Environmental Modelling & Software 72 (2015) 71–76

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Short communication

Feature-preserving interpolation and filtering of environmental time series

Gregoire Mariethoz [a, b, *], Niklas Linde [a], Damien Jougnot [a, c], Hassan Rezaee [d]

[a] Faculty of Geosciences and Environment, University of Lausanne, Switzerland
[b] School of Civil and Environmental Engineering, University of New South Wales, Sydney, Australia
[c] Sorbonne Universités, UPMC Univ Paris 06, CNRS, EPHE, UMR 7619 METIS, Paris, France
[d] Department of Civil, Geological and Mining Engineering, École Polytechnique de Montréal, Canada

ARTICLE INFO

Article history:
Received 29 January 2015
Received in revised form
3 June 2015
Accepted 2 July 2015
Available online xxx

Keywords:
Gap-filling
Geophysics
Geostatistics
Interferences
Multiple-point
Uncertainty

ABSTRACT

We propose a method for filling gaps and removing interferences in time series for applications involving continuous monitoring of environmental variables. The approach is non-parametric and based on an iterative pattern-matching between the affected and the valid parts of the time series. It considers several variables jointly in the pattern matching process and allows preserving linear or non-linear dependences between variables. The uncertainty in the reconstructed time series is quantified through multiple realizations. The method is tested on self-potential data that are affected by strong interferences as well as data gaps, and the results show that our approach allows reproducing the spectral features of the original signal. Even in the presence of intense signal perturbations, it significantly improves the signal and corrects bias introduced by asymmetrical interferences. Potential applications are wide-ranging, including geophysics, meteorology and hydrology.

1. Introduction

Time series are used universally in earth sciences as they constitute the most common output of measuring devices in disciplines such as hydrology, geophysics or meteorology. Almost as ubiquitous as the use of time series is the occurrence of periods with signal interference and measurement gaps. Such failures often constitute major impediments to quantitative analysis of environmental monitoring data. Examples include, among other fields, in-situ measurements of hydrological processes using geophysical methods (Smith-Boughner and Constable, 2012) or rainfall measurements (Kim and Ahn, 2009). This paper proposes a filtering method that addresses such contamination of time series data while preserving the underlying signal properties.

While well-established techniques exist to filter interferences and measurement errors, more challenging situations are intermittent interferences (e.g., measurement perturbations, earthquakes, magnetic storms) or data gaps (e.g., temporary measurement failure). Advanced time-series analysis, for instance time-frequency analysis with wavelets, often require continuous data for the entire measurement period, which implies that data gaps need to be filled by interpolation. In other applications, it is common to simply discard data from time-periods with significant interference. This excludes the possibility of using information in the underlying signal that could possibly be exploited. The focus of the present contribution is on signal interference and data gaps. To be effective, the ideal filtering approach should (a) maintain the same type of features as in measurement periods that are unaffected by data gaps or interferences; (b) recover information about the underlying signal that is hidden in the periods affected by interference; (c) provide uncertainty bounds relative to the filtered outputs and (d) preserve the coherence between several concurrent data sources. The last point can be illustrated by the example of two correlated time series, one of them showing low frequency variations and the other one high frequency fluctuations. If both time series are affected by a data gap, the reconstructed values need to preserve their joint statistical relationship, while maintaining the inherent variability specific to each time series.

The approaches commonly used for gap-filling and denoising include linear or spline interpolation and inverse weighted

* Corresponding author. Faculty of Geosciences and Environment, University of Lausanne, Switzerland.
E-mail address: gregoire.mariethoz@minds.ch (G. Mariethoz).

http://dx.doi.org/10.1016/j.envsoft.2015.07.001
1364-8152/

distances (Teegavarapu and Chandramouli, 2005). Spectral decomposition methods have been applied to time series with gaps (Schoellhamer, 2001), including multivariate data (Kondrashov and Ghil, 2006) and space-time domains (Wang et al., 2012). Parametric approaches such as autocorrelation models (Broersen, 2006) and kriging, which uses a variogram as statistical model of variability (Ruiz-Alzola et al., 2005), provide uncertainty estimates, but the resulting estimation is smooth and may not preserve the expected variability of the unsampled signal. Žukovič and Hristopulos (2013) introduced a directional gradient-curvature method to fill gaps in spatial remote sensing data. Recently, copula-based methods have been shown to outperform kriging for gap-filling problems (Bárdossy and Pegram, 2014). In another vein, Paparella (2005) formulates the gap-filling problem as an optimization that starts with a stitching of pieces from the observed signal.

In recent years, a family of non-parametric methods has emerged in geostatistics that are based on the recognition that parametric models may be poorly adapted to represent complex phenomena. Among these, multiple-point geostatistics use a statistical model of variability that consists of an example of the phenomenon studied, known as training data, from which high-order statistics or patterns are borrowed. The use of training data for filtering complex time series has not yet been investigated. We propose herein an approach based on training data consisting of the interference-free parts of the signal. Our approach performs well in the typical case where a large portion of the time series is unaffected by data gaps and interferences. In addition to populating the problematic periods with realistic data values, our approach is stochastic and allows for uncertainty characterization of the filtered results through multiple realizations.

2. Methodology

Our approach consists in iteratively replacing missing or interference-affected values by patterns sampled from the training data (i.e., recorded time-series that are unaffected by interferences or gaps). The sampling of the training data is achieved through the Direct Sampling algorithm (DS) that can perform multivariate simulation of continuous and/or categorical variables (Mariethoz et al., 2010). This algorithm was modified to perform iteratively, such that every iteration removes some of the interferences until convergence. The procedure adopted is illustrated in Fig. 1 and is described below. It is also available in a pseudocode form in Supplementary Table 1.

The input to the filtering algorithm is a noisy, possibly multivariate geophysical time series $\mathbf{S}_v(t)$, with t denoting the time stamp and $v = 1 \ldots V$ the variable considered. Firstly, all problematic time stamps affected by either interferences or data gaps are identified (denoted t_{SIM}), while the remaining time stamps are used as training data (denoted t_{TD}). The way of identifying noisy periods is case-specific (see application below). Each filtering iteration updates the values of the problematic time stamps and leaves unchanged the data that are unaffected by interference or gaps. These problematic time stamps can correspond to either gaps or interferences, however in practice all gaps are filled after the first iteration. All problematic time stamps, for all variables, are visited according to a random order that is initialized at each iteration. For each time stamp t_{SIM} to be simulated, the data pattern $\mathbf{N}$ formed by the n closest informed neighbors of t_{SIM} is identified (i.e., not including neighbors corresponding to interferences or gaps). The training data set is then searched in a random fashion for data patterns $\mathbf{N}_{TD}$ that are similar to $\mathbf{N}$. As soon as one pattern $\mathbf{N}_{TD}$ is found that is sufficiently similar to $\mathbf{N}$, its central value $S_v(t_{TD})$ is used to update the value currently considered. The random order in which time stamps are simulated and the random scanning of the training data ensure stochasticity in the process, hence allowing to generate multiple realizations.

In interference-affected periods, $\mathbf{N}$ will initially present features that do not correspond to the patterns found in the training data set. With additional iterations, the patterns present in those interference-contaminated parts of the data become increasingly similar to the training data. Because the initial state contains the interference-affected data, some of the embedded signal characteristics are preserved. This is in contrast with the commonly used approach that consists in removing the interference-affected data and resimulating it without any possibility to recover useful signals properties within those gaps.

In the case of data gaps, the pattern matching procedure is used to fill the gap at the first iteration (expanding the neighborhood until n informed values are found for each variable). In subsequent iterations this value is updated in the same way as for interference-affected values. In the sampling procedure, a training data value is accepted as the first occurrence where the distance $d(\mathbf{N}, \mathbf{N}_{\mathrm{TD}})$, is smaller than a user-defined threshold value τ. This "first matching wins" approach yields according to Shannon (1948) a statistically correct sampling of the conditional probability distribution corresponding to $\mathrm{Prob}\{S_v(i)|\mathbf{N}\}$. The parameter n defines the size of the data patterns considered, and also the order of the signal statistics to be reproduced. Values between 5 and 20 neighbors are recommended.

The data patterns $\mathbf{N}$ and $\mathbf{N}_{TD}$ can span over multiple variables, in which case $\mathbf{N}$ is a concatenation of the neighbors for each variable considered. Hence the data pattern comparison extends over many variables that are filtered simultaneously. This allows preserving linear or non-linear dependences between the variables present in the training data set.

Central to the method is the notion of similarity between patterns, defined by a distance function $d(.)$ that compares patterns from the problematic data with the training data. Since geophysical phenomena can present significant non-stationary behavior, it is

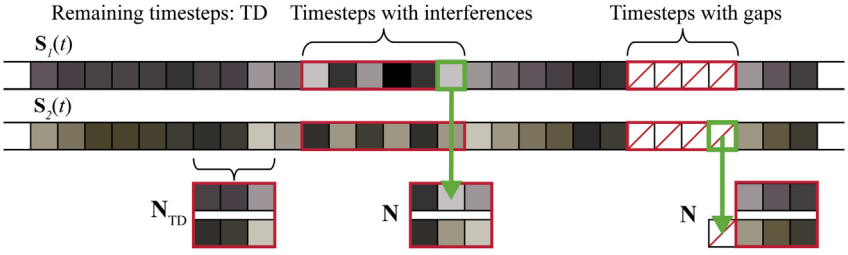

Fig. 1. Illustration of the filtering algorithm for a bivariate case with one area affected by interferences and another area affected by gaps. For each case, the time stamp currently simulated is highlighted in green. The neighborhood size is $n = 3$ for each variable, resulting in the 3 closest neighbors being chosen to simulate each time stamp (including the simulated time stamp itself in the case of interferences). The extracted neighborhood N is then compared to neighborhoods $\mathrm{N_{TD}}$ in the training data set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)