Transportation Research Part C 51 (2015) 29–40

Contents lists available at ScienceDirect

Transportation Research Part C

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

A hybrid approach to integrate fuzzy C-means based imputation method with genetic algorithm for missing traffic volume data estimation

Jinjun Tang[a], Guohui Zhang[b], Yinhai Wang[a,c,*], Hua Wang[a], Fang Liu[d]

[a] School of Transportation Science and Engineering, Harbin Institute of Technology, Harbin 150001, China
[b] Department of Civil Engineering, University of New Mexico, Albuquerque, NM 87131, USA
[c] Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195-2700, USA
[d] School of Energy and Traffic Engineering, Inner Mongolia Agricultural University, Hohhot 010018, China

ARTICLE INFO

Article history:
Received 13 February 2014
Received in revised form 7 November 2014
Accepted 7 November 2014
Available online 28 November 2014

Keywords:
Missing sensor data
Fuzzy C-means
Genetic algorithm
Imputation
Traffic volume

ABSTRACT

Although various innovative traffic sensing technologies have been widely employed, incomplete sensor data is one of the most major problems to significantly degrade traffic data quality and integrity. In this study, a hybrid approach integrating the Fuzzy C-Means (FCM)-based imputation method with the Genetic Algorithm (GA) is develop for missing traffic volume data estimation based on inductance loop detector outputs. By utilizing the weekly similarity among data, the conventional vector-based data structure is firstly transformed into the matrix-based data pattern. Then, the GA is applied to optimize the membership functions and centroids in the FCM model. The experimental tests are conducted to verify the effectiveness of the proposed approach. The traffic volume data collected at different temporal scales were used as the testing dataset, and three different indicators, including root mean square error, correlation coefficient, and relative accuracy, are utilized to quantify the imputation performance compared with some conventional methods (Historical method, Double Exponential Smoothing, and Autoregressive Integrated Moving Average model). The results show the proposed approach outperforms the conventional methods under prevailing traffic conditions.

1. Introduction

Various traffic sensors have been widely employed to generate a huge amount of traffic data to facilitate traffic control and operation, network-wide infrastructure management, and individual trip planning. Based on real-time network-wide traffic congestion conditions, Advanced Traveler Information System (ATIS) is able to provide travelers routing information to optimize their routing choices. Similarly, using real-time traffic sensor data, Advanced Traffic Signal Control System (ATS-CS) can optimize signal timing settings to improve arterial traffic operations for travel time savings and congestion mitigation. Currently, loop detector is one of the most commonly used traffic sensors in arterial and freeway networks compared with the other types of sensors, such as camera, infrared radar, microwave, and GPS navigation devices. Based on single loop detector outputs, traffic volume and lane occupancy can be obtained during certain time intervals. Length-based classification traffic volume and speed can be estimated. These data are essential for advanced network-wide traffic management and

* Corresponding author at: Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195-2700, USA.
E-mail address: yinhai@uw.edu (Y. Wang).

http://dx.doi.org/10.1016/j.trc.2014.11.003
0968-090X/© 2014 Elsevier Ltd. All rights reserved.

control. However, missing data are very common in reality because sensors and communication systems malfunction and substantially degrade information quality and application. Detection and communication failures can be attributed to many natural and man-made factors, such as malfunctioning devices, incomplete observations, and data transfer problems (Boyles, 2011; Ni et al., 2005; Patel, 1995; Qu et al., 2009). Consequently, how to estimate missing data based on loop detector outputs becomes a critical issue to facilitate advance traffic control and management strategy development and implementation.

Many research efforts have been undertaken to address this issue. The historical imputation algorithm was developed and used based on the average value of historical data in the same time interval to interpolate missing data (Chen and Shao, 2000; Nguyen and Scherer, 2003). Additionally, the factor approach calculates various impact factors from historical data and estimates the missing data using the average by considering these factors. Gold et al. (2000) proposed several methods for imputing abnormal traffic volume data collected in 5-min intervals. They applied a “factor-up and straight-line” interpolation strategy to estimate missing traffic volumes. Zhong et al. (2004) applied different factor methods, including Hourly Factors (HF), Daily Factors (DF), and Monthly Factors (MF), to predict missing data. Due to their simple structures, these approaches estimate missing data in a computationally efficient manner. However, these approaches are majorly based on the assumption that identical traffic patterns exist among days. Affected by many stochastic factors, traffic flow fluctuations and randomness from day to day have been ignored to some extent. Thus, it is difficult for these methods to achieve satisfactory performance under various traffic conditions by using historical or common factor approaches.

Simple interpolation strategies can be used to estimate missing data through interpolating missing ones based on observed data. Exponential smoothing and splines are major techniques in data interpolation (Baharaeen and Masud, 1986; Holt, 2004; de Boor, 1978). Regression imputation is another simple, effective method. These models estimate missing data based on a calibrated regression model. Chen et al. (2003) proposed a linear regression model based on the data from neighboring detectors and used the model to estimate missing volumes and occupancies. Al Deek and Chandra (2004) established various regression models to estimate missing data and they concluded that the quadratic regression model achieved the best imputation results. Boyles (2011) compared different approaches, such as simple linear regression model, multiple linear regression model, local and global regression model, and non-normal Bayesian linear regression model. Three types of missing data, including random, consecutive, and successive data, are applied to evaluate the accuracy of various algorithms. These regression models are easy to develop and use, but their performance is unreliable under different traffic conditions. Furthermore, many researchers used time series techniques to estimate missing data, such as Autoregressive Integrated Moving Average (ARIMA), or seasonal ARIMA (Ghosh et al., 2005; Redfern et al., 1993; Ramsey and Hayden, 1994; Williams and Hoel, 2003). Time series models were developed and trained by historical observations in the current time interval, and then applied to estimate missing data for the next one or multiple intervals. It assumes that the historical data provide an indication for unobserved ones in the temporal domain (Box and Jenkins, 1970). Currently, a novel algorithm using Kernel Probabilistic Principle Component Analysis (KPPCA) was introduced to impute missing data (Li et al., 2013). This approach extracted and utilized the temporal-spatial dependence in traffic volumes. Their study demonstrates the favorable performance compared with the Probabilistic Principle Component Analysis (PPCA)-based approach.

Based on these previous studies, we understand that an effective approach should consider not only historical information but also similarity and difference in data patterns from time interval to time interval. The existing imputation methods have been established based on the vector-based data structure, $V = \{v_1, v_2, \ldots, v_n\}$, in which missing data can be estimated based on historical data patterns. The vector data are comprised of the observations chronologically. However, vector data-based methods usually neglect similarities in data patterns from day to day. A large amount of field observations demonstrate that traffic volumes maintain similar variation patterns in weekdays among different weeks. Thus, a multi-vector or matrix-based data structure is necessary to complement the imputation process. The conventional methods tend to overlook the matrix-based dataset. Fuzzy C-Means (FCM) algorithm, has been widely used to address clustering problems with incomplete data (Di Nuovo, 2011; Hathaway and Bezdek, 2002; Liao et al., 2009; Li et al., 2010; Timm et al., 2004). Therefore, FCM algorithm is potentially applicable to analyze the data with multiple attributes. The hybrid approach developed in this study majorly consists of following four calculation steps: (1) analyze the weekly similarity of traffic volume data and then transform the vector-based data structure into the matrix-based data structure; (2) establish the FCM-based imputation model (3) optimize the membership degrees and cluster centroids based on GA; (4) estimate the missing data.

This paper is organized as follows: Section 2 introduces the research methodology and details missing traffic data characteristics, as well as working principles of FCM-based imputation algorithms. Section 3 describes the experimental tests and result comparisons with the conventional methods. Finally, the conclusions and recommendations are provided in Section 4.

2. Research methodology

2.1. Matrix-based missing data representation

Three types of missing data are commonly considered (Li et al., 2004; Qu et al., 2009): (1) Missing Completely at Random (MCR); (2) Missing Partially at Random (MPR); (3) Missing due to Systematical Errors (MSE). This study concentrates on MPR data estimation to verify the effectiveness of the proposed approach. Fig. 1 demonstrates randomly missing data characteristics collected during five weekdays. In order to fully extract weekly similarity patterns, the original vector-based data