On the Imputation of Missing Data for Road Traffic Forecasting: New Insights and Novel Techniques

Ibai Laña, Ignacio (Iñaki) Olabarrieta, Manuel Vélez, Javier Del Ser

PII: S0968-090X(18)30253-5
DOI: https://doi.org/10.1016/j.trc.2018.02.021
Reference: TRC 2004

To appear in: *Transportation Research Part C*

Received Date: 6 October 2017
Accepted Date: 26 February 2018

Please cite this article as: Laña, I., (Iñaki) Olabarrieta, I., Vélez, M., Ser, J.D., On the Imputation of Missing Data for Road Traffic Forecasting: New Insights and Novel Techniques, *Transportation Research Part C* (2018), doi: https://doi.org/10.1016/j.trc.2018.02.021

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], and 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*

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.

Keywords: Traffic forecasting; missing data; cluster analysis; data imputation.

*Corresponding author: ibai.lana@tecnalia.com (Ibai Laña). OPTIMA Unit. TECNALIA. P. Tecnologico, Ed. 700, 48160 Derio, Spain. Tl: +34 946 430 850.

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; Van 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 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, Lingras and Sharma, 2004). 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, Yu and Zhang, 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. (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. in (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, Li and Li, 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, Su, Zhang, Hu and Li, 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.(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. (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 hour 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, Lingras and Sharma, 2004; Zhong, Sharma and Lingras, 2004), which proposed to build ANNs optimized via genetic algorithms to obtain missing data estimations of up to 1 hour. 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 – specifically, a Stacked Denoising Autoencoder (SDAE) – to model relationships between sensors of each cluster. Once built, the model is able to impute missing values to all the sensors simultaneously. The performance of the model is tested over 6 days of data with 10% to 90% missing values. In a similar direction, (Duan et al., 2016) presented a SDAE that considers weekdays and non-weekdays, different selections of sensors, and up to 50% of missing data.

Along with all the above imputing methods, some authors derive robust models to cope with missing data and obtain forecasts without considering any imputation mechanism (Whitlock and Queen, 2000). Sun et al. (Sun, Yu and Zhang, 2004) introduced a sampling Markov chain method to carry out short-term traffic forecasting with incomplete data with no previous imputation phase. Later, this work was extended in (Sun et al., 2006) by using a Bayesian inference mechanism to obtain robust predictions with incomplete data, and complemented in (Sun and Zhang, 2007) by a selective random subspace predictor that leans on the information supplied by surrounding sensors that are correlated to the one under study. By exploiting this augmented and redundant information subsets with missing data can be dismissed.

1.2. Contribution

Despite these approaches, incomplete data can become a problem – even for data-oriented robust models – when the amount of missing values is high and spans long periods for which no useful information can be considered to obtain a model (Li et al., 2013). Surprisingly, despite this widely acknowledged statement the literature so far is scarce in what regards to empirical evidences of the comparative performance of imputation strategies under different yet realistically modeled distributions for missing data. Moreover, the implications of imputed data in the performance of predictive models for traffic forecasting have not been deeply studied and analyzed. This manuscript aims at presenting and discussing strategies to deal with missing data, as well as to obtain new insights and a comprehensive, global view on the relevance taken by data imputation methods in traffic forecasting scenarios. Specifically, the mayor contributions of our work can be summarized as follows:

- A review of the techniques for generating synthetic missing points and intervals (missing data), numerically exploring their implications on the quality of imputed data.

- An analysis of the impact of the distribution of missing data and the imputing methods on the performance of forecasting methods.
- Two novel imputing strategies to tackle long periods of missing data from two different perspectives: 1) a pattern clustering-classification (PCC) algorithm which incorporates external data that that are always available, such as days of the week, months or holiday information, and 2) an Extreme Learning Machine (ELM, (Huang et al., 2004)) model optimized with a genetic algorithm, that builds upon information obtained only from surrounding sensors, which we have called spatial context sensing complete (SSC) and optimized (SSO).
- The use of an extended 2-year dataset obtained from a sensor network deployed over the city of Madrid (Spain) and publicly available as open data (*Madrid Open Data Portal*, n.d.).

The rest of the paper is organized as follows: Section 2 describes the input data, the different artificial missing data generation techniques, the proposed imputing methods, the comparison methodology, and the results evaluation approaches. Section 3 presents and analyses the performance of the proposed methods in different missing data scenarios. Finally, Section 4 draws concluding remarks inferred from the obtained results and prescribes future research lines related to this work.

2. Materials and Methods

In order to extract informed conclusions from empirical findings our research work uses traffic data obtained from a public source. Over them, artificial missing data are created and our proposed imputation methods are applied. The following subsections describe the source and selection criteria for the input data, the missing data generation methods, our imputation models and the performance evaluation procedures.

2.1. Input data selection

Input data for this research have been collected from a public source maintained by the City Council of Madrid (Spain), which has more than 3600 ATRs deployed through its road network, some of them in purely urban context and others in urban freeways. Data provided by these sensors are published live every minute in its Open Data portal (*Madrid Open Data Portal*, n.d.), and historically in the form of 15-minute aggregated periods. Using one-minute resolution data would require a collecting process that would take as long as the time span of the desired dataset.

To overcome this issue, we instead focus on historic 15-minute data of complete years, which provides enough information to consider seasonality in a data driven approach: one year can be used as training data for the developed models, and any other as test data. This seasonality can be of great relevance depending on the traffic profile of a certain location: in a business area, a model trained with data collected in March would intuitively perform poorly when predicting values for the month of August. On the contrary, in a residential area with less fluctuating traffic profiles, winter data might be useful to obtain summer forecasts. A model trained with whole-year data can, on the other hand, learn seasonal patterns and apply them for the prediction.

By the time this research line was started aggregated published data were just available for the years 2014 and 2015, and three months of 2016. Therefore, input data are taken from a subset of sensors for 2014 and 2015. The choice of the sensors for further analysis was made under the following criteria: a location close to the city center, avoiding flat traffic profiles of residential areas (for which imputing missing data would be more straightforward, potentially misleading our conclusions); and the availability of data, required to assess the imputing performance after artificially generating missing data points and intervals. Figure 1 shows all ATRs located within a 2 kilometer radius of Puerta de Alcalá, one of the main business areas of the city, which represents a first filter for our imputing model. The color code portrays the available percentage of the total 35040 annual readings for each magnetic loop during 2014, which will be subsequently used as training data for our models. A considerable amount of sensors have less than 50% of data available in this year, and from 186 loops accessible in this area, only 21 have served data for more than 98% of the period. This noted fact emphasizes the actual need for robust imputation methods in this particular context of application.

PLACE FIGURE 1

The introduced spatial context imputing strategy is built upon past information of the studied ATR and past and current data coming from neighboring sensors. In an application context, our spatio-temporal strategy would rely on the neighboring sensors with the most complete information available. Hence, we have taken into account only those locations with more than 34500 observations available (more than 98%) for year 2014 and consider them as training data, yielding the set of 21 loops depicted in Figure 1 as the first of the categories. On the other hand, the testing of our spatial context model requires 2015 complete data from surrounding loops. Thus, data from aforementioned locations is examined for 2015, seeking the longest series of consecutive correct readings common to all locations. A shared subsequence of 8463 consecutive observations (ca. three months of data) has been

found for 13 of the sensors. One of these 13 sensors has been randomly selected as the target ATR, while the rest $N = 12$ are used as context sensors. In the test data from that sensor, artificially generated missing points and intervals will be introduced (modeled as later explained in the following Subsection), and imputation will be performed on those synthetic missing data. The rest will act as surrounding loops. These sensors are shown in Figure 1 highlighted with a star marker, while the loop under study is annotated as the *target*.

We denote the observation obtained from the i-th ATR at time index t as o_t^i where the time index t spans over years 2014 and 2015, and i takes integer values from the range $[0, N]$. The selected target ATR for which the imputation process will be performed corresponds to $i = 0$, and for notational convenience will be labeled henceforth as s. When referring only to the context ATRs index $j \in [1, N]$ is used. The subset of observations used for training the models, i.e. with time indexes corresponding to 2014 historic data, are denoted as $\mathcal{H}^i = \{o_t^i : t \in 2014\}$, while the subset of observations with 2015 time indexes, used for test, are denoted as $\mathcal{G}^i = \{o_t^i : t \in 2015\}$. As with the individual observations, these sets are instanced subsequently as $\mathcal{H}^s$ and $\mathcal{G}^s$ to specify observations taken from the target loop s, and $\mathcal{H}^j$ and $\mathcal{G}^j$ to refer to those of the context sensors.

2.2. Generative Models for Missing Data

Before delving into the models used for generating missing entries in the considered test dataset $\mathcal{G}^s$, it is insightful to note that some authors deal with incomplete datasets from a prediction perspective: instead of presenting a strategy to fill in the gaps, they rather propose models to obtain forecasts overcoming gaps (Li et al., 2013; Haworth and Cheng, 2012; Treiber and Helbing, 2002; Abadi et al., 2015). Consequently, their score to measure the effectiveness of their methods hinges on the performance of the chosen prediction models regardless of which data were declared as missing. On the contrary, this work focuses on comparing among imputing models by using a defined set of synthetically generated missing data, as well as determining to which extent an improvement of the imputed value yields an enhanced accuracy of subsequent traffic forecasting models.

This being said, three broad families of generative models for missing data can be found in the literature (Little and Rubin, 2014): the so-called *Missing Completely At Random* (MCAR), *Missing At Random* (MAR) and *Not Missing At Random* (NMAR). The first two imply that there is no mechanism underneath for generating the missing data, whereas the latter assumes a dependence of the distribution of missing data on the complete dataset. This general classification has been used as a reference by most researchers in the traffic context (Qu et al.,

2009; Tang et al., 2015; Henrickson et al., 2015). Van Lint et al. (Van Lint et al., 2005) describes three types of data failures: random, structural and intrinsic, where the first represents stochastic reading or transmitting errors, the second consists of gaps resulting of a sensor being offline, and the latter refers to noise, bias or errors caused by processing the data. A similar classification is proposed in (Haworth and Cheng, 2012), with two random (one with all independent and one with related missing points) and one structural model for missing data generation. Chiou et al. (Chiou et al., 2014) alludes to the unlikelihood of distinguishing the source or kind of the missing data, reducing them to two practical categories: *point-wise* and *interval-wise* missing data, representing MCAR and MAR respectively, and considering that intervals are groups that occur randomly. They also contemplate a mixture of both types in their datasets. Although with different names, these two approaches for creating missing data are common to most related contributions: some authors consider only point-wise random generation expecting that high percentages of missing data will create long intervals, whereas the rest tend to consider both methods, either combined or in isolation.

In this line of work, artificially generated gaps ranging from 25% to 65% of the dataset are considered in (Falge et al., 2001), using the rest of data as an input for their imputing methods. In (Chen et al., 2001) gaps are generated for 10%, 20% and 30%, and in (Van Lint et al., 2005) the percentage increases up to 40%. Moffat et al. (Moffat et al., 2007) produced up to 50 gap scenarios, defining 4 sizes of gaps and making 10 combinations of each size, plus 10 scenarios with mixed sizes, although in all cases the total amount of missing data amounted up to 10% of the entire dataset. In (Zhong et al., 2006) 12 successive hour gaps were introduced in different days, which allowed studying their effect and the effectiveness of their considered imputation techniques depending on the type and hour of the day. In (Ku et al., 2016) a clustering approach was applied to randomly generated gaps for up to 90% of the dataset, obtaining satisfactory results even with large portions of missing data. Interval-wise generation ranges from 24 consecutive points as in (Chiou et al., 2014) to 1 month as in (Henrickson et al., 2015). These extended range gaps can be regarded as a representative application of NMAR generative models for missing data, where a failure in the sensor or the communication hampers a proper collection of data for a long period.

The experiments in this work use a dataset $\tilde{\mathcal{G}}^s$, which is the result of artificially removing data from $\mathcal{G}^s$ by both of the approaches detailed below. The elements of $\tilde{\mathcal{G}}^s$ are denoted as $\tilde{o}_t^s$, namely $\tilde{\mathcal{G}}^s = \{\tilde{o}_t^s\}$. Each of these values, $\tilde{o}_t^s$, is either *well defined*, i.e. equal to the observation o_t^s, or is an artificially generated blank. For a

given value of t a function is defined such that

$$\delta(t) = \begin{cases} 1, & \text{if } \tilde{o}_t^s \text{ is well defined,} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

In the following subsections a detailed explanation of the artificially generated missing data methods is provided.

2.2.1. Point-wise generation

We have defined percentages ($\xi \in \{1, 10, 25, 50, 80, 100\}$ [%]) of missing observations in the whole test $\mathcal{G}^s$ sequence of data entries. The missing data points or *blanks* are placed individually at random. When the percentage is low, missing points are separated from each other naturally, i.e. consecutive blanks are rarely obtained for low values of ξ. When ξ is increased, clusters of gaps emerge and it results easier to find sequences of missing data. Due to the Gaussian distribution of holes, even for $\xi = 50\%$ and $\xi = 80\%$, there are no completely *empty* days (i.e. days with all-blank entries). The case when $\xi = 100\%$ is uncommon in previous works; its purpose is to test the effectiveness of methods introduced in this work under these circumstances (3 complete months of missing data).

2.2.2. Interval-wise generation

Traffic flow observations posted in (*Madrid Open Data Portal*, n.d.) for the urban network of Madrid have been preprocessed beforehand and, as in many other cases (Zhong, Lingras and Sharma, 2004), missing points could have already been imputed by the entity managing this repository. This means that a 15-minute reading integrates multiple shorter-term observations and also that, in this particular case, missing data could be mainly due to errors in the aggregation or processing stages. In order to generate intervals of missing data that actually reflect the behavior of gaps in our dataset, we have assessed the real distribution of missing data for all the year 2015, which contains the test set of our experiments, for each of the considered measuring points (Table 1).

PLACE TABLE 1

Despite their sparsity, the similarities found in all the considered locations, such as the almost identical percentage of missing data or the number of gaps, suggest that errors are probably produced in the aggregation stage and affect similarly to groups of sensors. The most frequent gap length is 96 positions, which corresponds exactly to one day worth of data; a further inspection of the data at hand reveals that these gaps usually match natural days, starting at 0:00 AM and

ending at 11:45 PM. Also the distribution of gap lengths has been examined for the target loop: besides the 96 length gap, the most frequent lengths are 48 and 192 positions (half a day and two days of data records, respectively).

According to these characteristics of the input data, it is expected that any missing data generation strategy not producing entirely empty days (e.g. any of the random point-wise generation percentages) will not properly represent the statistical distribution of real gaps in this particular scenario. Consequently we have defined 6 sets of target data, each of them with gaps of 24, 48, 72, 96, 144 and 196 consecutive positions respectively. For each set, gaps are placed randomly and amount 13% of the total test data.

2.3. Imputing Data Methods

In this work we introduce two new approaches for imputing missing data. One that depends on information gathered from other sensors and other approach depending on external factors that define clusters of days. The following subsections describe the details of these two methods.

2.3.1. Spatial Context Sensing

Traffic state data gathered by a sensor network supply spatio-temporal information, as vehicles often navigate through several detectors along their trajectories[1]. Intuitively, the traffic profile at a road segment should be very similar to that in an upstream segment a τ before, whenever τ equals the average travel time between both segments. Nonetheless, there are two main factors that put into question this intuitive statement, e.g. the lack of continuity due to road bifurcations or parking areas, and the speed dispersion. Features of gathered data, such as the distance between points of collection, or the location of sensors in an urban context can make the effect of previously mentioned factors more noticeable. Moreover, the available temporal resolution, renders it impractical to establish a direct relationship between the measurements taken by two neighboring sensors, even when they are in two adjacent segments. To illustrate this, we hypothesize two sensors placed in an urban street with synchronized readings at intervals of ΔT = 15 minutes. In this scenario, any direct correlation of their traffic profile would be most probably

[1]The use of non-stationary sensing devices such as probing vehicles would require analyzing spatial-temporal correlations over time among the information captured by such vehicles. As indicated later in the concluding section, depending on the availability of this particular kind of data the aforementioned research line should drive future research efforts aligned with this work.

spurious, as in 15 minutes great variations may occur in an urban context. Despite this noted relational uncertainty, plenty of contributions dealing with traffic prediction and missing data imputation (Van Lint et al., 2005; Li et al., 2013; Haworth and Cheng, 2012; Ku et al., 2016; Duan et al., 2016; Sun et al., 2006; Sun, Zhang, Yu, Lu and Xiao, 2004) have relied on spatio-temporal relationships, even in urban contexts and with coarse-grained data, on account of different techniques that allow researchers to find interrelation models among nearby located sensors (Chen et al., 2003).

This being said, it is noticeable in Figure 1 that distance between the location under study and the others is not necessarily short, i.e. they are not so *closely neighboring*. Our first proposed imputing method leans on the relationships between measurements of different, not necessarily nearby, sensors at the center area of a city. Missing data entries are imputed by means of a forecasting model that predicts values for a sensor by learning from the information provided by other sensors. As exposed in (Sun and Zhang, 2007), correlations between the traffic among two separate links produce better forecasting results disregarding the distance (Ku et al., 2016). Conceptually, the model retrieves data from locations (where available), being defined initially by a great deal of observations collected from each loop. This does not necessarily produce a good model, as some of the loops can be placed in locations with very different traffic profiles, and would constitute noise for the imputation procedure. For this reason, an optimization step is added to the predictive model to adjust the amount of information that each sensor contributes to the training dataset. Figure 2 displays the overall operation of this model.

For a specific reading o_t^s of the selected loop s at time t, a number w_j of observations prior to t are taken from each surrounding loop $j \in [1, N]$ towards defining a vector of features that ultimately constitutes the dataset with the observation o_t^s as target variable. The window size w_j of each loop can be different, suggesting a level of influence of the surrounding of loop j in the prediction of o_t^s. The forecast horizon h is defined as the number of time steps in the future for which the prediction is made, i.e. the difference between t and the most recent time of the samples collected for the surrounding sensors. For instance, by setting $h = 1$ forecasts of values taking place ΔT in the future.

It must be noted that this method requires a certain degree of completeness within the historical training data $\{\mathcal{H}^i\}$ from which the model is constructed (indeed the selection of surrounding loops has been made accordingly), but its training phase is robust to sporadic missing data in the historic dataset. Real missing data in the train time series are flagged and after the train dataset is built, instances containing flags are removed, still resulting in a relatively large dataset for the sce-

nario in hands (more than 32000 training samples available out of the initial 34500 entries in the retrieved repository). This is an important feature, as most imputing methods require complete historic data (Asif et al., 2016), becoming a practical issue in the majority of real life scenarios.

The selection of neighboring sensors hinges on the aforementioned aspects. Since the distance between the target sensor and its surrounding ones is not exploited anyhow by the model, the main selection criterion is set to the level of data completeness of the neighboring sensors. The number of selected loops is not fixed beforehand, and is a result of the balance between the minimization of the size of the dataset in terms of number of features, and the maximization of the variability of inputs provided by different sensors. All sensors available in the central area of Madrid were considered for this research, but after disregarding those with more than 2% of missing data, only 12 remained for subsequent processing. The flexibility of this method would allow selecting sensors from a larger area, should the initial ones not comply with the completeness constraint.

PLACE FIGURE 2

The combination of diverse kinds of Artificial Neural Networks (ANNs) and other machine learning methods with heuristic optimization algorithms has been extensively explored in previous works (Vlahogianni et al., 2014), often yielding more responsive results to changes in data than time-series forecasting. In this scenario, our imputing method is built upon the predictions obtained by an Extreme Learning Machine (ELM) model, which is trained with a dataset built by following the scheme depicted in Figure 2. The general operation of the model is similar to that in our previous works (Laña, Del Ser, Vélez and Oregi, 2017; Laña, Del Ser and Vélez, 2017), introducing in this case the inputs from surrounding sensors. A bio-inspired heuristic solver is introduced to find the optimal window sizes w_j of each sensor j. This procedure can reduce considerably the processing time, and provide insights on the importance of some of the sensors to predict and impute the missing values of the target location, if any of the optimized window sizes equals 0.

Initially, a maximum value of 50 steps (12.5 hours) is defined for the window size w_j of all 12 surrounding loops around the target loop, rendering a dataset of 600 features and around 32000 instances. The population of the bio-inspired heuristic solver is composed of the window sizes of each surrounding loop, and in each generation of the optimization algorithm, the prediction model is built, trained and validated, obtaining an RMSE performance metric. When the optimization process ends, the window sizes found in the best generation are the ones that potentially yield the best RMSE score. The ELM model is then trained con-

sidering these windows on the $\{\mathcal{H}^i\}$ dataset, and tested on the $\{\mathcal{G}^i\}$ dataset as a single hold-out. This produces forecasts for all values of the $\mathcal{G}^s$ series, as if the missing data were the 100%, making this method an interesting option under such circumstances. Values obtained for positions where a gap was generated are then compared to the actual observation, and assessed via the metrics discussed below.

2.3.2. Pattern Clustering and Classification

The second method proposed in this paper involves only data from the target loop. As in the previous technique, a set of samples prior to the period at hand is required for training. This method is designed to produce data in a complete day fashion, as opposed to point-wise filling counterparts. Several schemes in the literature (Zhong, Lingras and Sharma, 2004; Chiou et al., 2014; Ran et al., 2016; Li, Su, Zhang, Hu and Li, 2014; Ku et al., 2016) involve splitting the series of data in lots of data per day. The pattern clustering and classification imputing method, as well as two of the proposed comparison methods, perform this splitting of incoming data into day-wise vectors:

$$\mathbf{o}^{s,d} = \left[o^s_{t_d}, o^s_{t_d+1}, \ldots, o^s_{t_d+P-1}\right], \tag{2}$$

where $o^s_{t_d}$ is the value of the observation captured at sensor s and time t_d, being t_d the first time index of day d; and $P = 96$ is the number of observations obtained within a day for a capture period of $\Delta T = 15$ min. Following the dataset division criterion explained in Subsection 2.1, we have defined a training dataset with $\mathcal{H}^s$ data, $\boldsymbol{\mathcal{H}} = \{\mathbf{o}^{s,d} : t_d \in 2014\}$ and a dataset with $\tilde{\mathcal{G}}^s$ data $\tilde{\boldsymbol{\mathcal{G}}} = \{\tilde{\mathbf{o}}^{s,d} : t_d \in 2015\}$.

In general, as $\tilde{\mathcal{G}}^s$ includes artificially generated blanks, each vector $\tilde{\mathbf{o}}^{s,d}$ has $P^{s,d} \leq P$ valid (non blank) values given by

$$P^{s,d} \doteq \sum_{p=0}^{P-1} \delta(t_d + p). \tag{3}$$

Based on this definition we establish a metric of similarity between any vector from $\boldsymbol{\mathcal{H}}$ and any vector from $\tilde{\boldsymbol{\mathcal{G}}}$:

$$S(d, d') \doteq \frac{1}{P^{s,d'}} \sqrt{\sum_{p=0}^{P-1} \delta(t_{d'} + p) \left(o^s_{t_d+p} - \tilde{o}^s_{t_{d'}+p}\right)^2}, \tag{4}$$

where $d \in \boldsymbol{\mathcal{H}}$ and $d' \in \tilde{\boldsymbol{\mathcal{G}}}$.

2.3.2.1. Clustering. Once the input data are separated in days, a clustering algorithm is performed over $\mathcal{H}$, obtaining groups of days with similar set of measurements. Performing a clustering process over a space with such a large number of dimensions requires large computational resources. Furthermore, the overall process could be biased by localized, high-frequency noise, producing too many groups for the overall cluster space to be useful. To overcome this issue the dataset is preprocessed by averaging every K samples.This averaging process not only reduces the number of dimensions of the space over which to perform the clustering process from P down to $\lceil P/K \rceil$, but also smooths out any local disturbance the measurements may undergo, reducing the chances of producing too many clusters (more than the necessary to represent the actual traffic patterns).

Two clustering algorithms have been considered to produce groups within the feature space based on the above similarity metric: DBSCAN (Ester et al., 1996) and Affinity Propagation (Frey and Dueck, 2007). DBSCAN is a density based clustering algorithm which delimits clusters by regions where the density of samples is high, labeling points located in low-density regions as *outliers*. Affinity Propagation is a clustering algorithm based on exchanging messages between the different data points. It finds "exemplars", members of the input set that are representative of clusters. Both methods produce similar results when their parameters are chosen appropriately, therefore only one of them (DBSCAN) has been used for the experiments.

As opposed to other clustering techniques such as K-Means, DBSCAN does not require the number of clusters as an input. Instead, its parameters have been tuned through an off-line optimization process aimed to reduce the number of *noise* instances (outliers) while maximizing the number of well-conformed clusters. A technique such as K-Means would always assign all data instances to a cluster disregarding whether their dissimilarity to the closest cluster calls for another cluster to be spanned. Thus, this method is not suitable for a problem in which the number of clusters cannot be easily established. The parametrization of density-based clustering schemes such as DBSCAN allows for a fine-grained determination of clusters and the isolation of overly dissimilar data instances as noise clusters. A balance between these two extremes is sought: the cluster space should discriminate as many *relevant* data patterns in the dataset (cohesive clusters, with high intra-cluster and low inter-cluster similarity), and discard those instances (namely, daily traffic data traces) notably dissimilar to any other in the dataset. As a result, the DBSCAN clustering model is configured with $eps = 6$ (maximum Euclidean distance between samples belonging to the same cluster) and $minPts = 5$ (minimum number of data instances for a cluster to become meaningful, i.e. not noise).

The result of the clustering algorithm produces a partitioning of $\boldsymbol{\mathcal{H}}$ into C clusters $\{\boldsymbol{\mathcal{H}}_c\}_{c=1}^{C}$. Cluster $\boldsymbol{\mathcal{H}}_c$ is represented by its centroid $\mathbf{o}^{s,\odot_c} = [o_{t_d}^{s,\odot_c}, \ldots, o_{t_d+P-1}^{s,\odot_c}]$, which is computed by taking the average of its member observations, i.e. the p-th element of the centroid is computed by taking the average of the p-th elements of all the members of that cluster:

$$o_p^{s,\odot_c} = \frac{1}{|\boldsymbol{\mathcal{H}}_c|} \sum_d o_{t_d+p}^{s} \tag{5}$$

where $o^{s,d} \in \boldsymbol{\mathcal{H}}_c$ and $|\cdot|$ denotes cardinality of a set.

2.3.2.2. Classification. The previously defined clustering process would suffice for imputing missing values, and in fact it will be used as a comparison method in the experiments later discussed, choosing the closest $\mathbf{o}^{s,\odot_c}$ to the element of $\tilde{\boldsymbol{\mathcal{G}}}$ where the imputation is needed. However, when there is a particular day for which all measurements are missing – i.e. $P^{s,d} = 0$, the clustering process is not able to assign it to any of the clusters. In order to overcome this shortcoming, external information independent from the traffic data is incorporated to the dataset, and an algorithm is built over the C clusters obtained in the method explained in the previous subsection. A supervised learning classifier is trained with cluster indexes as classes, and over those features that do not depend on the actual traffic observations. We have designated as features the day of the week D, the month M and a binary feature bH to indicate whether a day is a bank holiday (Laña et al., 2016). These time-related features are very relevant to group traffic by days, as traffic patterns are mostly daily cyclical (Whitlock and Queen, 2000). Other external features such as the weather or the celebration of regular events could also be included to obtain a more precise classification. Thus, a dataset with 3 features and C classes is composed from $\boldsymbol{\mathcal{H}}$, which is used to train a supervised classifier to estimate the cluster assignment of a day belonging to $\tilde{\boldsymbol{\mathcal{G}}}$.

The supervised learning model utilized for the regression problem posed in this paper is the so-called Random Forest (RF), which relies on the *bagging* concept (Ho, 1995; Breiman, 2001) to create a diverse set of regressors by introducing randomness in the construction of an ensemble of tree learners. This procedure has been shown to decrease the variance of the model without increasing its bias, as weak learners are fed with different training sets that consequently decorrelate their structure and provide diversity to the ensemble. Imputation is finally done by equaling missing entries of the tested day to those of the centroid $\mathbf{o}^{s,\odot_c}$ of the cluster to which it is predicted to belong. Specifically, if $\tilde{\mathbf{o}}^{s,d}$ denotes a test day for which $P^{s,d}$ missing entries are to be imputed, the proposed method creates a

vector with components:

$$\hat{o}^s_{t_d+p} = \begin{cases} o_p^{s,\odot_c} & \text{if } \delta\left(t_d+p\right)=0, \\ \tilde{o}^s_{t_d+p} = o^s_{t_d+p} & \text{otherwise.} \end{cases} \tag{6}$$

The whole clustering classification process is graphically summarized in Figure 3.

PLACE FIGURE 3

2.4. Methods for Comparison

A selection of the most common methods have been used to appraise the performance of the ones here proposed. Early research in this field (Chen et al., 2001; Smith et al., 2003) adopted some of the basic imputing methods that have been used ever since for comparison: historical average, average over surrounding locations, average over close timestamps, or Expectation-Maximization (EM) methods. The diversity of imputation methods reported in the literature has grown lately, achieving high levels of complexity, but in general they continue to be benchmarked against the portfolio of imputing techniques mentioned above on account of their good performance when missing data entries are not profuse. All comparison methods are based on data taken from the target loop ($\mathcal{H}^s$ and $\mathcal{G}^s$), and do not use any context sensor information. Following this common practice, we have compared our proposed methods to 5 techniques of increasing complexity:

- *Basic Imputation (BASIC):* this naïve approach consists of imputing a constant value for all missing data, usually 0 (Chen et al., 2001) or a value based on statistical characteristics of the dataset, commonly the average of non-missing observations (Chen et al., 2001; Van Lint et al., 2005; Sun, Yu and Zhang, 2004; Asif et al., 2016). Although imputed values would probably differ from the actual ones, this method provides effortlessly a dataset without missing data, allowing the application of forecasting techniques in a straightforward manner.

- *Linear Interpolation (INT):* a reliable technique when missing data are scarce and individual (Van Lint et al., 2005). Imputing gaps of a small amount of positions with linear interpolation is fast, easy, produces fairly accurate values, and provides a smooth traffic profile. However, it degrades severely when the length of intervals with missing data increases.

- *Mean Day Variation (MDV):* this is one of the most common techniques to impute missing data, which resorts to averages of the available observations at the

same time index of the day to compute the value to fill in the missing entry (Moffat et al., 2007). In order to quantify the impact of the ratio ξ of missing values in the performance of this method, we have considered 2 possible input datasets: $\mathcal{H}^s$ dataset (corresponding to data captured in 2014 without any missing values) (MDV14) and $\tilde{\mathcal{G}}^s$ (corr., 2015 with artificial missing data) (MDV). The latter case depends on the quantity of missing data, and the performance of MDV is expected to degrade as the ratio ξ of missing values increases.

- *1-Nearest Neighbor (1NN):* this method, similarly to the clustering-classification scheme proposed in this paper, relies on the day-splitting paradigm described in Subsection 2.3.2. Conceptually similar to the method proposed in (Zhong et al., 2006), days with missing data from $\tilde{\mathcal{G}}$ are compared to those in the dataset with complete days $\mathcal{H}$, looking for the closest one under the measure of similarity given in Expression (4). Missing values within the incomplete sample $\tilde{\mathfrak{o}}^{s,d}$ are filled with those of its closest instance in $\mathcal{H}$.

  This procedure is simple to implement and computationally efficient, but presents several potential problems: 1) when $\mathcal{H}$ is large finding the minimum through a exhaustive search can be time demanding; 2) when trying to impute values for a day with completely different measurements than any of the days in $\mathcal{H}$ the process will produce values far from the *real* ones; 3) this procedure might be highly influenced by high frequency noise in the data sample.

- *Clustering (CL):* the use of clustering techniques has become frequent in the field of missing data imputation (Tang et al., 2015; Ku et al., 2016). For comparison purposes we consider a clustering algorithm defined analogously to the one described in Subsection 2.3.2. In this standard clustering, instances from the $\tilde{\mathcal{G}}$ dataset are mapped directly to the clusters defined with the $\mathcal{H}$ dataset, instead of creating a proxy classifier, which in turn represents the core contribution of our clustering method. To this end, clusters are selected based on the minimum distance – as per (4) – between the instance to be imputed and the cluster centroids. Missing data of a particular day are filled with the averaged values of the cluster it belongs to. As other methods that rely on partitions of the dataset on a per day basis, this technique is expected to fail when entire days of data are missing.

2.5. Quantifying the Imputation Performance

Missing data imputation should not be regarded as an end in itself, but a necessary step to reconstruct data and perform forecasts. In contrast with some authors

that develop robust techniques to predict traffic regardless the missing data (Sun, Yu and Zhang, 2004; Whitlock and Queen, 2000; Sun et al., 2006; Sun and Zhang, 2007), most authors validate the imputing results by measuring their distance to real data, but they do not test the prediction accuracy of methods that use imputed data as inputs. In some cases, an imputation strategy can provide a marginal improvement over traditional approaches, but at a high computational cost. This efficiency trade-off could be worthless in practice should the differences between prediction performances with and without imputed data be negligible.

For this reason, besides the usual imputation error analysis, we propose an alternative methodology for assessing the prediction performance for each method. Once missing observations have been imputed over $\tilde{\mathcal{G}}^s$, an *imputed* dataset $\hat{\mathcal{G}}^s$ is produced for each value of ξ (ratio of missing observations) and L (interval length). Each dataset $\hat{\mathcal{G}}^s$ is split in two chunks: one containing the first 80% of observations ($\hat{\mathcal{G}}^{trn}$, *training*) and the second with the remaining 20% (corr. $\hat{\mathcal{G}}^{tst}$, *test*), for which imputed data $\hat{o}_t^s$ are replaced with their respective real values o_t^s. Thus, $\hat{\mathcal{G}}^{tst}$ is for all cases the same chunk of real test data, whose observations belong to $\mathcal{G}^s$. A Random Forest regression model is then trained on each of the $\hat{\mathcal{G}}^{tst}$ datasets, after which predictions are obtained and tested against $\hat{\mathcal{G}}^{tst}$. The prediction scores achieved with each input set of data (with original or imputed values) are compared to each other, providing insights on how the accuracy of every technique propagates to the prediction score of predictive models when the imputation is used to reconstruct datasets for traffic forecasting.

PLACE FIGURE 4

As depicted in Figure 4 the model is built analogously to the one presented in Subsection 2.3.1: a window of w observations (fixed to 20) predicts the value of the observation h = 1 steps in the future.

3. Results and Discussion

After conducting the experiments described in previous sections, we examine the performance of the proposed missing data imputing techniques in compared with the methods enumerated in Subsection 2.4. For the sake of statistical characterization, 20 independent runs have been completed for each percentage and missing interval distribution with different random positions of the missing points and intervals. Thus, each missing data imputation method is evaluated against 20 different sets of missing data ratios ξ and length interval L, except for ξ = 100, as no different combinations of missing data can be performed when all the data

are missing. The score utilized to evaluate results is the Root Mean Squared Error (RMSE), defined as:

$$\mathrm{RMSE} \doteq \sqrt{\frac{1}{N} \sum_{\forall t:\delta(t)=0} (o_t^s - \hat{o}_t^s)^2}, \tag{7}$$

where $N \doteq \sum_{t\epsilon 2015} (1 - \delta(t))$ denotes the number of imputed observations in $\tilde{\mathcal{G}}^s$.

Also, the coefficient of determination R^2, which shows the likelihood of real values to fall within the predicted ones, is calculated according to:

$$R^2 \doteq 1.0 - \frac{\sum_{\forall t:\delta(t)=0} (o_t^s - \hat{o}_t^s)}{\sum_{\forall t:\delta(t)=0} (o_t^s - \bar{o_t^s})} \tag{8}$$

These two evaluation metrics are averaged over the 20 experiments, obtaining averages and standard deviations reported in Tables 2 and 3 for RMSE and Tables 4 and 5 for R^2. The imputing methods are identified as defined in Subsection 2.4, using also SSC for Spatial context sensing complete (without optimization of window size), SSO for Spatial context sensing optimized and PCC for Pattern clustering and classification. Estimations are shown considering 1 significant figure, following the criteria described in (Hildebrand, 1987).

PLACE TABLE 2

PLACE TABLE 3 AROUND HERE

PLACE TABLE 4 AROUND HERE

PLACE TABLE 5 AROUND HERE Results displayed in both sets of tables lead to similar conclusions. When missing data consist of percentages of random missing points, most methods perform reasonably well even when percentage of gaps ξ reaches 80%. As expected, when all data are missing ($\xi = 100\%$), any technique relying in the availability of these data fails. Basic imputing does not yield good results in any case, as we are comparing values obtained from the distribution defined by real traffic to a constant. Mean-based methods produce an outcome with stability along different values of ξ and L, due to the averaging process that uses all the measurements available for a certain time frame of the day, disregarding the differences among types of days, which are remarkable in these central locations of the city. Linear interpolation outperforms the rest of methods in almost any scenario, which reflects the main inconvenient of this random missing data generation method: empty positions are distributed uniformly, and rarely in lengthy groups, allowing a simple linear interpolation method to produce good

estimations. But this randomly uniform distribution does not commonly represent the real disposition of gaps.

This issue is observed more clearly when inspecting Tables 3 and 5, where results for any of the comparison methods (except for the mean-based ones and the basic imputing) degrade severely as the length of gaps L is increased. With linear interpolation, this effect is specially noticeable: having 48 missing entries (half a day), the interpolation is made between two points within which great traffic variations may occur, hence traffic data in between cannot be modeled by a line. Above that size of gap, this technique is unable to produce acceptable values. Clustering and 1NN similarity interpolations behave similarly in both situations: with point-wise missing data generation, they are able to model fairly well traffic with the data they have available for each day, even with a 80% of missing data. When interval-wise gaps are 1 day long (L = 96 positions) their performance decays; no data are available to establish their similarity, thus they are assigned to random days. Figure 5 shows a boxplot for both kinds of gap generation scenarios, considering an interval of L = 96 positions, being this the most common, and a 10% of missing data, being this percentage the most similar to the total amount of missing data in the interval-wise mode (13%). This figure displays clearly that for a very similar volume of missing data, the way in which gaps are generated affects severely the RMSE results, except for the more robust techniques, such the ones proposed in our work (SSC, SSO, PCC).

PLACE FIGURE 5

In contrast with the performance deterioration that all comparison methods suffer when long intervals or complete absence of data are introduced, our proposed methods achieve a stable operation independent of the abundance or size of data gaps. Among them, spatial context sensing, based in measurements from surrounding sensors is more resilient to the unavailability of data during a certain time frame. Two versions of this scheme have been tested: one with "complete" sets of measurements (50 sized time windows for each surrounding loop, resulting in a dataset of 600 features), and one optimized by using the bio-inspired heuristic wrapper described in Subsection 2.3.1. In the latter, optimized window sizes are never greater than 6, and for some of the loops are 0, indicating in these cases the null importance of those sensors for the imputation and subsequent prediction procedures. This reduces the number of features to 40 on average, and speeds up the prediction algorithm used for imputing data: in an Intel i7 machine, the running time of SSO is 50% of the time required to run SSC.

Besides this lower computational cost, the outcomes of both versions of the method in terms of imputing performance are very similar; therefore, a Wilcoxon

test has been run in order to examine if such a difference is statistically significant. The p-values of this test are shown in Table 6. They demonstrate that within a 95% confidence interval, only in some of the point-wise percentage scenarios the results provided by SSO are significantly different from those obtained with SSC. Even in those cases, maximum error differences amount up to only 6 vehicles per hour. Thus, the complexities involved in the optimization of the algorithm could be avoided when the computation capacity is not a practical limitation.

PLACE TABLE 6

PCC also performs robustly in all considered situation. Moreover, its operation is more efficient in terms of computational complexity, with runtimes in the order of 94% relative to the time taken by SSC for the same scenario. A Wilcoxon test has been also performed to compare its results to those of spatial context sensing without optimization, rendering p-values shown in Table 7. As for the previous two methods, the statistically significant differences are found for the point-wise missing data, for which PCC performs better. Aside from this particular performance gain, it is remarkable that this method can be further improved by adding more traffic-agnostic features that in theory could improve the classification process.

PLACE TABLE 7

3.1. Prediction-wise imputation performance

According to the results discussed above, naïve similarity or interpolation methods should be used when missing data consist of short gaps, while the longer sized gaps would require alternative methods. Nonetheless, when the imputed data are analysed from the perspective of their ability to obtain accurate predictions, this intuition may change as discussed next:

PLACE TABLE 9

PLACE TABLE 8

Tables 8 and 9 shows the averaged RMSE and its standard deviation for 20 forecasts performed for each method and missing data model, following the process described in Section 2.5. With this evaluation approach, the most relevant matter is the abundance of missing data, as for low percentages the rest of data are enough to build a good forecasting model, regardless the quality of imputed values. This is clearly observed in Table 9, which represents 13% missing data for all cases. Although up to 2-day long gaps are created in this experiment, and even if we know that imputed values are inaccurate for some of the methods, all of them perform well, because the 87% remaining real data are sufficient to train properly the regression model. The same algorithm applied to a dataset with no missing data returns a score equal to $\mathrm{RMSE} = 74 \pm 2$, very similar to those presented in

Tables 8 and 9. Undoubtedly, this good performance owes much to the way in which machine learning algorithms operate and model data. When imputed data are "bad", they are essentially noise for the training, hence the forecasting performance will also depend on the algorithm's ability to deal with noise. Similarity methods (1NN and Clustering) impute a default day when they are not able to find a similarity pattern. This default day comes from the same loop and it is probable that its profile is similar to the missing data, therefore the acceptable results of completely missing dataset. On the contrary, for great amounts of missing data, performance degrades for naïve methods, but our proposed algorithms generate fairly robust predictions.

4. Conclusions and Future Work

In this work we have investigated into missing data imputation strategies for traffic data, covering all stages of the procedure, from the creation of datasets with artificial gaps to the evaluation of the obtained results. We have also presented two novel missing data imputing methods based in contextual and historic information of a certain traffic measuring point, aimed to obtain accurately imputed values when missing data are abundant or presented in long intervals. Real data obtained from the traffic sensor network deployed in the city of Madrid (Spain) reveal that a noticeable proportion of the locations present lengthy intervals of missing data, making it necessary to design an efficient imputation strategy for scenarios with these missing data characteristics.

First, the missing data model has been approached from the perspective of the allocation of artificially created missing points. Missing data distribution can vary substantially among different ATRs in a city, and much more among cities, thence a proper examination of the real gap distribution of a dataset should be performed prior to the generation of artificial gaps. For this particular dataset we have found out that simple random point-wise generation strategies fail to represent the true nature of the gaps. In fact, some of the evaluated imputation techniques perform reasonably good with high missing data percentages, but decay quickly when the length of missing data is increased, even if the total amount of gaps is low. A fair amount of the literature reviewed in this manuscript does not test validate models with long intervals, so it is not possible to know how they would perform if that were the case.

Furthermore, the evaluation of imputation results has been discussed. Beyond the similarity measurement of performance that is found in most of the literature, (based on comparing the imputed value to the actual one), we have gauged the

imputation performance by considering the forecasting ability of a dataset built with those imputed data. As a result, we have concluded that if the prediction technique is based on a machine learning approach capable (to a certain extent) of dealing with noisy data, and the amount of imputed data is low (less than 10%), the imputed values will have a low impact in the final forecasting performance. Even when the imputed values are very dissimilar to real ones (due to the use of a naïve imputing technique), they will represent a low amount of noise in the whole dataset, but a well trained forecasting method still can obtain accurate predictions. Being these forecasting techniques currently popular, a previous analysis of the distribution and profusion of missing data could save a significant processing time in future contributions dealing with this topic.

The presented imputation methods have produced steady outcomes regardless the distribution and size of missing data, and indifferently to the evaluation procedure. Our spatial context sensing algorithm has proven that with enough contextual information, it is possible to impute missing values even when no data is available. This reinforces the notion of relations among traffic profiles registered over an entire city, even though they are not directly upstream or downstream correlated. On the other hand, our clustering approach highlights again the relevance of external features to obtain traffic patterns, an easily obtained input that is scarcely used in traffic forecasting and imputing research. Three simple temporal characteristics have been enough to obtain a good performance of the clustering-classification algorithm, but in theory, other external features such as sports events, demonstrations or spatial information about locations of interest could enhance further the predictive power of these patterns. Both introduced techniques require big amounts of previous or spatial context data that are not always available, but their operation is flexible enough to take data from any existing sources.

In light of the experiments, we consider that adaptive techniques should be implemented for data imputing, using a mixture of computationally efficient methods, such as linear interpolations for individual or short interval missing values, and more complex algorithms for filling in long intervals of missing data. This hybridization will lie at the core of future research lines stemming from this work. Furthermore, traffic prediction techniques are gradually steering towards online learning models over data streams, due to the presence of non-stationarities within the data and the need for incremental learning models capable of learning efficiently from fast-arriving data streams. The added difficulty yielded by this change of paradigm calls for the study of *online* imputation schemes suited to deal with concept drifts and adaptive windowing. In addition, the consideration and fusion of spatial-temporal information provided by probing vehicles will stimulate further

developments around online learning, with emphasis on how to dynamically infer the predictive context of the traffic at a certain location from the captured data.

Acknowledgments

This work has been supported by the Basque Government through the ELKARTEK program (ref. KK-2015/0000080 and the BID3ABI project), as well as by the H2020 programme of the European Commission (grant no. 691735).

Bibliography

Abadi, A., Rajabioun, T. and Ioannou, P. A. (2015), 'Traffic flow prediction for road transportation networks with limited traffic data', *IEEE Transactions on Intelligent Transportation Systems* **16**(2), 653–662.

Abdel-Aty, M. A., Kitamura, R. and Jovanis, P. P. (1997), 'Using stated preference data for studying the effect of advanced traffic information on drivers' route choice', *Transportation Research Part C: Emerging Technologies* **5**(1), 39–50.

Arteaga, F. and Ferrer, A. (2002), 'Dealing with missing data in mspc: several methods, different interpretations, some examples', *Journal of chemometrics* **16**(8-10), 408–418.

Asif, M. T., Mitrovic, N., Dauwels, J. and Jaillet, P. (2016), 'Matrix and tensor based methods for missing data estimation in large traffic networks', *IEEE Transactions on Intelligent Transportation Systems* **17**(7), 1816–1825.

Bie, Y., Wang, X. and Qiu, T. Z. (2016), 'Online method to impute missing loop detector data for urban freeway traffic control', *Transportation Research Record: Journal of the Transportation Research Board* (2593), 37–46.

Breiman, L. (2001), 'Random forests', *Machine learning* **45**(1), 5–32.

Chen, C., Kwon, J., Rice, J., Skabardonis, A. and Varaiya, P. (2003), 'Detecting errors and imputing missing data for single-loop surveillance systems', *Transportation Research Record: Journal of the Transportation Research Board* (1855), 160–167.

Chen, C., Wang, Y., Li, L., Hu, J. and Zhang, Z. (2012), 'The retrieval of intra-day trend and its influence on traffic prediction', *Transportation research part C: emerging technologies* **22**, 103–118.

Chen, H., Grant-Muller, S., Mussone, L. and Montgomery, F. (2001), 'A study of hybrid neural network approaches and the effects of missing data on traffic forecasting', *Neural Computing & Applications* **10**(3), 277–286.

Chiou, J.-M., Zhang, Y.-C., Chen, W.-H. and Chang, C.-W. (2014), 'A functional data approach to missing value imputation and outlier detection for traffic flow data', *Transportmetrica B: Transport Dynamics* **2**(2), 106–129.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the royal statistical society. Series B (methodological)* pp. 1–38.

Duan, Y., Lv, Y., Liu, Y.-L. and Wang, F.-Y. (2016), 'An efficient realization of deep learning for traffic data imputation', *Transportation research part C: emerging technologies* **72**, 168–181.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X. et al. (1996), A density-based algorithm for discovering clusters in large spatial databases with noise., *in* 'Kdd', Vol. 96, pp. 226–231.

Falge, E., Baldocchi, D., Olson, R., Anthoni, P., Aubinet, M., Bernhofer, C., Burba, G., Ceulemans, R., Clement, R., Dolman, H. et al. (2001), 'Gap filling strategies for long term energy flux data sets', *Agricultural and Forest Meteorology* **107**(1), 71–77.

Frey, B. J. and Dueck, D. (2007), 'Clustering by passing messages between data points', *science* **315**(5814), 972–976.

Haworth, J. and Cheng, T. (2012), 'Non-parametric regression for space–time forecasting under missing data', *Computers, Environment and Urban Systems* **36**(6), 538–550.

Henrickson, K., Zou, Y. and Wang, Y. (2015), 'Flexible and robust method for missing loop detector data imputation', *Transportation Research Record: Journal of the Transportation Research Board* (2527), 29–36.

Hildebrand, F. B. (1987), *Introduction to numerical analysis*, Courier Corporation.

Ho, T. K. (1995), Random decision forest, *in* 'Proceedings of the 3rd International Conference on Document Analysis and Recognition', pp. 278–282.

Huang, G.-B., Zhu, Q.-Y. and Siew, C.-K. (2004), Extreme learning machine: a new learning scheme of feedforward neural networks, *in* 'Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on', Vol. 2, IEEE, pp. 985–990.

Kondrashov, D. and Ghil, M. (2006), 'Spatio-temporal filling of missing points in geophysical data sets', *Nonlinear Processes in Geophysics* **13**(2), 151–159.

Ku, W. C., Jagadeesh, G. R., Prakash, A. and Srikanthan, T. (2016), A clustering-based approach for data-driven imputation of missing traffic data, *in* 'Integrated and Sustainable Transportation Systems (FISTS), 2016 IEEE Forum on', IEEE, pp. 1–6.

Laña, I., Del Ser, J. and Olabarrieta, I. I. (2016), Understanding daily mobility patterns in urban road networks using traffic flow analytics, *in* 'Network Operations and Management Symposium (NOMS), 2016 IEEE/IFIP', IEEE, pp. 1157–1162.

Laña, I., Del Ser, J. and Vélez, M. (2017), A novel fireworks algorithm with wind inertia dynamics and its application to traffic forecasting, *in* 'Evolutionary Computation (CEC), 2017 IEEE Congress on', IEEE, pp. 706–713.

Laña, I., Del Ser, J., Vélez, M. and Oregi, I. (2017), Joint feature selection and parameter tuning for short-term traffic flow forecasting based on heuristically optimized multi-layer neural networks, *in* 'International Conference on Harmony Search Algorithm', Springer, pp. 91–100.

Li, L., Li, Y. and Li, Z. (2013), 'Efficient missing data imputing for traffic flow by considering temporal and spatial dependence', *Transportation research part C: emerging technologies* **34**, 108–120.

Li, L., Su, X., Zhang, Y., Hu, J. and Li, Z. (2014), Traffic prediction, data compression, abnormal data detection and missing data imputation: An integrated study based on the decomposition of traffic time series, *in* 'Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on', IEEE, pp. 282–289.

Li, Y., Li, Z. and Li, L. (2014), 'Missing traffic data: comparison of imputation methods', *IET Intelligent Transport Systems* **8**(1), 51–57.

Little, R. J. and Rubin, D. B. (2014), *Statistical analysis with missing data*, John Wiley & Sons.

Lv, Y., Duan, Y., Kang, W., Li, Z. and Wang, F.-Y. (2015), 'Traffic flow prediction with big data: a deep learning approach', *IEEE Transactions on Intelligent Transportation Systems* **16**(2), 865–873.

Madrid Open Data Portal (n.d.), `http://datos.madrid.es`. Accessed: 2017-03-31.

Moffat, A. M., Papale, D., Reichstein, M., Hollinger, D. Y., Richardson, A. D., Barr, A. G., Beckstein, C., Braswell, B. H., Churkina, G., Desai, A. R. et al. (2007), 'Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes', *Agricultural and Forest Meteorology* **147**(3), 209–232.

Qu, L., Li, L., Zhang, Y. and Hu, J. (2009), 'Ppca-based missing data imputation for traffic flow volume: A systematical approach', *IEEE Transactions on intelligent transportation systems* **10**(3), 512–522.

Ran, B., Tan, H., Wu, Y. and Jin, P. J. (2016), 'Tensor based missing traffic data completion with spatial–temporal correlation', *Physica A: Statistical Mechanics and its Applications* **446**, 54–63.

Sainani, K. L. (2015), 'Dealing with missing data', *PM&R* **7**(9), 990–994.

Schafer, J. L. (1997), *Analysis of incomplete multivariate data*, CRC press.

Schimbinschi, F., Nguyen, X. V., Bailey, J., Leckie, C., Vu, H. and Kotagiri, R. (2015), Traffic forecasting in complex urban networks: Leveraging big data and machine learning, *in* 'Big Data (Big Data), 2015 IEEE International Conference on', IEEE, pp. 1019–1024.

Shrive, F. M., Stuart, H., Quan, H. and Ghali, W. A. (2006), 'Dealing with missing data in a multi-question depression scale: a comparison of imputation methods', *BMC medical research methodology* **6**(1), 57.

Smith, B., Scherer, W. and Conklin, J. (2003), 'Exploring imputation techniques for missing data in transportation management systems', *Transportation Research Record: Journal of the Transportation Research Board* (1836), 132–142.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M. and Carpenter, J. R. (2009), 'Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls', *Bmj* **338**, b2393.

Sun, S., Yu, G. and Zhang, C. (2004), Short-term traffic flow forecasting using sampling markov chain method with incomplete data, *in* 'Intelligent Vehicles Symposium, 2004 IEEE', IEEE, pp. 437–441.

Sun, S. and Zhang, C. (2007), 'The selective random subspace predictor for traffic flow forecasting', *IEEE Transactions on intelligent transportation systems* **8**(2), 367–373.

Sun, S., Zhang, C. and Yu, G. (2006), 'A bayesian network approach to traffic flow forecasting', *IEEE Transactions on intelligent transportation systems* **7**(1), 124–132.

Sun, S., Zhang, C., Yu, G., Lu, N. and Xiao, F. (2004), 'Bayesian network methods for traffic flow forecasting with incomplete data', *Machine Learning: ECML 2004* pp. 419–428.

Tan, H., Wu, Y., Cheng, B., Wang, W. and Ran, B. (2014), 'Robust missing traffic flow imputation considering nonnegativity and road capacity', *Mathematical Problems in Engineering* **2014**.

Tang, J., Zhang, G., Wang, Y., Wang, H. and Liu, F. (2015), 'A hybrid approach to integrate fuzzy c-means based imputation method with genetic algorithm for missing traffic volume data estimation', *Transportation Research Part C: Emerging Technologies* **51**, 29–40.

Treiber, M. and Helbing, D. (2002), 'Reconstructing the spatio-temporal traffic dynamics from stationary detector data', *Cooperative Transportation Dynamics* **1**(3), 3–1.

Van Hinsbergen, C., Van Lint, J. and Sanders, F. (2007), Short term traffic prediction models, *in* 'Proceedings of the 14th World Congress on Intelligent Transport Systems (ITS)'.

Van Lint, J., Hoogendoorn, S. and van Zuylen, H. J. (2005), 'Accurate freeway travel time prediction with state-space neural networks under missing data', *Transportation Research Part C: Emerging Technologies* **13**(5), 347–369.

Vlahogianni, E. I., Karlaftis, M. G. and Golias, J. C. (2007), 'Spatio-temporal short-term urban traffic volume forecasting using genetically optimized modular networks', *Computer-Aided Civil and Infrastructure Engineering* **22**(5), 317–325.

Vlahogianni, E. I., Karlaftis, M. G. and Golias, J. C. (2014), 'Short-term traffic forecasting: Where we are and where we're going', *Transportation Research Part C: Emerging Technologies* **43**, 3–19.

Whitlock, M. E. and Queen, C. M. (2000), 'Modelling a traffic network with missing data', *Journal of Forecasting* **19**(7), 561–574.

Zhong, M., Lingras, P. and Sharma, S. (2004), 'Estimation of missing traffic counts using factor, genetic, neural, and regression techniques', *Transportation Research Part C: Emerging Technologies* **12**(2), 139–166.

Zhong, M., Sharma, S. and Lingras, P. (2004), 'Genetically designed models for accurate imputation of missing traffic counts', *Transportation Research Record: Journal of the Transportation Research Board* (1879), 71–79.

Zhong, M., Sharma, S. and Lingras, P. (2006), 'Matching patterns for updating missing values of traffic counts', *Transportation Planning and Technology* **29**(2), 141–156.

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.

Figure 1: Automatic traffic recorders (ATRs) in the center of Madrid, colored by their data availability during 2014 (in % of valid 15-minute intervals over the year).

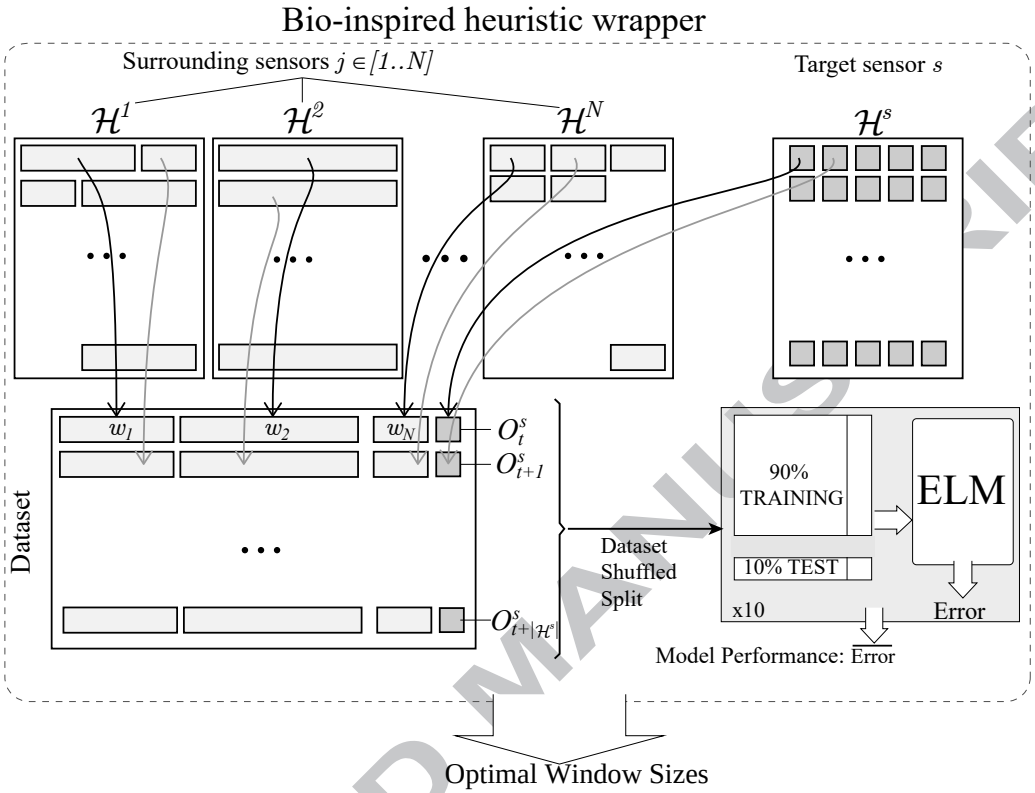

Figure 2: Training of a forecasting model through window-size optimization and Extreme Learning Machine (ELM) regression models. Optimized parameters of this model are used after to obtain predictions that act as imputed values.

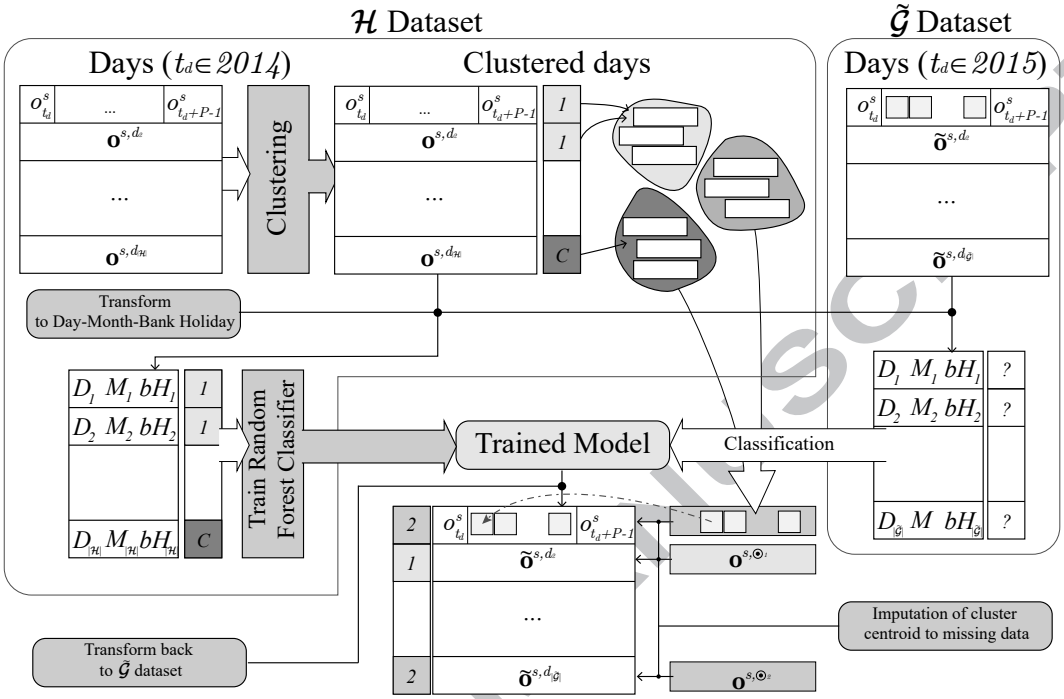

Figure 3: Clustering-Classification process.

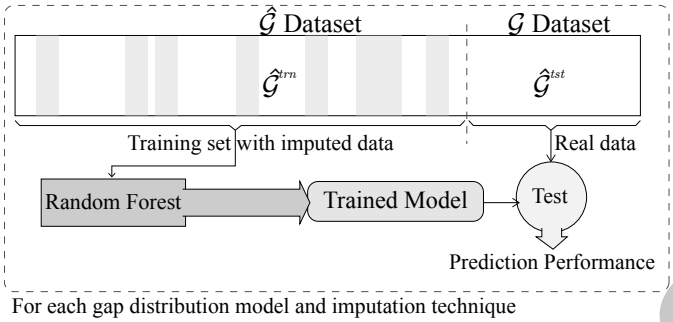

Figure 4: Proposed method for evaluating the performance and quality of the values imputed by every technique in a context of traffic forecasting.

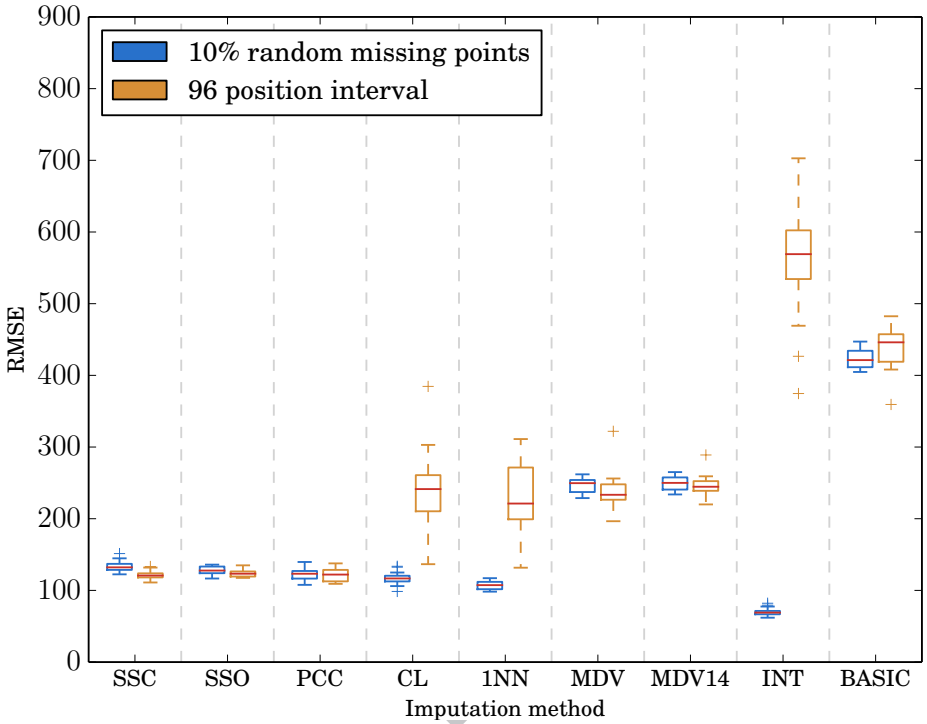

Figure 5: Boxplot of imputation performances for both gap generation models.

ID	Avg. flow	# Missing entries	Missing data	# intervals	$\langle L \rangle$	mode(L)
10006	420	4626	13.2%	27	135.70	96
10018	134	4576	13.1%	26	139.12	21
10023	143	4550	13.0%	25	143.83	96
10030	82	4713	13.5%	28	133.89	96
13026	131	4559	13.0%	26	138.44	96
13032	474	4575	13.1%	26	139.08	21
18018	332	4622	13.2%	27	135.54	96
19011	683	4791	13.7%	28	136.78	21
21007	204	4556	13.0%	26	138.32	96
90033	317	4640	13.2%	28	131.19	96
90034	236	4640	13.2%	28	131.19	96
90035	198	4639	13.2%	28	131.15	96
90041	233	4636	13.2%	28	131.04	96

Table 1: Analysis of *actual* missing data distribution. Column names stand for loop ID (as per the naming convention of the repository), yearly average flow of cars measured in vehicles/hour, total number of missing points, percentage of total missing data (considering 35040 samples), number of intervals grouping missing points, average gap length measured in samples $\langle L \rangle$, statistical mode of the length of the gaps L

Method	**1 %**	**10 %**	**25 %**	**50 %**	**80 %**	**100 %**
BASIC	420 ± 30	420 ± 10	438 ± 9	489 ± 5	586 ± 2	661
INT	$\mathbf{65 \pm 6}$	$\mathbf{70 \pm 5}$	$\mathbf{75 \pm 4}$	$\mathbf{92 \pm 5}$	175 ± 9	500
MDV	260 ± 30	250 ± 10	248 ± 6	250 ± 5	251 ± 3	661
MDV14	260 ± 30	250 ± 10	249 ± 6	249 ± 3	248 ± 1	249
1NN	110 ± 20	107 ± 6	107 ± 4	109 ± 3	$\mathbf{121 \pm 3}$	390
CL	110 ± 20	117 ± 8	117 ± 5	116 ± 2	$\mathbf{121 \pm 3}$	400
SSC	140 ± 20	133 ± 7	134 ± 4	133 ± 2	132 ± 1	132
SSO	130 ± 20	128 ± 6	127 ± 3	126 ± 2	126.6 ± 0.6	126
PCC	120 ± 20	122 ± 8	124 ± 5	122 ± 2	122 ± 1	**122**

Table 2: RMSE results for different percentages ξ of point-wise missing data. In this and following tables, results are shown as *mean*±*standard deviation*, and statistically best results (determined by a Wilcoxon test with 95% confidence interval) are highlighted in bold.

Method	**24**	**48**	**72**	**96**	**144**	**192**
BASIC	420 ± 40	410 ± 20	420 ± 30	440 ± 30	420 ± 40	420 ± 30
INT	300 ± 50	430 ± 50	470 ± 70	560 ± 80	490 ± 60	530 ± 70
MDV	260 ± 30	240 ± 30	240 ± 30	240 ± 20	260 ± 20	240 ± 20
MDV14	250 ± 20	240 ± 20	240 ± 20	250 ± 10	260 ± 20	250 ± 20
1NN	130 ± 20	130 ± 10	170 ± 30	230 ± 50	260 ± 50	310 ± 60
CL	120 ± 20	140 ± 10	160 ± 30	240 ± 60	270 ± 40	320 ± 50
SSC	130 ± 20	130 ± 20	$\mathbf{120 \pm 10}$	$\mathbf{121 \pm 6}$	122 ± 8	$\mathbf{120 \pm 8}$
SSO	123 ± 7	124 ± 4	124 ± 5	123 ± 5	123 ± 6	124 ± 7
PCC	$\mathbf{120 \pm 10}$	$\mathbf{120 \pm 10}$	$\mathbf{120 \pm 10}$	122 ± 9	$\mathbf{120 \pm 10}$	122 ± 9

Table 3: RMSE results for different length L intervals of missing data.

Method	**1 %**	**10 %**	**25 %**	**50 %**	**80 %**	**100 %**
BASIC	-0.009 ± 0.007	-0.018 ± 0.008	-0.1 ± 0.01	-0.38 ± 0.02	-0.98 ± 0.01	-1.52
INT	0.975 ± 0.006	0.972 ± 0.004	0.968 ± 0.003	0.951 ± 0.005	0.82 ± 0.02	-0.17
MDV	0.60 ± 0.09	0.65 ± 0.02	0.65 ± 0.02	0.64 ± 0.02	0.635 ± 0.009	-1.52
MDV14	0.61 ± 0.06	0.65 ± 0.02	0.64 ± 0.01	0.64 ± 0.009	0.644 ± 0.003	0.64
1NN	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.935 \pm 0.005}$	$\mathbf{0.934 \pm 0.004}$	$\mathbf{0.931 \pm 0.003}$	$\mathbf{0.916 \pm 0.005}$	0.11
CL	0.92 ± 0.03	0.923 ± 0.008	0.922 ± 0.006	0.922 ± 0.003	0.915 ± 0.004	0.10
SSC	0.89 ± 0.04	0.899 ± 0.009	0.896 ± 0.007	0.898 ± 0.004	0.899 ± 0.002	0.90
SSO	0.90 ± 0.03	0.91 ± 0.01	0.907 ± 0.005	0.908 ± 0.003	0.907 ± 0.001	$\mathbf{0.91}$
PCC	0.91 ± 0.04	0.915 ± 0.008	0.912 ± 0.006	0.914 ± 0.004	0.914 ± 0.002	$\mathbf{0.91}$

Table 4: R^2 results for different percentages ξ of point-wise missing data.

Method	24	48	72	96	144	192
BASIC	-0.03 ± 0.04	-0.03 ± 0.04	-0.04 ± 0.03	-0.04 ± 0.02	-0.03 ± 0.04	-0.03 ± 0.02
INT	0.4 ± 0.1	-0.2 ± 0.3	-0.4 ± 0.4	-0.7 ± 0.4	-0.4 ± 0.3	-0.7 ± 0.4
MDV	0.6 ± 0.1	0.6 ± 0.1	0.6 ± 0.1	0.7 ± 0.1	0.6 ± 0.1	0.64 ± 0.09
MDV14	0.61 ± 0.08	0.64 ± 0.07	0.64 ± 0.08	0.66 ± 0.08	0.60 ± 0.08	0.64 ± 0.06
1NN	0.90 ± 0.02	0.89 ± 0.02	0.82 ± 0.07	0.7 ± 0.1	0.6 ± 0.2	0.4 ± 0.2
CL	0.91 ± 0.02	0.89 ± 0.02	0.83 ± 0.07	0.7 ± 0.1	0.6 ± 0.1	0.4 ± 0.20
SSC	0.90 ± 0.03	0.88 ± 0.04	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.91 \pm 0.01}$
SSO	0.91 ± 0.02	0.90 ± 0.01	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.91 \pm 0.01}$
PCC	$\mathbf{0.92 \pm 0.02}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.91 \pm 0.01}$	$\mathbf{0.91 \pm 0.01}$

Table 5: R^2 results for different length L intervals of missing data.

Methods	**1%**	**10%**	**25%**	**50%**	**80%**	**100%**	**24**	**48**	**72**	**96**	**144**	**192**
SSC vs SSO	0.09	0.02	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.41	0.1	0.05	0.11	0.63	0.09

Table 6: Wilcoxon test p-values showing statistical significance of differences between RMSE results of SSC and SSO.

Methods	**1%**	**10%**	**25%**	**50%**	**80%**	**100%**	**24**	**48**	**72**	**96**	**144**	**192**
SSC vs PCC	0.01	$< 10^{-3}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.07	0.09	0.15	0.58	0.82	0.28

Table 7: Wilcoxon test p-values showing statistical significance of differences between RMSE results of SSC and PCC.

Method	**1 %**	**10 %**	**25 %**	**50 %**	**80 %**	**100 %**
BASIC	80 ± 2	103 ± 4	154 ± 10	240 ± 20	350 ± 20	618
INT	$\mathbf{77 \pm 2}$	$\mathbf{78 \pm 2}$	$\mathbf{76 \pm 2}$	$\mathbf{78 \pm 1}$	85 ± 1	400
MDV	81 ± 3	91 ± 4	107 ± 4	123 ± 10	130 ± 10	618
MDV14	82 ± 5	92 ± 3	107 ± 4	128 ± 7	135 ± 6	171
1NN	78 ± 2	82 ± 4	85 ± 4	87 ± 3	$\mathbf{83 \pm 3}$	**76**
CL	78 ± 2	81 ± 5	81 ± 3	84 ± 3	86 ± 3	95
SSC	80 ± 3	86 ± 4	91 ± 4	95 ± 3	97 ± 3	89
SSO	79 ± 2	81 ± 3	85 ± 3	87 ± 3	87 ± 2	86
PCC	78 ± 2	79 ± 4	81 ± 3	86 ± 4	94 ± 8	89

Table 8: RMSE prediction results for different percentages of point-wise missing data.

Method	24	48	72	96	144	192
BASIC	78 ± 2	78 ± 2	78 ± 2	77 ± 2	76 ± 2	78 ± 3
INT	75 ± 2	75 ± 2	76 ± 3	77 ± 2	77 ± 2	77 ± 2
MDV	76 ± 2	75 ± 1	75 ± 2	75.0 ± 0.8	76 ± 1	76 ± 1
MDV14	$\mathbf{74 \pm 1}$	$\mathbf{74 \pm 1}$	74 ± 1	$\mathbf{74 \pm 1}$	$\mathbf{73.7 \pm 0.8}$	$\mathbf{74 \pm 1}$
1NN	79 ± 3	79 ± 2	76 ± 2	76 ± 2	76 ± 2	76 ± 3
CL	77 ± 1	76 ± 2	$\mathbf{73 \pm 2}$	74 ± 2	74 ± 2	73 ± 2
SSC	76 ± 2	77 ± 1	77 ± 1	77 ± 1	77 ± 1	76 ± 1
SSO	78 ± 2	78 ± 3	78 ± 2	78 ± 2	78 ± 2	77 ± 2
PCC	78 ± 2	78 ± 2	78 ± 2	78 ± 2	78 ± 2	78 ± 2

Table 9: RMSE prediction results for different length intervals of missing data.

FIGURE CAPTIONS

Figure 1 Automatic traffic recorders (ATRs) in the center of Madrid, colored by their data availability during 2014 (in % of valid 15-minute intervals over the year).

Figure 2 Training of a forecasting model through window-size optimization and Extreme Learning Machine (ELM) regression models. Optimized parameters of this model are used after to obtain predictions that act as imputed values.

Figure 3 Clustering-Classification process.

Figure 4 Proposed method for evaluating the performance and quality of the values imputed by every technique in a context of traffic forecasting.

Figure 5 Boxplot of imputation performances for both gap generation models.

TABLE CAPTIONS

Table 1 Analysis of *actual* missing data distribution. Column names stand for loop ID (as per the naming convention of the repository), yearly average flow of cars measured in vehicles/hour, total number of missing points, percentage of total missing data (considering 35040 samples), number of intervals grouping missing points, average gap length measured in samples $\langle L \rangle$, statistical mode of the length of the gaps L

Table 2 RMSE results for different percentages ξ of point-wise missing data. In this and following tables, results are shown as *mean±standard deviation*, and statistically best results (determined by a Wilcoxon test with 95% confidence interval) are highlighted in bold.

Table 3 RMSE results for different length L intervals of missing data.

Table 4 R^2 results for different percentages ξ of point-wise missing data.

Table 5 R^2 results for different length L intervals of missing data.

Table 6 Wilcoxon test p-values showing statistical significance of differences between RMSE results of SSC and SSO.

Table 7 Wilcoxon test p-values showing statistical significance of differences between RMSE results of SSC and PCC.

Table 8 RMSE prediction results for different percentages of point-wise missing data

Table 9 RMSE prediction results for different length intervals of missing data.

KEYWORDS

- Traffic forecasting;
- missing data;
- cluster analysis;
- data imputation.

Highlights

Manuscript ID TRC-D-17-00868 - Review

- Review of the techniques for generating artificial missing data.
- Impact analysis of missing data imputing methods on forecasting models performance.
- Two novel imputing methods to tackle long periods of missing data.