1. Abstract

In contrast to great advances on stochastic simulation techniques in hydrology and their importance on water management and uncertainty assessment studies, operational software packages for generating synthetic data are limited and hardly accessible. This limits their adoption to a narrow audience, excluding the vast majority of researchers and practitioners. In an effort to bridge this gap, we introduce CastaliaR, a software package that constitutes the R-based, open-source implementation of a state-of-the-art methodology for multivariate stochastic simulation. Its background builds upon the works of Koutsoyiannis and Manets (1996), Koutsoyiannis (1999, 2000) and Efstratiadis et al. (2014). Briefly, the overall scheme reproduces the statistical characteristics of the historical data at three temporal scales (annual, monthly and daily).

The generation procedure relies upon a symmetric moving average process for the annual scale and a periodic autoregressive process for the finer scales, while a Monte Carlo disaggregation approach re-establishes consistency across the three temporal scales.

2. Methodological framework at a glance

The synthetic time series are generated through a three-level multivariate scheme, by moving from the coarser (annual) to the finer (daily) time scale.

- The desirable stochastic structure of the annual process is implemented through a theoretical, user-specified generalized autocovariance function, given by:

\[ R(k) = \gamma_0 + \gamma_1 \cdot k + \cdots + \gamma_p \cdot k^p \]

where \( R \) is the annual variance, \( \gamma \) is a scale parameter, and \( P \) is a shape parameter, associated with the changing statistics and long term persistence (Hurst-Kalmanovgou behavior) of the process, for \( \gamma > 0 \) the process is ARMA-type (Hurst exponent \( H > 0.5 \)), while as \( \gamma \) increases it becomes HK-type.

- For the given structure, annual data are generated through a symmetric moving average (SMA) scheme:

\[ z_t = \phi_0 + \phi_1 \cdot z_{t-p} + \cdots + \phi_p \cdot z_{t-p} \]

where \( z \) are white noise, and \( \phi \) are numerical coefficients that are analytically determined from the sequence of \( \gamma \).

- The preservation of cross-correlations is ensured by generating correlated white noise, i.e. \( \gamma \neq 0 \).

- For the monthly and daily time scales, auxiliary data are initially provided by a multivariate periodic autoregressive (MPAR) scheme with gamma-distributed noise, i.e.: \( \gamma \neq 0 \).

- Where \( \gamma \), \( \phi \) are monthly-sampling parameter matrices, \( \gamma \) is the while noise, and \( t \) is the index of the month.

- Key property of daily processes is intermittency, which is preserved by employing a hybrid procedure, involving the sequential application of three empirical rules.

- A Monte-Carlo disaggregation scheme is employed to establish statistical consistency between the three scales; first the monthly data are adjusted to the known annual ones, and next the daily data are adjusted to the disaggregated monthly ones, through a multivariate coupling scheme.

3. The R software

CastaliaR consists of the open-source, R implementation of a synthetic time series generator originally developed in Delphi.

Regarding the formulation of the theoretical autocorrelation structure of the annual processes, the user can: (a) either manually parameterize the target autocorrelation function of each modelled process, or (b) employ an automatic optimization routine to fit the theoretical function to the sample autocorrelation, by minimizing the mean square error between the theoretical and empirical autocorrelations.

CastaliaR is fully automated and simplified, aiming to provide a user-friendly and practical synthetic time series generation package for engineers and researchers.

4. Case study 1: Generation of monthly data for Boeotian Kephisos basin

- Historical data: Monthly rainfall, PET, and runoff data at the basin outlet – the longest hydrological records in Greece (130 years; 1907-2017).

- Problem statement: Multivariate generation of monthly time series of 2000 years length, preserving the Hurst behavior of the observed data, through appropriate formulation of the theoretical autocovariance function.

- Key challenge: Substantially changing statistics (by means of trends and long-term fluctuations) of historical data at the 30-year average (i.e. climatic) scale, resulting in very high Hurst coefficients.

5. Case study 2: Generation of daily rainfall data across Achelous river basin

- Historical data: Daily rainfall at 17 meteorological stations across the river basin of Achelous (largest river in Greece, in terms of runout production) and the neighboring closed basins of Tripolis and Lysimachia lakes, also drained to Achelous.

- Problem statement: Multivariate generation of daily synthetic data of 2000 years length, to be used for flood simulations.

- Key challenge: Computational drawback in estimation of variance-covariance matrices at all temporal scales; the problem is handled through a nonlinear optimization approach that requires the determination of 17–17 unknown elements.

As shown in the charts, the preservation of the statistical characteristics of the observed rainfall data becomes more difficult by moving to higher-order statistics and finer time scales; in particular, the empirical rules that are applied at the daily scale for preserving the probability dry mainly affect the preservation of the observed correlation structures, resulting in systematically underestimated auto-correlations and generally underestimated cross-correlations.

Theoretical background of CastaliaR (and beyond)


Contact info: I.Koutsoyiannis@wur.nl, P.Gioumou@central.ntua.gr. Presentation available at: http://www.ltes.ntua.gr/1782/