CLASSIFICATION OF DAILY STREAMFLOW DATA: A STUDY ON REGIME CHANGES

Corduas Marcella [1], Domenico Piccolo[1]

[1] Department of Political Sciences, University of Naples Federico II
(e-mail: marcella.corduas@unina.it, domenico.piccolo@unina.it)

ABSTRACT: This contribution presents a classification strategy, based on widely available statistical tools, for detecting time series that have changed flow regime in recent years. The results from the analysis of 221 time series of unregulated streamflows in the United States is discussed.

KEYWORDS: time series, classification, flow regime, AR metric.

1 Introduction

The climate change often affects the variability and persistence of river discharges that may show an alterated balance between snow and rainfall and an intensification of extreme hydrological events. Such climate-induced hydrologic changes may have relevant consequences on the freshwater ecosystem (Dhungel *et al.* , 2016). The search of simple but effective tools for river regime classification is still a topic of interest in order to investigate variations in flow regimes and evaluate future climate impact (Yang & Olivera, 2023). In this article, we present a procedure for classifying streamflow time series according to their underlying dynamic structures. We illustrate our approach analyzing streamflow data from 221 unregulated catchments in the United States (Newman & al., 2015).

2 Methods

Streamflow time series are typically characterized by a marked seasonal pattern, due to the alternating of wet and dry periods, and a persistent or long term component. The seasonality often appears as a deterministic component in the spectrum. This makes the time series unsuitable for stochastic modelling, because the marked seasonal pattern obscures the other dynamic components. At this stage, we assume that the effect of data skewness, calendar effects, outliers and missing value have already been removed by preliminary analysis and transformations and that the time series W_t has zero mean. Thus, W_t is described by the harmonic regression model:

$$W_t = \sum_{j=1}^{[s/2]} [\alpha_{wj} \sin(2\pi jt/s) + \beta_{wj} \cos(2\pi jt/s)] + Z_t \qquad (1)$$

where s denotes the seasonal period, and Z_t follows a stationary Autoregressive model, *AR(p)*:

$$\varphi(B) Z_t = a_t, \qquad (2)$$

where a_t is a Gaussian White Noise (WN) process with constant variance σ^2_{aw}. It is well known that any process with an absolutely continuous spectrum can be adequately approximated by an Autoregressive model, then (2) describes both short and long memory stationary components. The order p can be selected by BIC criterion, so that parsimonious models are preferred. Thus, the time series W_t is characterized by the coefficients estimated by GLS: $\widehat{\delta}_w = (\hat{\alpha}_{w1}, ..., \hat{\alpha}_{wk}, \hat{\beta}_{w1}, ..., \hat{\beta}_{wk})'$ and $\widehat{\varphi}_w = (\hat{\varphi}_{w1}, ..., \hat{\varphi}_{wp})'$.

Given two independent time series W_t and Y_t, the dissimilarity will be measured by comparing the seasonal and non-seasonal coefficients separately because, as already mentioned, the two components (seasonality and inertia) have a very different weight in determining the dynamics of the series.

Seasonal components are compared by evaluating the Mahalanobis distance: $M_{wy} = (\widehat{\delta}_w - \widehat{\delta}_y)'(\sigma^2_{aw}\Omega_w + \sigma^2_{ay}\Omega_y)^{-1}(\widehat{\delta}_w - \widehat{\delta}_y)$, where $\sigma^2_{a\bullet}\Omega_\bullet$ is the covariance matrix of $\widehat{\delta}_\bullet$. The dissimilarity between the residual components is measured by means of the *AR* metric (Piccolo, 1990; Corduas & Piccolo, 2008): $D_{wy} = \sqrt{\sum_{j=1}^{\infty} (\varphi_{wj} - \varphi_{yj})^2}$.

Then, the corresponding distance matrices $\mathcal{M}$ and $\mathcal{D}$ are objects of a clustering algorithm in order to identify groups of time series having similar seasonal pattern and different level of inertia. Here, we use the complete linkage method because it does not require the preliminary specification of the cluster number and produces compact clusters.

3 Results

The analysis has been conducted on 221 time series of mean daily discharge (feet^3/sec) of unregulated streamflows in the United States (available from the US Geological Survey at https://waterdata.usgs.gov/nwis/). Two non overlapping reference periods have been considered: from 1930.10.01 (or later, depending on data availability) to 1974.09.30 and from 2000.10.01 to 2021.09.30.

The complete link clustering of the Mahalanobis distance matrices, $\mathcal{M}$, evaluated in the two reference periods, leads to the identification of six clusters.

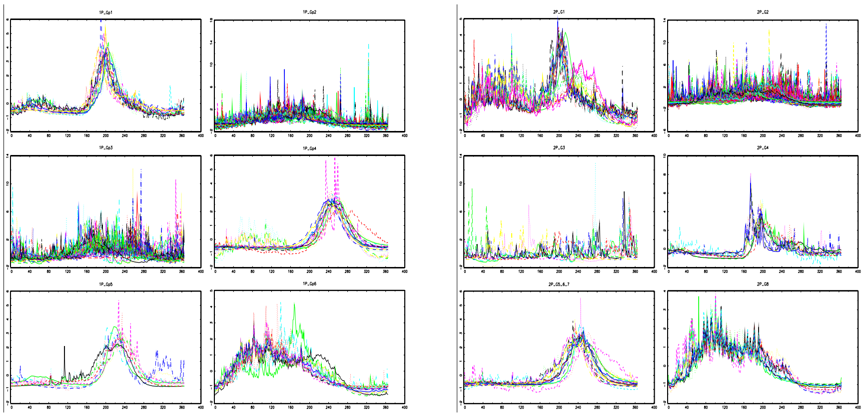

Figure 1: Average daily discharge of clustered time series (1st period-left panels; 2nd period-right panels)

In particular, the clusters describe: strong fall/spring regime (G1: mostly in the North Atlantic and Pacific NW coast); intermittent winter/spring regime (G2: mid Atlantic coast and central valleys); intermittent regime (G3: Gulf coast); weak winter regime (G4: upper Great lakes and Northern Great Planes); melt regimes (G5: mostly in the Rocky mountain and Northern Great planes); strong winter regime (G6: mostly in the NW coast). Fig.1 illustrates the average daily discharge of the series belonging to each cluster in the two reference periods. The fundamental features of the seasonal patterns are rather stable in the two periods, but a number of series (33%) have changed their class memberships. Changes are due to various factors: the anticipation of the seasonal peak due to early snow-melt, the increase of the winter rainfall, the increase of dry periods and flashy peaks. Moreover, the analysis of residual components by means of the AR metric identifies three clusters of series with increasing level of inertia (low, moderate, high). The parametric spectral densities of the cluster centroids help to define these level of inertia. However, the long term dynamics does not change remarkably in the two periods. This may be due to the fact that the residual components are heavily affected by specific physio-characteristics of the basins (for example, the slope).

At the end of the procedure, each time series is characterized by two labels specifying the seasonal regime and the level of inertia. These features can be summarized in a two-way table. In the period 2000-2021, there are 13 clusters

(Table 1) and most rivers show an intermittent regime with peaks in winter or spring and a low/medium level of inertia. The intermittent regime gather numerous series that have changed their class memberships in recent years.

Table 1: Final classification for the dataset observed the period 2000-2021

Seasonality	Inertia		
	high	medium	low
strong fall/spring	2	17	6
intermittent winter/spring	0	72	61
intermittent	0	6	6
weak winter	0	10	5
snow-melt	0	15	5
strong winter	0	15	1

4 Final remarks

The results that we have achieved using widely applicable statistical tools provide a useful basis for further discussion about the relationships of streamflow regimes with physiographic and climate indices, and for determining the future regime changes according to simulated scenarios from models driven by climate data.

References

CORDUAS, M., & PICCOLO, D. 2008. Time series clustering and classification by the autoregressive metric. *Computational Statistics & Data Analysis*, **52**, 1860–1872.

DHUNGEL, S., TARBOTON, D.G., JIN, J., & HAWKINS, C.P. 2016. Potential effects of climate change on ecologically relevant streamflow regimes. *River Research and Applications*, **32**, 1827–1840.

NEWMAN, A.J., & AL. 2015. Development of a large-sample watershed-scale hydrometeorological dataset for the contiguous USA: dataset characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, **19**, 209–223.

PICCOLO, D. 1990. A distance measure for classifying ARIMA models. *Journal of time series analysis*, **11**, 153–164.

YANG, M., & OLIVERA, F. 2023. Classification of watersheds in the conterminous United States using shape-based time-series clustering and Random Forests. *Journal of Hydrology*, **620**, 129409.