Transportation Research Part C 90 (2018) 18–33

Contents lists available at ScienceDirect

Transportation Research Part C

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

On the imputation of missing data for road traffic forecasting: New insights and novel techniques

Ibai Laña[a,*], Ignacio (Iñaki) Olabarrieta[a], Manuel Vélez[b], Javier Del Ser[a,b,c]

[a] OPTIMA Unit, TECNALIA, P. Tecnologico Bizkaia, Ed. 700, 48160 Derio, Spain
[b] Dept. of Communications Engineering, University of the Basque Country UPV/EHU, Alameda Urquijo S/N, 48013 Bilbao, Spain
[c] Basque Center for Applied Mathematics (BCAM), 48009 Bilbao, Spain

ARTICLE INFO

Keywords:
Traffic forecasting
Missing data
Cluster analysis
Data imputation

ABSTRACT

Vehicle flow forecasting is of crucial importance for the management of road traffic in complex urban networks, as well as a useful input for route planning algorithms. In general traffic predictive models rely on data gathered by different types of sensors placed on roads, which occasionally produce faulty readings due to several causes, such as malfunctioning hardware or transmission errors. Filling in those gaps is relevant for constructing accurate forecasting models, a task which is engaged by diverse strategies, from a simple null value imputation to complex spatio-temporal context imputation models. This work elaborates on two machine learning approaches to update missing data with no gap length restrictions: a spatial context sensing model based on the information provided by surrounding sensors, and an automated clustering analysis tool that seeks optimal pattern clusters in order to impute values. Their performance is assessed and compared to other common techniques and different missing data generation models over real data captured from the city of Madrid (Spain). The newly presented methods are found to be fairly superior when portions of missing data are large or very abundant, as occurs in most practical cases.

1. Introduction

Road traffic forecasting methods have been under active research, development and implementation for more than 40 years, a history that has hitherto involved time-series analysis and prediction models with a wide diversity of algorithmic variants and processing enhancements. More recently, machine learning techniques have acquired momentum by virtue of the large amount of successful methodologies, algorithms and optimization procedures (Abdel-Aty et al., 1997; Vlahogianni et al., 2007; Hinsbergen et al., 2007; Vlahogianni et al., 2014), further propelled by the advent of Big Data technologies (Schimbinschi et al., 2015; Lv et al., 2015).

In this context, the most relevant traffic variables (i.e. flow, speed, travel time, occupancy) have been predicted using data captured by magnetic loops, cameras, plate readers and floating car data, among many other sources. Within them, inductive loops or Automatic Traffic Recorders (ATR) are one of the most frequently selected data sources for traffic forecasting (Vlahogianni et al., 2014). ATRs count each vehicle passing through a particular point in the network, but they often undergo situations in which the output data are faulty, to the extreme of existing long periods of time with no captured data due to prolonged reading, recording or transmission errors. In some cases, organizations that manage the sensors and provide data remove measurements that are considered

[*] Corresponding author.
E-mail address: ibai.lana@tecnalia.com (I. Laña).

https://doi.org/10.1016/j.trc.2018.02.021
Received 6 October 2017; Received in revised form 20 February 2018; Accepted 26 February 2018
0968-090X/ © 2018 Elsevier Ltd. All rights reserved.

to be samples with invalid values, like miscounts, sensor calibration errors or round-off errors (Van Lint et al., 2005). In other cases, the same managers aggregate or process data before publishing, a mechanism that sometimes entails errors (Zhong et al., 2004a). These eventualities result in data streams with missing portions of data of diverse sizes, having a negative effect on the forecasting models (Van Lint et al., 2005; Chen et al., 2001; Sun et al., 2004; Li et al., 2013).

Evidently, missing data unchain problems not only in traffic forecasting, but in any prediction, regression or data analysis based on data obtained from diverse sources (Schafer, 1997). Thus, researchers from many fields have devoted significant efforts towards new imputation methods for missing data. As such, one of the most straightforward approaches is to fill in the gaps with artificially created data (Moffat et al., 2007; Kondrashov and Ghil, 2006; Shrive et al., 2006; Sainani, 2015; Arteaga and Ferrer, 2002; Sterne et al., 2009). Although these fields are related to atmospheric, meteorological or geophysical variables, they relate to time series and some of their typical issues are common to traffic time series. For instance, a thorough review of imputation techniques for CO_2 flux time series is contributed in Moffat et al. (2007), most of which are applicable to a traffic context. Strategies for imputing missing data can be of paramount relevance also in traffic datasets. As a matter of fact, the quality of data, defined as the *fullness of data*, has been lately identified as one of the major challenges of road traffic forecasting, including data-driven approaches (Vlahogianni et al., 2014).

1.1. Related work

In the traffic forecasting domain, elaborated missing data imputing methods were first reported in the early 2000s, when a few approaches were introduced in Chen et al. (2001) and later categorized by Smith et al. (2003) in two main groups: (1) statistical, considering Expectation Maximization (Dempster et al., 1977) and Data Augmentation algorithms; and (2) heuristic methods, comprising various averaging techniques over historic data. A more recent classification by Li et al. (2013) divides imputation strategies into those based on prediction, interpolation and statistical learning. The inclusion of a prediction category brings many more methods based on considering missing data as values to be predicted. Among the representative literature related to this category it is worth to highlight the seminal work in Chen et al. (2001), where a simple historical mean imputation was shown to outperform *no-substitution* and *substitution-by-zero* methods when used in combination with an Auto Regressive Integrated Moving Average (ARIMA) and an Artificial Neural Network (ANN) as prediction models. Remarkably for the scope of our research, this early study considered missing data densities of up to 30%, generated uniformly at random. Authors also showed that ARIMA models are more sensitive to missing values than their ANN counterparts.

In general, a model that relies on the time dimension of a dataset is prone to be sensitive to missing data, as these models typically require an uninterrupted time series as their input. On the other hand, when a dataset has a substantial extension with very few corrupted/missing data entries, a simple strategy of removing instances affected by gaps or imputing a constant value to them may suffice for the forecasting method to model the traffic conditions (Vlahogianni et al., 2014). Van Lint et al. (2005) consider null imputation, linear interpolation and ARIMA as filling methods prior to a State Space Neural Network predictive model, dealing with up to 40% of randomly located missing data occurring successively in intervals of length up to 30 samples. In their scenario, simple, non-parametric imputation methods were shown to handle missing data efficiently. Henrickson et al. (2015) introduce a statistical approach that performs successfully even with 1-month-long missing data. Their so-called predictive mean matching method draws random values to impute from a distribution obtained from the present values, considering one measuring station. Probabilistic Principal Component Analysis (PPCA) method was also proposed in Qu et al. (2009), addressing some commonly made assumptions about missing data. Methods relying on component analysis have been widely used ever since (Li et al., 2013; Chen et al., 2012; Chiou et al., 2014; Asif et al., 2016; Ran et al., 2016; Li et al., 2014) and, to the date of this survey, they embody one of the most popular processing approaches for imputing missing data. In a comparison among 6 methods performed by Li et al. (2014) authors conclude that PPCA is the most efficient imputing technique within their sample not only in terms of performance, but also in ease of implementation and speed. Other numerical approaches include (1) Bie et al. (2016), where an online imputation method is proposed consisting of a multiple linear regression based on data from loops that are part of the same measuring station; and (2) the similarity-based imputation technique proposed by Zhong et al. (2006), where daily curves with gaps are compared to candidate curves without gaps, using the closest one – under a measure of similarity – to impute. The missing intervals reached 12 h length, but they only considered one type of day pertaining to a particular season of the year. Tensor based methods have been exploited recently to deal with missing data introducing spatial context relations (Ran et al., 2016; Asif et al., 2016; Tan et al., 2014). These methods model the interactions between multiple traffic variables into multi-dimensional arrays (tensors), thus allowing for the combination of multiple correlations between the different variables to impute missing data.

Machine learning methods are also becoming prominent in recent years, most of them falling in the aforementioned *prediction* category. Kernel regression in combination with k-Nearest Neighbors (KNN) was used in Haworth and Cheng (2012) to obtain forecasts of missing values using information from neighboring stations. The study only covered input data generated on Tuesdays, but they performed an analysis of the missing data characteristics present in the dataset in order to generate gaps that realistically mimic the real ones. Imputation of missing data was also tackled as predictions in Zhong et al. (2004a,b), which proposed to build ANNs optimized via genetic algorithms to obtain missing data estimations of up to 1 h. Clustering approaches have been recently explored in Tang et al. (2015) and Ku et al. (2016). The former introduces the widely neglected distinction between days of the week, representing the input data as values taken on a time step of a certain day of the week. This helps the model to distinguish patterns in different days. A Fuzzy C-means algorithm is then used to group known days, and a genetic algorithm to estimate missing data by minimizing errors between imputation and actual values of clusters. Likewise, Ku et al. (2016) considers a large group of sensors of a network and uses a K-means algorithm to cluster them based on their average daily traffic; then they use a deep learning method –

Download English Version:

https://daneshyari.com/en/article/6935962

Download Persian Version:

https://daneshyari.com/article/6935962

Daneshyari.com