Application of Change-Point Detection to a Structural Component of Water Quality Variables

A. Manuela Gonçalves[*,†] and Marco Costa[**,‡]

[*] *Departamento de Matemática e Aplicações, Universidade do Minho, Portugal*
[†] *CMAT - Centro de Matemática da Universidade do Minho*
[**] *ESTGA - Escola Superior de Tecnologia e Gestão de Águeda, Universidade de Aveiro, Portugal*
[‡] *CMAF - Centro de Matemática e Aplicações Fundamentais da Universidade de Lisboa*

Abstract. In this study, methodologies were developed in statistical time series models, such as multivariate state-space models, to be applied to water quality variables in a river basin. In the modelling process it is considered a latent variable that allows incorporating a structural component, such as seasonality, in a dynamic way and a change-point detection method is applied to the structural component in order to identify possible changes in the water quality variables in consideration.

Keywords: water quality, state-space modelling, hydro-meteorological factor, change-point detection
PACS: 05.45.Tp, 02.50.Ey

Introduction

This study focuses on a rather extended data set relative to the River Ave basin in Northwest Portugal and consists mainly of monthly measurements of physical-chemical and microbiological variables in a network of water quality monitoring sites and of monthly precipitation in an udometric network of meteorological monitoring sites. The methodology is applied to dissolved oxygen concentrations levels (DO) (mg/l) in 8 monitoring sites in the River Ave basin over a 12-year period (1998-2009). The proposed methodology starts by using a multivariate statistical approach–cluster analysis–to classify the water quality monitoring sites into homogeneous space-time groups based on the DO quality variable which was selected and considered relevant to characterize the water quality. A hydro-meteorological factor is constructed, by using ordinary Kriging method for each quality monitoring site (totalling 8 sites) based on the analysis of the space-time behaviour of the precipitation (monthly total) observed in a udometric network constituted by a total of 19 meteorological sites located in the geographical area of the River Ave basin. For each cluster, a linear state-space model was fitted to modelling the DO concentration quality variable by taking into account the seasonal variation throughout the year and the estimated hydro-meteorological factor. By means of the Kalman filter algorithm ([3]), are obtained filtered predictions of states which allow separating the DO evolution in both hydro-meteorological conditions and structural or endogenous components. The separation of these two sources that contribute to DO concentration allows evaluating the existence of any change-point in time considering only the endogenous component evolution. For this, it is applied a change-point detection procedure of *maximum type* in order to assess if, by excluding the hydro-meteorological factor, there is a change-point in DO concentration possibly due to large investments in infrastructure or to government inspections. There is a huge literature on change-point detection in climate or environmental settings. For instance, [7] makes a review of this topic for models with independent and identically distributed (IID) errors and [6] develops a test for undocumented change-points for periodic and autocorrelated time series. Nevertheless, the present work considers basic statistical tests by applying *maximum type* statistics because they have been considered useful and cover a large set of situations, ([5]).

Cluster analysis

Hierarchical agglomerative CA was performed on the raw data set by means of Ward's method. It used a dissimilarity measure that corresponds to the average of this distance over all months t where there is observed value of the DO

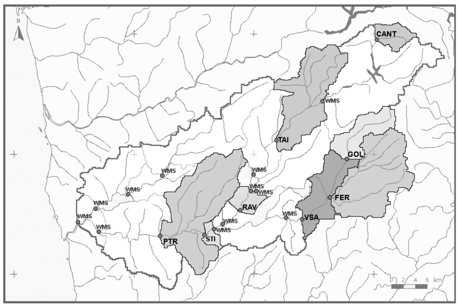

FIGURE 1. The limits of the hydrological basins as defined by the water quality monitoring sites (WMS).

quality variable with measurements in the two sites i and j, i.e.

$$d_{ij} = \frac{1}{\#M_{ij}} \sum_{t \in M_{ij}} \left| x_{it} - x_{jt} \right|, \quad i,j = 1,\ldots,8,$$

where M_{ij} is the set of all months with DO measured in both sites i and j. By using Ward's method, two well-differentiated clusters were observed and the results confirm previous knowledge about the effluents discharge according to the economic activities located along the River Ave basin. Also, the effects of these discharges in the water quality vary according to natural and geographical/economical reasons. Cluster I is composed by the monitoring sites CANT, TAI, GOL, FER, and VSA. There is a set of locations which have the best water quality indicators (the highest values obtained from the DO concentration), including sites situated upstream the Rivers Ave and Vizela (CANT corresponds to the source of River Ave). In Cluster II, comprised of the three monitoring sites RAV, STI, and PTR located in the River Ave near the most polluted area of the Ponte Trofa and Santo Tirso Municipalities, there is a growing urban population and a high concentration of industrial activity, and it is also where the Ave receives similarly polluted waters from its adjacent rivers (Selho and Vizela), and, consequently, these sites present the worst water quality.

The hydro-meteorological factor

For each water monitoring site, the monthly mean area precipitation was computed in its influence region based on the average point prediction by using Kriging stochastic methodology. In this context, the influence regions of each water monitoring site were defined by the INAG (Portuguese Institute of Water) technicians and they are corroborated by the region's topography and the land's drainage dynamics. Figure 1 shows the River Ave hydrological basin with its influence areas delineated. The precipitation amount of the sub-basin A_i associated to the water monitoring site i in month t is estimated by $p_t^{*(i)} = \widehat{Z}_t(A_i) \times a_i$, where a_i is the sub-basin's area in Km^2. For re-scale purposes, it is considered the proportional value $p_t^{(i)} = p_t^{*(i)} \times 10^{-3}$. In order to construct a hydro-meteorological factor with a stronger linear correlation to the response variable (DO concentration), after an exploratory analysis the hydro-meteorological factor $h_t^{(i)}$ is defined as the logarithm of a linear combination of the covariates $p_{t-1}^{(i)}$ and $p_{t-2}^{(i)}$, as follows $h_t^{(i)} = \log\left(0.7\ p_{t-1}^{(i)} + 0.3\ p_{t-2}^{(i)}\right)$.

The linear state-space model

In order to accommodate the temporal correlation structure, linear state-space (LSS) models are considered as follows. Suppose there are measures of the water quality variable classified in k clusters of sample sites where cluster i has k_i water monitoring sites, with $i = 1,2,..,k$. The state-space model for cluster i is:

$$\begin{aligned} \mathbf{Y}_t &= [\mathbf{h}_t|\mathbf{s}_t]\,\boldsymbol{\beta}_t + \mathbf{e}_t \qquad (1) \\ \boldsymbol{\beta}_t &= \boldsymbol{\mu} + \boldsymbol{\Phi}\left(\boldsymbol{\beta}_{t-1} - \boldsymbol{\mu}\right) + \boldsymbol{\varepsilon}_t \qquad (2) \end{aligned}$$

TABLE 1. Parameters estimates of linear state-space models for Clusters I and II.

Cluster	$\boldsymbol{\mu}$	$\boldsymbol{\Phi}$		$\boldsymbol{\Sigma_\varepsilon}$		σ_e^2				
I h_t	-0.73	0.53	0.01	1.23	-0.09	CANT	FER	TAI	GOL	VSA
s_t	1.04	1.01	0.53	-0.09	0.01	0.46	0.91	0.36	0.26	0.86
II h_t	0.02	0.27	-1.45	0.34	0.02	RAV	STI	PTR		
s_t	1.01	-0.03	0.21	0.02	0.01	1.07	1.37	0.57		

where the measurement equation, Eq. 1, relates the observable water quality variable $\mathbf{Y}_t = [\ Y_{1,t} \quad Y_{2,t} \quad \cdots \quad Y_{k_i,t}\ ]'$ in the k_i sites in cluster i with the vector of unobservable variables, $\boldsymbol{\beta}_t = [\ \beta_{h,t} \quad \beta_{s,t}\ ]'$, called states. The $k_i \times 2$ matrix $\mathbf{A}_t = [\mathbf{h}_t|\mathbf{s}_t]$ is a matrix of known values and accommodates the hydro-meteorological factor and the seasonal component, respectively. Thus, matrices $\mathbf{h}_t$ and $\mathbf{s}_t$ are column matrices with the form $\mathbf{h}_t = [\ h_{1,t} \quad h_{2,t} \quad \cdots \quad h_{k_i,t}\ ]'$ and $\mathbf{s}_t = \mathbf{1}_{k_i} s_t$. For simplification, the seasonal coefficients are taken by the mean of the monthly averages of water quality variable inside each cluster. The error term $\mathbf{e}_t$ is a white noise $k_i \times 1$ vector, called the measurement error, with a covariance matrix $\boldsymbol{\Sigma}_{\mathbf{e}}$, which may be a diagonal covariance matrix $\boldsymbol{\Sigma}_{\mathbf{e}} = diag\{\sigma_1^2, \sigma_2^2, \cdots, \sigma_{k_i}^2\}$, for simplification. The state process $\{\boldsymbol{\beta}_t\}$ follows a stationary VAR(1) according to Eq. 2, the state equation, with a mean given by the 2×1 vector $\boldsymbol{\mu}$. To secure the stationarity of the state equation, it is assumed that the eigenvalues of the autoregressive matrix $\boldsymbol{\Phi}$ are inside the unit circle and that $\boldsymbol{\varepsilon}_t$ is a white noise vector with covariance matrix $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$. Furthermore, the noises $\mathbf{e}_t$ and $\boldsymbol{\varepsilon}_t$ are serially uncorrelated, i.e., $E(\mathbf{e}_t\boldsymbol{\varepsilon}_r') = \mathbf{0}$ for all t and r. The vector of unknown parameters $\boldsymbol{\Theta} = \{\boldsymbol{\mu}, \boldsymbol{\Phi}, \boldsymbol{\Sigma}_{\mathbf{e}}, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\}$ must be estimated from the data. Parameters are estimated by distribution-free estimators based on the generalized method of moments, as a generalization of [2], and they are presented in Table 1. Considering the parameters estimates previously obtained, the recursive equations of the Kalman filter are performed and filtered predictions of $\beta_{h,t}$ and $\beta_{s,t}$ are computed at each time t for Cluster I and II. These values are represented in Figure 3.

Change-point detection

In order to discover a sudden change in the behaviour of the structural component, based on the calibrations factors of seasonality, a basic statistical test applying maximum type statistics to detect a change in location is performed. Not considering the hydro-meteorological impact in DO concentration, and if there are no structural changes in the water quality sources, it is expected that the calibration factor $\beta_{s,t}$ has no changes over time in its mean. It is applied a statistical test to the filtered predictions of $\beta_{s,t|t}$ by taking into account that $\beta_{s,1|1}, \beta_{s,2|2}, ..., \beta_{s,140|140}$ are an AR(1) process. [1] shows that when random variables $X_1, X_2, ..., X_n$ are not independent but form an ARMA sequence then the asymptotic critical values of the test statistics considering independence have to be multiplied by $\sqrt{2\pi f(0)/\gamma}$, where $\gamma = varX_t$ and $f(\cdot)$ denote the special density of the corresponding ARMA process. Especially for an AR(1) sequence, the critical values should be multiplied by $\left[(1+\rho)(1-\rho)^{-1}\right]^{1/2}$ where ρ is the first autoregressive coefficient, ([4]). Briefly, it is exposed the test of *maximum type*

H_0: $\beta_{s,t|t}$ are an $AR(1)$ sequence normally distributed variables according to the same $N(\mu, \sigma^2)$

vs

H_1: there is a time point $k \in \{1, ..., n-1\}$ such that $\beta_{s,1|1}, ..., \beta_{s,k|k}$ are distributed according to $N(\mu_1, \sigma^2)$ and $\beta_{s,k+1|k+1}, ..., \beta_{s,n|n}$ are distributed according to $N(\mu_2, \sigma^2)$.

Supposing σ^2 is unknown, then the test statistic $T(n)$ is the maximum of the absolute values of two sample t-test statistics

$$T(n) = \max_{1 \le k < n} |T_k| = \max_{1 \le k < n} \sqrt{\frac{(n-k)k}{n}} \; \left|\overline{\beta}_{s,k|k} - \overline{\beta}^{*}_{s,k|k}\right| \; \frac{1}{s_k}$$

where

$$\overline{\beta}_{s,k|k} = \frac{1}{k}\sum_{i=1}^{k} \overline{\beta}_{s,i|i}, \; \overline{\beta}^{*}_{s,k|k} = \frac{1}{n-k}\sum_{i=k+1}^{n} \overline{\beta}_{s,i|i} \text{ and } s_k = \sqrt{\frac{1}{n-2}\left[\sum_{i=1}^{k}\left(\beta_{s,i|i} - \overline{\beta}_{s,k|k}\right)^2 + \sum_{i=k+1}^{n}\left(\beta_{s,i|i} - \overline{\beta}^{*}_{s,k|k}\right)^2\right]}.$$

The null hypothesis can be rejected if the statistic $T(n)$ is greater than the critical value. The exact distribution of $T(n)$ for IID sequence is so complex and was derived by [8], but was able to calculate the critical values only

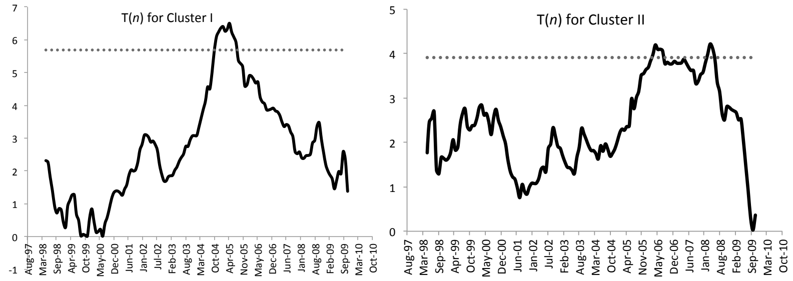

FIGURE 2. Statistics $\{T(n)\}$ corresponding to $\beta_{s,t|t}$ processes for Clusters I and II.

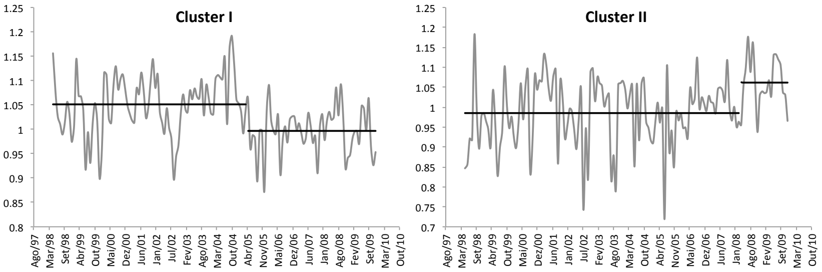

FIGURE 3. Representations of $\beta_{s,t|t}$ and its means before and after change-point detected, for each cluster.

for the number of observations n less than 10. Alternatively, approximate critical values can be computed by other methods, namely, the Bonferroni inequality, simulation or the asymptotic distribution. The present work uses critical values obtained by simulation achieved in [4], where more details can be obtained about approximation methods of the critical values. Statistics values $\{T(n)\}$ were computed considering the sequences of $\beta_{s,t|t}$ for each Cluster I and II and they are plotted in Figure 2. Based on [4], the interpolated value to $n = 140$ is 3.172. In order to obtain 5% critical values, the value 3.172 is multiplied by $\left[(1+\widehat{\rho})(1-\widehat{\rho})^{-1}\right]^{1/2}$, where $\widehat{\rho}$ is obtained in Table 1. Thus, 5% critical values for Clusters I and II are 5.691 and 3.913, respectively. In the case of Cluster I, the maximum of the statistics $\{T(n)\}$ occurred for $k = 85$, i.e. in May 2005, and was 6.505. For Cluster II, $T(n) = 4.218$ and occurs for $k = 120$, i.e. in April 2008. Both values are greater then 5% critical values, so, in both cases, the null hypothesis is rejected at the 0.05 significance level. Having detecting the changes, the question may arise if there is some other change in addition to the two ones detected. In order to confirm this, the test was separately applied to the series before the change-point and to the series after the change, but the test did not discover inhomogeneities in the series. Figure 3 shows the averages for each part of the series considering change-points detected.

REFERENCES

1. J. Antoch, M. Hušková and Z. Práškova, Effect of dependence on statistics for determination of change, *Journal of Statistical Planning and Inference* **60**, 291–310 (1997).
2. M. Costa and T. Alpuim, Parameter estimation of state space models for univariate observations, *Journal of Statistical Planning and Inference* **140**, 1889–1902 (2010).
3. A. C. Harvey, *Forecasting, structural time series models and Kalman filter*, University Press, Cambridge, 1996.
4. D. Jarušková, Change-point detection meteorological measurement, *Monthly Weather Review* **124**, 1535–1543 (1996).
5. D. Jarušková, Some problems with application of change-point detection methods to environmental data, *Environmetrics* **8**, 469–483 (1997).
6. R. Lund, X. L. Wang, Q. Lu, J. Reeves, C. Gallacher and Y. Feng, Changepoint Detection in Periodic and Autocorrelated Time Series, *Journal of Climate* **20**, 5178–5190 (2007).
7. J. Reeves, J. Chen, X.L. Wang, R.B. Lund and Q. Lu, A review and comparison of changepoint detection techniques for climate data, *Journal of Applied Meteorology and Climatology* **46**, 900–915 (2007).
8. K. J. Worsley, On the likelihood ratio test for a shift in location of normal populations, *Journal of the American Statistical Association* **74**, 365–367 (1979).