Time Series Analysis by State Space Models Applied to a Water Quality Data in Portugal

A. Manuela Gonçalves[1,a)], Olexandr Baturin[2,b)] and Marco Costa[3,c)]

[1]*CMAT-Centre of Mathematics, DMA-Department of Mathematics and Applications, University of Minho, Portugal*
[2]*DMA-Department of Mathematics and Applications, University of Minho, Portugal*
[3]*CIDMA-Centre for Research and Development in Mathematics and Applications, University of Aveiro, Portugal*

[a)]mneves@math.uminho.pt
[b)]olexandr.baturin@gmail.com
[c)]marco@ua.pt

Abstract. Time series analysis by state space models provide a very flexible tool for analysing dynamic phenomena and evolving systems, and have significantly contributed to extending the classical domains of application of statistical time series analysis. In this study, in the context of a surface water quality monitoring problem in a river basin, it is proposed an approach for the structural time series analysis based on the state space models associated to the Kalman filter. The main goals are to analyse and evaluate the temporal evolution of the environmental time series, and to identify trends or possible changes in the water quality on a dynamic monitoring procedure. The data concerns the River Ave's hydrological basin located in the Northwest of Portugal, where monitoring has become a priority in water quality planning and management because its water has been in a state of obvious environmental degradation for many years. As a result, the watershed is now monitored by seven monitoring sites distributed along the River Ave and its main streams. For the modeling process we consider the monthly dissolved oxygen concentration dataset between January 1999 and January 2014. The framework of the state space models shows versatility to incorporate unobserved components, such as trends, cycles and seasonals, that have a natural interpretation and represent the salient features of the environmental time series under investigation. From the environmental point of view, the proposed approach allows to obtain pertinent findings concerning water surface quality interpretation and change point, thus highlighting the potential value of this type of analysis, and it is also relevant to identify unanticipated changes that are important in the management process and for the assessment of water quality.

INTRODUCTION AND DESCRIPTION OF THE DATA

In recent years there has been an increasing interest in the application of state space models in time series analysis. State space models consider a time series as the output of a dynamic system perturbed by random disturbances. They allow a natural interpretation of a time series as the combination of several components, such as trend, seasonal or regressive components [7]. A structural model can therefore not only provide forecasts, but also, through estimates of the components, present a set of stylised facts and this formulation will allow making some useful interpretations [2], [3] and [4]. In this study, it is proposed a dynamic modeling procedure based on the state space approach (associated to the Kalman filter) in time series of water quality variables.

This work focuses on data concerning the River Ave's hydrological basin located in the Northwest of Portugal, and its main adjacent streams: the rivers Este, Selho, and Vizela. The surface water of River Ave has high pollution levels and the water quality measurements failed to comply with the objectives of minimum quality for surface waters prescribed by the Portuguese legislation. The Northern Regional Directory for the Environment and Natural Resources (DRARN) and the National Institute of Water (INAG) have been collecting various water quality variables (monthly physicochemical and microbiological analyses) from 7 quality monitoring sites (Figure 1). There are more than 23 water quality variables available, and they are selected according to their importance in the evaluation of river's water quality (point sources: industry, domestic wastewater, agriculture, wastewater treatment plants).

At this time, we focus on the Dissolved Oxygen (DO) measured in mg/l. DO refers to the level of free, non-compound oxygen present in water or other liquids. In this modeling process we considered series from 7 monitoring sites (TAI, RAV, STI, PTR, GOL, FER, and VSA) and the data derives from monthly observations between January

International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2017)
AIP Conf. Proc. 1978, 470101-1–470101-4; https://doi.org/10.1063/1.5044171
Published by AIP Publishing. 978-0-7354-1690-1/$30.00

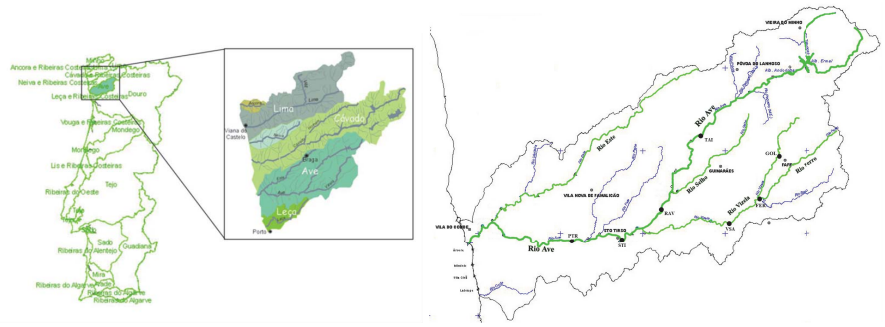

FIGURE 1. (Left) Geographical location of River Ave hydrological basin. (Right) Spatial distribution of the 7 water quality monitoring sites in the River Ave's basin and its main adjacent stream (River Vizela).

1999 and January 2014, ($n = 181$ months). An exploratory analysis was performed (Table 1). The data exhibited seasonal behavior (by analyzing the monthly averages) and sometimes a moderate temporal dependency [1], [2], and [3]. The linearity of the possible trend that was suggested in [2] by a graphical analysis of the 12 times series corresponding to each month was found suitable for the time series. In this paper we consider data series from one water monitoring site FER (FERRO) on pretence of example of the modeling approach.

TABLE 1. Descriptive statistics of Dissolved Oxygen (DO) in the 7 water monitoring sites.

Monitoring site	Abbrev.	River	N. obs.	Range	Average	St. dev.	Missing data
Taipas	TAI	Ave	176	6.60-11.72	9.334	1.108	5
Riba d'Ave	RAV	Ave	180	1.80-11.70	8.580	1.660	1
Santo Tirso	STI	Ave	180	1.67-12.00	8.318	2.024	1
Ponte Trofa	PTR	Ave	180	2.00-11.70	8.088	1.882	1
Ferro	FER	Vizela	175	6.10-11.70	9.492	1.113	4
Golães	GOL	Vizela	176	6.60-11.70	9.426	1.082	5
Vizela (Santo Adrião)	VSA	Vizela	176	5.40-12.40	9.526	1.156	5

Modeling Approach

As the series of observations present intrinsic environmental data proprieties, the initial model is very versatile since it can accommodate several statistical properties often presented in environmental data, such as a stochastic local level, a stochastic slope, stochastic seasonality (allowed to vary over time) and temporal correlation. It is proposed an application of a structural time series model by taking into account this data structure. So, the monthly time series are modeled by equations

$$Y_t = \mu_t + \beta_t + \gamma_t + e_t \tag{1}$$

$$\mu_{t+1} = \mu_t + v_t + \varepsilon_{1,t}$$

$$v_{t+1} = v_t + \varepsilon_{2,t}$$

$$\beta_{t+1} = \phi\beta_t + \varepsilon_{3,t}$$

$$\gamma_{1,t+1} = -\gamma_{1,t} - \gamma_{2,t} - \gamma_{3,t} - \gamma_{4,t} - \gamma_{5,t} - \gamma_{6,t} - \gamma_{7,t} - \gamma_{8,t} - \gamma_{9,t} - \gamma_{10,t} - \gamma_{11,t} + \varepsilon_{4,t}$$

$$\gamma_{2,t+1} = \gamma_{1,t}, \gamma_{3,t+1} = \gamma_{2,t}, \cdots, \gamma_{11,t+1} = \gamma_{10,t}$$

with $e_t \sim N(0, \sigma_e^2)$, $\varepsilon_{1,t} \sim N(0, \sigma_{\varepsilon_1}^2)$, $\varepsilon_{2,t} \sim N(0, \sigma_{\varepsilon_2}^2)$, $\varepsilon_{3,t} \sim N(0, \sigma_{\varepsilon_3}^2)$ and $\varepsilon_{4,t} \sim N(0, \sigma_{\varepsilon_4}^2)$.

The first equation represents the structure for the observed data, the observation equation, where Y_t is the monthly environmental variable, in this case OD variable with $t = 1, 2, ..., 181$. The observations are driven by three state equations where the local linear trend contains two state equations: one for modeling the level, and one for the modeling the slope. The third state equation represents the seasonal effect by adding a seasonal component (requires $(s-1)$ state equations) where s is given by the periodicity of the seasonal, we have $s = 12$. And to deal with the temporal correlation structure it is considered β_t following a 1st order autoregressive process, AR(1). Since the model has a state space representation, it allows obtaining forecast or other predictions of interest (filtered or smoother predictions). The parameters of the state space models must be estimated for each environmental series, and they are estimated by Gaussian maximum likelihood estimation in the state space framework. Since the state process is unobserved, both forecast and filtered predictions are obtained through the Kalman filter algorithm. The referred methodology was applied in R environment (http://www.r-project.org/), [5] and [6].

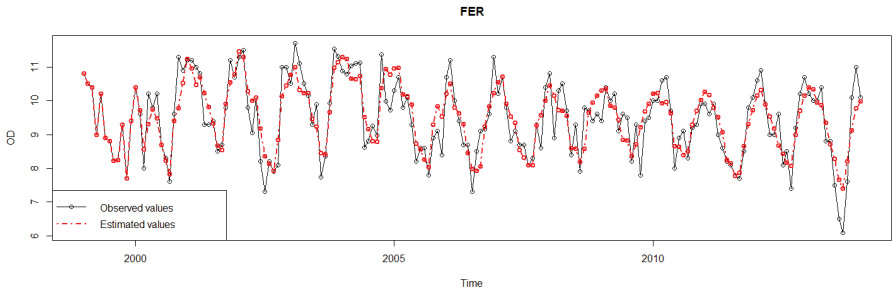

FIGURE 2. Observed values and filtered state estimates of DO in FER (Ferro) water monitoring site.

As starting point, we fitted a state space model by incorporating the several components for the DO time series in FER. A first analysis of the adjustment model was done, and we concluded for this monitoring site that the innovations are uncorrelated, so it was not necessary to consider a 1st order autoregressive process in the model. All the other structural components were considered. Figure 2 shows the observed DO concentration and the filtered predictions in FER for the final model. The maximum likelihood estimates of the state variances are given by $\sigma_e^2 = 4.7268 \times 10^{-1}$, $\sigma_{\varepsilon_1}^2 = 1.7854 \times 10^{-2}$, $\sigma_{\varepsilon_2}^2 = 1.1780 \times 10^{-18}$, and $\sigma_{\varepsilon_4}^2 = 3.0829 \times 10^{-9}$, respectively. The empirical root mean square errors (RMSE) is $RMSE = 0.539$ in the final model.

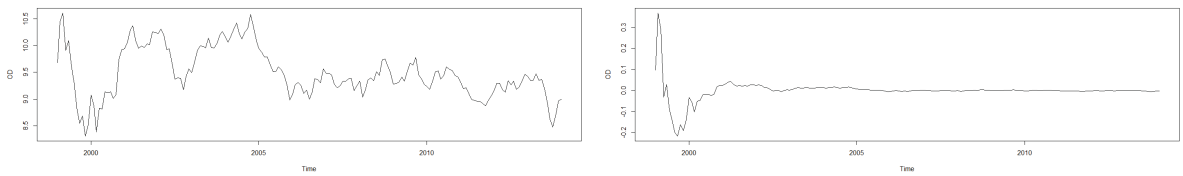

FIGURE 3. (Left) Filtered state of the stochastic level component. (Right) Filtered state of the stochastic slope component.

Plots of the stochastic components obtained from this analysis are displayed in Figures 3 and 4. As the graphics show, the model captures changes in a dynamic that way overlaps the default behavior evidenced by the several components and provides a useful tool to evaluate the behavior and changes in real time concerning DO concentration in each month. The filtered level predictions μ_t indicate that there have been some significant level changes by 2005, and, from that date on there is a level with less variability, but with an indication for reduction (to be confirmed through methodologies of nonparametric statistics for trend analysis: for example, the Mann-Kendall test). The plot of the filtered predictions of the stochastic slope suggests that is a change of the slope signal (positive for negative) before 2000 corresponding to a water quality deterioration and the slope tends to stabilize around mid-2001, assuming, however, negative values. This means that the DO concentration in FER tends to decrease at the end of the reporting period, indicating a slow but continuous water quality degradation with respect to this physicochemical indicator.

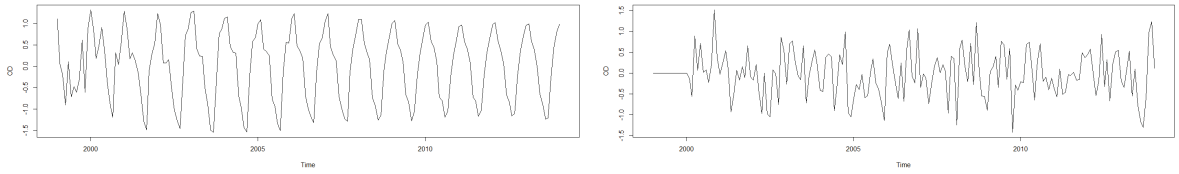

FIGURE 4. (Left) Filtered state of the stochastic seasonal component. (Right) Irregular component.

The filtered state of the stochastic seasonal component indicates, as expected given the previous works [1], [2] and [3], that the DO is related to the hydrometeorological conditions, by verifying a strong seasonality. On the one hand, there is an annual pattern: in the winter months (lower temperatures and wettest months) DO presents higher values, while in the summer months (hot and dry) DO concentration is lower. The plot of the irregular component shows that the residuals do not present a correlation structure (also verified by analysis of the FAC and FACP), thus validating the main assumptions associated with the adjusted space state model. The model verifies the assumption of normality which was also checked by the residuals distribution.

CONCLUSIONS

The analysis performed in this study shows that the structural time series models (the state space models associated to the Kalman filter) are suitable to model DO concentration series at the FER water monitoring site, and they allow to obtain pertinent findings concerning water surface quality interpretation and change point of view [2] and [4], thus highlighting the potential value of this type of analysis. In addition, the state-space approach allows doing an online monitoring procedure to detect DO concentration values that are statistically unexpected. The next step is to analyze the filtered and smoothed predictions (forecasts) of states given by the Kalman filter, which allows an interesting analysis of the structural components, and further research is in progress to improve the modeling process and the results obtained.

ACKNOWLEDGMENTS

A. Manuela Gonçalves was supported by the Research Centre of Mathematics of the University of Minho with the Portuguese Funds from the FCT-Fundação para a Ciência e a Tecnologia, through the Project PEstOE/MAT/UI0013/2014. Marco Costa was supported by Portuguese funds through the CIDMA-Centre 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.

REFERENCES

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

[2] A. M. Gonçalves and M. Costa, Predicting seasonal and hydro-meteorological impact in environmental variables modelling via Kalman filtering, *Stochastic Environmental Research and Risk assessment* 127(5):1021–1038 (2013).

[3] M. Costa and A. M. Gonçalves, "Combining Statistical Methodologies in Water Quality Monitoring in a Hydrological Basin - Space and Time Approaches," In *Water Quality Monitoring and Assessment*, edited by Kostas Voudouris and Dimitra Voutsa (InTech Published, 2012) pp. 121–142.

[4] A. M. Gonçalves and M. Costa, "Application of Change-Point Detection to a Structural Component of Water Quality Variables," In *AIP Conference Proceedings* Vol 1389, edited by Simos, T.E., Psihoyios, G., Tsitouras, Ch. (American Institute of Physics Publishing, 2011), pp. 1565–1568.

[5] D. R Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2015)

[6] G. Petris, S. Petrone and P. Campagnoli, *Dynamic Linear Models with R* (Springer, New York, 2009).

[7] A. C. Harvey, *Forecasting, structural time series models and Kalman Filter* (Cambridge University Press, Cambridge, 1996).