Environmental Modelling & Software 102 (2018) 172–184

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

A comparison of data imputation methods using Bayesian compressive sensing and Empirical Mode Decomposition for environmental temperature data

D. Alexandra Williams [a], Benjamin Nelsen [b], Candace Berrett [a], Gustavious P. Williams [b, *], Todd K. Moon [c]

[a] Statistics, 223 TMCB, Brigham Young University, Provo, UT 84602, USA
[b] Civil and Environmental Engineering, 368 CB, Provo, UT 84602, USA
[c] Electrical & Computer Engineering, Utah State University, Logan, UT 84322-4120, USA

ARTICLE INFO

Article history:
Received 5 July 2017
Received in revised form
21 November 2017
Accepted 13 January 2018

Keywords:
Environmental data imputation
Bayesian compressive sensing
Empirical mode decomposition

ABSTRACT

We present two Bayesian compressive sensing (BCS) imputation methods, BCS-on-Signal and BCS-on-IMF, and compare to temporal and spatio-temporal methods. We build sparse BCS models using available data, then use this sparse model for imputation. Most BCS applications have the sparse data distributed across the computational space, in our adaptation the "sparse" data are outside the reconstruction space. We used 30 years of temperature data and created gaps of 1% (~110 days), 5% (~1.5 years), 10% (~3 years), and 20% (~6 years). Performance was not sensitive to gap size with RMSE slightly above 6 °C for the BCS-on-Signal and Temporal models, the two best methods. The methods which only required data from the target station performed as well as, or better than, the spatio-temporal model which requires data from surrounding stations. Visually the BCS-on-IMF results seem to better represent longer-period random temporal fluctuations while having poorer performance metrics.

1. Introduction

Various data compression or reduction methods can accurately represent full data sets using very sparse data representations. To our knowledge, the ability to regenerate a signal from a sparse data set has not been exploited to impute missing data in the environmental sciences, although that is essentially what many of these sparsity-based compression/reduction methods are doing. The main difference between using a sparse representation to reconstruct the original signal and using this approach for data imputation is that data used for reconstruction are distributed throughout the signal while data used for imputation are complete (or dense) outside gaps and no data within the gaps. This paper explores whether these sparsity-based approaches can be used for data imputation given this difference.

We propose two novel imputation methods using sparsity-based compression sensing (CS) approach. We find that one method provides predictions as accurate as commonly-used imputation methods, while the other provides promise in accurately capturing trends and variation in the data. This work is of interest and beneficial because sparsity-based compressive/data reduction methods are fast, rely only on the signal of interest (do not need data from other locations or sources), and can be implemented relatively automatically, that is a user does not need to provide additional information about a signal. Most other commonly used methods for environmental data imputation, such as the temporal and spatio-temporal models compared in this paper, require more application- and data-specific information for modeling. For example, we know our temperature data have a strong annual trend and use that feature in both the temporal and spatio-temporal models. In contrast the CS methods did not require domain-specific knowledge to fit and model the data; we did not need to explicitly use the annual periodic nature of the data. This study provides evidence that these sparsity-based methods, both the CS methods presented and other compressive techniques, can be of value for environmental data imputation.

Environmental sciences use time series data, such as

* Corresponding author.
E-mail addresses: williams.d.alex17@gmail.com (D.A. Williams), Benjamin.w.nelsen@gmail.com (B. Nelsen), cberrett@stat.byu.edu (C. Berrett), gus.williams@byu.edu (G.P. Williams), todd.moon@usu.edu (T.K. Moon).

https://doi.org/10.1016/j.envsoft.2018.01.012
1364-8152/© 2018 Elsevier Ltd. All rights reserved.

streamflow, temperature, wind speed, and solar radiation, to describe the environment. These data are used to better understand, plan, manage, or control a wide variety of important hydrologic and environmental processes (Khalil et al., 2001). Unfortunately, while there are numerous environmental data archives, nearly all the available data series have gaps or periods of missing data (Gill et al., 2007; Di Piazza et al., 2011; Gyau-Boakye and Schultz, 1994; Sorjamaa et al., 2010; Mariethoz et al., 2015). Studies have shown that these missing data can affect data analysis and modeling. To effectively use these data series, the gaps must be filled using imputation methods (i.e., data estimation) (Gill et al., 2007; Sorjamaa et al., 2010; Gilroy, 1970; Henn et al., 2012; Raman et al., 1995; Oriani et al., 2016; Wang, 2008).

Data imputation in the earth sciences has a significant amount of reported research with methods ranging from simple replacement, to using spatially-correlated data, to various interpolation schemes, to complex statistical models. Simple examples include Battaglia and Protopapas (2012) and Auer et al. (Aueret al, 2007) who estimated missing temperature values using nearby stations with a simple offset; Craigmile and Guttorp (2011) who used only data from the target site and for a single missing value averaged observed values before and after and for longer gaps used an average of the values from the previous and following year; and Benth et al. (Benth et al., 2007) and Lemos et al. (Lemos et al., 2007) who reported similar approaches. More complex approaches include spatio-temporal models (Jeffrey et al., 2001), pattern matching (Mariethoz et al., 2015), and data modeling (Romanowicz et al., 2006). A large part of the published imputation work addresses streamflow or precipitation data and includes approaches such as spatial correlation with nearby sites (Gilroy, 1970; Beard, 1962; Fiering, 1962; Moran, 1974; Giustarini et al., 2016; Serrano-Notivoli et al., 2017), multivariate statistics (Kuczera, 1987; Vogel and Stedinger, 1985; Grygier et al., 1989); Bayesian modeling (Wang, 2008), and models such as neural networks (Khalil et al., 2001; Coulibaly and Evora, 2007; Kim and Ahn, 2009), chaos theory (Elshorbagy et al., 2002a), and a Markov-chain Monte Carlo algorithm within a Bayesian modeling framework (Lemos et al., 2007).

Researchers have reported comparisons between different methods, examples include Beauchamp et al. (Beauchamp et al., 1989) who compared a regression approach to a time-series approach and found that the time series approach performed better; Gyau-Boakye and Schultz (1994) who evaluated 10 methods and reported the best method varied depending on location and data; Raman et al. (Raman et al., 1995) who compared regression methods; and Hirsch (1982) who compared two regression methods and two maintenance of variance extension (MOVE) methods and found the MOVE approaches better. Other researchers have extended MOVE techniques to include other variables (Grygier et al., 1989). Researchers have reported that nearest neighbor methods were better than ARMA models for stream flows (Jayawardena and Lai, 1994), that ANN approaches were better than ARMA models (Hsu et al., 1995), and that ANNs performed better than linear regression (Elshorbagy et al., 2000) and nonlinear regression (Elshorbagy et al., 2002b). Besides stream flow, using ANN models for data imputation for other spatio-temporal data types has been reported (Diamantopoulou, 2010). This is not a comprehensive literature survey as the field is very large with a long tradition, but is intended to show the range of work in this area and also that reported "best" methods are dependent on data specifics, with no clear "best" method.

This paper presents two data imputation methods using Bayesian compressive sensing (BCS). In the first we apply BCS to the original signal to impute data to fill gaps in the data series. In the second, we first decompose the signal into a series of IMFs using EMD then apply BCS to each IMF in turn to impute data to fill the gaps, then recreate the signal by summing the filled IMFs. We hypothesized that as IMFs are less complex than the parent signal it would be easier to use BCS to model each IMF individually then use the IMFs to recreate the full signal. We present some background on BCS and EMD methods along with the formulation of our two models.

Using three different performance metrics and a number of different gap sizes and locations, we compare these BCS models to three other approaches: simple linear interpolation, a temporal model, and a spatio-temporal model. These methods were selected to represent a range of complexity and data requirements. For context we present information on the data series, the performance metrics, and the three imputation models used for comparison. We discuss the results and some features of the models and present our conclusions.

2. Background

2.1. Approach

CS exploits sparsity and is capable of representing a signal using a sampling frequency significantly less than the Nyquist frequency (Candès and Wakin, 2008; Candès, 2006; Shannon, 1949). BCS casts the resulting optimization problem in a Bayesian framework for computational advantages. CS and BCS are most often used for data reduction or compression by taking an original signal and discovering a sparse representation then using this sparse representation to recreate the original signal when required. We use this process for data imputation by building a sparse CS model with existing data then "recreate" missing data using the sparse model (Gemmeke et al., 2010; Gemmeke and Cranen, 2008). Commonly CS uses the sparse signal model to recreate the original signal between sparse measurement points that are distributed throughout the signal space. This allows a minimal number of measurement points to either be taken (reducing measurement time) or stored (reducing data size). We adapted this approach for data imputation by retaining all the original data, then recreating the data across the gap. As noted above, the main difference between using a sparse representation to reconstruct the original signal and using a sparse model for data imputation is that data used for reconstruction are distributed throughout the reconstruction space while data used for imputation are complete (or dense) outside gaps and no data within the reconstruction space or gaps.

Empirical Mode Decomposition (EMD) (Huang et al, 1998) is a signal deconvolution method that works with non-stationary, nonlinear, semi-periodic data; typical of environmental data series. EMD is data driven and does not assume or require a stationary or linear signal. EMD decomposes a signal into a series of intrinsic mode functions (IMFs) each representing independent components of the signal and a residual. Summing the IMFs and the residual exactly reproduces the original signal.

We evaluated these two BCS methods using historic temperature data from the Salt Lake City Airport, a long-term record, to characterize their behavior. We created gaps of various lengths in the signal, used the different methods to fill the gaps, and then compared the imputed data with the original signal. We compared the BCS methods with commonly used data imputation methods; a temporal spline, a spatial-temporal spline, and linear interpolation. The two BCS methods and the temporal spline only require data from the target station, the spatio-temporal spline requires data from surrounding stations.

While we used temperature data for this example, mostly because of the available long continuous data set for verification, this approach could be applied to any semi-periodic or cyclic earth