Journal of Hydrology 553 (2017) 384–397

Contents lists available at ScienceDirect

Journal of Hydrology

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

Research papers

Application of a novel hybrid method for spatiotemporal data imputation: A case study of the Minqin County groundwater level

Zhongrong Zhang [a,b,*], Xuan Yang [c], Hao Li [c], Weide Li [c], Haowen Yan [d], Fei Shi [e]

[a] School of Mathematics and Physics, Lanzhou Jiaotong University, China
[b] School of Environmental and Municipal Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China
[c] School of Mathematics and Statistics, Lanzhou University, Lanzhou 730000, China
[d] Faculty of Geomatics, Lanzhou Jiaotong University, Lanzhou 730070, China
[e] Gansu Minqin Supply of Water Supplies Limited Liability Company, Wuwei 733300, China

ARTICLE INFO

Article history:
Received 6 March 2017
Received in revised form 23 July 2017
Accepted 25 July 2017
Available online 19 August 2017
This manuscript was handled by Corrado Corradini, Editor-in-Chief, with the assistance of Felipe de Barros, Associate Editor

Keywords:
Spatiotemporal
Imputation
Missing data
Least squares support vector machine
Self-organizing feature map
Cross validation

ABSTRACT

The techniques for data analyses have been widely developed in past years, however, missing data still represent a ubiquitous problem in many scientific fields. In particular, dealing with missing spatiotemporal data presents an enormous challenge. Nonetheless, in recent years, a considerable amount of research has focused on spatiotemporal problems, making spatiotemporal missing data imputation methods increasingly indispensable. In this paper, a novel spatiotemporal hybrid method is proposed to verify and imputed spatiotemporal missing values. This new method, termed SOM-FLSSVM, flexibly combines three advanced techniques: self-organizing feature map (SOM) clustering, the fruit fly optimization algorithm (FOA) and the least squares support vector machine (LSSVM). We employ a cross-validation (CV) procedure and FOA swarm intelligence optimization strategy that can search available parameters and determine the optimal imputation model. The spatiotemporal underground water data for Minqin County, China, were selected to test the reliability and imputation ability of SOM-FLSSVM. We carried out a validation experiment and compared three well-studied models with SOM-FLSSVM using a different missing data ratio from 0.1 to 0.8 in the same data set. The results demonstrate that the new hybrid method performs well in terms of both robustness and accuracy for spatiotemporal missing data.

1. Introduction

Missing value imputation (using a reasonable value to estimate or replace the missing value) (Abellan et al., 2003; Feng et al., 2014; Kondrashov and Ghil, 2006; Jerez et al., 2010; Sim et al., 2015; Galan et al., 2016) has become a major research problem in data analysis, because the data preprocessing phase involves the crucial processes of searching and replacing missing values. Missing values are ubiquitous (Myneni et al., 2017), but most data analysis or mathematical models assume a complete data matrix (Feng et al., 2014). Since the presence of missing values may compromise the integrity of a data set and therefore limit the use of the data for various applications and research (Kornelsen and Coulibaly, 2012), it is crucial to establish an effective and robust imputation model to deal with missing values. Many researchers have proposed various imputation methods for missing time series data (Malek et al., 2008; Yozgatligil et al., 2013), missing panel data (Young and Johnson, 2015) and other forms of missing data. However, when the missing data occur in a spatiotemporal data set (spatiotemporal air quality data sets (Junninen et al., 2004), spatiotemporal meteorological data (Yozgatligil et al., 2013) and spatiotemporal wind data sets (Poloczek et al., 2014), few studies, (for example, Abellan et al., 2003, Feng et al., 2014 and Poloczek et al., 2014) have attempted to find a way to impute this type of missing values. Fig. 1 presents an example to illustrate the type of spatiotemporal data set for which missing data occur.

Feng et al. (2014) stated that mathematical properties or logical relationships are applied as the two cores of the relationship serving as the basis for imputation models. Traditional imputation methods for missing values can be roughly split into simple imputation (SI) (Linacre, 1992; Chen et al., 2000), interpolation imputation (II) (Junninen et al., 2004), regression imputation (RI) (Cooper et al., 1991; Schneider, 2001) and nearest neighbor imputation

* Corresponding author at: School of Environmental and Municipal Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China.
E-mail addresses: glszzhangzhongr@126.com (Z. Zhang), yangxuan14@lzu.edu.cn (X. Yang), haoli14@lzu.edu.cn (H. Li), weideli@lzu.edu.cn (W. Li), haowen2010@gmail.com (H. Yan), 185182758@qq.com (F. Shi).

http://dx.doi.org/10.1016/j.jhydrol.2017.07.053
0022-1694/

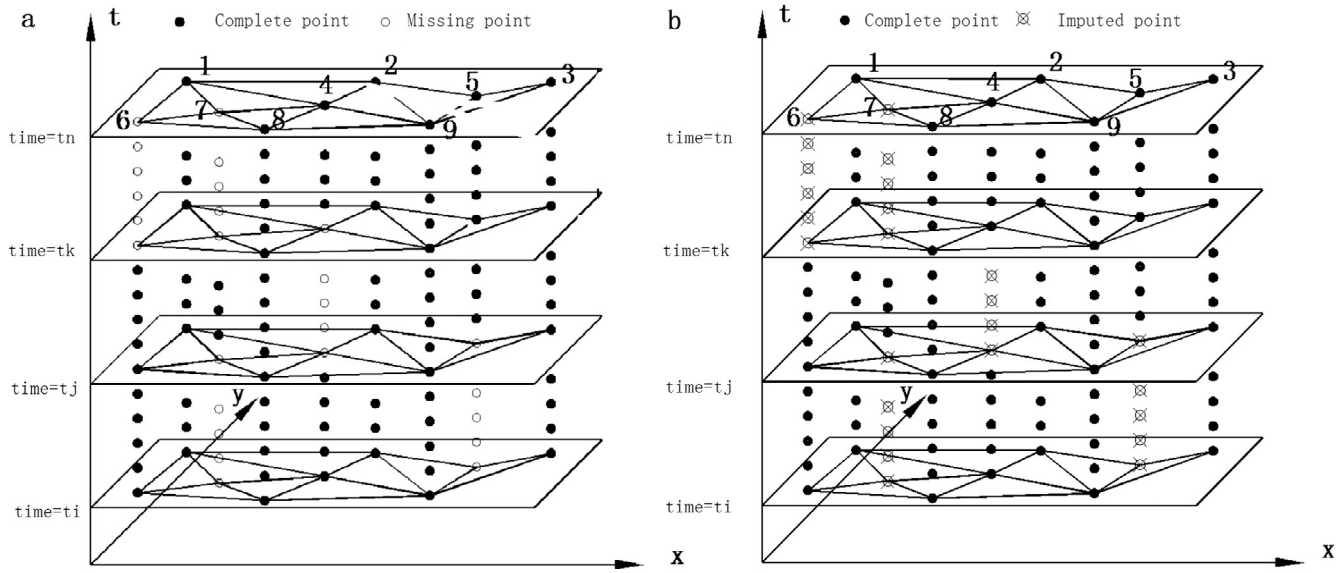

Fig. 1. Diagrammatic sketch of the missing data in a spatiotemporal date set and the result of the imputed missing point. a: nine space observation stations in four different time sections with the blank circle representing the missing point. b: all of missing points shown in panel a have been imputed in this image.

(NNI) (Chen and Shao, 2011). In the early research, SI used a direct deletion method which deletes the missing values from the original data sets. As a result, we obtain a complete data set but lost some implicit information that will affect subsequent data analysis results (Cismondi et al., 2013). Linacre (1992) computed the mean values to substitute the missing values (mean imputation) and Chen et al. (2000) presented a simple adjusted random imputation method that eliminates the imputation variance of the estimator of a mean or total. The II method includes many interpolation techniques, such as linear, splines and cubic interpolation. These methods replace missing values with the interpolation values. Junninen et al. (2004) describes linear interpolation and cubic spline and illustrated how these methods work. Zainudin et al. (2015) conducted an extendable examination into investigating the potential used of cubic Bezier technique and cubic Said-Ball method as estimator tools. The RI method is commonly used (Schneiderman et al., 1993; Raghunathan et al., 2001; Xian et al., 2006), and Cooper et al. (1991) developed two loss functions for RI method to estimate missing values. Schneider tested an iterated analysis of linear regression with the expectation maximization (EM) algorithm for solving missing data Schneider (2001). The NNI method is one of the most popular nonparametric hot deck imputation method used to compensate for nonresponses in sample surveys (Chen and Shao, 2011). All in all, these simplified methods are not a satisfactory way to solve the missing data problems which occur in spatiotemporal data sets, particularly in terms of their accuracy, since these simple methods not take influence between sites into account.

Other methods have been proposed. For example, Smith et al. (1996) employed a least squares method to fit of empirical orthogonal functions (EOFs) in the integrated data and this can be considered another form of optimal interpolation. Moreover, other new methods have been proposed in EOF space, such as Kalman filtering (Kalman, 1960) and optimal smoothing (Kaplan et al., 1997). All these methods need to use a priori information which contains the spatiotemporal covariance or others in the data set. By 2007, singular spectrum analysis (SSA) and multi-channel SSA (MSSA) were used to fill the gaps in several types of data sets (Kondrashov and Ghil, 2006); however, the accuracy and reliability of this technique will be influenced by the type of missing data.

Some machine learning methods for spatiotemporal imputation have been developed; Jerez et al. (2010) state that machine learning techniques can do well in the imputation of missing values. One major difference between these and statistical procedures is that they can enhance the accuracy of the imputation. In recent years, more machine learning techniques have been developed, such as multi-layer perception (MLP), SOM and the k-nearest neighbor (KNN). Sharpe and Solly (1995) concluded that MLP can be a useful method for dealing with missing values. Rustum and Adeloye (2007) used SOM to fill missing values in the time series, and confirmed that this method was appropriate for activated sludge data in Edinburgh, UK. Decision trees (Quinlan, 1986) and random forests (Breiman, 2001) were proposed and missing values were imputed using the similarity and correlations imputed missing values by Rahman and Islam (2013). Feng et al. (2014) compared KNN, SVD, multiple imputation and random forest with CUTOFF imputation performance in three spatiotemporal data sets, and the results showed that the CUTOFF method performs well. KNN regression was proposed as a geo-Imputation method and applied successfully to spatiotemporal wind data (Poloczek et al., 2014). Carvalho et al. (2016) proposed a spatiotemporal model and used it for daily rainfall imputation. The results show that the spatiotemporal model performed better than ordinary kriging.

To overcome the shortcomings of the single imputation method, researchers proposed more hybrid imputation methods after 2011. Narravula and Vadlamani (2011) illustrated a novel soft computing hybrid for data imputation, which involves a 2-stage soft computing approach, and the method was shown to be successful. Aydilek and Arslan (2013) combined fuzzy c-means, support vector regression and a genetic algorithm as a hybrid imputation model and the new imputation method performed well. Tian et al. (2014) carefully conducted a review of imputation methods and proposed a hybrid model called the multiple imputation, which blended gray-system-theory and entropy based on clustering MIGEC. Initially, MIGEC extracts the non-missing values and split them into several classes, then, analyses and imputes missing data. An imputation method called AR-ANN was proposed (Shukur and Lee, 2015) for daily wind speed data. The proposed AR-ANN method was compared with three basic models (linear, nearest neighbor, and state space methods) and was found to sur-