A State-Space and Clustering Approach for Analyzing the Water Quality in a River Basin

Arminda Manuela Gonçalves[1] and Marco Costa[2]

[1] Department of Mathematics for Science and Technology, University of Minho Campus de Azurém, 4800-058 Guimarães, Portugal
e-mail: mneves@mct.uminho.pt
[2] Higher School of Technology and Management of Águeda, University of Aveiro, Apartado 473, 3754-909 Águeda, Portugal
e-mail: marco@ua.pt

Abstract: The aim of this contribution is to apply the state-space models to identify homogeneous groups of water quality monitoring sites based on comparison of temporal dynamics of the concentration of pollutants in the surface water of a river basin. This comparison is performed using the Kullback information, adapting the approach used in Bengtsson and Cavanaugh (2007). The purpose of our study is to identify spatial and temporal patterns.

Keywords: hydrological basin, water quality, state-space modelling, Kalman filter, classification.

1 Introduction

The aim of this study is to identify homogeneous regions, based on similarities in the temporal dynamics of variables of water quality measured patterns, by observing hydrological series (recorded in time and space) in a river basin (in a geographical region) with the purpose to evaluate the surface water quality. This research follows the work done in done in Gonçalves (2006) and Costa (2006). We considered the Ave river hydrological basin located in the north-west of Portugal, with an approximate basin area of $1390Km^2$ and its main stream length of $101Km$.
In this work we intend to continue with the goal of contributing to the discussion and understanding of an environmental issue of such a high importance to the community, as is the case of the quality control of the surface water of river Ave basin. The water is a precious asset as well as a potential inducer of riches. In a region such as the Ave valley, with its economic ground highly dependent on industry (predominantly textile, there are 340 registered factories), water plays without any doubt a determining role in assigning industry to this valley. The water streams of this region have been in a situation of obvious environmental degradation, for many years. The worsening of the environmental situation of this basin has led, from

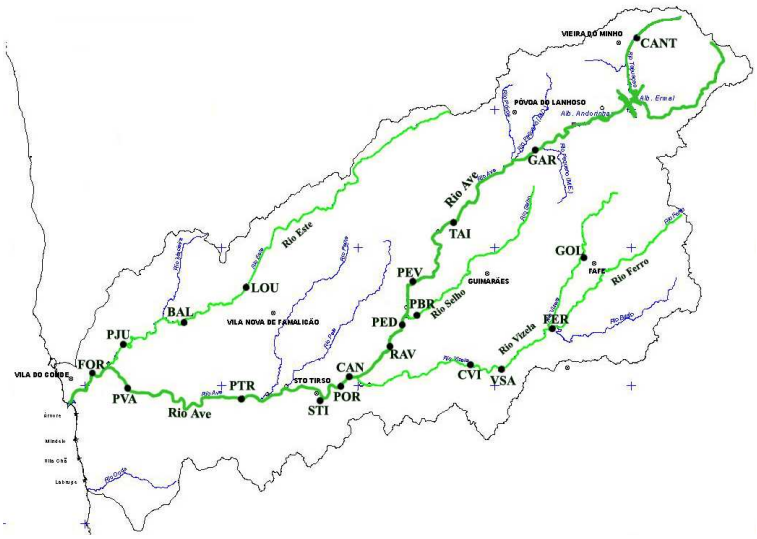

FIGURE 1. Spatial distribution of the water quality monitoring sites of the Ave river hydrological basin.

the seventies onwards, the government authorities to be concerned with the increase of the water pollution in this basin. Since 1988, and as part of a national plan, the Central Administration, through the Regional Directory for the Environment and Natural Resources Northern, and the Institute of Water monitored the quality of surface water periodically (monthly) along the Ave river and its main adjacent streams.

2 Data set description

The Regional Directory for the Environment and Natural Resources Northern and the Institute of Water has been collecting various water quality variables (monthly physical-chemical and microbiological analysis) from 19 quality monitoring sites. The data sets of 19 water quality monitoring sites, comprising 11 water quality variables have been measured monthly between 1988 and 2006. Although there are more than 23 water quality variables available, only 11 variables are selected due to their continuity in measurement at all selected water quality monitoring sites and their importance in the evaluation on river water quality (point sources: industry, domestic wastewater, agriculture, wastewater treatment plants). We intend, with an alternative approach, develop new statistical methodologies to classify geographically homogeneous groups of water quality monitoring sites based

on similarities in the temporal dynamics of the quality variables. The proposed methodology intends to classify the water quality monitoring sites into spatial and temporal homogeneous groups, based on the quality variables, which have been selected and considered relevant to characterize the quality of the water.

3 Methods

We start with a preliminary space-time analysis of the monthly water quality variables. The complete records contain many missing observations; we have noticed many irregularities in the frequency of data collection. Applying and adapting Bengtsson and Cavanaugh (2007) methodologies we fit the monthly data using an additive, structural state-space model. As starting point, we establish the following model:

$$\begin{aligned} Y_{it} &= \mu_i + s_{it} + \beta_{it} + e_{it} \\ \beta_{it} &= \phi_i \beta_{i,t-1} + \epsilon_{it} \end{aligned}$$

where $i = 1, 2, ..., k$ denotes the monitoring sites, $t = 1, 2, ..., n_k$ the month and Y_{it} denotes the observation from the stochastic process observed, i.e, the observed quality variable at water quality monitoring site i and in month t. For the water quality monitoring site i, the model represents the quality variable as a sum of an overall constant mean, a seasonal component, a monthly quality variable anomaly and a white noise error. The deterministics components are denoted respectively by μ_i and s_{it}. The stochastic components are β_{it} and the white noise e_{it}, where β_{it} is a latent process. The state-space models, associated to the Kalman filter algorithm, are applied in several areas, in particular to environmental problems (for example in radar area rainfall estimation (Alpuim et al., 1999)).
Differing measures of the quality variables levels among the various water monitoring sites could easily be used to delineate different homogeneous regions in the Ave river hydrological basin. However, since our main interest is in the dynamics of the monthly series, we remove the overall mean and the seasonality of each series.
The monthly quality variable anomalies are an AR(1) processes where ϵ_{it} is white noise with normal distribution with zero mean and variance $\sigma^2_{\epsilon_i}$.
The error e_{it} is viewed as contributing variability that is unexplained by the structural components that we assume i.i.d. normal with zero mean and variance $\sigma^2_{e_i}$, and uncorrelated to ϵ_{it} process, i.e., $cov(\epsilon_{it}, e_{is}) = 0$, for all i, t and s.
The latent process β_{it} is an unobserved variable but can be predicted by the Kalman filter equations that produces the best linear predictor.
Using the ML-parameter estimates which were obtained using the EM algorithm, a discrepancy measure is formulated, calculated and used for clustering the quality variables data. This discrepancy measure is based on the

autoregressive anomaly process β_{it}. Considering the Kullback information (Kullback, 1968) we define a discrepancy measure based on the monthly anomaly state variable for classification of state-space processes and used to cluster monthly quality variables records from monitoring sites across Ave river hydrological basin. To more clearly identify potential clusters, the discrepancy matrix by evaluation of the discrepancy measure was subjected to clustering procedures (e.g., Hartigan 1975; Gordon 1999) and other multivariate analysis procedures (e.g., Johnson, R., Wichern, D., 1992). The results confirm the expected behaviour of temporal dynamics of concentration of pollutants (along the river and its main streams) and agree with those produced by the different classifications performed. This study illustrates the usefulness of the methodologies we implement for analysis and interpretation of complex data sets: water quality assessment, identification of pollution sources/factors and understanding temporal/spatial variations in water quality for effective river water quality management.

References

Alpuim, T. and Barbosa, S. (1999). The Kalman filter in the estimation of area precipition. *Environmetrics*, **10**, 377-394.

Bengtsson, T. and Cavanaugh, J.E. (to appear in Environmetrics). State-Space Discrimination and Clustering of Atmospheric Time Series Data Based on Kullback Information Measures.

Costa, M. (2006). *Estimaço dos Parâmetros de Modelos em Espaço de Estados*, PhD Thesis. Faculty of Sciences of University os Lisbon.

Gonçalves, A.M. (2006). *Modelação Estatística da Qualidade das Águas de Superfície da Bacia Hidrográfica do rio Ave*, PhD Thesis. Faculty of Sciences of University os Lisbon.

Gordon, A.D. (1999). *Classification*. London: Chapman and Hall/CRC.

Hartigan, J.A. (1975). *Clustering Algorithms*. New York: Wiley.

Johnson, R. and Wichern, D. (1992). *Applied Multivariate Statistical Analysis* (3 ed.). Prentice-Hall.

Kullback, S. (1968). *Information Theory and Statistics*. Dover.