Environmental Modelling & Software 82 (2016) 308–320

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

A user-driven case-based reasoning tool for infilling missing values in daily mean river flow records

Laura Giustarini*, Olivier Parisot, Mohammad Ghoniem, Renaud Hostache, Ivonne Trebs, Benoît Otjacques

Luxembourg Institute of Science and Technology (LIST), Environmental Research & Innovation Department (ERIN), 5, avenue des Hauts-Fourneaux, L-4362 Esch/Alzette, Luxembourg

ARTICLE INFO

Article history:
Received 19 November 2015
Received in revised form
23 March 2016
Accepted 12 April 2016

Keywords:
Gap filling
Hydrological time series
Case-based reasoning
Knowledge database
Human in the loop

ABSTRACT

Missing data in river flow records represent a loss of information and a serious drawback in water management. In this work, we introduce gapIt, a user-driven case-based reasoning tool for infilling gaps in daily mean river flow records. Given a set of flow time series, gapIt builds a database of artificial gaps for which it computes several flow estimates, to find the best combinations of infilling algorithm and automatically selected donor station(s), according to state-of-the-art performance indicators. We obtained satisfactory results with Nash-Sutcliffe >0.7 for more than half of the ~5000 synthetic gaps of various lengths and positions, randomly created along the available records. gapIt was evaluated on 24 daily river discharge time series recorded in Luxembourg over seven years from 01/01/2007 to 31/12/2013. We also discuss the benefits of coupling this approach with user-expertise for an improved infilling of real data gaps.

Software availability

Name of software: gapIt
Developer: Olivier Parisot (olivier.parisot@list.lu)
Programming language: Java
Required hardware: 4 GB RAM minimum
Supported systems: Windows, Unix, Linux, Mac
Required softwares: Maven (≥3.0.2), JDK (≥1.7)
Availability: https://github.com/ERIN-LIST/gapIt
License: GNU General Public License version 3

1. Introduction

Long uninterrupted hydrological time series are often not available for many of the stream gauges in the world. Rather, time series of hydrological data are often affected by data gaps, which are discontinuities in the record of data. They are an inevitable consequence of factors such as station maintenance, equipment malfunctioning, human errors, changes in instrumentation and data processing issues (Harvey et al., 2010). Missing data in river flow records represent a loss of information and a serious drawback in water management. The existence of gaps results in difficulties in data interpretation and is a large source of uncertainty in data analysis. Specifically, the presence of discontinuities precludes the computation of hydrological statistics and physiographic indices. It also limits the use of such data for hydrological or hydrodynamic model calibration/validation purposes. A consequence of these issues is the need of data infilling methods to reconstruct missing data, when appropriate and before hydrological time series can be used in a number of applications.

From a technical point of view, a wide choice of data analysis tools is nowadays offered to hydrologists. For instance, specific user friendly software tools are already available or can be developed in platforms like R[1] or Matlab[2] to interpolate missing data and/or address hydrological problems. But most of these tools require some data mining and machine learning expertise, as well as fine-tuning in order to meet user needs and be properly exploitable by end-users (Serban et al., 2013). As a result, hydrologists have access to a collection of usable tools, but they still need to deal with several technical issues (like *data wrangling, tuning predictive algorithms*)

* Corresponding author.
E-mail address: laura.giustarini@gmail.com (L. Giustarini).

[1] http://www.r-project.org/.
[2] http://www.mathworks.com/products/matlab/.

http://dx.doi.org/10.1016/j.envsoft.2016.04.013
1364-8152/© 2016 Elsevier Ltd. All rights reserved.

before solving their initial problem, i.e. infilling missing values.

Data infilling is a challenging task that has been addressed by previous research work.

2. Related work

For infilling gaps in hydrological time series, classical methods of data analysis have long been applied (Salas, 1980) and recent studies have proposed more efficient techniques (Harvey et al., 2010; Mwale et al., 2012). Most of the methods proposed in the literature are based on data transfer from one or more donor stations (gauges) to a target station. Among all possible infilling methods, the choice of the most appropriate one is not a trivial task. The same holds true for the selection of a set of donor stations. Moreover, results greatly depend on the context and, in non automated techniques, also on the user expertise.

Recently, Harvey et al. (2010) tested different infilling techniques, simulating an entire target flow record, for several stations in the UK. For each target station, two donor stations were selected a priori, based on the hydrological knowledge of the region and catchment metadata. Their work focused on the performance analysis of gap infilling techniques. In a follow-up study, Harvey et al. (2012) assessed a wide range of target-donor combinations, trying at the same time to improve data infilling performance by either seasonally grouping flows or excluding known inhomogeneity.

Gyau-Boakye and Schultz (1994) presented a Decision Support System (DSS) for selecting the most appropriate infilling model, as a function of gap length, season, climatic region and data characteristics of the records. The main disadvantage of their approach is that all rules are *hard-coded* and specific to a given region, namely West Africa. The same idea was applied by Johnston (1999) to build a DSS that helps experts select an estimation method for missing rainfall data in the United States. More recently, Griffioen et al. (2006) proposed a Case-Based Reasoning system (CBR) to intercompare water stress among different catchments in Europe. In their work, CBR was presented as a retrieval method to offer large amounts of filtered information to the end-user. In a broader context, Matthies et al. (2007) provided a review on environmental DSS, showing a general tendency towards integration and visualization of temporal and spatial results.

Despite the numerous studies available in the literature, a standardized procedure for gap filling in hydrological time series is still missing. One of the main limitations of many of the currently available approaches is their incomplete level of automation. Generally, donor stations are often determined a priori and tend to be specific to only a given region of interest. The user expertise is fundamental for this type of settings but it also limits the level of automation and the transferability of the approach to different areas.

In this work, we present a first attempt towards standardization, providing an interactive tool that allows performing gap filling in a consistent and traceable manner, bridging the gap between data-driven and user-expertise approaches.

gapIt is an interactive and visual data-driven tool that offers several infilling techniques, coupled with different sets of donor stations. It assesses the performance of all possible configurations, i.e. combinations of infilling method and set of donor station(s), to fill a given gap in a consistent way, eventually providing the best data-driven solution according to performance indicators. The visual interface allows users to select different infilling methods and/or donor station(s) than those automatically proposed by the tool, according to their expertise and specific knowledge of the region of interest. The fact that users can interactively inject their knowledge allows an iterative refinement of the results, while keeping track of all modifications.

In the general practice, infilling techniques require both a strong methodological background and a significant knowledge of the application domain (Maimon and Rokach, 2005; Domingos, 2012). In this paper, we show how gapIt can provide a bridge between a purely data-driven approach and an infilling method based on user expertise only. The automated approach, coupled with a visual inspection system for user-defined refinement, allows for standardized infilling, where subjective expert decisions can easily be incorporated in a traceable manner.

In the remainder of this paper, we will present a case study and the related data sources (section 3). Then we describe the proposed gapIt algorithm in section 4, which is followed by an analysis of the results obtained for both synthetic and real gaps in section 5. Advantages and limitations of the method are summarized at the end.

3. Case study

The dense river network of hydrometric stations in Luxembourg offers an excellent opportunity to test the proposed tool. The gauge network considered here is composed of 24 stations, displayed in Fig. 1, including both very responsive and groundwater-fed rivers. The region has a temperate, semi-oceanic climate. Precipitation is relatively uniform throughout the year, although strong seasonality in low flow exists due to higher evapotranspiration from July to September. High discharge values are recorded in winter (maximum January–February), sometimes leading to inundations, while low flows are observed particularly in September. The influence of snow can be considered negligible.[3]

We use discharge data, originally available as 15-min time series and subsequently aggregated using gapIt itself to daily values, covering the period from 01/01/2007 to 31/12/2013. A total number of 28 gaps are present in the dataset; most of them have been observed in winter.

4. Methods and tools

In this section, algorithm implementation and input data requirements for gapIt will be described. It has to be noted that this approach is based on a single variable, discharge, provided as input to the software. This loosens dependency on other types of variables, for instance catchment rainfall, which may not always be available (Harvey et al., 2010). In the following, we designate as target station (respectively, donor station(s)) the station characterized by a gap to be infilled (respectively, the station(s) whose data is used to derive infilled data for the target). The underlying hypothesis of the presented tool is the availability of a sufficiently dense river network that provides continuous measurements. gapIt infills gaps in discharge time series, providing the final user with the best solution that is possible to obtain, given the available donor stations. The best solution is individuated based on performance measures. As we are dealing with both synthetic and real gaps, two different strategies will be proposed to compute performance measures, depending on the type of gap. The insertion of estimated values in the database in lieu of gaps is subject to the acceptance by the end-user.

All infilled data are consistently flagged, for the sake of traceability of the reconstructed values. Moreover, the configuration used for infilling each gap is stored in the database, for the sake of reproducibility.

[3] http://www.hydroclimato.lu.