Journal of Hydrology 407 (2011) 73–80

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Clustering streamflow time series for regional classification

Marcella Corduas *

Department of Statistical Sciences, University of Naples Federico II, Via L. Rodinò 22, 80138 Naples, Italy

ARTICLE INFO

Article history:
Received 29 October 2010
Received in revised form 17 April 2011
Accepted 5 July 2011
Available online 23 July 2011
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Luis E. Samaniego, Associate Editor

Keywords:
Time series clustering
Autoregressive metric
Hydrologic regionalization
Streamflow series
ARMA models

SUMMARY

The article aims to show how some dissimilarity criteria, the Mahalanobis distance between regression coefficients and the Euclidean distance between Autoregressive weights, can be applied to hydrologic time series clustering. Specifically, the temporal dynamics of streamflow time series are compared through the estimated parameters of the corresponding linear models which may include both short and long memory components. The performance of the proposed technique is assessed by means of an empirical study concerning a set of daily streamflow series recorded at sites in Oregon and Washington State.

1. Introduction

Hydrologic regionalization is a typical example, where statistical techniques are needed in order to group river flow series with similar behavior. The analysis is aimed at delineating homogeneous regions, where watersheds are similar with respect to several attributes such as physical, climatic, and hydrologic features in order to transfer information or models from gauged watershed to ungauged ones.

Several approaches have been considered in literature. Traditionally, multivariate techniques, such as clustering and discriminant analysis, have been applied to find groups of sites which can be assimilated (Tasker, 1982; Nathan and Mcmahon, 1990; Burn and Boorman, 1993). Ungauged sites, sharing similar features with one of the identified clusters, are then treated in an analogous manner as any other member of the group. The selection of relevant variables for grouping is a critical step in this approach. For this reason, most contributions are, in fact, focused on specific regional problems and on the search for the best subset of catchments' descriptors. In such a context, hydrological features are typically represented by means of synthetic indices related either to the overall behavior of the stream or to the seasonal pattern (Castellarin et al., 2001; Barberis et al., 2003; Wagener et al., 2007). Geographical proximity of the grouped sites represents a further constraint which has often been considered. A further critical issue is the choice of a threshold value of the dissimilarity measure evaluated from catchment descriptors because most techniques proposed in literature are not developed in an inferential framework.

Other approaches rely on regression analysis which helps to study the relationships between the parameters of hydrologic models and watershed characteristics. The clustering of gauged sites into homogeneous regions allows the extension of the corresponding estimated models to any ungauged river belonging to a certain region (Post and Jakeman, 1996; Merz and Bloeschl, 2004; Wagener and Wheater, 2006; Yadav et al., 2007; Li et al., 2010).

Finally, special flow properties, such as monthly mean flows or peak flows (Burn, 1990; Cooper, 2005), or flow duration and flood frequency curves (Singh, 1971; Burn, 1997), have been investigated as a means for hydrologic regionalization.

Recently, attention has been paid to the use of river flows time series in order to supplement landscape characteristics and catchment features with other data describing river flow dynamics. For example, Zoppou et al. (2002) applied wavelet analysis to extract a signature of catchment response, whereas Chiang et al. (2002a,b) discussed an articulated methodology which adds the estimated parameters from time series modeling to the set of variables useful for regional classification.

This article moves from the latter study and aims to show how some well-established dissimilarity criteria, the Mahalanobis

* Tel.: +39 0812537465; fax: +39 0812537466.
E-mail address: marcella.corduas@unina.it

0022-1694/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2011.07.008

distance (Mardia et al., 1979) between regression coefficients and the *AR* metric (Piccolo, 1990; Corduas and Piccolo, 2008), can be applied to hydrologic time series clustering. Specifically, the *AR* metric helps to locate time series comparison in the context of hypothesis testing and to add further strength to the final results. The performance of the proposed technique will be assessed by means of an empirical study concerning a set of unregulated daily streamflow series recorded at sites in Oregon and Washington (Sanborn and Bledsoe, 2006).

The article is organized as follows. Firstly, the methodology proposed in the present paper is outlined by introducing the dissimilarity measures and discussing the problem of time series clustering. Secondly, a brief description of the available data and the results from the empirical case study are illustrated. Final remarks and some insights into further methodological developments conclude the article.

2. Background and methodology

First of all, the concept of similarity between time series has to be further qualified since different aspects of temporal behavior can be the object of comparison. This can in fact concern the observed trajectory of a time series, the unobserved components, such as seasonal or long term trend, or the generating stochastic process (see Liao, 2005, for a review).

The first approach is typical of data mining techniques and has originated a vast amount of literature (Keogh and Kasetty, 2003) which has, recently, attracted the attention of hydrologists (Spate et al., 2003; Spate et al., 2006). The aim is to discover series that move similarly or closely follow certain given patterns. The proposed techniques are numerous and vary from very simple Euclidean distance between observations to more refined Dynamic Time Warping which allows non-linear alignments of observed trajectories (Sakoe and Chiba, 1974; Berndt and Clifford, 1994).

The second approach set up the comparison in frequency domain. Specifically, dissimilarity measures involving spectral densities or periodograms or smoothed periodograms have been developed (Caiado et al., 2006; Shumway, 1982). This allows for detection of resemblances or differences on selected bands of frequencies, or in other words, with respect to specific dynamic components (Ng and Huang, 1999). In the same vein, wavelet analysis has become a valuable alternative (Priestley, 1996; Chan and Fu, 1999; Zoppou et al., 2002; Smith et al., 1998).

Finally, the last approach focuses on structural dissimilarity. This is measured defining the distance between adequate parametric formulation or synthesis of the data generating processes. The latter are generally assumed to be Gaussian, linear and stationary though some extension are possible depending on the selected type of metric.

In the following sections, the temporal dynamics of streamflow time series are compared through the parameters of the corresponding linear models. For this purpose, we consider a wide class of models given by *ARIMA* and *ARFIMA* models, that often constitutes the basis for the generation of synthetic series (Bras and Rodríguez-Iturbe, 1985; Montanari et al., 2000; Grimaldi, 2004).

The use of model parameters for classification purposes is motivated by the following consideration. The pattern of observed streamflow series is, to a great extent, a product of genetic sources of flow (rain, glacier or snow melt, groundwater), climatic and landscape characteristics. The effects of the many external factors and variables affecting the dynamics are merged in the past records of the observed stream. Then, univariate time series models will take all potentially acting factors into account by relating "present values" to "past". Consequently, it is reasonable to expect that comparing streamflow series by means of the underlying regression *ARIMA* (*ARFIMA*) model may help to produce a meaningful classification. This consideration is strengthened by considering that at low timescale the influence and the interaction of driving factors with hydrological response cannot be easily recognized.

The importance of using linear model parameters as variables for streamflow classification was already underlined by Chiang et al. (2002a,b) who stated that using the estimated parameters from a seasonal regression model with *MA*(2) autocorrelated errors was more appropriate for regionalization than using only watershed features. In this article, we widen the class of models and extend the comparison to general structures allowing for operators of different nature and with varying orders.

3. The need for preadjustment

Seasonality has a prominent role in determining the pattern of monthly or daily streamflow time series. This is often represented by a strong deterministic component, with period s, which has to be preliminary modeled in order to enhance the further temporal dynamics of the phenomena under study. Wold's decomposition theorem constitutes the foundation of such a modeling step, being any stationary process represented by the sum of two mutually uncorrelated processes: one is linearly deterministic and the other purely stochastic (see Brockwell and Davis, 1991, pp. 187). Specifically, in the case of hydrologic time series, in order to take the deterministic seasonality into account, suppose that the centered time series W_t is described by the harmonic regression model:

$$W_t = \sum_{j=1}^{[s/2]} \left[a_{wj} \sin\left(\frac{2\pi jt}{s}\right) + b_{wj} \cos\left(\frac{2\pi jt}{s}\right) \right] + Z_t \tag{1}$$

where Z_t follows an *ARMA* model:

$$\varphi(B) Z_t = \vartheta(B) a_t \tag{2}$$

and a_t is a zero mean White Noise (WN) process with constant variance σ^2_{aw}, B is the backshift operator such that $B^k Z_t = Z_{t-k}$, $\forall k = 0, \pm 1, \ldots$. Moreover, the polynomials $\varphi(B) = \phi(B)\Phi(B^s) = (1 - \phi_1 B - \ldots - \phi_p B^p)(1 - \Phi_1 B^s - \ldots - \Phi_P B^{sP})$ and $\vartheta(B) = \theta(B)\Theta(B^s) = (1 - \theta_1 B - \ldots - \theta_q B^q)(1 - \Theta_1 B^s - \ldots - \Theta_Q B^{sQ})$, for any $s \geqslant 0$, have no common factors, and all the roots of $\varphi(B)\vartheta(B) = 0$ lie outside the unit circle.

In general, the two step modeling strategy, implied by the Wold's decomposition theorem, also helps to treat other special features of the time series, such as deterministic trend, purely periodic components, outliers and missing values which can be removed by regression. In addition, at this stage, we assume that the effect of data skewness have already been removed from W_t by preliminary transformations.

Having observed a time series $\{w_t, t = 1, \ldots, n_w\}$ the above model is written in a matrix form as: $\mathbf{w} = \mathbf{X}\boldsymbol{\alpha}_w + \mathbf{z}_\mathbf{w}$, where $\mathbf{X}$ is the matrix with columns the deterministic regressors, and $\boldsymbol{\alpha}_\mathbf{w} = (a_{w1}, \ldots, a_{wk}, b_{w1}, \ldots, b_{wk})'$ is the vector of coefficients.

A natural way to measure the dissimilarity between two independent time series is to compare the estimated coefficients from the corresponding harmonic regression models by means of the Mahalanobis distance. In particular, according to standard theory and under Normality assumption, the generalized least square estimated coefficients $\hat{\alpha}_\mathbf{w} \sim MN(\boldsymbol{\alpha}_\mathbf{w}, \sigma^2_{aw}(\mathbf{X}'\boldsymbol{\Omega}_\mathbf{w}^{-1}\mathbf{X})^{-1})$. The matrix $\sigma^2_{aw}\boldsymbol{\Omega}_w$ is the covariance matrix of the disturbance vector $\mathbf{z}_\mathbf{w}$. Similarly, let Y_t be a zero mean Gaussian process, independent from W_t, and represented by model (1). Using an obvious notation, denote with $\boldsymbol{\alpha}_\mathbf{y}$ the regression coefficients. Then, estimation leads to $\hat{\boldsymbol{\alpha}}_\mathbf{y} \sim MN(\boldsymbol{\alpha}_\mathbf{y}, \sigma^2_{ay}(\mathbf{X}'\boldsymbol{\Omega}_\mathbf{y}^{-1}\mathbf{X})^{-1})$. Moreover, the difference $(\hat{\boldsymbol{\alpha}}_\mathbf{w} - \hat{\boldsymbol{\alpha}}_\mathbf{y}) \sim MN(\boldsymbol{\alpha}_\mathbf{w} - \boldsymbol{\alpha}_\mathbf{y}, \sigma^2_{aw}(\mathbf{X}'\boldsymbol{\Omega}_\mathbf{w}^{-1}\mathbf{X})^{-1} + \sigma^2_{ay}(\mathbf{X}'\boldsymbol{\Omega}_\mathbf{y}^{-1}\mathbf{X})^{-1})$.