ARCHIVES OF TRANSPORT *ISSN (print):* *0866-9546*
Volume 54, Issue 2, 2020 ***e-ISSN (online):*** *2300-8830*
DOI: *10.5604/01.3001.0014.2968*

AN ADAPTIVE K NEAREST NEIGHBOUR METHOD FOR IMPUTATION OF MISSING TRAFFIC DATA BASED ON TWO SIMILARITY METRICS

Yang WANG[1], Yu XIAO[2], Jianhui LAI[3], Yanyan CHEN[4]

[1, 2, 3] *Beijing Engineering Research Centre of Urban Transport Operation Guarantee, Beijing University of Technology, Beijing, China*

[4] *Beijing Key Laboratory of Traffic Engineering, Beijing University of Technology, Beijing, China*

Abstract:

Traffic flow is one of the fundamental parameters for traffic analysis and planning. With the rapid development of intelligent transportation systems, a large number of various detectors have been deployed in urban roads and, consequently, huge amount of data relating to the traffic flow are accumulatively available now. However, the traffic flow data detected through various detectors are often degraded due to the presence of a number of missing data, which can even lead to erroneous analysis and decision if no appropriate process is carried out. To remedy this issue, great research efforts have been made and subsequently various imputation techniques have been successively proposed in recent years, among which the k nearest neighbour algorithm (kNN) has received a great popularity as it is easy to implement and impute the missing data effectively. In the work presented in this paper, we firstly analyse the stochastic effect of traffic flow, to which the suffering of the kNN algorithm can be attributed. This motivates us to make an improvement, while eliminating the requirement to predefine parameters. Such a parameter-free algorithm has been realized by introducing a new similarity metric which is combined with the conventional metric so as to avoid the parameter setting, which is often determined with the requirement of adequate domain knowledge. Unlike the conventional version of the kNN algorithm, the proposed algorithm employs the multivariate linear regression model to estimate the weights for the final output, based on a set of data, which is smoothed by a Wavelet technique. A series of experiments have been performed, based on a set of traffic flow data reported from serval different countries, to examine the adaptive determination of parameters and the smoothing effect. Additional experiments have been conducted to evaluate the competent performance for the proposed algorithm by comparing to a number of widely-used imputing algorithms.

Keywords: *missing traffic data, similarity metrics, k nearest neighbour method, stochastic characteristics*

To cite this article:

Wang, Y., Xiao, Y., Lai, J., Chen, Y., 2020. An adaptive k nearest neighbour method for imputation of missing traffic data based on two similarity. Archives of Transport, 54(2), 59-73. DOI: https://doi.org/10.5604/01.3001.0014.2968

Contact:

1) wang_yang@bjut.edu.cn [https://orcid.org/0000-0003-4507-951X], 2) xiaoy@emails.bjut.edu.cn, 3) laijianhui@bjut.edu.cn [https://orcid.org/0000-0003-3189-8872] -corresponding author, 4) cdyan@bjut.edu.cn [https://orcid.org/0000-0002-7068-4669]

1. Introduction

Perceiving traffic flow parameters through detectors facilitate an accurate estimate of traffic state, which can be used at various aspects, such as dynamic routing or signal control. Therefore, investment has intensively put into the construction of traffic detectors in recent years. However, in the data collection process, it is inevitable that various unpredictably malfunctions, such as communication interruption, power outages or storage equipment damage, can occur even for the advanced detectors, resulting in a number of data missing. For example, the solar-powered sensors, which are recently promoted in China, cannot function properly sometimes during the period of rainy season. On the other hand, the data transmission problems can be interrupted for short time when an update is performed for the communication system (e.g., from 3rd generation to 4th generation).

The missing data problem and associated effects have been constantly reported in recent years. Missing data problem, where some subsets of traffic data become missing, has greatly hindered the collection and subsequent analysis, estimation and prediction of traffic flow data (Wang and Mao, 2019). In Texas Transportation Institute, the rate of missing data is between 16% and 93% (Li, Zhang, Wang, et al., 2019). Tan et al. found more than 5% are missing from the PeMS traffic flow database (Tan, Feng, Feng et al., 2013). Similarly, the missing ratio of the traffic data reported in (Xu, Li and Shi, 2010) can be as high as 90%, with the average ratio of 50% for the period of 7 years. An analysis on the traffic flow data obtained from the microwave sensors amounted on the Beijing ring expressways reports up to 50% of data missing (Ma, Luan, Du, et al., 2017). It is no doubt that the missing data build an intangible barrier for understanding and modelling the traffic phenomenon due to incomplete information.

To remedy the undesired effects caused by the presence of missing data, a number of imputing approaches have been proposed in last few decades. These imputation methods have taken different procedures to provide plausible estimations for those missing values given other observed data, based on the assumption that the real traffic system usually possess inherent structure due to its flow dynamic characteristics. Although these existing approaches can be classified from various aspects (e.g., the angle of data construction (Tan, Feng, Feng et al., 2013)), the mechanism adopted by each method to infer the values for missing data mostly determines the performance of imputation and application scope. Therefore, we have grouped the existing imputing approaches into three categories, prediction-based, interpolation-based and statistical learning-based imputing approaches. However, it should be aware that there is a possibility for a few approaches which can fit into the different categories from different view of point.

1.1. Prediction-based imputing approaches

The approaches which fall in this category usually take advantage of sample data to build a model to account for the mapping relationship between historical data and future data and the model is subsequently used to predict the values for the missing data. Typical examples of this category include Bayesian networks (BNs) (Zhang, Sun, and Yu, 2004; Ghosh, Basu, and O'Mahony, 2007), auto-regressive integrated moving average (ARIMA) (Zhong, Sharma, and Lingras, 2004), Adaptive Spatial-Temporal Correlations (Wang, Zhang, Piao, et al., 2019), feed-forward neural network (Vlahogianni, Karlaftis, and Golias, 2005), Convolutional Neural Network (Zhuang, Ke and Wang, 2019), spatio-temporal cokriging (Bae, Kim, Lim, et al., 2018), Fuzzy C-Means Method (Tang, Wang, Zhang, et al., 2015), and support vector regression (Castro-Neto, Jeong, Jeong et al., 2009). In addition, the regression techniques, such as linear or polynomial regression, can also be used to estimate the values for the missing data. Although these approaches can even model the highly non-linear relationship between independent and dependent variables for the prediction purpose, the use of continuous chunks in the time series can significantly degrade the imputation performance due to inefficient use of all observed data.

1.2. Interpolation-based imputing approaches

The simplest interpolation method is the nearest-neighbour interpolation, which uses the value at the previous time instant nearest to the instant of a missing data or at the same time instant of the previous day (Chen, and Shao, 2000). This imputing approach is often a favourable choice if the real time is one of the rigid requirements. A slightly complicated approach is to fill the missing data by averaging the observed data which are close to the missing data in

time generally. More complicated interpolation approaches include linear, spline, and polynomial (e.g., Lagrange's interpolation formula (Stoeck, and Prajwowski, 2010)), etc. The kind of imputing approaches can be further divided into two groups, interpolation and extrapolation. Although the phrase, 'interpolation', is generally used to refer these two situations, there is a difference between interpolation and extrapolation according to whether the missing data are in between or outside of the observed data. As these approaches force the interpolating function passes exactly through the given data points, spurious features in the region of missing data may be generated.

1.3. Statistical learning-based imputing approaches

One of representative approaches in this category is k Nearest neighbours (kNN) (Sliva, H. D., and Perera, A. S., 2017; Esawey, M. E., and Sayed, T., 2012) which attempts to derive the missing data from a number of similar patterns. Local least squares (LLS) (Kim, Golub, and Park, 2005) also takes advantage of the information from similar patterns to infer the possible values for missing data. The principal component analysis (PCA) based methods, such as probabilistic principal component analysis (PPCA) and Bayesian principal component analysis (BPCA), extract the statistical characteristics of the observed data and map the relationship between the observed data and latent variables by constructing a number of principle components (Qu, Li, Zhang et al., 2009; Li, Li, and Li, 2014). PPCA employs the expectation-maximization algorithm to determine the projection matrix for the later variables, but BPCA applies the Bayesian estimation approach. As these approaches make use of both global and local information of the observed data, better performance than the other conventional imputation methods (e.g., the nearest-neighbour interpolation, the spline interpolation, and the mean historical methods) has been reported for the traffic flow data. However, it has been shown that there is no significant difference with regards to the imputation accuracy between the different PCA-based approaches (Li, Li and Li, 2013). In recent years, a number of tensor-based approaches have been proposed to impute the missing data by taking advantage of traffic spatial–temporal information (Chen, He and Sun, 2019). In theory, more correlation information considered in the imputation process can generally produce more accurate estimation for the missing data. However, numerous detectors are deployed outside urban areas and most of them are sparsely spaced, resulting in a very weak spatial correlation.

Without loss of generality, the work presented in this paper ignores the spatial correlation information in an attempt to deliver an imputation method which can also applied to the rural areas where the spatial correlation is generally weak. Based on the traffic temporal information, we propose a novel imputation approach based on the kNN method which has been reported to be one of competent imputation approaches in terms of accuracy and efficiency (Loukopoulos, Sampath, Pilidis, et al., 2016). The motivation for the improvements of the kNN approach is presented in the next section. The proposed imputation approach is introduced in Section 3. Section 4 presents the experiments and the corresponding results. The concluding remarks are provided in Section 5.

2. Motivation

In this section, the imputation problem of traffic flow data considered in this paper is firstly formulated, followed by a short introduction to the kNN algorithm. Based on the analysis of the stochastic characteristic of traffic flow, the difficulties to apply the kNN algorithm for traffic flow imputation are discussed, which motivates us to make an improvement.

2.1. The traffic flow imputation problem

As stated above, the spatial information is not considered in the imputation process in this paper and, consequently, the imputation problem can be modelled as follows. Suppose that we have a set of data **X** for D days. That is:

$$\mathbf{X}_D = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_D]^{\mathrm{T}} \tag{1}$$

in which the *d*th day's traffic flow data is a vector, whose elements are sequentially arranged in time order, as follows:

$$\mathbf{x}_D = [x_d^1, x_d^2, \dots, x_d^M]^{\mathrm{T}};\ \ d = 1, 2, \dots, D \tag{2}$$

Note that *M* is the number of recordings of traffic flow and determined based on the recording frequency adopted in the detector system.

The last column in the data set **X** contains a number of missing data, whose values are required to estimate. In this paper, it is assumed that values in the data set are missing at random, which means the missing data are independent to each other.

Note that it is not necessary that the set of data **X** contains the traffic flow recordings from consecutive days. In order to take advantage of the observed data information, the data set may reject the data containing missing values in the previous days, but instead, include the fully observed data even in earlier days. Furthermore, it is natural to exclude the data which contain the different traffic patterns due to obvious reasons (e.g., the traffic patterns for weekdays and weekends are generally different).

2.2. kNN Algorithm

As a non-parametric method, the kNN algorithm has enjoyed great popularity in many classification and regression applications due to its simplicity. The idea of the kNN algorithm for the missing data imputation is based on the assumption of local similarity in data space: the missing data embedded in a time series are expected to have the similar values, as those observed in the same time instants but contained in the historical time series if the counterparts in those time series are similar.

Figure 1 illustrates the imputation process for a number of missing data by using the kNN algorithm. Suppose we have a set of data, which has been arranged as a matrix **X** with each row and column denoting the time of day and day, respectively. Due to the fact that the traffic data are normally sensed at the fixed time instant with a constant interval, we simply use the index to indicate the time instant and day, respectively, when the data are collected. The last column vector, $\mathbf{x}_D$, contains a number of missing data, each of which is denoted by a question mark in Figure 1. For convenience, we separate the missing data from the observed data and the corresponding parts are denoted by $\mathbf{x}_D^{\text{miss}}$ and $\mathbf{x}_D^{\text{obse}}$, respectively. That is:

$$\mathbf{x}_D = \left[\mathbf{x}_D^{\text{miss}}, \mathbf{x}_D^{\text{obse}}\right]^{\text{T}} \tag{3}$$

The first step is to find k nearest neighbours for the observed part $\mathbf{x}_D^{\text{obse}}$, which can be mathematically expressed as:

$$\mathbf{I} = \underset{\mathbf{x}_d^{\text{obse}} \in \mathbf{H}}{\operatorname{argmin}} \sum_{i=1}^{k} s\left(\mathbf{x}_d^{\text{obse}}, \mathbf{x}_D^{\text{obse}}\right) \tag{4}$$

where **H** contains the historical data $\mathbf{x}_d^{\text{obse}}$ (d = 1, 2, …, D-1) and s is a function to estimate the distance for $\mathbf{x}_d^{\text{obse}}$ and $\mathbf{x}_D^{\text{obse}}$. A number of metrics, such as Manhattan, Chebychev, Levenshtein (Abbasifard, Ghahremani, and Naderi, 2014), have been proposed in literature. However, the Euclidean distance has been widely chosen as the similarity metric since a large number of problems can be defined in the Euclidean space.

The number of neighbours, k, is usually specified by users based on their experiences or trail-and-error. As shown in Figure 1, the column vectors marked with green colour are the k neighbours nearest to the observed part of the last column. Subsequently, the weight of ith neighbour can be obtained by normalizing the distance s_i over the summation of the distances of the k neighbours, as follows:

$$w_i = \frac{s_i}{\sum_{j=1}^{k} s_j}; \quad i = 1, 2, \ldots, k \tag{5}$$

Finally, the missing value x^m at mth time instant can be estimated as:

$$x^m = \sum_{i=1}^{k} w_i x_i^m \tag{6}$$

For easy reference in the sequel, we summarize the primary steps to perform the kNN algorithm in Figure 2. Although the implementation of the kNN algorithm is straightforward, a number of elements, such as the parameter k and similarity metric, are critical to the success of imputation of missing data. To determine these critical elements, domain knowledge is frequently employed, thus resulting in a number of variants of the kNN algorithm. An excellent review on the variants of the kNN algorithm can be found in the works (Bhatia, and Vandana, 2010). However, to the best of our knowledge, there are few attempts to consider the traffic stochastic characteristic to improve the imputation accuracy of the kNN algorithm. The next sub-section presents an analysis of the traffic stochastic characteristic with a discussion of possible difficulties that may arise from the stochastic characteristic.

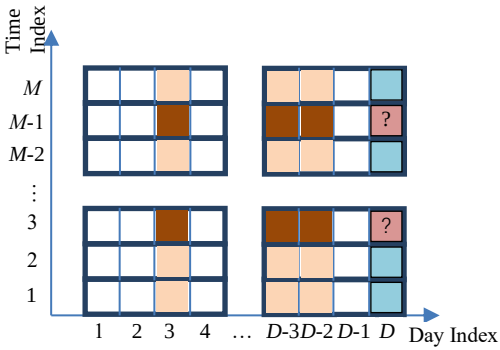


Fig. 1. Illustration of the kNN algorithm used to impute the data

2.3. The Effect of Stochastic Characteristics

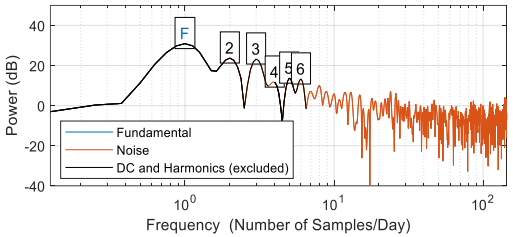
Traffic flow often exhibits a strong stochastic characteristic, which can be attributed to various factors such as stochastic travel demands or drivers' bounded rationality. To examine the stochastic characteristic of traffic flow, the signal-to-noise ratios (SNRs) have been estimated for three sets of traffic flows collected in the Portland metropolitan area, the freeways in California, and the sub-urban area of Beijing. For easy reference, the three data sets used in the following experiments are denoted as 'Portal' (i.e., the Portland Oregon Regional Transportation Archive Listing (http://portal.its.pdx.edu, 2018)), 'PeMS' (i.e., the freeway Performance Measurement System (http://pems.dot.ca.gov, 2017)), and 'Beijing' (i.e., a provincial road in the sub-urban area of Beijing). The aggregated 5-minute flow data of 5 weekdays in the three sets were used to examine the stochastic characteristics of traffic flow.

Figure 3 shows the power spectrums of the traffic flows with the SNR values obtained for the three data sets. From the results presented in Figure 3, it is evident that the traffic flows have a significant stochastic characteristic. It should be aware that the term "noise" is used here to reflect the stochastic of traffic flow, but it does not mean the noise that reflects nothing about the intrinsic characteristics of traffic flow. Due to the stochastic characteristic of traffic flow, k nearest neighbours, determined by the kNN algorithm with the similarity metric of Euclidean distance, can vary with different data missing. To examine such sensitivity, the three data sets were used with the missing rate ranging from 0.05 to 0.9 with the interval of 0.05. For each missing rate, 100 independent runs were performed and the missing data were randomly selected for each run. With the parameter k set to be 4 for all tests, the order of k nearest neighbours determined for each run was recorded.

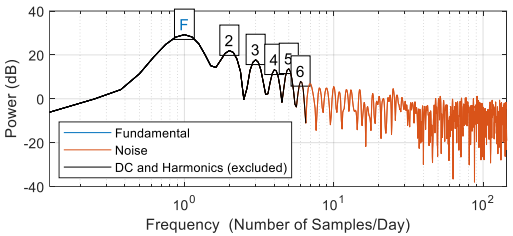
To reflect the fluctuation of the order of k nearest neighbours, we define a ratio, named as Difference Ratio, as the number of the unique order of k nearest neighbours over 100 independent runs. The results obtained for the three data set are illustrated in Figure 4. From Figure 4, it can be seen that the uncertainty in the determination of k nearest neighbours increases with missing rates for all test data. Such uncertainty resulted from the stochastic characteristic of traffic flow can obviously affect the imputation performance. To remedy this issue, we propose a mild solution, which is described in Section 3, instead of striving hard for the exact answer.

Specify the parameter *k* and similarity metric;
Prepare the data set to include the historical time series;
For d = 1, 2, …, D-1,
  Calculate the distance of the observed data between *d*th and *D*th day;
End
Find *k* nearest neighbours based on the distances;
Estimate the weight for each neighbouring time series;
Impute the missing values by weighting the corresponding data of the k nearest neighbours.

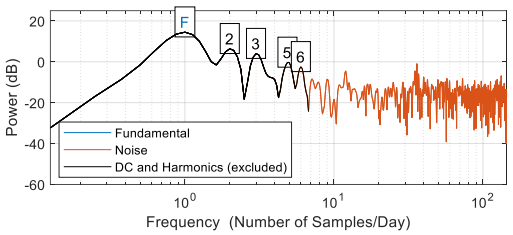
Fig. 2. The structure of the kNN algorithm



(a) Portal data set (SNR = 8.35 dB)



(b) PeMS data set (SNR = 8.54 dB)



(c) Beijing data set (SNR = 3.53 dB)

Fig. 3. The power spectrums of the traffic flows for the three data sets

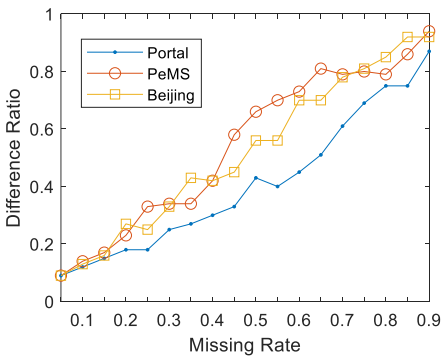


Fig 4. Difference ratios (the number of the unique order of k nearest neighbours over 100 independent runs) for each missing rate for the three data sets

3. Methodology

This section begins with an introduction of the structure of the improved version of kNN, followed by an explanation for the modified implementation.

As shown in Figure 5, the algorithm starts with the preparation of data set which includes the traffic flow with missing data and the historical data without missing values. For conveniences, we use "current" and "historical" data in the sequel to refer the flow data of a day with and without missing values. Obviously, the current data containing the missing values are required to impute. Next, the similarity is measured for each historical time series data with respect to the current flow data, based on two metrics, namely interweaving degree and Euclidean distance, which are introduced in the following sub-section. According to the similarities estimated, the historical flow data are classified into different groups, which can be used to identify k nearest neighbors. Subsequently, the k nearest neighbors are subject to a smoothing process, before those data are employed to build a regression model. Finally, the values for the missing data can be obtained by applying the regression model constructed.

Prepare the data set to include the historical time series;
For d = 1, 2, …, D-1,
  Measure the similarity for the observed data between ith and Dth day based on two metrics (i.e., interweaving degree and Euclidean distance)
End
Cluster the historical time series into different groups for each metric;
Identify k nearest neighbours based on the two metrics;
Smooth the k nearest neighbours;
Construct a multivariate linear model based on the smoothed data;
Apply the model to impute the missing data

Fig. 5. The structure of the improved kNN algorithm

4. The Identification of k Nearest Neighbours

As explained in Section 2, the stochastic characteristic of traffic flow makes it difficult to correctly identify the nearest neighbours when using the conventional similarity metrics. This difficulty is stemmed from the fact that the random fluctuation of traffic flow makes the nearest neighbours indistinguishable. Instead of strenuously discriminating the neighbouring ranks for those that are highly interwove, we decide to classify those indistinguishable neighbours into one group and the remaining into

another group. For convenience, those two groups are called "neighbouring group" and "remote group", respectively, in this paper. As shown in Figure 5, two similarity metrics, namely interweaving degree and Euclidean distance, are employed here to separate the historical time series data into the corresponding groups. Consequently, four groups, denoted as "NN", "NR", "RN", and "RR", can be formed by combining neighbouring and remote groups for the two metrics (e.g., "NR" representing neighbouring group for the first metric, interweaving degree, and remote group for the second, Euclidean distance). The elements in the "NN" group are selected as the k nearest neighbours to be used in the subsequent process. In such way, the parameter, k, can be adaptively determined without the need to be predefined by users. Note that it is possible that empty set can be obtained for "NN" group and, if so, the historical data in the neighbouring group with the metric of Euclidean distance will be used as nearest neighbours.

The interweaving degree proposed in this paper is defined as follows:

$$r = \frac{N_c}{N} \tag{7}$$

where *N* and *Nc* denote the number of points of a time series and the number of crossing points between two time series, respectively. After linearly interpolating each successive pair of data points for a time series, it is straightforward to determine the crossing points for each pair of line segments.

After obtaining the interweaving degrees, a simple clustering process is performed by classifying each neighbouring series with respect to the two centres, which are defined as the two extreme values of the interweaving degrees calculated, and it will be classified into the neighbouring group if its interweaving degree is closer to the largest interweaving degree (i.e., the centre of neighbouring group). The main reason to adopt such simple classification is to reduce possible computation overhead. The same procedure is also employed to classify the neighbouring series for the metric of Euclidean distance.

5. The Smoothing Process

The indistinguishable neighbours determined previously are all subject to a smoothing process to eliminate the random effect on the following process. While a wide spectrum of smoothing techniques, such as moving average filter (Arce, 2005), Butterworth filter (De Boor, 2001), and smoothing spline (Bianchi, and Sorrentino, 2007) etc, has been reported in literature, the smoothing technique based on wavelet transform (Misiti, and Misiti, 2007) is employed here to eliminate the random effect of traffic flow as it is able to localize the characteristics in the temporal and frequency domains by the hierarchically organized decompositions (Chui, 1992). Except for the theoretical soundness of the wavelet based smoothing technique, the decomposition level in practice is one of the critical factors, which has significant impact on the smoothing performance (El-Dahshan, 2011). If the smoothing process performs successfully, the part removed from the original time series data should manifest the stochastic process which can be modelled as a white noise. According to the Wiener–Khinchin theorem (Zbilut, and Marwan, 2008), the corresponding correlation function $A_N(\tau)$ can be derived as follows:

$$A_N(\tau) = \frac{C_0}{4\pi}\int_{-\infty}^{\infty} e^{jw\tau}dw = \frac{1}{2}C_0\delta(\tau) \tag{8}$$

where w is the frequency, τ is the time shift, C_0 is a constant and $\delta(\tau) = 1$ for $\tau = 0$, and 0, otherwise.

Then, the autocorrelation coefficient is:

$$a_N(\tau) = \frac{A_N(\tau)}{A_0(\tau)} \tag{9}$$

Ideally, the part removed after smoothing process should be a random noise-like time series and its corresponding coefficient of autocorrelation should be either 1 or 0 for τ=0 or τ>0. However, it is unlikely to have such pure noise-like time series by the smoothing operation in general. Therefore, the approximate 95% confidence interval for a noise-like time series is also computed with the sample autocorrelation. With the dependence structure for a set of time lags, the number of coefficients that fall within the confidence bounds are counted. That is, the more the coefficients fall into the confidence bounds, the more likely it is a random noise. Such calculation is performed for each decomposition level, which is incremented by one at each time before a pre-defined maximum level is reached. The decomposition level with which the random part of a time series can be maximally removed will be used as the final decomposition level to smooth the time series data.

6. The Estimation of Missing Values

As explained in Section 2, it is likely that the stochastic characteristics of traffic flow result in an error in the similarity measurement for the imputation by the kNN algorithm. As a consequence, such error can be embedded to the weights when the weighted sum of k nearest neighbours is used to produce the final estimation. Therefore, we adopt a regression model in this work to estimate the missing values in order to avoid the possible weighting errors. Although a large variety of regression models is available, we employ the multivariate linear regression model in this work to estimate the missing values with consideration in compromising imputation accuracy and computational overhead.

Multivariate linear regression is a generalization of simple linear regression to the case where two or more explanatory variables and a response variable can be modelled by a linear function (Wichura, 2006). The basic model for multivariate linear regression can be represented as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i \qquad (10)$$

for each p dimensional observed data indexed as $i = 1, \ldots, n$. In our case, the set of observed data are those that k neighbouring time series have the observed responses at the corresponding time instants, with dimension p being equal to k. That is to say that we use the regression technique rather than the similarity distance to determine the weights for the final estimation.

7. Experiments

In order to evaluate the imputation performance of the improved kNN algorithm, we have designed and performed a series of experiments for the three data sets, as described in Section 2.3. Before conducting the experiments, the traffic flow data have been prepared as follows. The aggregated 5-minute flow data of 9 weekdays were chosen for each data set and a number of data in the traffic flow of the last day was selected randomly and removed as the missing data. Note that it is not necessary to choose traffic data of successive days. In the experiments presented in the following, a set of missing rates, ranging from 0.05 to 0.9 with an interval of 0.05, was used to evaluate the imputation performance for different missing situations. For each missing rate, 100 independent runs were performed with a number of missing data randomly chosen for each run. Therefore, the performance averaged over the 100 runs for each experiment are reported in the section after introducing the performance indicators adopted for the subsequent performance evaluations.

7.1. Performance Indicators and Prediction Examples

The performance indicators, namely mean absolute percentage error (MAPE) and variance of absolute percentage error (VAPE) (Zhang, and Liu, 2011), have been chosen to evaluate the imputation performance. While MAPE calculates the average relative error between the estimated values and actual observed data, VAPE represents the performance stability.

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{y}(i) - y(i)}{y(i)}\right| \times 100\% \qquad (11)$$

$$\text{VAPE} = \text{var}\left(\left|\frac{\hat{y}(i) - y(i)}{y(i)}\right|\right) \times 100\% \qquad (12)$$

where $y(i)$ and $\hat{y}(i)$ are *i*th true and predicted values and N is the number of data.

Figure 6 shows the results imputed by our proposed algorithm for traffic flow data set ‘PeMS’, ‘Portal’ and ‘Beijing’, respectively, with the missing rate of 0.7 and Table 1 summaries the parameter k and the coefficients of the multivariate linear regression determined for the experiments shown in Fig 6. As the parameter k is determined in an adaptive way, the values of k are different for the three tests.

The sold lines in Figure 6 indicate the true data of traffic flow and the data imputed for a number of missing data are marked with “o”. Note that a number of data are randomly selected from the original data (as the solid lines in Figure 6) and discarded according to the missing rate specified, before the proposed algorithm is applied to impute these missing data. As show in Figure 6, traffic flow from the data set ‘Portal’ is relatively larger in magnitude than those from ‘Portal’ and ‘Beijing’. Furthermore, it can also be seen that the missing data can be estimated with a reasonable accuracy, even though the traffic flows fluctuate significantly over time of day.

Table 1. The coefficients of multivariate linear regression and parameter k determined for the results shown in Fig 6.

Data set	k	Coefficients of multivariate linear regression
PeMS	3	{0.113, 0.535, 0.316}
Portal	4	{0.077, 0.073, 0.701, 0.077}
Beijing	2	{0.421, 0.566}

7.2. Comparison to the Original kNN Algorithm

The first experiment has been designed to evaluate the performance of the proposed imputing algorithm by comparing it to the original kNN algorithm which has been used as a basis to make specific improvements. As the parameter k is critical for both the original kNN algorithm and the proposed one, a fair comparison can be made with the same parameter k for the both imputing algorithms. For the original kNN algorithm, the parameter k is generally specified by users before performing the imputation for missing data and the lack of sufficient domain knowledge can induce a subjective decision on the parameter. On the contrary, the improved kNN algorithm automatically determine the parameter with the assistance of two similarity metrics. Therefore, we first use the proposed algorithm to determine the parameter k which is subsequently employed by the original kNN algorithm. Table 2 lists the MAPE and VAPE values obtained by the original kNN algorithm and improved version, denoted as "kNN" and "IkNN", respectively.

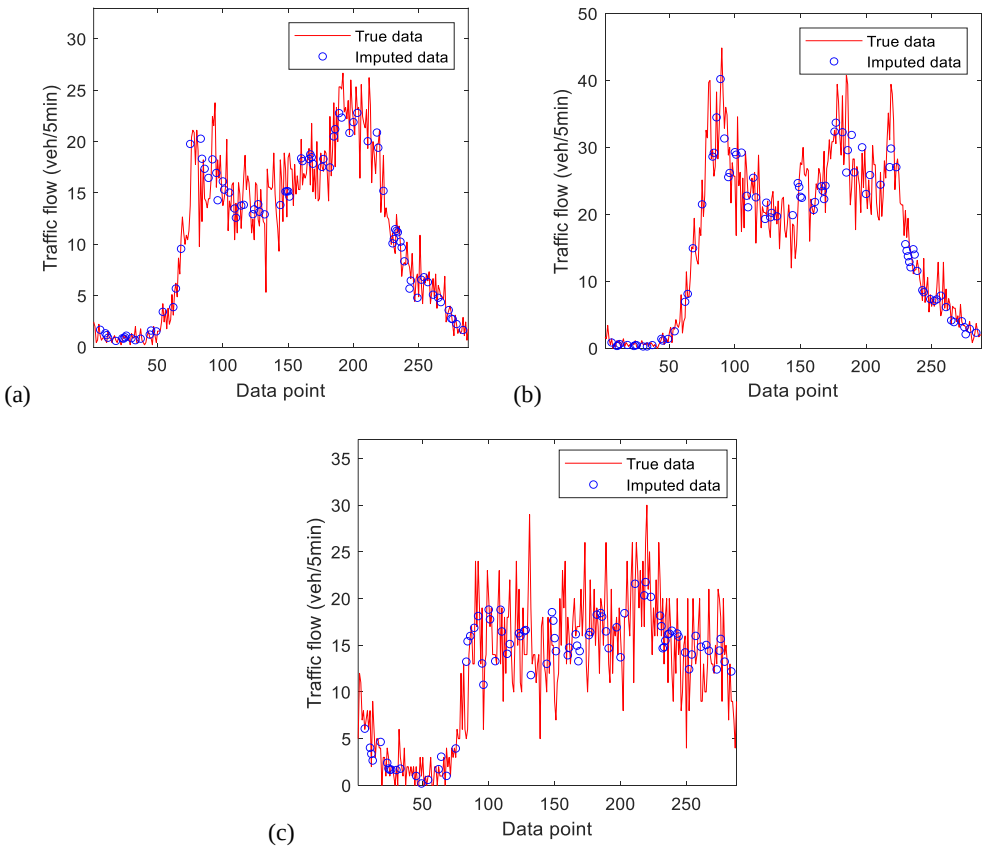


Fig. 6. The imputed results for (a) PeMS data set, (b) Portal data set, (c) Beijing data set, with missing rate of 0.7

Table 2. The performance of imputation by the original and improved kNN algorithms (denoted as kNN and IkNN) over 100 independent runs for different missing rates with the same parameter k

Missing Rate	PeMS					Portal					Beijing				
	KNN		IKNN		Mean k	KNN		IKNN		Mean k	KNN		IKNN		Mean k
	MAPE	VAPE	MAPE	VAPE		MAPE	VAPE	MAPE	VAPE		MAPE	VAPE	MAPE	VAPE	
0.05	36.30	77.90	30.01	34.29	2.14	26.18	10.34	24.15	9.14	4.99	34.31	18.58	30.26	13.61	2.07
0.10	37.38	71.91	31.55	37.60	2.16	25.57	9.81	23.76	8.71	4.95	35.37	18.32	30.42	12.71	2.20
0.15	35.28	52.36	30.11	30.30	2.34	25.65	10.00	23.63	8.82	4.88	34.80	18.39	31.13	14.19	2.37
0.20	36.33	63.15	30.89	33.47	2.26	25.20	8.86	23.34	8.22	4.86	35.12	18.37	31.20	14.73	2.44
0.25	36.19	65.79	30.77	34.48	2.29	25.87	10.71	23.68	9.13	4.72	34.69	17.44	31.02	14.22	2.43
0.30	36.60	65.94	31.11	34.68	2.36	25.75	10.14	23.38	8.48	4.74	34.37	17.30	30.87	14.33	2.49
0.35	36.49	71.09	30.68	35.80	2.38	26.01	10.24	23.86	8.95	4.89	35.00	17.67	31.12	14.96	2.48
0.40	36.81	67.77	31.16	36.15	2.45	25.97	9.88	23.71	8.83	4.46	35.21	18.22	32.28	16.13	2.61
0.45	35.94	58.57	30.65	33.10	2.47	25.98	10.49	23.61	9.08	4.47	35.28	18.59	31.88	15.51	2.61
0.50	35.84	53.68	30.76	32.33	2.48	26.00	10.34	23.59	8.69	4.50	34.90	17.92	31.65	15.01	2.57
0.55	37.16	64.96	31.12	34.51	2.34	26.37	11.13	23.93	9.28	4.57	35.34	18.37	31.95	15.74	2.51
0.60	36.99	62.19	31.39	34.76	2.36	26.59	11.15	24.21	9.52	4.32	34.78	17.94	31.71	14.95	2.72
0.65	36.70	60.43	31.60	35.80	2.48	26.60	10.87	24.52	9.92	4.12	35.19	18.36	31.61	15.08	2.48
0.70	37.58	63.27	31.77	36.44	2.29	26.40	10.59	24.49	9.82	4.26	35.00	18.18	32.03	15.44	2.59
0.75	37.19	61.80	31.70	36.19	2.36	26.33	10.53	24.42	9.75	4.30	34.98	18.34	32.20	15.60	2.58
0.80	37.13	61.67	32.38	37.43	2.19	26.44	10.76	24.83	10.10	4.04	35.02	17.65	32.38	15.75	2.44
0.85	36.80	60.09	31.95	37.45	2.42	26.69	10.87	25.25	10.32	3.99	35.78	18.81	32.37	15.31	2.30
0.90	38.02	61.11	32.94	39.45	2.26	26.98	11.71	**27.09**	**12.41**	3.89	35.69	18.80	32.56	15.10	2.21

As shown in Table 2, the values of MAPE and VAPE obtained by IkNN are lower than those by kNN in most cases, indicating the proposed algorithm outperforms the original kNN algorithm when the parameter k is same. The reductions in MAPE and VAPE are most profound for the PeMS data set, as compared to other two sets. However, by a close inspection to the performance of the proposed algorithm, there are increasing trends for the values of MAPE and VAPE for the test data sets with the missing rates, implying that the imputation performance is inversely affected with the increase of missing rate. The parameter k, determined by the proposed algorithm, are similar for the data set of PeMS and Beijing, but different for the data set of Portal. For the data set of PeMS and Beijing, the parameter k, determined by the proposed algorithm, indicates that only 2 or 3 historical time series can be classified as the neighbours. On the other hand, the more historical time series are similar in the data set of Portal, as the parameter k determined ranges from 3 to 5.

Figure 7 shows the dependence and adaptation of parameter k on the missing rate. Note that we use different colours, as shown in colour bar, to indicate the value of k determined by the proposed method described in Section 3.1. From Figure 7, it seems that the parameter k fluctuates slightly when missing rate is low, but there is an increasing tendency in fluctuation of the parameter k when the number of missing data becomes large. When there are a few missing data, the information used to estimate the similarity is relatively adequate and, therefore, this may partially explain the relatively small fluctuation in the parameter k over 100 independent runs. On the other hand, only small part of the time series is available for the similarity measure and different part may provide different local information, resulting in a large fluctuation of the parameter k over 100 runs.

7.3. Evaluation of Smoothing Effect

The experiments presented in this sub-section aim to evaluate the imputation performance enhanced with the smoothing process. To this end, the three sets of traffic flow data with the missing rates ranging from 0.05 to 0.9 were imputed by the proposed algorithm without and with the smoothing process and the values of MAPE and VAPE obtained are listed in Table 3, along with the decomposition level automatically determined. Obviously, the results (see Table 3) indicate the smoothing process can effectively improve the imputation accuracy and reduce the fluctuations over the 100 independent runs for each missing rate. In addition, it can also be seen that the imputation performance, in terms of averaged accuracy and fluctuation over 100 independent runs, is slightly degraded with the increase of missing rate. The decomposition levels determined for the data sets of PeMS and Beijing are close to 3 and 4, respectively, while the decomposition levels for the data set of Portal is blew 3. This implies that the stochastic degree of the traffic flow in the data set of Portal is less than those in the data sets of PeMS and Beijing.

7.4. Comparison to Other Typical Methods

Additional experiment was performed to make a comparative study to a set of existing imputation methods, namely interpolation-based techniques, ARIMA, PPCA, and nearest historical recording methods (denoted as 'NH' in the followings) which are typical imputing algorithms falling into the prediction, interpolation, and statistical-learning based imputation algorithm categories. There are three interpolation techniques, i.e., linear, spline, and pchip interpolation, used for comparison and denoted as 'Linear', 'Spline', and 'Pchip', respectively, in the followings, for easy reference. The parameters of ARIMA, p and q, were determined by calculating the autocorrelation and partial autocorrelation coefficients, while the augmented Dickey–Fuller test was used to help determine the differential parameter, d. For the imputation by PPCA, three principle components, determined by trail-and-error, was used for reconstruction. The simplest imputing method adopted here for comparison is the nearest historical recordings method that the recordings for the same time period in the previous day are used to impute the missing values.

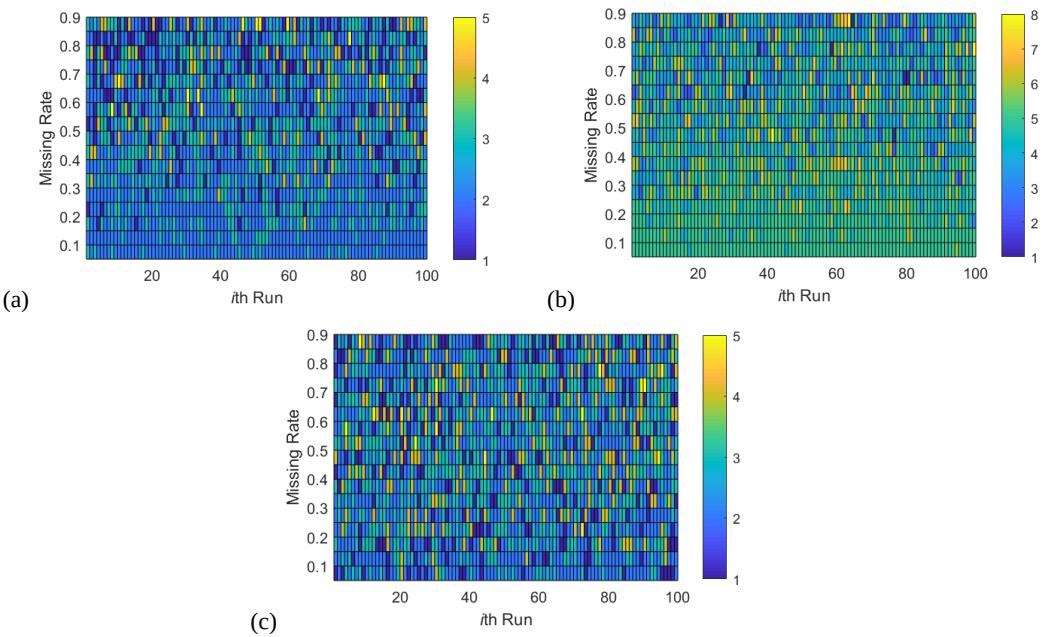


Fig. 7. Parameter k determined for each independent run with different missing rate. (a) PeMS data set, (b) Portal data set, (c) Beijing data set

Table 3. The performance of imputation by the proposed algorithm with and without smoothing process over 100 independent runs for different missing rates

Missing Rate	PeMS					Portal					Beijing				
	Without		With			Without		With			Without		With		
	MAPE	VAPE	MAPE	VAPE	Level	MAPE	VAPE	MAPE	VAPE	Level	MAPE	VAPE	MAPE	VAPE	Level
0.05	34.38	71.32	30.01	34.29	3.00	26.11	10.45	24.15	9.14	2.39	33.11	15.96	30.26	13.61	3.90
0.1	35.42	71.44	31.55	37.60	3.00	26.33	10.58	23.76	8.71	2.38	33.74	16.29	30.42	12.71	3.85
0.15	33.35	53.05	30.11	30.30	2.99	26.18	10.54	23.63	8.82	2.38	33.07	15.57	31.13	14.19	3.74
0.2	33.79	55.66	30.89	33.47	2.95	25.82	9.42	23.34	8.22	2.37	33.87	16.75	31.20	14.73	3.67
0.25	33.81	59.29	30.77	34.48	2.95	26.45	11.40	23.68	9.13	2.35	33.42	16.07	31.02	14.22	3.67
0.3	34.28	62.72	31.11	34.68	2.96	26.23	10.53	23.38	8.48	2.35	33.59	16.38	30.87	14.33	3.58
0.35	34.28	65.49	30.68	35.80	2.93	26.67	10.95	23.86	8.95	2.33	33.96	16.52	31.12	14.96	3.50
0.4	34.37	62.18	31.16	36.15	2.95	26.34	10.23	23.71	8.83	2.41	33.88	16.05	32.28	16.13	3.54
0.45	33.84	54.21	30.65	33.10	2.91	26.64	11.18	23.61	9.08	2.38	34.39	17.04	31.88	15.51	3.54
0.5	33.86	51.40	30.76	32.33	2.91	26.55	10.93	23.59	8.69	2.37	33.72	16.49	31.65	15.01	3.45
0.55	34.54	55.84	31.12	34.51	2.93	26.78	11.65	23.93	9.28	2.37	34.22	16.78	31.95	15.74	3.36
0.6	34.92	58.01	31.39	34.76	2.91	27.10	11.67	24.21	9.52	2.40	33.58	15.65	31.71	14.95	3.35
0.65	34.80	54.20	31.60	35.80	2.94	27.19	11.48	24.52	9.92	2.37	34.39	17.22	31.61	15.08	3.33
0.7	35.29	58.38	31.77	36.44	2.96	27.05	11.52	24.49	9.82	2.38	34.22	16.40	32.03	15.44	3.34
0.75	35.28	57.61	31.70	36.19	2.92	27.09	11.65	24.42	9.75	2.42	34.52	17.02	32.20	15.60	3.22
0.8	36.14	57.41	32.38	37.43	2.99	27.28	11.52	24.83	10.10	2.45	34.66	16.92	32.38	15.75	3.34
0.85	35.49	55.08	31.95	37.45	2.89	27.92	11.99	25.25	10.32	2.45	35.20	17.48	32.37	15.31	3.17
0.9	36.62	56.68	32.94	39.45	2.71	29.26	14.24	27.09	12.41	2.44	35.39	17.74	32.56	15.10	3.22

Figure 8 (a), (b), and (c) present the values of MAPE obtained by the test imputing algorithms for the three data sets. It is evident that the proposed algorithm can produce more accurate imputation than other test algorithms for different missing rates, even though a similar performance can be obtained by ARIMA for the data sets, 'PeMS' and 'Portal', when the missing rate is small. In general, the 'Spline' and 'Pchip' interpolation techniques perform poorly, while the 'Linear' interpolation can produce similar accurate imputation as 'ARIMA' and 'PPCA'. Amongst all the existing algorithms used for comparison, 'PPCA' performs better than others when a number of missing data becomes large. In addition, it is interesting to note that the simplest imputing strategy, 'NH', can produce medium accuracy for the imputation with various missing rate among all test algorithms and its performance seems to vary slightly for different missing rate.

On the other hand, the values of VAPE obtained by the test algorithms are shown in Figure 9. For the data sets, 'Portal' and 'Beijing', the proposed algorithm can produce least fluctuation when different data are removed as missing data. However, ARIMA performed more stable than the other algorithms when it was used to impute for data set 'PeMS'. Again, it can be seen that poor performances in terms of stability are generated by the 'Spline' and 'Pchip' interpolation techniques. In addition, a reasonable stability can be achieved when the 'Linear' interpolation algorithm was used to impute missing data. Furthermore, a relatively medium stability can be achieved by the 'NH' method, and there is no significant difference in the stability for different missing rate, indicating it is insensitive to the number of missing data.

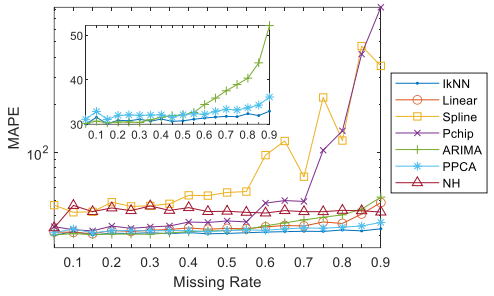
8. Conclusions

The work presented in this paper attempts to improve the imputation performance by proposing a modified version of the kNN algorithm based on the analysis of the stochastic characteristic of traffic flow which results in a number of difficulties to determine the critical elements of the original kNN algorithm. The improvements have been motivated by the intention to eliminate both the uncertainties resulted from the stochastic characteristic of traffic flow and the requirement to predefine parameters.

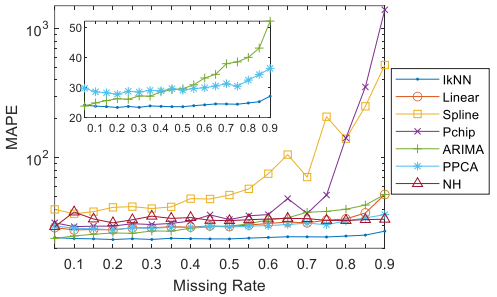
A series of experiments for a set of traffic flow data has been performed to evaluate the imputation improvements for the proposed algorithm from various aspects. The comparative study indicates the proposed algorithm outperforms the other conventional approaches in terms of imputation errors and corresponding fluctuations in general. Furthermore, no need to predefine parameters is a unique advantage

of the developed imputing algorithm over the other commonly-used algorithms. One of the critical parameters can be automatically determined in an adaptive fashion to fit different traffic patterns.

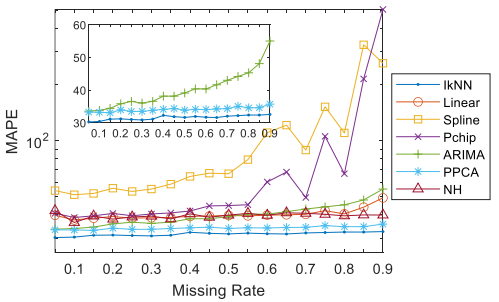
On the other hand, the experimental results also imply a number of weaknesses, which is required to improve in the further research. The current strategy to adaptively determine the critical parameter cannot guarantee an optimal imputation in terms of accuracy, even though the current version outperforms the original kNN algorithm with same parameter settings. Also, it is interesting to investigate the imputation performance if other regression techniques instead of multivariate linear model are employed.



(a)

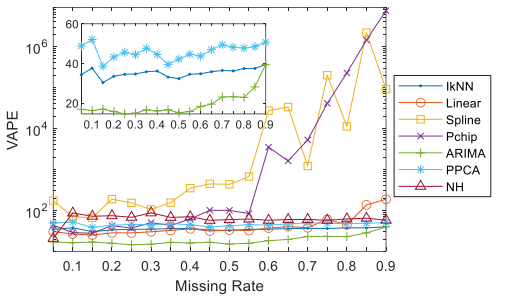


(b)

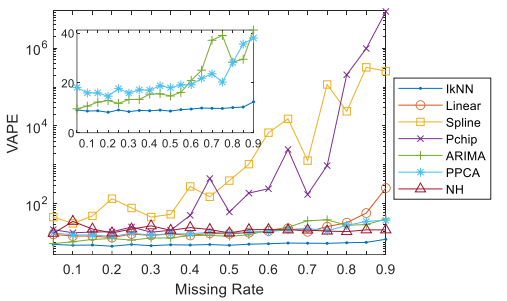


(c)

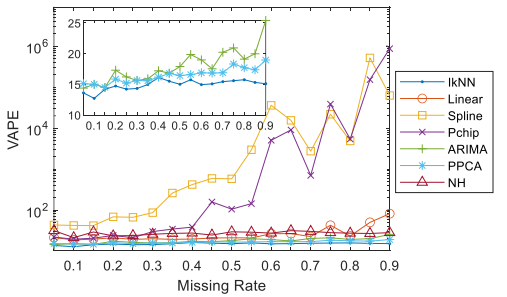
Fig. 8. The values of MAPE obtained by a set of imputing algorithms for (a) PeMS data set, (b) Portal data set, (c) Beijing data set



(a)



(b)



(c)

Fig. 9. The values of VAPE obtained by a set of imputing algorithms for (a) PeMS data set, (b) Portal data set, (c) Beijing data set

Acknowledgments
The project has been funded by National Key R andD Program of China (2016YFE0206800, 2017YFC0803903), Scientific Research Common Program of Beijing Municipal Commission of Education (KM201710005030), and the Foundation Research Fund Project of Beijing University of Technology (038000514315501).

References

[1] Abbasifard, M. R., Ghahremani, B., Naderi, H., 2014. A survey on nearest neighbor search methods. Int J Comput Appl, 95(25), 39-52.
[2] Arce, G. R., 2005. Nonlinear Signal Processing: A Statistical Approach (Wiley: New Jersey, USA).
[3] Bae, B., Kim, H., Lim, H., et al., 2018. Missing data imputation for traffic flow speed using spatio-temporal cokriging[J]. Transportation Research Part C Emerging Technologies, 88, 124-139.
[4] Bianchi, G., Sorrentino, R., 2007. Electronic filter simulation and design (McGraw-Hill Professional, 2st edn), 17-20.
[5] Bhatia, N., Vandana., 2010. Survey of nearest neighbor techniques. Int. J. Comput. Sci. Inf. Secur, 8(2), 302-305.
[6] Castro-Neto, M., Jeong, Y.-S., Jeong, M.-K., Han, L.D., 2009. Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions. Expert Syst. Appl., 36(3), 6164-6173.
[7] Chen, J., Shao, J., 2000. Nearest neighbour imputation for survey data. J. Off. Stat., 16(2), 113-131.
[8] Chen, X., He, Z. Sun, L., 2019. A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation. Transp. Res. C, Emerg. Technol., 98, 73-84.
[9] Chui, C.K., 1992. An Introduction to Wavelets (Academic Press, 1st edn).
[10] De Boor, C., 2001. A Practical Guide to Splines (Springer, Rev edn.), 207-214.
[11] El-Dahshan, E.S.A., 2011. Genetic algorithm and wavelet hybrid scheme for ECG signal denoising. Telecommun Syst., 46, 209-215.
[12] Esawey, M. E., Sayed, T., 2012. Neighbour corridors travel time estimation: Concept and a case study[J]. Advances in Transportation Studies, 28(28):81-96.
[13] Ghosh, B., Basu, B., O'Mahony, M., 2007. Bayesian time-series model for short-term traffic flow forecasting. ASCE J. Transp. Eng., 133(3), 180-189.
[14] Kim, H., Golub, G.H., Park, H., 2005. Missing value estimation methods for DNA microarrays gene expression data: local least squares imputation. Bioinformatics, 21(2), 187-198.
[15] Li, L. C., Zhang, J., Wang, Y. G., et al., 2019. Missing Value Imputation for Traffic-Related Time Series Data Based on a Multi-View Learning Method. IEEE Transactions on Intelligent Transportation Systems, 20(8), 2933-2943.
[16] Li, Y., Li, Z., Li, L. et al, 2013. Comparison on PPCA, KPPCA and MPPCA Based Missing Data Imputing for Traffic Flow. Proc. Int. Conf. Transportation Information and Safety, Wuhan, China, 1151-1156.
[17] Li, Y., Li, Z., Li, L., 2014. Missing traffic data: Comparison of imputation methods', IET Intell. Transp. Sy., 8(1), 51-57.
[18] Loukopoulos, P., Sampath, S., Pilidis, P. et al, 2016. Dealing With Missing Data for Prognostic Purposes. Proc Conf. Prognostics and System Health Management, Chengdu, China, 1-5.
[19] Ma, X., Luan, S., Du, B. et al, 2017. Spatial copula model for imputing traffic flow data from remote microwave sensors. Sensors, 17(10), 2160.
[20] Misiti, M., Misiti, Y., Oppenheim et al, 2007. Wavelets and their Applications (Wiley-ISTE, 1st edn).
[21] Performance Measurement System (PeMS). http://pems.dot.ca.gov/, accessed 15 February 2017.
[22] Portland Oregon Regional Transportation Archive Listing (PORTAL). http://portal.its.pdx.edu, accessed 27 September 2018.
[23] Qu, L., Li, L., Zhang, Y. et al., 2009. PPCA-based missing data imputation for traffic flow volume: a systematical approach. IEEE T Intell. Transp, 10(3), 512-522.
[24] Silva, H. D., Perera, A. S.,2017. Missing data imputation using Evolutionary k- Nearest neighbor algorithm for gene expression data[C]. Sixteenth International Conference on Advances in Ict for Emerging Regions.
[25] Stoeck, T., Prajwowski, K., 2010. Application of Interval Interpolation for the Description of

Compression-Ignition Engine Performance Characteristics[J]. Archives of Transport, 22(3).

[26] Tang, J., Wang, Y., Zhang, S., et al., 2015. On Missing Traffic Data Imputation Based on Fuzzy C-Means Method by Considering Spatial-Temporal Correlation[C]. Transportation Research Board Meeting.

[27] Tan, H., Feng, G., Feng, J. et al, 2013. A tensor-based method for missing traffic data completion. Transport Res C-Emer, 28, 15-27.

[28] Vlahogianni, E. I., Karlaftis, M. G., Golias, J. C, 2005. Optimized and meta-optimized neural networks for short-term traffic flow prediction: a genetic approach. Transp. Res. C, Emerg. Technol., 13(3), 211-234.

[29] Wang, S. B. Mao, G. Q., 2019. Missing Data Estimation for Traffic Volume by Searching an Optimum Closed Cut in Urban Networks. IEEE Transactions on Intelligent Transportation Systems, 20(1), 75-86.

[30] Wang, Y., Zhang, Y., Piao, X., et al., 2019. Traffic Data Reconstruction via Adaptive Spatial-Temporal Correlations. IEEE Transactions on Intelligent Transportation Systems, 20(4), 1531-1543.

[31] Wichura, M. J., 2006. The coordinate-free approach to linear models (Cambridge Series in Statistical and Probabilistic Mathematics; Cambridge University Press, 1st edn).

[32] Xu, J., Li, X., Shi, H., 2010. Short-term traffic flow forecasting model under missing data. Journal of Computer Applications, 30(4), 1117-1120.

[33] Zbilut, J. P., Marwan, N., 2008. The wiener-khinchin theorem and recurrence quantification. Phys Lett A, 372(44), 6622-6626.

[34] Zhang, C. S., Sun, S., Yu, G., 2004. A Bayesian network approach to time series forecasting of short-term traffic flows. Proc. IEEE Conf. Intelligent Transportation Systems, Washington, D.C., 216-221.

[35] Zhang, Y., Liu, Y., 2011. Analysis of peak and non-peak traffic forecasts using combined models. J Adv Transport, 45, 21-37.

[36] Zhong, M., Sharma, S., Lingras, P., 2004. Genetically designed models for accurate imputations of missing traffic counts. Transp. Res. Rec., 1879(1), 71-79.

[37] Zhuang, Y., Ke, R. Wang, Y., 2019. Innovative method for traffic data imputation based on convolutional neural network. IET Intelligent Transport Systems, 13(4), 605-613.