International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7 Issue-5s, January 2019

A Study on Imputation Methods for Vehicle Traffic Data

S. Narmadha, V. Vijayakumar

Abstract: ***Quality traffic data is an essential for traffic related researches and developing transport applications. These data is collected from inductive loops, sensors, radars, cameras, mobile GPS (Global positioning system) and microwave sensors etc. Harsh environment, malfunctions of detectors, hardware, software and communication device failures leads the problem of traffic data loss. It will greatly reduce the predicting performance of the traffic volume data. It affects the important process of intelligent transportation systems. Imputation methods are used to find the missing data and produce as a complete data. Many imputation methods have been proposed in existing to estimate traffic analysis. In this paper imputation methods were discussed and analyzed.***

Keywords: Imputation, Missing data, Quality, Prediction, Deep learning.

I. INTRODUCTION

Cloud computing, internet of things, Floating car devices, social networks are the new technologies leads to era of big data for the past few years. State and provincial highway departments collect and maintain daily traffic data. The traffic data contains flow, volume, speed, occupancy, travel time, density, vehicle information, emission level.Transport agencies often reporting both health percentage of loop detectors and range of missing values [8]. Transportation agencies are required to report on various annual traffic statistics such as average annual daily traffic (AADT), speed, vehicles miles travelled (VMT) and missing rate of data. Many researches are available on how transportation practitioners handled missing values in their traffic data [5]. Missing values are usually represented by zero or blanks in traffic counter files. There are many reasons for the machine failures, including power surges, lighting, loss of battery power, solar panel failure, vandalism, and environmental effects such as storm, fog, frost heave and rain [8]. Traffic information management and analysis systems are currently suffering from poor quality data and missing data. Missing data brings great troubles for further utilization of traffic systems [16]. The traffic flow, speed, density, weather, weekdays, weekends, holidays are important parameters and essential for the planning, design and operation of urban traffic systems. If any data is missing or poor quality, further process of analysismust be degraded. Imputation methods are used to reduce the impact of incomplete data on utilization.

Manuscript published on 30 January 2019.
***Correspondence Author(s)
S. Narmadha, Research Scholar, Sri Ramakrishna College of Arts and Science, Coimbatore & India.
Dr. V. Vijayakumar, Professor, Sri Ramakrishna College of Arts and Science, Coimbatore & India.

The research on missing and poor quality data management in intelligent transportation systems has been emerged due to extensive use by traffic engineers and researchers [1]. The advanced traffic signal control system designed with optimal control schemes to improve traffic conditions and relieve traffic congestion. Global positioning system and electronic tolling devices also applied in traffic data collection. The data, volume and occupancy of particular time can be obtained, and average speed can also be estimated subsequently by vehicle count and occupancy rate with loop detector. The detecting data are not only the use base for achieving the functions of the intelligent transportation system but are also important to transportation management and planning [7].

The remaining of the paper is discussed as follows. Section II discusses about missing traffic data. Section III presents about the imputation methods. Section IV discusses performance measure used for compare the models. Section V compares the methods of imputation. Conclusions are described in Section VI.

II. MISSING TRAFFIC DATA

Missing traffic data can be affected in two categories. First one is loss of data at certain locations and time periods. Complete data is important for analysis in transportation modelling and prediction. For example if speed or volume of traffic data are missing in heavy congested road at peak hours, the vehicle emission levelwill be under estimated. Second one is statistical information loss i.e primary assumptions of statistical methods used in an imputation process are violated by missing traffic patterns which is giving optimal solutions [2].Generally there are three type of missing data are occurring [10],

1. Data missing are completely at Random (MCR), which specifies the missing data entirely independent. So the missing data appear as isolated points and distributed randomly [10].
2. Missing data at random (MR) are related with their neighbouring points. So missing data appear as a small cluster of successive points and data lost at the unique time. But this is a random distribution [10].
3. Not missing at Random (NMR) is the reason of long time failure of detectors. It gives less availability of data [10].

Dynamic and ability to impute large data set is a challenge for imputation methods.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

III. IMPUTATION METHOD

Imputation method can be classified into three categories [16]. It is represented in fig (1).

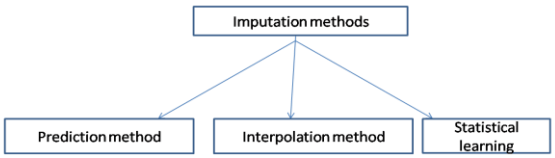


Fig 1: Imputation methods

Prediction method replaces the missing value by historical data. Interpolation method replaces by history or neighboring data points. Statistical learning fills by statistical dependencies [14, 10].

In existing methods of imputation Auto-Regressive Integrated Moving Average (ARIMA) and feed forward prediction based methods (FFNN) was used [3]. Past historical value is taken as a input parameter to predict missing data. If traffic data missed continuously, prediction based methods is not sufficient to provide the accurate results for imputation [3].

Interpolation method is replaced by an average and weighted average of neighbour for handling missing value. Two types of neighbouring data are there, which is temporal neighbouring and pattern neighbouring. Temporal neighbouring datais collected in the nearby days of same location and same time period.In pattern neighbouring data is collected from similar data flow variation of the days at the same time and same period[3].

Statistical learning methods normally used observed data to identify the corrupted data or missing data in a repetitive manner. From the observed data probability distribution is calculated for best value. It will be used to fill the missing values [10].

- Vector based methods are deals with small amount of missing data and handles spatial temporal correlation [6]. Spline or regression imputation methods, and autoregressive integrated moving average (ARIMA) models are some of the historical methods are repeatedly used in traffic data imputation [6].
- Matrix based method is combined with two mode traffic date such as Bayesian Principal Component Analysis and probalistic principal component analyses. This method has been proved to be a more effective with higher accuracy than vector based methods [6].
- Tensor based method, traffic volume data can be represented as multidimensional tensor pattern to keep the multidimensional characteristics and spatial temporal correlations.It can capture global information than matrix completion methods due to the intrinsic multi way characteristics of tensor model [6].

The inherent relations of the traffic flow are used in the above imputation methods. Traffic data with spatial and temporal correlation is the base and most important for many imputation algorithms.

A. Principal Component Analysis

Huachun Tan et al. [6] recommended tensor based method from the characteristics of multi correlations and principle components of the traffic data. Tensor based imputation algorithms such as EM-Tucker3 and CP-Wopt were chosen to authenticate the performance of all methods with PEMS dataset. It shown tensor based method is superior to all others. To validate the traffic volume data quantitatively, Pearson correlation coefficient and similarity coefficient is used. In this work the required number of components or the rank in each mode to reconstruct the original data has not fully explored from the view of principal components. It needs to be taken into future. [6].

Ningyu Zhao et al. [10] used Probabilistic Principal Component Analysis (PPCA) and Mixed Probabilistic Principal Component Analysis (MPPCA) methods to impute traffic data. MPPCA meets the non linear characteristics of traffic imputation data. The spatial and temporal information for traffic data collected from multiple neighbouring detectors to improve data imputation accuracy. The temporal and spatial information needs to process more traffic data but takes more time. So we must determine the best time in advance. If so, spatial temporal information from multiple detectors is the good choice for improving the accuracy of imputation process.

Li Qu et al. [8] proposed a method probabilistic principal component analysis (PPCA) to find the missing traffic volume data based on historical data. PPCA method had predominantly use of periodicity, local predictability, as well as statistical information of traffic flow. And also it doesn't require strict daily flow data and large enough database. Check whether non liner PCA can apply, and how to accurately combine PCA and PPCA for simultaneous abnormal detection In future.

B. Tensor Completion

Bin Ran et al. [1] were deal the missing data by traffic speed data instead of traffic volume data. Traffic speed data is same as traffic volume data both has same characteristics of spatial and temporal features. Imputation methods of traffic volume data can be directly applied to the traffic speed data. Tensor completion method is used to deal with noisy entries of traffic speed data. To estimate the missing entries in the traffic speed data, a high accuracy low rank tensor completion algorithm called HaLRTC is used, which can deal with noisy observed entries [1].

C. Fuzzy Method

Qiang Shang et al. [11] proposed a hybrid method which using fuzzy C-means (FCM) by a combination of particle swarm optimization (PSO) algorithm and support vector machine (SVR). Patterns of missing traffic data of urban road were analysed and matrix based structure is used to represent the missing traffic data. The "day pattern," "week pattern," "link pattern," and "section pattern" of traffic flow data are taken into account, and matrix-based structure is used to represent the missing traffic data. The results can be achieved by spatial temporal correlation of traffic flow data. Jinjun Tang et al. [7] developed an imputation method based on Fuzzy C-means. This method imputed missing traffic data in loop detectors.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

The vector based data structure of spatial temporal correlation transformed into matrix based data pattern which represent multi attributes. Genetic algorithm is applied to optimize the parameters of clusters size and weighting factor in FCM model. Each missing value of a particular member function has more than one cluster centroid in the FCM. It leads to successful imputation results.

D. Cluster Method

Wei Chiet Ku et al. [14] proposed a data-driven imputation method that makes full use of the spatial and temporal relationships existing between the traffic flows of multiple road segments and correlated each other. The K-means clustering technique is used to cluster road segments with similar traffic flow patterns. A deep-learning based stacked denoiseautoencoder was constructed for each group of road segments to extract their spatial-temporal relationships. After the model is trained in a layer-wise greedy fashion, it is able to collectively estimate the missing data from multiple locations under a unified framework.

HossamAbdelgawad et al. [5] presented an iterative approach based on the integration method (an imputation algorithm and a time series model) to impute missing data of permanent data collection stations (PDCV), Ontario, Canada. Traffic volume data collected using PDCS stations are often missing for periods of time ranging from an hour to several days, or may be a weeks or months. Imputation algorithms is used to impute the small gaps of data and form a complete dataset, then time serious models is used to fill the large gaps. This method imputes the missing values with seasonal variation curves.

E. Support Vector Machine

Yang Zhang et al [15] introduced a least squares support vector machines (LS-SVMs) to predict missing traffic flow based on spatio-temporal analysis in urban arterial streets. Computational intelligence technique is incorporated with state space approach in missing traffic data imputation. The regression model is applied to impute false and missing traffic data occurring at some time-spots in time-series detector data.But this method is effective only for limited input data. In future, need a potential tool for large traffic data imputation.

F. Cokrige Method

BumjoonBae et al. [2] proposed cokriging methods. The interpolated values are modeled by a Gaussian process, using multiple data sources (independent) and microwave sensors. The interpolation value is calculated as a weighted sum of their known neighbour's values.Unobserved location vehicle speed is estimated using four different kriging methods: simple kriging (SK), ordinary kriging (OK), ordinary cokriging (OCK), and simple cokriging (SCK). Further the work can be carried out using simulated data. it has no original missing values.

G. Deep Learning

Traffic data imputation is the process of corrupted data recovering and it should fill in the corrupted or missing data points automatically [16]. Deep learning is used for prediction of traffic flow, but also used for imputation process. More layers of architecture can learn the latent representation of deep network. Auto encoder, recurrent neural network are some of the deep learning networks used for imputation purpose [13, 10].

Wei Li et al. [13] proposed a time serious processing method that can construct a two dimensional pattern which can be easily used for traffic data imputation and state identification. The Recurrent Neural Network type of long short term memory used to train the time serious data as it is capable of learning long-term dependencies and memorizing long historical input data. This time serious pattern fuses both spatial and temporal data.

YanjieDuan et al. [16] developed deep learning based an imputation approach. This approachrepresents the traffic data as a whole data item which includes the observed data and missing and restores the complete data with the deep structural network. Denoiseautoencoder is used for removing noise and fill the missing values. Pretraining and fine tuning methods are used for obtaining complete data. The traffic data structures and imputation patterns can be various in real times. The complex structures of deep network can be expected in future for imputation.

YanjieDuan et al. [17] proposed a denoising stacked autoencoder model (DSAE) for imputation. It contains autoencoder(AE) and denoiseautoencoder(DAE). Autoencoder can extract the features from input data. DAE can capture the statistical dependencies between theinputs. Different locations for spatial and weekdays, weekends used for temporal information. California based pems dataset used for training and testing process. Input with missing values mapped into corrupted version, in which both CC (continuous corruption) and RC (Random corruption) may occur in corrupted vectors. Deep learning can deserves the further investigation on developing new structures of deep learning models for imputation.

IV. PERFORMANCE MEASURE

To evaluate the imputation approach normally used some criterions to measure the error of the imputed data.

(i) Root Mean squared Error:

$$RMSE = \frac{1}{n}\sum_{i=1}^{n}(x_i^r - y_i)$$

(ii) Mean absolute error:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_i^r - y_i\right|$$

(iii) Mean relative Error

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|x_i^r - y_i\right\|}{x_i^r}$$

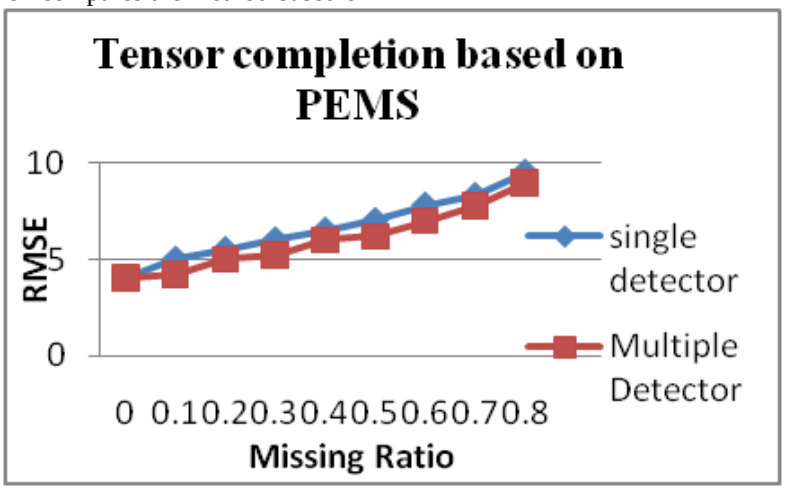
These criteria's are performed at three stages, such as missing at Random, Missing completely at random, Not Missing at Random. Week days, weekends, both weekend and week days, Station wise detectors data are used for performing the imputation. Above the all methods deep learning provides the good performance results than others.

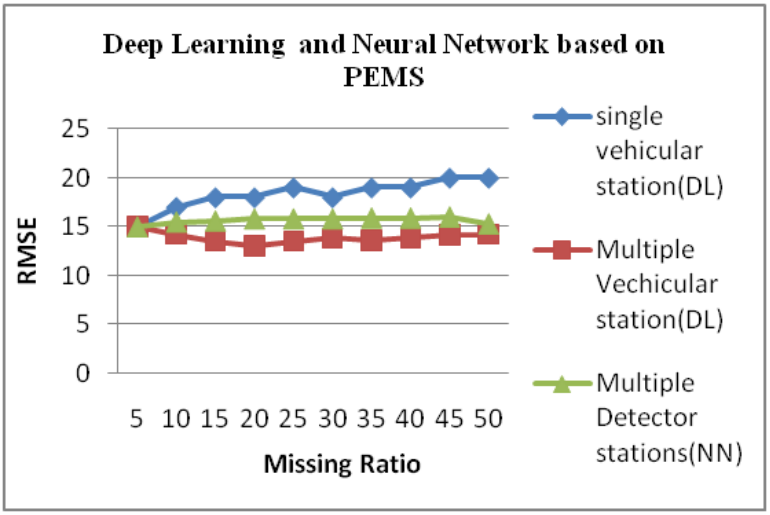
Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

V. COMPARISION OF METHODS

The characteristics of road transportation frequently changes due to external factors like vehicle flow, rain, dew, wind etc. Imputation method must be chosen carefully, because accurate and reliable data is important for ITS services such as travel time estimation, prediction, forecasting etc. Without reliable dataset traffic prediction and forecasting models will be useless and also less effective one [17]. Data can be collected by some transport management systems like PEMS which is aggregate the collected data into 5 or 10 mins. Else may collect as a raw data from permanently fixed sensors. It doesn't have an aggregate value. This section compares the method based on dataset and type of imputation methods used in the imputation process for traffic data.

Most of the research works done using California related PEMS (performance measurement system) dataset. The data missing rate of the freeway Performance Measurement System (PeMS) in Los Angeles County was found to be 15% [4]. The comparison chart is based on PEMS dataset. In this weekday is taken as a temporal. Coefficient and single detector station and multiple detector station are taken as a spatial factor for testing the imputation performance. Based on this multiple detector and deep learning based techniques provide the best result.





VI. CHALLENGES AND ISSUES

- Imputing large data with accuracy[1]
- Dynamic data imputation[16][17]
- Consideration of spatio temporal feature is a challenging in imputation[18]
- Multiple location at various time data imputation[17]
- Current imputation methods performed only in loop detector data. Others sources like microwave data is a crucial one.[7]

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

S. No	Imputation Method	Dataset	Advantages	Disadvantages & Future study
1	PPCA-Based Missing Data (Li Qu et al.(2009)	Traffic volume data collected in Beijing	Doesn't need strictly daily flow similarity, large database, no continuous flow of data points	Want to check whether non linear PCA should applied, how to combine PCA and Robust
2	Tensor Completion (Bin Ran et al.(2015)	Pems(Performance measurement systems) http://pems.dot.ca.gov	Multiple correlations of the traffic data improve the performance.	Can apply for large and dynamic network
3	FCM and GA (Jinjun Tang et al. (2015)	Section of the motorway in Harbin, China.	Estimate the missing values in Multi attribute data.	The MCR and NMR patterns need to be considered. Handle Incomplete training data. Other Detectors such as micro wave and infrared missing data should be imputing in future.
4	Tensor based traffic imputation method (Huachun Tan et al.(2013)	Pems(Performance measurement systems) http://pems.dot.ca.gov	Higher multi mode correlations produce better performance.	essential number of components to reconstruct the original data has not fully explored
5	Imputation based on FCM(Qiang Shang et al.(2018)	Collected from loop detectors. it is located in North and South Elevated Expressway, Shanghai, China.	Improves the accuracy	Other traffic flow features such as traffic flow speed, travel time, occupancy and other factors data such as accidents and weather want to be used for proposed method.
6	Clustering based deep learning approach. Wei Chiet Ku et al.(2016)	Dataset obtained from Quantum Inventions, Singapore. http://cms.quantuminventions.com	Imputation performance is robust under high missing rates.	Larger and complex network will be evaluated. Determine the optimal parameters of the autoencoder such as no of neuron and no of hidden layer.
7	Deep learning based approach. YanjieDuan et al. (2014)	Pems(Performance measurement systems) http://pems.dot.ca.gov	Improves the accuracy	Need multi pattern imputation method(i.e 3 dimensional-multiple period and multi location)
8	Deep learning based imputation. Yanjieduan et al.(2016)	Pems(Performance measurement systems) http://pems.dot.ca.gov	Improves the accuracy	Develop new structures of deep learning model for imputation.
9	Kriging based data imputation method Hongtai Yang et al. (2018)	Raw data rather than aggregated, collected from ong Ellington Parkway in Nashville, Tenness	Higher imputation accuracy	Need to compare with other methods except K-NN and historical average method.
10	Modified K-Nearest neighbor method (SehyunTak et al. (2016)	Data collected from highways, korea.	Give better performance for all missing types, missing ratios, traffic states and day types.	In future want to test the performance of imputation method with accident data which have uncertainty , how to apply for arterial road

Table 1:Comparison of imputation methods for traffic data

VII. CONCLUSION AND FUTURE DIRECTIONS

Noise is the one of the challenging issue in the design and analysis of Intelligent Transportation systems. Many methods were proposed to impute the missing traffic flow data. In this paper the recent methods of noise removal for traffic data are studied and made comparison. Though much of technologies developed missing data issues is still continues and this leads to the problem of traffic data based applications such as traffic forecasting, dynamic route guidance and real time incident detection. Powerful tools need to be developed to handle imputation for large data of dynamic road network, because existing handles only limited input data. Recent studies imputes data comes from loop detectors only. In future may impute data from microwave and infrared facilities. Nowadays deep learning is an emerging technique to impute the traffic data effectively. More complex and powerful deep structures need to be investigated in the field of data imputation.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

ACKNOWLEDGEMENT

We thank Sri Ramakrishna College of arts and science for giving support for doing this research work.

REFERENCES

1. Bin Ran, Huachun Tan, JianshuaiFeng, Ying Liu, and WuhongWang, “Traffic Speed Data Imputation Method Based on Tensor Completion”, Computational Intelligence and Neuroscience, Article ID 364089, 9 pages, Hindawi, 2015.
2. BumjoonBae,, Hyun Kim, Hyeonsup Lim, Yuandong Liu, Lee D. Han, Phillip B. Freeze, “Missing data imputation for traffic flow speed using spatiotemporal cokriging”, Transportation Research Part C, 124–139, ELSEVIER, 2018.
3. Gang Chang, TongminGe, “Comparison of Missing Data Imputation Methods for Traffic flow”, International Conference on Transportation, Mechanical, and Electrical Engineering (TMEE), IEEE Xplore, 2011.
4. Hongtai Yang, Jianjiang Yang, Lee D. Han, Xiaohan Liu , Li Pu , Shih-miao Chin , Ho-ling Hwang, “A Kriging based spatiotemporal approach for traffic volume data imputation”, . PLoS ONE 13(4): e0195957, 2018.
5. HossamAbdelgawad, Tamer Abdulazim, BaherAbdulhai, AlirezaHadayeghi, and William Harrett, “Data imputation and nested seasonality time series modelling for permanent data collection stations: methodology and application to Ontario”, Can. J. Civ. Eng. **42**: 287–302, NRC research press, 2015.
6. Huachun Tan, Zhongxing Yang, Guangdong Feng, Wuhong Wang, Bin Ran, "Correlation Analysis for Tensor-based Traffic Data Imputation Method", 13th COTA International Conference of Transportation Professionals (CICTP 2013), ELSEVIER, 2013.
7. Jinjun Tang, Yinhai Wang, Shen Zhang, Hua Wang, Fang Liu, and Shaowei Yu, “On Missing Traffic Data Imputation Based on Fuzzy *C*-Means Method by Considering spatial–Temporal Correlation” ,Journal of the Transportation Research Board, pp.86–95, Washington, D.C., 2015.
8. Li Qu, Li Li, Yi Zhang, and Jianming Hu, “PPCA-Based Missing Data Imputation for Traffic Flow Volume: A Systematical Approach”, IEEE transactions on intelligent transportation systems, vol. 10, no. 3, september 2009
9. Ming Zhong, and Satish Sharma, “Development of Improved Models for Imputing Missing Traffic Counts”, The Open Transportation Journal, 2009, *3*, 35-48.
10. Ningyu Zhao, Zhiheng Li, Yuebiao Li, “Improving the Traffic Data Imputation Accuracy Using Temporal and Spatial Information”, International Conference on Intelligent Computation Technology and Automation. 978-1-4799-6636-3/14 IEEE, 2014.
11. Qiang Shang ,Zhaosheng Yang , Song Gao, and Derong Tan, “An Imputation Method for Missing Traffic Data Based on FCM Optimized by PSO-SVR”, Journal of Advanced Transportation, Volume 2018, Article ID 2935248, Hindawi, 2018.
12. SehyunTak, Soomin Woo, and Hwasoo Yeo, Data-Driven Imputation Method for Traffic Data in Sectional Units of Road Links, IEEE Transactions On Intelligent Transportation Systems, IEEE, 2016.
13. Wei Li , Jianming Hu , Zuo Zhang ,and Yi Zhang, “A Novel Traffic Flow Data Imputation Method for Traffic State Identification and Prediction Based on Spatio-Temporal Transportation Big Data”, 17th COTA International Conference of Transportation Professionals, July 7–9, Shanghai, China, 2017.
14. Wei Chiet Ku, George R. Jagadeesh, AlokPrakash, ThambipillaiSrikanthan, “A Clustering-Based Approach for Data-Driven Imputation of Missing Traffic Data”, 2016 IEEE Forum on Integrated and Sustainable Transportation Systems (FISTS), Beijing, China, IEEE Xplore Digital library, 2016.
15. Yang Zhang, Yuncai Liu, “Missing Traffic Flow Data Prediction using Least Squares Support Vector Machines in Urban Arterial Streets”, 2009 IEEE Symposium on Computational Intelligence and Data Mining,IEEE Xplore Digital library, 2009.
16. YanjieDuan, YishengLv, Wenwen Kang, Yifei Zhao, “A Deep Learning Based Approach for Traffic Data Imputation”, 17th International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2014.
17. YanjieDuan, YishengLv , Yu-Liang Liu, Fei-Yue Wang, “An efficient realization of deep learning for traffic data imputation”, Transportation Research Part C 72, 168–181 Elsevier, 2016.
18. Yuebiao Li, Zhiheng Li, Li Li, “Missing traffic data: comparison of imputation methods”, IET Intelligent Transport Systems, , Vol. 8, Iss. 1, pp. 51–57, 2013.

AUTHORS PROFILE

S. Narmadha, Research Scholar, Sri Ramakrishna College of Arts and Science, Coimbatore & India.

Dr. V. Vijayakumar, Professor, Sri Ramakrishna College of Arts and Science, Coimbatore & India.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)