Transportation Research Part C 88 (2018) 124–139

Contents lists available at ScienceDirect

Transportation Research Part C

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

Missing data imputation for traffic flow speed using spatio-temporal cokriging

Bumjoon Bae[a,*], Hyun Kim[b], Hyeonsup Lim[c], Yuandong Liu[a], Lee D. Han[a], Phillip B. Freeze[d]

[a] Department of Civil and Environmental Engineering, University of Tennessee, Knoxville, 851 Neyland Drive, Knoxville, TN 37996, USA
[b] Department of Geography, University of Tennessee, Knoxville, 1000 Phillip Fulmer Way, Knoxville, TN 37916, USA
[c] Center for Transportation Analysis, Oak Ridge National Laboratory, 2360 Cherahala Blvd, Knoxville, TN 37932, USA
[d] Tennessee Department of Transportation, James K. Polk Building, 505 Deaderick Street, Suite 300, Nashville, TN 37243, USA

ARTICLE INFO

Keywords:
Missing data
Imputation
Spatio-temporal kriging
Cokriging
Missing patterns

ABSTRACT

Modern transportation systems rely increasingly on the availability and accuracy of traffic detector data to monitor traffic operational conditions and assess system performance. Missing data, which occurs almost inevitably for a number of reasons, can lead to suboptimal operations and ineffective decisions if not remedied in a timely and systematic fashion through data imputation. A review of literature suggests that most traffic data imputation studies considered the temporal continuity of the data but often overlooked the spatial correlations that exist. Few of the studies explored the randomness of the patterns of the missing data. Therefore, this paper proposes two cokriging methods that exploit the existence of spatio-temporal dependency in traffic data and employ multiple data sources, each with independently missing data, to impute high-resolution traffic speed data under different data missing pattern scenarios. The two proposed cokriging methods, both using multiple independent data sources, were benchmarked against classic simple and ordinary kriging methods, which use only the primary data source. An array of testing scenarios were designed to test these methods under different missing rates (10–40% data loss) and different missing patterns (random in time and location, random only in location, and non-random blocks of missing data). The results suggest that using multiple data sources with the spatio-temporal simple cokriging method effectively improves the imputation accuracy if the missing data were clustered, or in blocks. On the other hand, if the missing data were randomly scattered in time and location, the classic ordinary or simple kriging method using only the primary data source can be more effective. Our study, which employs empirical traffic speed data from radar detectors and vehicle probes, demonstrates that the overall predictions of the kriging-based imputation approach are accurate and reliable for all combinations of missing patterns and missing rates investigated.

1. Introduction

Traffic detector data collected from transportation facilities are essential inputs for modern transportation systems to monitor traffic conditions and assess system performance. A challenge for using the data is 'missingness' in the data collection processes of the systems (Buuren, 2012; Carpenter and Kenward, 2013). This includes (but is not limited to) the malfunctions of hardware or software,

* Corresponding author.
E-mail address: bbae1@utk.edu (B. Bae).

https://doi.org/10.1016/j.trc.2018.01.015
Received 30 May 2017; Received in revised form 15 January 2018; Accepted 15 January 2018

communication network problems, restricted power supply conditions, scheduled maintenance, and so on. As Orchard and Woodbury (1972) remarked, it is obvious that not to have missing data is the best way to address the missing data issue; however, this ideal circumstance rarely happens.

The effects of missing data and imputation methods have been examined in other disciplines, such as statistics, sociology, and epidemiology, because analysis results are considered rough when data are missing (Buuren, 2012). Unfortunately, this issue has not been well addressed in transportation studies (Smith et al., 2003; Ni and Leonard Ii, 2005). Measuring the effects of missing data and treatments to impute them are rarely investigated, even though the issue of handling missing data has been addressed to some degree in transportation modeling. Meanwhile, the need to measure the performance of transportation systems such as delay, travel time reliability, and emissions has been underlined in transportation systems management and operations. In this context, the appropriate methods to impute missing data should be explored, otherwise the results of such performance measures will be biased.

The effects of missing traffic flow data on transportation modeling and prediction can be divided into two categories (Bennett et al., 1984): First, it causes information loss for certain locations and time periods, which may be important to the objective of an analysis in transportation modeling and prediction. For instance, if traffic speed and traffic volume data are missing for a severely congested road segment during peak hours, the total vehicle emission will be underestimated. Second, it causes statistical information loss. In general, a sample size that is smaller due to missing data, i.e., smaller degrees of freedom, may lead to overfitting problems in the modeling process. More importantly, underlying assumptions of statistical methods used in an imputation analysis are violated by different missing patterns, resulting in biased solutions.

Therefore, to avoid erroneous statistical inference, understanding missing patterns and missing mechanisms from the datasets used in a statistical analysis is as important as determining how to sample from population. Rubin (1976) points out that distributional inferences on the parameters of data are generally conditional on the observed missing patterns. According to recent works by Buuren (2012) and Carpenter and Kenward (2013), a typology of missing patterns associated with the impact on statistical analysis are identified with three types: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Few of previous studies explored the randomness of the patterns of the missing traffic flow data, but not fully investigated these missing patterns (Qu et al., 2009; Li et al., 2013; Tang et al., 2015; Duan et al., 2016).

According to Li et al. (2014), missing data imputation methods can be classified into three types, including prediction, interpolation, and statistical learning methods. Imputation of traffic data using the interpolation method aims to estimate unobserved value at a specific location and time to improve the accuracy of further analyses (traffic speed prediction, traffic incident detection, and so on). Unlike the prediction methods, interpolation based imputation utilizes observations within the surrounding areas for an unobserved value in a spatio-temporal domain. Several recent transportation studies paid attention to a spatial interpolation approach, called kriging, to estimate or predict traffic variables for unobserved locations (Eom et al., 2006; Wang and Kockelman, 2009; Zou et al., 2012; Selby and Kockelman, 2013; Shamo et al., 2015). Considering that traffic data have spatio-temporal dependency, kriging has an advantage over other statistical approaches for improving imputation accuracy. This is because the method takes the observed neighboring data correlated with a missing value into account in space-time dimension. A recent kriging study extends the modeling dimension from a single spatial dimension to a spatio-temporal dimension to impute traffic speed data, arguably suggesting that spatio-temporal kriging outperforms the historical average and k-nearest neighborhood (KNN) methods (Yang et al., 2016).

1.1. Aim of the study

This paper aims to extend the spatio-temporal kriging approach for high-resolution traffic detector data imputation in the literature to a multivariate framework considering the following three factors: (a) kriging types with use of a secondary data source; (b) missing patterns; and (c) missing rates.

Cokriging is inherently the multivariate extension of kriging (Marcotte, 1991). It allows to use secondary data sources to complement observed primary data. In this study, ordinary and simple cokriging methods are employed to use a secondary traffic dataset for imputing the missing detector data. The imputation results of both cokriging methods are compared to those of two benchmark methods, which are ordinary and simple kriging. Because available traffic data resources are abundant, using the information from multiple data sources is anticipated to improve the imputation results of the spatio-temporal cokriging approach.

The effectiveness of cokriging relies on the pattern of missing data. To address this issue, we investigated the prediction performance of it with four different kriging methods based on three missing patterns (MCAR, MAR, and MNAR) in the traffic speed data. In addition, the imputation accuracy is further investigated over varying missing rates.

The next section presents a comprehensive literature review on imputation techniques and kriging in transportation studies, and describes the data used in this study. The following section explains the kriging and cokriging methods. The last two sections provide a case study result of applying the spatio-temporal cokriging approach to impute traffic speed data, then conclusions follow.

2. Literature review

Missing data imputation methods can be either single imputation or multiple imputation (Buuren, 2012; Carpenter and Kenward, 2013). Hot-deck, average, and regression are commonly used as single imputation methods. Most of the imputation studies in transportation examine single imputation methods because of their fast-computational speed for real-time analysis. The historical average, expectation maximization (EM) algorithm (Smith et al., 2003), pairwise regression (Al-Deek and Chandra, 2004), moving average, ARIMA, and regression model with genetic algorithm (Zhong et al., 2004) have been explored for imputing five or 10 min loop detector data.