IET Intelligent Transport Systems

Special issue Call for Papers

Be Seen. Be Cited.
Submit your work to a new IET special issue

Connect with researchers and experts in your field and share knowledge.

Be part of the latest research trends, faster.

Read more

The Institution of Engineering and Technology

IET Intelligent Transport Systems

Research Article

IET Journals
The Institution of Engineering and Technology

Innovative method for traffic data imputation based on convolutional neural network

ISSN 1751-956X
Received on 4th April 2018
Revised 25th September 2018
Accepted on 29th October 2018
E-First on 26th November 2018
doi: 10.1049/iet-its.2018.5114
www.ietdl.org

Yifan Zhuang[1], Ruimin Ke[1], Yinhai Wang[1] ✉

[1]*Smart Transportation Application and Research Laboratory, Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, USA*

✉ *E-mail: yinhai@uw.edu*

Abstract: The quality of traffic data is crucial for modern transportation planning and operations. However, data could be missing for various reasons. Hence, the data imputation approaches which aim at predicting/replacing the missing data or bad data have been considered very important. The traditional traffic data imputation approaches mainly focus on using different probability models or regression methods to impute data, and they only take limited temporal or spatial information as inputs. Thus, they are not very accurate especially for data with a high missing ratio. To overcome the weaknesses of previous approaches, this study proposes an innovative traffic data imputation method, which first transforms the raw data into spatial-temporal images and then implements a deep-learning method on the images. The key idea of this approach is developing a convolutional neural network (CNN)-based context encoder to reconstruct the complete image from the missing source. To the best of the authors' knowledge, this is the first time a CNN method has been incorporated for traffic data imputation. Experiments are conducted on three months of data from 256 loop detectors. Through comparison with two state-of-the-art approaches, the results indicate that this new approach increases the imputation accuracy greatly and has a stable error distribution.

1 Introduction

Traffic sensors are the key component of intelligent transportation systems (ITS) which lays the foundation for modern traffic operation and control. With the rapid development of information technologies and the automobile industry, sensors have transitioned to a much larger scale with respect to both their type and the amount of data they record. As a result, a greater amount of traffic data is being produced than ever before [1–6]. The sufficiency and accuracy of traffic data are important for both individuals' personal use and agency application in traffic management [7]. However, bad or missing data can lead to inefficiency or inaccuracy in analysis tasks, and data are often missing due to various reasons such as detector damage and transmission loss [8–10]. Inductive loop detectors are likely the most widely deployed traffic sensors at this moment, and the normal missing data ratio for them is between 5 and 10%; unfortunately, it can reach up to 25% in some cases [11]. According to the study by Tan *et al.* [12], in extreme situations, a 90% missing ratio could be possible. The simplest way to solve this problem is removing data from all sensors at the instant data are missing or removing data for the whole period at the missing sensor, either of which will cause cracks in spatial and temporal dimensions. Data imputation, which fills the missing region based on both the time-neighbouring and location-neighbouring traffic data, plays an important role in improving the performance of the whole transportation system.

This paper proposes a traffic data imputation method based on the convolutional neural network (CNN) to improve the imputation accuracy, especially when in the high missing ratio situation. Compared to previous methods, a CNN-based method can analyse features in both macroscopes and microscopes. First, raw traffic data is transformed into a two-dimensional image where the horizontal axis represents the sensor number, and the vertical axis represents the time stamp. The value of valid pixels in this image represents the observed traffic volume, and the value of missing points is set to −1. In this way, the traffic data imputation problem is reformed into an image inpainting problem which focuses on making up the missing regions. The CNN-based context encoder–decoder pipeline is adopted to inpaint the spatial-temporal image, which will decompose the image with a missing region into a feature map first and then reconstruct a complete image with the same size as the original one. To test the effectiveness of this approach, different missing ratios from 5 to 50% with 5% interval are considered, and random missing regions of different sizes and positions are taken as inputs. Two other state-of-the-art approaches – Bayesian principal component analysis (BPCA) and denoising stacked autoencoder (DSAE) – are adopted for comparison under the same experiment condition. BPCA is known to be better than most traditional approaches, and DSAE is a deep learning approach showing high imputation accuracy. The results indicate that the proposed CNN-based approach has better accuracy than BPCA and DSAE under different error criterions.

The rest of paper is organised as follows: Section 2 briefly reviews the related work in traffic data imputation and introduces basic methods of image inpainting; Section 3 describes the CNN-based data imputation approach; Section 4 shows the experimental results and analysis; and Section 5 concludes this paper and presents some ideas for future work.

2 Literature review

Traditional traffic data imputation approaches can be categorised into three types – prediction, interpolation, and statistical methods [13]. The basic idea of the prediction approach is building a relationship between historical and unknown future data. Thus, the missing data can be imputed by using the previous data in the same time series. The representative methods include Bayesian networks [14], autoregressive integrated moving average models [15], support vector regression [16], and so on. One great strength of these approaches is predicting short-term traffic flow with high efficiency. When the data missing ratio is low, they can produce a relatively accurate result in a short time. However, their shortcomings are also obvious. The first one of which is that they only make use of data preceding to the missing data in the temporal dimension. The traffic data collected following the imputed point is ignored. Another one is that such methods only focus on relationships in the temporal dimension without considering

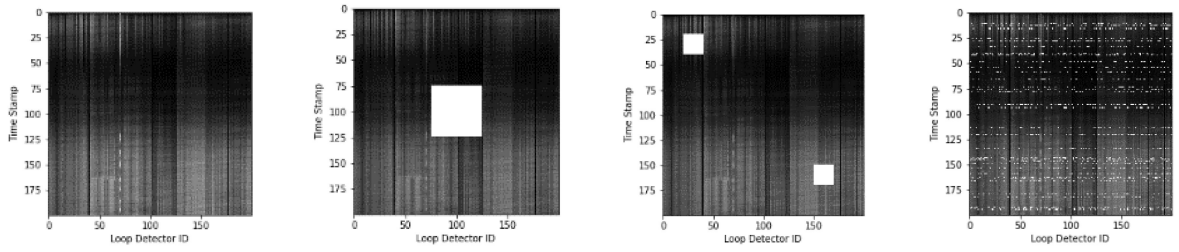


Fig. 1 *Traffic volume image with different missing status*
(a) Complete data without any missing region, ***(b)*** Data with missing central regions, ***(c)*** Data with randomly missing blocks, ***(d)*** Data with randomly missing regions

relationships in the spatial dimension which can also provide crucial information for imputation.

The interpolation methods mainly include two branches – temporal-neighbouring method and pattern-similar method. The temporal-neighbouring method [17] imputes the missing data point by using data from the same sensor and the same time period in neighbouring days or on the same day in neighbouring periods. In most cases, the computation method takes the average value of known data [18]. This method is widely used in highway traffic data imputation because its basic assumption is that traffic flow has strong regularity and keeps almost the same pattern every day. It puts a tight constraint on data and does not make use of data from other time periods. For example, this method does not work for holidays or weekends because the highway traffic volume on holidays or weekends is often very different from that of weekdays. The pattern-similar method imputes missing data by using historical data at the same sensor for different days, which have similar data patterns to that day on which data are missing. Two typical examples are the k-nearest neighbours (k-NN) [19] method and local least squares [20]. Compared to the temporal-neighbouring method, this type of methods considers both traffic flow data for the day on which data are missing and traffic flow data with similar patterns on other days. Thus, they overcome the limitations of focusing only on nearby data in the temporal-neighbouring method.

Different from the previous approaches that impute data at the microscopic level, statistical approaches focus on the macroscopic view by analysing its statistical regularity, such as Markov Chain Monte Carlo [21] and BPCA [11, 22]. The BPCA approach is adopted as one comparison approach in this study, which is built on the traditional principal component analysis by incorporating prior knowledge that the fluctuations of traffic flow are the Gaussian type. The statistical approach can be achieved in two steps. First, the probability distribution model is well developed. Then, the historical data works as the model input to train parameters and impute the missing part at the same time. However, this type of methods may lose detailed information because of capturing overall patterns. Their accuracy is highly dependent on the probability distribution model, which may be totally different in different situations and become tough for determining the most suitable distribution model.

In recent years, deep learning methods have been widely used in various fields, including data imputation. They help improve the imputation accuracy significantly through a large amount of training data. Therefore, deep learning methods solve the aforementioned problems by capturing both temporal and spatial information from historical and neighbouring data. Duan *et al.* [23] first applied the deep learning method in traffic data imputation with DSAE approach. The basic idea of this research is putting the raw data with the missing part into the encoder. Then, the output of the decoder is the imputation result. The encoder is built on $h = a(Wx + b)$ and decoder is built on $\hat{x} = a_h(W'h + b_h)$. The DSAE approach is also adopted as another comparison method in the Experiment section.

In another perspective, the traffic data can be seen as images with the horizontal axis of spatial position and the vertical axis of the temporal stamp. The value of each pixel represents traffic volume data measurements while the missing part is set to −1. Thus, in such cases, the traffic data imputation problem turns into an image inpainting problem. Tak *et al.* [24] did data imputation in sectional units of road links via transforming raw data into several two-dimensional images using k-NN, which made use of both neighbouring temporal and spatial data. Compared to the traditional k-NN method, their work also took spatial data into considerations; however, it still ignored the historical data. To make up the above weakness, methods in the computer vision field [25, 26] provide other possible solutions with neural networks. CNN is suitable for image processing by extracting high-level features when considering that there may exist features beyond texture regularity. Thus, it is mainly applied in object recognition. In 2016, Pathak *et al.* [27] proposed a context encoder trained by CNN to generate missing parts of inputs. This paper examines the feasibility of a CNN in traffic data imputation and is developed based on Pathak's method. The structure of CNN-based imputation method is similar to the DSAE method. However, the great difference lies in ConvoPooling layers which can analyse more complex features. We also improve the CNN network structure in order to gain higher accuracy and stable error distribution.

3 Methodology

This paper puts forward an innovative method for missing traffic data imputation which solves the imputation problem from the perspective of the spatial-temporal 2D image using a deep-learning based image inpainting approach. Traditionally, the traffic data imputation approach applied the pure mathematical methods which cannot make full use of both spatial and temporal information. The basic idea of our approach is first transforming raw data into a two-dimensional image and then applying a CNN-based image inpainting approach to impute the missing data. This study focuses on imputing traffic volume data. Therefore, the raw data is traffic volume data. This method has good scalability which means it can apply to other types of data, such as traffic velocity data, without little modification.

3.1 Data transformation

The first step of data imputation is transforming the raw traffic volume data into two-dimensional images. The horizontal axis (or x-axis) denotes the loop detectors' IDs which are selected continuously from one road or several intersecting roads. In this way, the selection of loop detectors in this way ensures all loop data is spatially related because two adjacent columns are spatially neighbouring. The vertical axis (or y-axis) denotes the detection time stamp. Thus, the image size is $N \times M$, where N represents the loop detector number and M represents the time stamp number. The value of each pixel represents the traffic volume obtained from loop detectors during a specific time period. When displaying the traffic volume in an image, measurements are linearly mapped into a range between 0 and 254 in grey scale. Thus, the traffic volume can be intuitively observed in both the spatial and temporal dimensions. An example 2D traffic volume image is shown in Fig. 1*a*.

When part of data is missing, the corresponding pixel value is set to −1 in the spatial-temporal matrix, and the corresponding missing pixel is mapped into 255 in the data volume image (pure white colour). The dark parts in Fig. 1 indicate the existence of traffic volume data. The darker the pixel is, the lower the traffic volume is. As mentioned, white pixels indicate a lack of data at

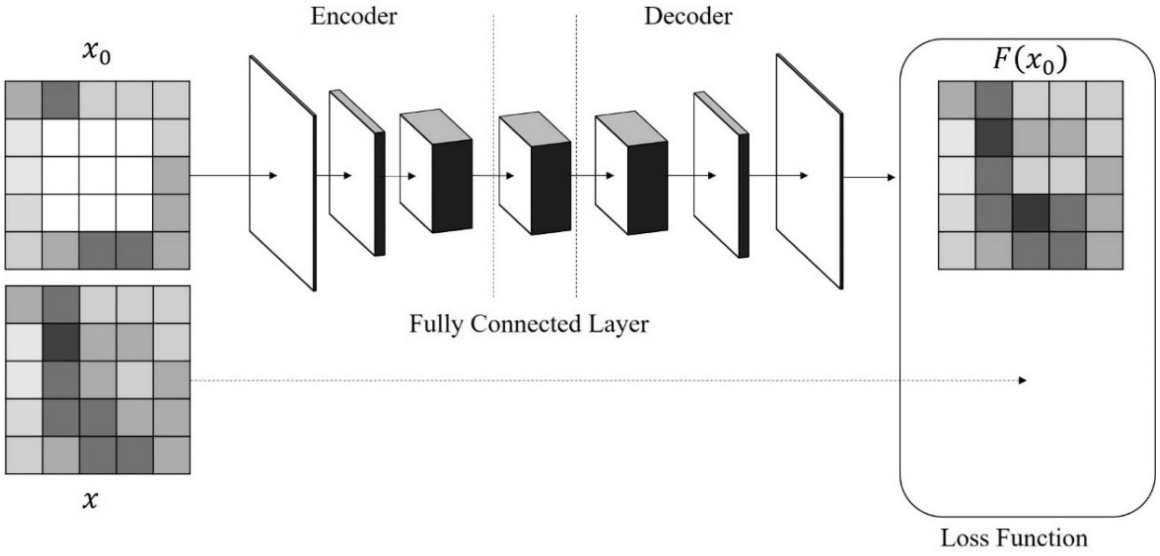


Fig. 2 *CNN-based imputation approach structure*

such a spatial/temporal location (i.e. a location where data imputation is needed). There are three types of missing data in an image – the central region case, the random block case, and the random region case. Cases involving the central region indicate that the missing part is one entire rectangle region located in the centre of the image. This means that several successive loop detectors lost data during the same long period. The central region case is shown in Fig. 1*b*. The random block case occurs when there are several small missing rectangle regions randomly located in the image; it is shown in Fig. 1*c*. The random region case, which is shown in Fig. 1*d*, occurs when the shape and location of the missing regions are both random. The last two situations represent cases for which there are several loop detectors losing data randomly over a short period. The difference between the last two cases is the length of the missing period.

3.2 CNN-based imputation approach

The CNN is widely used in the object classification and recognition tasks. It can extract high-level features which can filter some disturbances such as light and position. Thus, applying a CNN to analyse traffic volume image can obtain inner correlations among different loop detectors and time stamps.

The basic idea is combining features of the encoder–decoder pipeline and the loss function of the generative adversarial network (GAN) [28]. The main architecture is an encoder–decoder pipeline. The encoder at first transforms the input image with missing parts into the feature map. Then, decoder recovers the feature map back to a complete image. The inpainting regions can thus be extracted by putting the mask of missing parts on the output image. The encoder and decoder are connected by a fully connected layer. The structure of the CNN-based imputation approach is shown in Fig. 2.

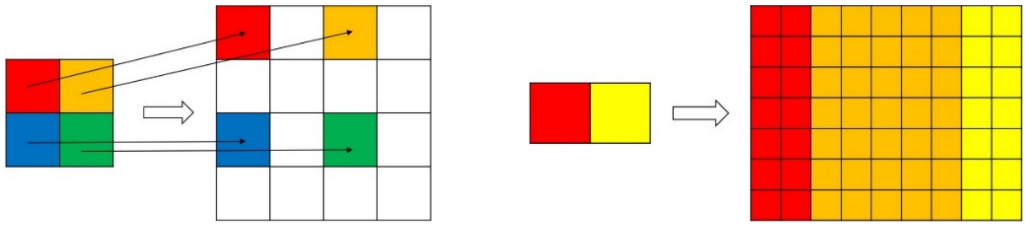
The encoder adopts the AlexNet architecture [29] which was put forward in 2012 and became one of the most popular architectures. CNN comes from regular neural networks which receive input and transform it through a series of hidden layers. Each hidden layer is made up of a set of neurons, where each neuron is fully connected to all neurons in the previous layer. The neurons in a single layer function are completely independent and do not share any connections. The last fully connected layer is also called the 'output layer'. When this network is used for classifications, it represents the class scores. Compared to regular neural network, the CNN takes advantage of that the input is an image and arrange neurons in three dimensions – width, height, and depth. To be more specific, the hidden layers in CNN typically consist of convolutional layers, pooling layers, and fully connected layers in order, instead of only fully connected layers in regular neural networks. In this case, the input image size is 200×200, and the output feature map is $6 \times 6 \times 180$.

The convolutional layer computes the output of neurons which are connected to local regions in the input volume. After the convolutional layer, the pooling layer will perform a downsampling operation in the spatial dimensions. There are three convolutional layers, and each one is followed by one pooling layer. To make the following explanation clear, one convolutional layer and one pooling layer are referred to as a ConvoPooling layer. There are three ConvoPooling layers in total. Different layer sizes are tested, and the following combination is the optimised one. For the first layer, there are 20 filters with size 9×9. The pooling size is 4×4. Thus, the output dimension is $48 \times 48 \times 20$. The filter size for the second layer is 9×9, and there are 60 of them. The pooling size is 2×2. The output dimension is $20 \times 20 \times 60$. The last convolutional layer has 180 filters with size 9×9, and the pooling size is 2×2. The output dimension of $6 \times 6 \times 180$.

The fully connected layer intends to spread information with activations of each feature map. The input layer has 180 feature maps of size 6×6, and it will output 180 feature maps of dimension 6×6. Different from other fully connected layers, it has no parameter connecting different feature maps and only spreads information within feature maps. Thus, the number of parameters of this layer is $180 \times 6 \times 6$ rather than $180^2 \times 6 \times 6$ which is the dimension in the other fully connected layers.

The rest of the pipeline is the decoder following the fully connected layer [30]. The core of decoder is three up-convolutional layers with the rectified linear unit (ReLU) activation function. The definition of ReLU is $f(x) = \max(0, x)$. Contrast to traditional convolution operation, the up-convolution is still a convolution but produces a higher resolution image, which can be learned as a convolution following the unpooling (i.e. increasing the spatial span). The unpooling is just replacing each pixel of the feature map by a $m \times m$ block with the pixel value in the top left corner. The rest is set to 0. The basic principle of the decoder is shown in Fig. 3. The different colour means the different pixel values in the image. The orange part in Fig. 3*b* is the weighted sum of red and yellow parts in convolution operation.

The context encoder–decoder pipeline is trained by regressing the content with missing regions to the originally complete content. The only difference between these two kinds of contests is that missing region while other areas are the same. Thus, the loss function is important for optimising the pipeline. The loss function selected in this paper is jointly comprised of two functions – reconstruction (L2) loss and adversarial loss. The reconstruction loss will capture the overall structure of the missing region and correlations about its context. It will average the influence of multiple modes in predictions. The adversarial loss tends to choose



a *b*

Fig. 3 *Principle/setup of decoder*
(a) 2 × 2 Unpooling, ***(b)*** 2 × 2 Unpooling and 5 × 5 convolution

a particular mode from the distribution to make the prediction look realistic.

For the input image x_0 with missing regions, the context encoder–decoder pipeline F produces an output $F(x_0)$. The actual input image without missing regions is defined as x. The joint loss function only computes the loss of x and $F(x_0)$ in the missing regions. To explain the loss function clearly, the missing region mask $\hat{M}$ is defined. The corresponding values of the missing region are set to one while others are set to zero. The reconstruction loss function is derived from a Euclidean distance which is shown in (1) as follows:

$$L_{\text{rec}}(x) = \hat{M} \odot (x - F(x_0))^{\frac{2}{2}} \tag{1}$$

where $\odot$ is a convolution operation for a specific when combining with mask $\hat{M}$. After applying this operation, the loss function effect range is limited in the missing region. This loss function makes the pipeline tend to produce the rough outline of the inpainting image.

The adversarial loss function comes from GAN. One important parameter of GAN, called D, is the adversarial discriminator which takes both the prediction result and originally complete object into consideration, and tries to distinguish them. The adversarial loss function is shown in the following equation:

$$L_{\text{adv}}(x) = \max_{D} E_{x \in \chi}[\log(D(x)) - \log(1 - D(F(x_0)))] \tag{2}$$

where both F and D are optimised by stochastic gradient descent. This loss function reduces the imputation errors and makes the output of the pipeline closer to the actual data. Therefore, the joint loss function is shown in (3) as follows. The joint loss is the weighted sum of both reconstruction loss and adversarial loss. The loss function is important for the final accuracy. The optimised weights are $\lambda_{\text{rec}} = 0.35$ and $\lambda_{\text{adv}} = 0.65$

$$L = \lambda_{\text{rec}} L_{\text{rec}} + \lambda_{\text{adv}} L_{\text{adv}} \tag{3}$$

4 Experiment

The dataset used in this study was collected from loop detectors on the Interstate 5 (I-5) freeway in Washington State. In total, three months of data from January 2016 to March 2016 were collected from 256 loop detectors and used for the experiment in our study. With records collected at a 5 min time interval, this generates more than 50,000 records, which is considered a sufficient sample size for the case study.

Based on the literature review, the missing ratio is normally between 10 and 15%, but the ratio could be as high as 90% in some extreme cases. This paper takes the whole traffic volume dataset as the raw data with no missing information. Manual operations have been performed on the raw data to remove part of it with certain ratios ranging from 5 to 50% (5% gain). This is better than using raw data with originally missing parts because the comparison between estimated values and ground truth data can be performed. The dimension of one training data image is set as 200×200, which means there are 200 detectors in the spatial dimension and 200 records in the temporal dimension. It is possible to make the images with other dimensions, such as 100×100. To avoid a lack of features, the dimension should not be small. The training process and testing process may take a longer time when the dimension is too large.

The sample size of the training dataset is 14,000. There are ten groups of test data images, each of which has the same dimension as the training images, but with different missing ratios. Each group has ten images with the same missing ratio, and the error of each group is calculated by averaging the estimated errors of all images in the group. Besides the proposed CNN-based approach, another two state-of-the-art approaches are implemented for comparison purposes: a DSAE-based approach and a BPCA approach. The comparison results highlight the strengths of our approach. Three criteria for the evaluation are described in the Evaluation Criteria section, and the experiment results are shown in the next section.

4.1 Evaluation criteria

The evaluations of the imputation approaches are measured by the error of the imputed data. Three widely used criteria are adopted in this paper which focus on different perspectives of error. The first criterion considered is the root mean square error (RMSE), which is shown in (4). It represents the sample standard deviation of the differences between predicted values and observed values. Further, it just compares forecasting errors of different models for particular data, but not the whole dataset because it is scale dependent

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i^r - y_i)^2} \tag{4}$$

The second criterion used is the mean absolute error (MAE) shown in (5), and it is the summary of two components – quantity and allocation disagreement. The quantity disagreement is defined as the absolute value of the mean error, and the allocation disagreement is the negative quantity disagreement. The MAE calculates the average absolute difference

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i^r - y_i| \tag{5}$$

The last evaluation criterion used is the mean percentage error (MPE) shown in (6). It computes the average of absolute percentage errors by which predictions of a model differ from actual values of the quantity being predicted

$$\text{MPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_i^r - y_i|}{x_i^r} \tag{6}$$

where n is the total number of the missing data, y_i is the ith imputed data, and x_i^r is the y_i's corresponding raw data.

4.2 Results and analysis

This model is built on TensorFlow via Python. Thus, the training and testing work are both done on the TensorFlow framework. The training data is divided into 200 batches with each batch containing 70 images before being put into the network for the training process.

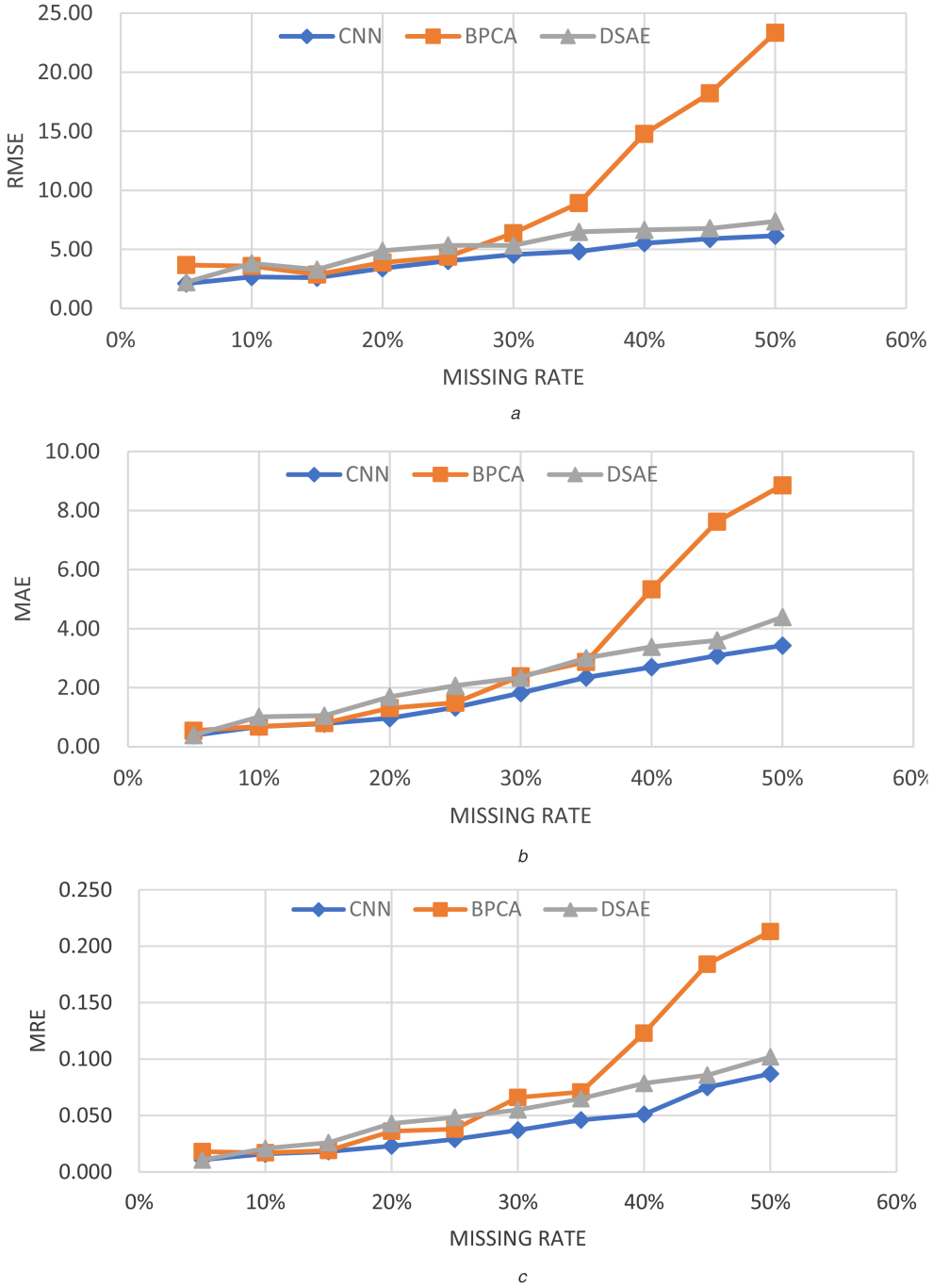


Fig. 4 *Error comparison with different criterions*
(a) RMSE result, ***(b)*** MAE result, ***(c)*** MPE result

The RMSE, MAE and MPE curves of the three approaches are shown in Fig. 4. In order to reduce the influence of the different selections of missing regions, the mean operation is done on ten evaluation results for the same method and the same missing ratio. However, the missing regions are randomly chosen for each evaluation. According to Fig. 4, the overall trends of the three curves are similar. When the missing ratio is <30%, it is difficult to tell which approach is the best and they have similar errors. Also, it appears that random disturbances can greatly influence the final results. In our experiment, the BPCA has a better performance than the DSAE. This is because BPCA mainly focuses on previous information at the same detector and will not be affected by nearby detector data. Additionally, the Bayesian model can correct errors on a general level. When the missing ratio is low, the temporal information plays a major role in imputation. The spatial information plays a minor role and might even become a disturbance. Although the proposed approach also makes use of spatial information, it can also extract higher-level features than the DSAE, which means that it can filter out portions of useless information. As the missing ratio increases, the error of the BPCA increases much faster than the other two approaches. The temporal information itself is not sufficient for data imputation when there is a lot of data missing. Therefore, the DSAE and the CNN start to show the strength of their data imputation approaches due to them combining both temporal and spatial information. Ultimately, CNN still performs better than the DSAE in almost all cases.

The imputation results using the CNN-based approach are shown in Fig. 5 with three missing ratios of 0, 10, and 40%. To make these figures easy for configuration, the loop detector number is from 0 to 49 and the time stamp number of 50 on 20 January 2016. The *x*-axis indicates the loop detector number corresponding to the spatial information and the *y*-axis indicates the time stamp. These figures use different colours to show the traffic volume, where the bright colour means that the traffic volume is high, and the dark colour means that the traffic volume is low. The three images in Fig. 5 are very similar in appearance, which indicates that our approach imputes the data well on a general level.

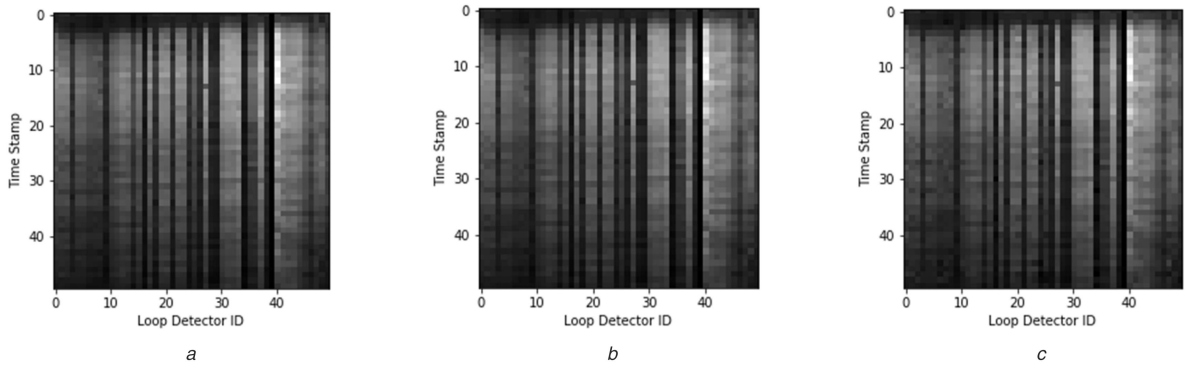


Fig. 5 *Imputation results with CNN-based approach*
(a) Actual traffic volume data, ***(b)*** Imputation traffic volume data with 10% missing ratio, ***(c)*** Imputation traffic volume data with 40% missing ratio

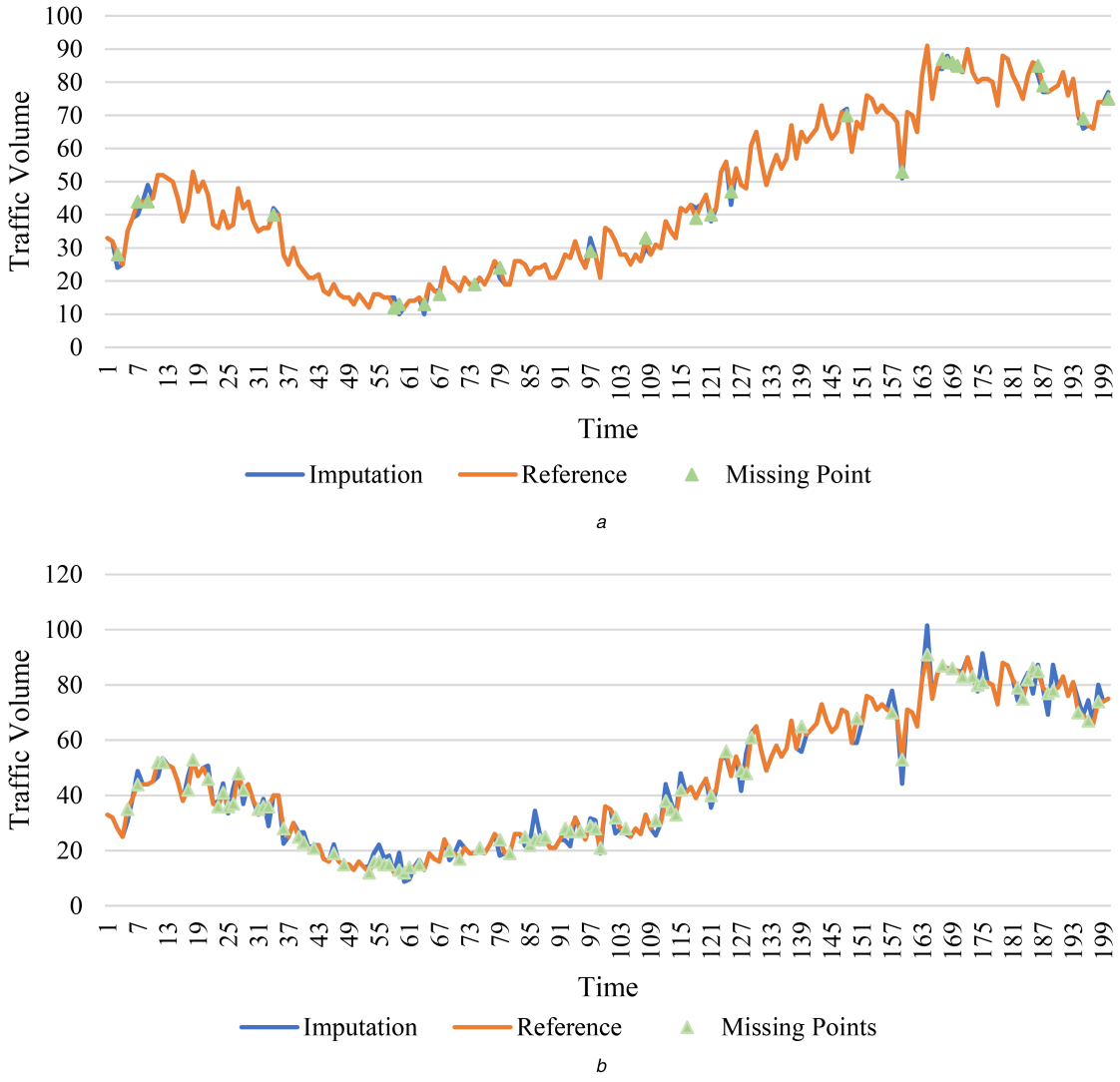


Fig. 6 *Imputation results of CNN-based approach*
(a) 10% Missing ratio, ***(b)*** 40% Missing ratio

To see the detail of the imputation results, the data of loop detector No. 50 with time stamp number of 100 is extracted from one column of the preceding imputation results. Fig. 6 shows the imputation results when the missing ratio is 10 and 40%. The blue line represents the imputed data points, and the orange line represents the original data points. Similarly, the imputation results applying the DSAE and BPCA approaches are shown in Figs. 7 and 8. Hence, the *x*-axis shows the time stamp whose interval is 5 min. The *y*-axis shows the traffic volume and its unit is *vehicles* (observed in a 5 min interval). The imputation results of the CNN-based approach show that it can impute the missing portion of the data well and recover it in a manner such that values are close to the actual ones observed in the ground truth data. Further, the distribution of errors on each missing point is stable, and there are no outliers.

Similar to the CNN-based approach, the DSAE approach also shows good stability in error distribution. However, the average and variance of the error are larger than those for our approach, partly because the DSAE approach can only extract low-level features. The texture of the traffic volume image has a strong pattern which means the texture will change regularly. Therefore, the DSAE is more likely to be influenced by noise and its error will be greater than those for the CNN-based approach on average.

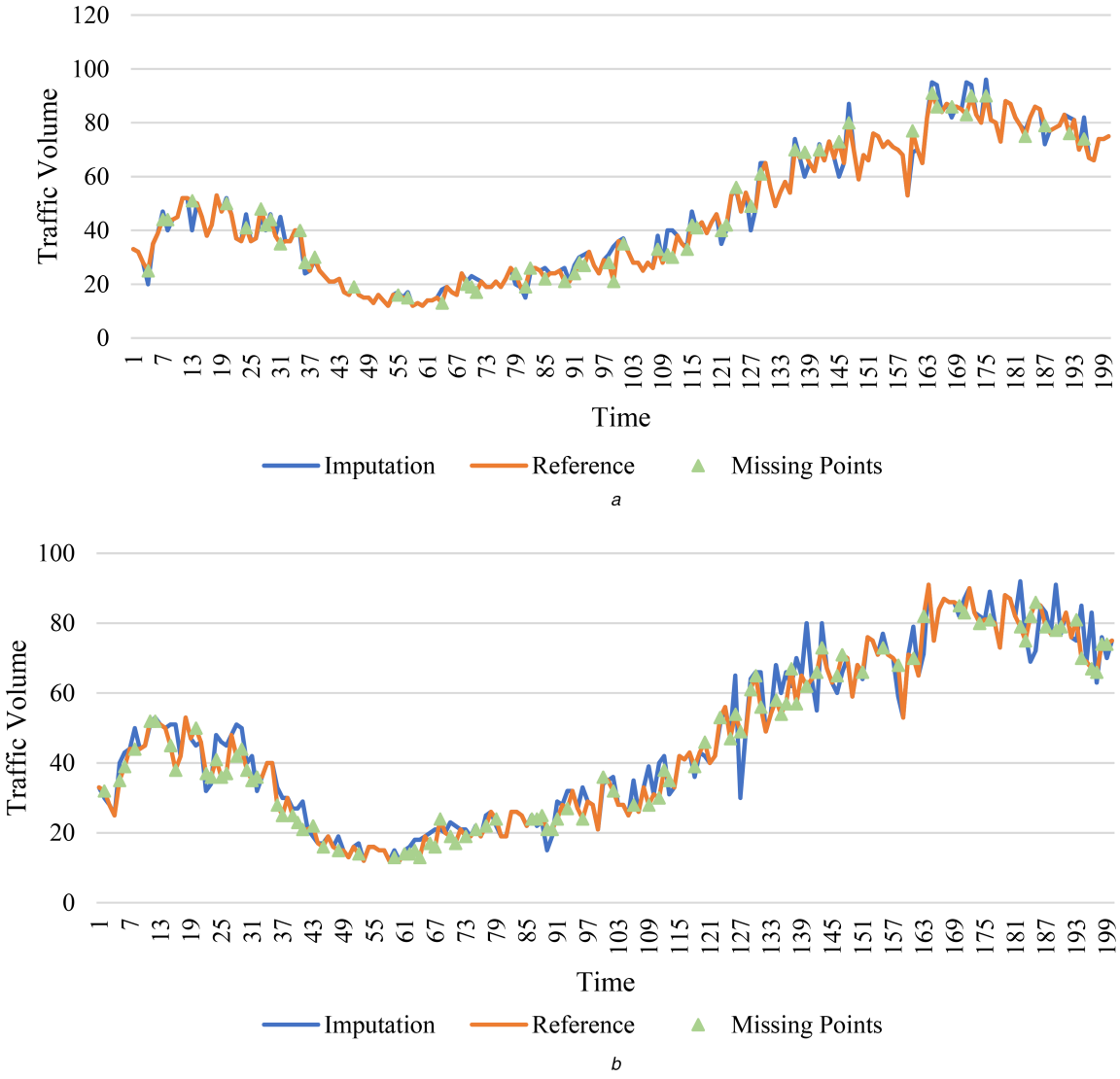


Fig. 7 *Imputation results of DSAE approach*
(a) 10% Missing ratio, ***(b)*** 40% Missing ratio

The BPCA shows greater error value earlier in the time series because the method mainly depends on previous data. In the beginning, the amount of previous data is small, leading to a lack of information available from which to impute the next data point. As the time series moving forward, the amount of previous data increases and helps to improve the imputation accuracy of the Bayesian model. However, when there is a large amount of missing data in the dataset, this approach still does not perform well. Also, its imputation accuracy also depends on the Bayesian model. It is hard to decide whether this model is most suitable for one specific situation or can be generalised. Ultimately, when the situation changes, the Bayesian model may change as well, and thus, it is not generally applicable.

5 Conclusions

This paper proposed an innovative method to solve the traffic data imputation problem with different missing ratios from 5 to 50% based on a deep learning method, which displayed an impressive result especially when the missing ratio is higher than 30%. This method considered the traffic volume data as spatial-temporal images and missing data as the blank region in the image. An image inpainting method based on CNN was developed to reconstruct the spatial-temporal data matrix. The basic idea was extracting high-level features of incomplete data and then reconstructing the complete data with the same size based on these features. The CNN-based approach could gain high-level features and filter out disturbances. Comparisons between the proposed CNN-based approach and the state-of-the-art approaches, i.e. the BPCA and another deep learning method known as DSAE have been made on the same dataset under the same condition. The results showed that the CNN-based method achieved the best performance with the highest accuracy and the most stable error distribution under three error criteria. When the missing ratio was low, these three methods had similar error rates. However, when the missing ratio reaches over 30%, the deep learning methods worked significantly better, and our proposed method performed even better than the DSAE method. As a result, the proposed method was able to play an important role in the field of traffic data imputation fields and benefit a variety of ITS applications. In future studies, in addition to traffic data imputation, the authors will extend this approach to work for more challenging tasks such as short-term data prediction. Based on historical data and nearby detector data, this approach is believed to have the potential to forecast traffic flow with high accuracy.

6 Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (grant no. 51329801) and the Shenzhen Science and Technology Planning Project (grant no. GJHZ20150316154158400). The authors appreciate John Ash's effort in polishing the languages of this paper.

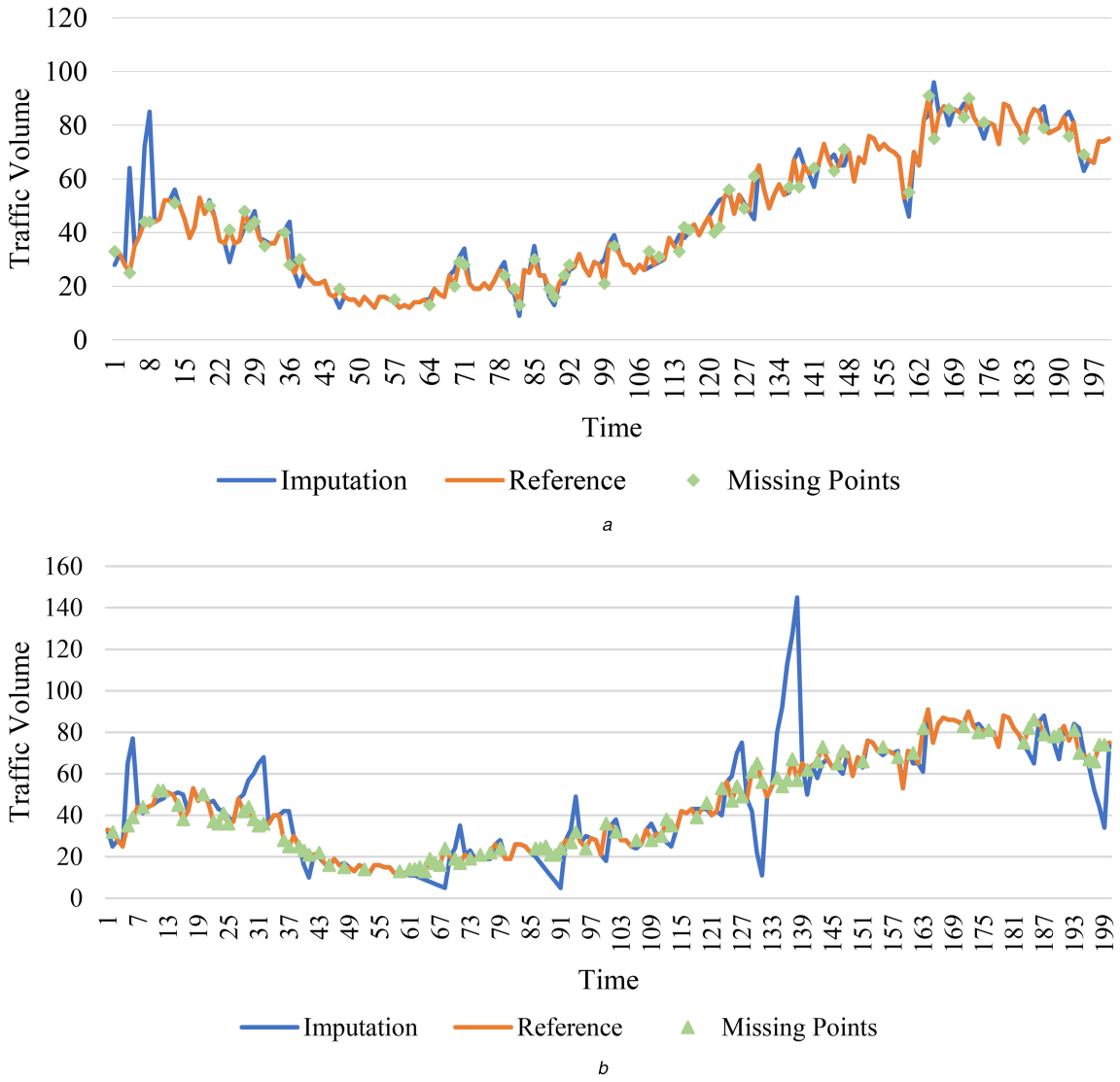


Fig. 8 *Imputation results of BPCA approach*
(a) 10% Missing ratio, ***(b)*** 40% Missing ratio

7 References

[1] Laharotte, P.A., Billot, R., Come, E., *et al.*: 'Spatiotemporal analysis of bluetooth data: application to a large urban network', *IEEE Trans. Intell. Transp. Syst.*, 2015, **16**, (3), pp. 1439–1448

[2] Rodrigues, J.G.P., Aguiar, A., Vieira, F., *et al.*: 'A mobile sensing architecture for massive urban scanning'. Int. IEEE Conf. Intelligent Transportation Systems, Washington, DC, USA, 2011, pp. 1132–1137

[3] Płaczek, B.: 'Selective data collection in vehicular networks for traffic control applications', *Transp. Res. C*, 2012, **23**, (7), pp. 14–28

[4] Ke, R., Li, Z., Kim, S., *et al.*: 'Real-time bidirectional traffic flow parameter estimation from aerial videos', *IEEE Trans. Intell. Transp. Syst.*, 2017, **18**, (4), pp. 890–901

[5] Ke, R., Lutin, J., Spears, J., *et al.*: 'A cost-effective framework for automated vehicle-pedestrian near-miss detection through onboard monocular vision'. The IEEE Conf. Computer Vision and Pattern Recognition (CVPR) Workshops, Honolulu, HI, USA, 2017, pp. 25–32

[6] Li, L., Su, X., Zhang, Y., *et al.*: 'Trend modeling for traffic time series analysis: an integrated study', *IEEE Trans. Intell. Transp. Syst.*, 2015, **16**, (6), pp. 3430–3439

[7] Wang, F.Y.: 'Parallel control and management for intelligent transportation systems: concepts, architectures, and applications', *IEEE Trans. Intell. Transp. Syst.*, 2010, **11**, (3), pp. 630–638

[8] Turner, S., Albert, L., Gajewski, B., *et al.*: 'Archived intelligent transportation system data quality: preliminary analyses of San Antonio TransGuide data', *Transp. Res. Rec. J. Transp. Res. Board*, 2001, **1719**, (1), pp. 77–84

[9] Chen, C., Kwon, J., Rice, J., *et al.*: 'Detecting errors and imputing missing data for single-loop surveillance systems', *Transp. Res. Rec. J. Transp. Res. Board*, 2003, **1855**, (1855), pp. 53–57

[10] Smith, B., Conklin, J.: 'Use of local lane distribution patterns to estimate missing data values from traffic monitoring systems', *Transp. Res. Rec. J. Transp. Res. Board*, 2002, **1811**, (1), pp. 50–56

[11] Qu, L., Li, L., Zhang, Y., *et al.*: 'PPCA-based missing data imputation for traffic flow volume: a systematical approach', *IEEE Trans. Intell. Transp. Syst.*, 2009, **10**, (3), pp. 512–522

[12] Tan, H., Feng, G., Feng, J., *et al.*: 'A tensor-based method for missing traffic data completion', *Transp. Res. Part C Emerg. Technol.*, 2013, **28**, (3), pp. 15–27

[13] Li, Y., Li, Z., Li, L.: 'Missing traffic data: comparison of imputation methods', *IET Intell. Transp. Syst.*, 2014, **8**, (1), pp. 51–57

[14] Ghosh, B., Basu, B., O'Mahony, M.: 'Bayesian time-series model for short-term traffic flow forecasting', *J. Transp. Eng.*, 2007, **133**, (3), pp. 180–189

[15] Lee, S., Fambro, D.: 'Application of subset autoregressive integrated moving average model for short-term freeway traffic volume forecasting', *Transp. Res. Rec. J. Transp. Res. Board*, 1999, **1678**, (1), pp. 179–188

[16] Castro-Neto, M., Jeong, Y.S., Jeong, M.K., *et al.*: 'Online-SVR for short-term traffic flow prediction under typical and atypical traffic conditions', *Expert Syst. Appl.*, 2009, **36**, (3), pp. 6164–6173

[17] Yin, W., MurrayTuite, P., Rakha, H.: 'Imputing erroneous data of single-station loop detectors for nonincident conditions: comparison between temporal and spatial methods', *J. Intell. Transp. Syst.*, 2012, **16**, (3), pp. 159–176

[18] Zhong, M., Sharma, S., Liu, Z.: 'Assessing robustness of imputation models based on data from different jurisdictions: examples of Alberta and Saskatchewan, Canada', *Plos One*, 2005, **1917**, (1), pp. 116–126

[19] Troyanskaya, O., Cantor, M., Sherlock, G., *et al.*: 'Missing value estimation methods for DNA microarrays', *Bioinformatics*, 2001, **17**, (6), p. 520

[20] Chang, G., Zhang, Y., Yao, D.: 'Missing data imputation for traffic flow based on improved local least squares', *Tsinghua Sci. Technol.*, 2012, **17**, (3), pp. 304–309

[21] Ni, D., Leonard, J.II: Markov chain Monte Carlo multiple imputation using Bayesian networks for incomplete intelligent transportation systems data. *Transp. Res. Rec. J. Transp. Res. Board*, 2005, **1935** (1): pp. 57–67

[22] Li, L., Li, Y., Li, Z.: 'Efficient missing data imputing for traffic flow by considering temporal and spatial dependence', *Transp. Res. C, Emerg. Technol.*, 2013, **34**, pp. 108–120

[23] Duan, Y., Lv, Y., Kang, W., *et al.*: 'A deep learning based approach for traffic data imputation'. IEEE, Int. Conf. Intelligent Transportation Systems. 2014, pp. 912–917

[24] Tak, S., Woo, S., Yeo, H.: 'Data-driven imputation method for traffic data in sectional units of road links', *IEEE Trans. Intell. Transp. Syst.*, 2016, **17**, (6), pp. 1762–1771
[25] Telea, A.: 'An image inpainting technique based on the fast marching method', *J. Graph. Tools*, 2004, **9**, (1), pp. 23–34
[26] Yang, C., Lu, X., Lin, Z.*, et al.*: 'High-resolution image inpainting using multi-scale neural patch synthesis'. The IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2017
[27] Pathak, D., Krahenbuhl, P., Donahue, J.*, et al.*: 'Context encoders: feature learning by inpainting'. The IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2536–2544
[28] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M.*, et al.*: 'Generative adversarial nets'. Int. Conf. Neural Information Processing Systems, Montreal, Canada, 2014, pp. 2672–2680
[29] Krizhevsky, A., Sutskever, I., Hinton, G.E.: 'Imagenet classification with deep convolutional neural networks'. Int. Conf. Neural Information Processing Systems, Stateline, NV, USA, 2012, pp. 1097–1105
[30] Dosovitskiy, A., Springenberg, J.T., Brox, T.: 'Learning to generate chairs with convolutional neural networks'. Computer Vision and Pattern Recognition, Boston, MA, USA, 2015, pp. 1538–1546