The Known Sub-Sequence Algorithm (KSSA) for making optimal decisions about imputation methods for multiple missing data in time series for environmental research

Iván Felipe Benavides PhD.
Post-doctoral Researcher
Instituto de Estudios del Pacífico, Universidad Nacional de Colombia, sede Tumaco

Track: 1, Environmental Science

Abstract: Missing data in time series is a frequent problem for the environmental sciences. This is a serious limitation for statistical analysis and therefore, imputation (the process of filling missing data) is a keystone task. Several imputation methods have been proposed and implemented in programming software, however, their efficiency is data-dependent. There is no universal imputation method best for all time series, but instead, each method suits the structure of particular groups of time series. ¿Which imputation method is best to fill a time series? the main problem is that the target time series (of interest for imputation) already contains missing data, so validation of methods cannot be performed directly on it. Instead it needs a full time series (no missing data) to simulate missing data, perform imputations and compare actual to imputed. However, the best imputation method for the full time series is not necessarily the best for the target time series. The Known Sub-Sequence Algorithm (KSSA) is a novel approach to solve this problem by validating imputation methods directly on target time series. It uses the information contained within sub-sequences between missing data gaps to produce an optimal decision about a best imputation method for any particular target time series, no matter the structure it has. This is done by means of a process of iterative bootstraping that randomly samples sub-sequences of the target time series in order to learn from them to find a best method form a set of candidates. This is a promising machine learning algorithm that will help environmental scientists and decision makers working with time series. KSSA will soon be implemented as the 'kssa' R-package in CRAN and is currently available in GitHub.

Key words: KSSA, missing data, environmental time series, validation