Environmental Modelling & Software 86 (2016) 264–276

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Missing data simulation inside flow rate time-series using multiple-point statistics

Fabio Oriani [a, e, *], Andrea Borghi [b, c], Julien Straubhaar [a], Grégoire Mariethoz [d], Philippe Renard [a]

[a] *Centre for Hydrogeology and Geothermics, Université de Neuchâtel, Neuchâtel, Switzerland*
[b] *École Nationale Supérieure de Géologie, Vandoeuvre-lès-Nancy, France*
[c] *Swiss Federal Office of Topography, Wabern, Switzerland*
[d] *Institute of Earth Surface Dynamics, Université de Lausanne, Lausanne, Switzerland*
[e] *Department of Hydrology, Geological Survey of Denmark and Greenland, Copenhagen, Denmark*

ARTICLE INFO

Article history:
Received 8 December 2015
Received in revised form
19 August 2016
Accepted 6 October 2016
Available online 21 October 2016

Keywords:
Time-series
Flow rate
Missing data
Non-parametric
Resampling
ARMAX
Multiple-point statistics

ABSTRACT

The direct sampling (DS) multiple-point statistical technique is proposed as a non-parametric missing data simulator for hydrological flow rate time-series. The algorithm makes use of the patterns contained inside a training data set to reproduce the complexity of the missing data. The proposed setup is tested in the reconstruction of a flow rate time-series while considering several missing data scenarios, as well as a comparative test against a time-series model of type ARMAX. The results show that DS generates more realistic simulations than ARMAX, better recovering the statistical content of the missing data. The predictive power of both techniques is much increased when a correlated flow rate time-series is used, but DS can also use incomplete auxiliary time-series, with a comparable prediction power. This makes the technique a handy simulation tool for practitioners dealing with incomplete data sets.

Software availability

The following information is about the software implementation of the simulation technique used in this paper:

Algorithm name: Direct Sampling (Mariethoz et al. (2010)).
Implementation name: DeeSse (Straubhaar (2015)).
Program language: C.
Developer: University of Neuchâtel, Julien Straubhaar (julien.straubhaar@unine.ch).
Year first available: 2015.
Minimal requirements: Windows/UNIX OS.
Availability: free license on request for research purposes, available on purchase for commercial use – for any request please contact Philippe Renard (philippe.renard@unine.ch).

A tutorial of the application shown in the paper is available upon request.

1. Introduction

The reconstruction of missing data portions inside time-series is a critical topic in applied hydrology since a large number of the numerical simulation techniques, used to model the hydrological processes, need continuous data records as input. Sometimes, technical failures of measurement instruments produce missing or unreliable data for long time periods for which the uncertainty about the observed phenomena is high. For this reason, a technique capable of generating realistic simulations of the missing data, reflecting the complex structures of the signal, and possibly making use of auxiliary information, is needed.

Many different approaches have been proposed for time-series gap filling in earth sciences: techniques based on mean diurnal variation or regression (Falge et al., 2001; Moffat et al., 2007), autoregression (Bennis et al., 1997; Wang, 2008), singular spectrum analysis (Schoellhamer, 2001; Kondrashov et al., 2014), self-organizing maps (Wang, 2003; Lamrini et al., 2011), look-up tables (Bamberger et al., 2014), rough sets (Dumedah et al., 2014), and artificial neural networks, widely used in recent years (Aminian

[*] Corresponding author. Department of Hydrology, Geological Survey of Denmark and Greenland, Øster Voldgade 10, DK-1350 Copenhagen K, Denmark.
E-mail address: fabio.oriani@protonmail.com (F. Oriani).

http://dx.doi.org/10.1016/j.envsoft.2016.10.002
1364-8152/© 2016 Elsevier Ltd. All rights reserved.

and Ameri, 2005; Dastorani et al., 2009; Diamantopoulou, 2010; Nkuna and Odiyo, 2011; Bahrami et al., 2011; Nourani et al., 2012; Dumedah et al., 2014). In this paper, we propose a non-parametric method to simulate missing data inside flow rate time-series based on the Direct Sampling (DS) technique (Mariethoz et al., 2010) belonging to multiple-point statistics (MPS). Already tested on gap filling in multivariate data sets representing natural heterogeneities (Mariethoz et al., 2012, 2015) and on rainfall time-series simulation (Oriani et al., 2014), DS can simulate the outcome of a complex natural process by reproducing similar patterns to the ones found in the available data without imposing a specific statistical model. More particularly, missing data are simulated by sampling the available data set where a sufficiently similar pattern is found. High-order statistical relations in the variable of interest are preserved by respecting the similarities in the neighborhood at multiple scales. The approach is almost entirely data-driven and fairly simple, but its efficiency largely depends on finding the good ensemble of auxiliary variables suitable to the current application. We present a multivariate standard setup for missing data simulation inside hydrological flow rate time-series using a correlated time-series as auxiliary variable. The setup is tested on the gap filling of a high-resolution karst flow rate time-series using different auxiliary variables. To make the test systematic and relevant for application, a gap size varying from a few hours to 20 days and total missing data percentage up to 30% are considered. Finally, a last group of tests focuses on the comparison of the proposed technique with a classical time-series model of type ARMAX. The general methodology, the setup, as well as the data set used, are illustrated in Section 2, the results are presented in Sections 3 And 4, while Section 5 is dedicated to the conclusions.

2. Methodology

2.1. The data set

The data set used to test the proposed technique is the 1990–2013 flow rate record from two karst springs of the Jura mountains (Switzerland) provided by the Swiss Federal Office for the Environment (FOEN). This paleozoic karst system is characterized by flashy spring discharges (Painter et al., 2008). Three high-resolution (10-min) time-series are used: the Areuse creek measured at St. Sulpice station (Ar) is used as a target variable, while the same water flow measured at Boudry station (Ar2) and the Seyon creek measured at Valangin station (Se) are used as auxiliary variables. The two river basins are contiguous (Fig. 1) and their regimes have been both classified as Jurassian pluvial and nivo-pluvial (FOEN). Ar station (443 m a.s.l.) lies at a distance of about 20 km from Ar2 (750 m a.s.l.) and 30 km from Se (628 m a.s.l.). Measuring from the same river, Ar and Ar2 are highly correlated (Pearson's correlation coefficient PCC = 0.96), whereas Ar and Se show a medium to weak correlation (PCC = 0.72). The considered time-series do not contain any missing data, but Ar and Ar2 show isolated sharp fluctuations around the local trend due to instrumental errors. To remove this kind of artifact, the following preprocessing treatment is applied (Oriani, 2015): given a time-series Z(t) and computing the differential operator $\delta Z(t) = Z(t) - Z(t-1)$, the artifacts are identified with the portions of Z(t) presenting $\sigma(t,a) > b$, where $\sigma(t,a)$ is the local standard deviation of $\delta Z(t)$, computed on the time interval $[t \pm a]$ and b is a user-defined threshold. The appropriate value for a and b depending on the smoothness of the signal and the magnitude of the artifacts, can be manually set by visually checking the results. In this paper, the chosen values are $a = 19$, $b = 0.3$ for $Z(t) = $ Ar and $b = 0.05$ for $Z(t) = $ Ar2. The data detected as artifacts are replaced by a cubic spline interpolation.

2.2. The Direct Sampling technique

Multiple-point statistics (MPS) techniques are based on the concept of training data set (TI): a representative sample of the target variable or conceptual model which is used to estimate the probability of occurrence of each event inside the simulation. MPS methods (Guardiano and Srivastava, 1993; Strebelle, 2002; Allard et al., 2006) generally consider a catalog of neighboring data patterns found in the TI to impose high-order conditioning in the simulation and thus reproduce similar structures to the ones found in the TI. This requires the estimation of the conditional probability

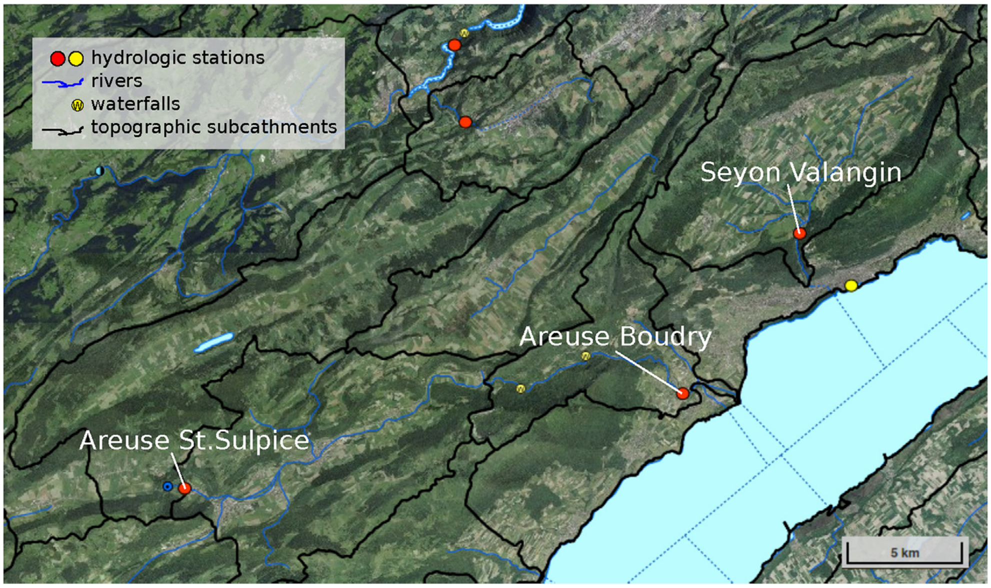

Fig. 1. Aerial photo of the study region (Jura mountains and Neuchtel lake), with location of the measure stations and topographic basin subdivision (modified from Swiss Federal Office of Topography, map. geo.admin.ch).