.

2 Methods

The informal approach is a general model selection technique that can be adapted to a diverse set of situations [2]. Hirotugu Akaike introduced in 1973 the Akaike information criterion (AIC) for model selection in statistics ([1]). The general formulation of the *AIC* to select among *M* models can be expressed by

$$AIC_j = -2\ln L(\hat{\Theta}_j) + 2p_j, \quad j = 1,2,...,M,$$

where $L(\hat{\Theta}_j)$ is the maximum likelihood function for model(j) and p_j is the number of parameters to be estimated for model j. Based on Akaike's work, many authors introduced modifications and other information criteria. One of the modifications is the Schwarz Information Criterion (SIC) ([6]). The SIC is defined as

$$SIC_j = -2\ln L(\hat{\Theta}_j) + p_j \ln n, \quad j = 1,2,...,M.$$

where n is the number of observations. Apparently, the difference between AIC and SIC is in the penalty term. However, SIC gives asymptotically consistent estimates of the order of the true model and makes use of the sample information ([5]). If we want to test if a shift occurred in the mean and in the variance at the same time, we will have to compare a model with a constant mean and variance, and a model with a shift in the mean and variance.

The data set $X_1, X_2, ..., X_n$ has to be a sequence of independent normal random variables, with means $\mu_1, \mu_2, ..., \mu_n$, and variances $\sigma_1^2, \sigma_2^2, ..., \sigma_n^2$, respectively. Assuming that all parameters are unknown, we want to test the following hypothesis

$$H_0: \ \mu_1 = \mu_2 = ... = \mu_n = \mu \quad \wedge \quad \sigma_1^2 = \sigma_2^2 = ... = \sigma_n^2 = \sigma^2$$

versus

$$H_1 \ : \ \mu_1 = ... = \mu_k \neq \mu_{k+1} = ... = \mu_n \quad \wedge \quad \sigma_1^2 = ... = \sigma_k^2 \neq \sigma_{k+1}^2 = ... = \sigma_n^2$$

where μ and σ^2 are unknown common parameters when there are no changes, k, with $2 \leq k \leq n-2$, is the unknown position of the change-point, [4].

Under H_0, the *SIC* is denoted by $SIC(n)$ and it is obtained as

$$\begin{aligned} SIC(n) &= -2\ln L_0(\hat{\mu}, \hat{\sigma}^2) + 2\ln n \\ &= n\ln 2\pi + n\ln \sum_{i=1}^{n} (X_i - \overline{X})^2 + n + (2-n)\ln n \end{aligned}$$

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 $SIC(k)$ for fixed k, $2 \leq k \leq n-2$, is obtained as

$$SIC(k) = -2\ln L_1(\hat{\mu}_1, \hat{\mu}_n, \hat{\sigma}_1^2, \hat{\sigma}_k^2) + 4\ln n$$

$$= n\ln 2\pi + k\ln \hat{\sigma}_1^2 + (n-k)\,\hat{\sigma}_n^2 + n + 4\ln n,$$

where $L_1(\hat{\mu}_1, \hat{\mu}_n, \hat{\sigma}_1^2, \hat{\sigma}_n^2)$ is the maximum likelihood function under H_1, [15]. The decision to accept H_0 or H_1 is based on the principle of minimum criterion [3]. That is, we reject H_0 if $SIC(n) > min_{2\leq k\leq n-2} SIC(k)$, and estimate the position of the change-point k by $\hat{k}$, such that

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

To assess significance, a critical value c_α can be included in the decision rule for a significance level α, where $c_\alpha \geq 0$. We reject H_0 if $SIC(n) > min_{2\leq k\leq n-2} SIC(k) + c_\alpha$. The approximate critical values for different series lengths that were obtained through the asymptotic distribution are presented in [4].

3 Results and discussion

We used the dissolved oxygen concentrations levels (DO) (mg/l) in some monitoring sites in the River Ave basin during 1999-2011. The observations are monthly measured and the seasonality component was removed, so we applied the informational approach to the series without it. We wanted to study whether there was a change in the average and variance simultaneously.

For the monitoring site of Santo Tirso, $SIC(n) = 523.379$ and $min_{2\leq k\leq 154} SIC(k) = SIC(89) = 493.801$ we used the table of critical values presented in [4] and $SIC(89) + c_{0.05} < SIC(n)$. Hence, a change-point at $t = 89$, corresponding to May 2006 was detected in both mean and variance. For another monitoring site, Golães, we also obtained a change-point in mean and variance for $k = 77$. Thus, the change-point was detected in May 2005. The values obtained for Schwarz Information Criterion was $SIC(n) = 348.51$ and $min_{2\leq k\leq 154} SIC(k) = SIC(77) = 312.436$. The representation of the series without seasonality and the existing changes can be seen in Figure 1 with the respective 95% confidence intervals

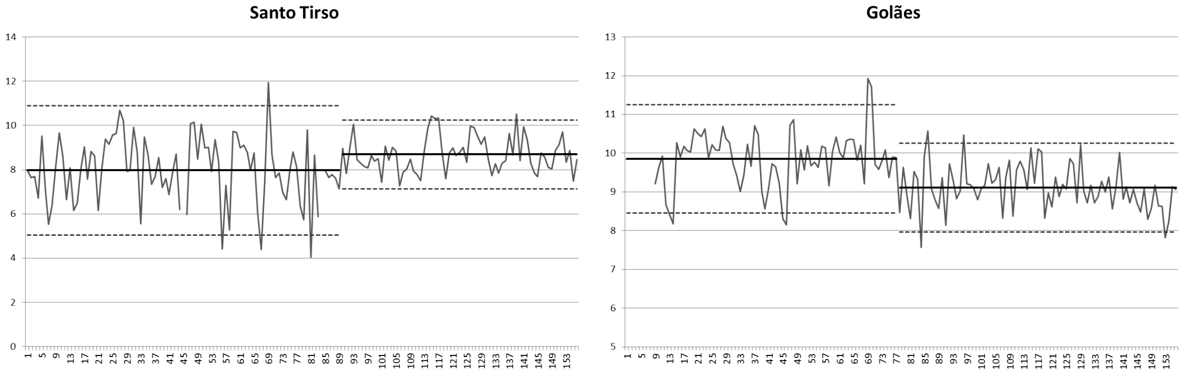

Figure 1: Santo Tirso and Golães series without seasonality, with its means before and after change-point and respective 95% confidence intervals.

Table 1 shows the estimates of both mean and standard deviation before and after the change-point detected. On the one hand, in Santo Tirso was detected a water quality improvement with a significant increase of the DO concentration mean in about 0.725, with a decrease of the variability at same time. On the other hand, in the monitoring site Golães, the water quality assessed in this context with the DO concentration has deteriorated from May 2005 in about of 0.740, while also reducing the variability.

	Change-point	$\hat{\mu}_1$	$\hat{\mu}_n$	$\hat{\sigma}_1^2$	$\hat{\sigma}_n^2$
Santo Tirso	May-2006	7.968	8.693	2.223	0.623
Golães	May-2005	9.851	9.111	0.507	0.340

Table 1: Parameters estimates before and after the change-points.

In order to investigate if the assumptions of the approach are verified, both ACF and PACF graphics of series $X_i - \hat{\mu}$ were drawn with the respective estimates of μ before and after change-points for Santo Tirso and Golães. This analysis shows that the data has a weak temporal correlation which must be considered in future work.

Acknowledgments. A. Manuela Gonçalves was partially financed by FEDER Funds through *"Programa Operacional Factores de Competitividade - COMPETE"* and by Portuguese Funds through FCT - *"Fundação para a Ciência e a Tecnologia"*, within the Project Est-C/MAT/UI0013/2011. Marco Costa was partially supported by Fundação para a Ciência e a Tecnologia, PEst OE/MAT/UI0209/2011.

References

[1] Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEE Trans. Auto. Control* **19**, 716–723.

[2] Beaulieu, C., Chen, J., Sarmiento, J.L. (2012). Change-point analysis as a tool to detect abrupt climate variations. *Phil. Trans. R. Soc. A 2012 370*, 1228–1249.

[3] Chen, J., Gupta, A.K. (1997). Testing and locating variance change points with application to stock prices. *J. Am. Stat. Assoc.* **92**, 739–747.

[4] Chen, J., Gupta, A.K. (1999). Change point analysis of a Gaussian model. *Stat. Papers* **40**, 323–333.

[5] Chen, J., Gupta, A.K. (2001). On change-point detection and estimation. *Comm. Statist. Simulation Comput*, **30(3)**, 665–697.

[6] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464.