Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 184 (2021) 84–91

Procedia Computer Science

www.elsevier.com/locate/procedia

The 12th International Conference on Ambient Systems, Networks and Technologies (ANT)
March 23-26, 2021, Warsaw, Poland

Imputation of Missing Traffic Flow Data Using Denoising Autoencoders

Boyuan Jiang[a, b], Muhammad Danial Siddiqi[b], Reza Asadi[b], Amelia Regan[a, b, *]

[a]*Institute of Transportation Studies (ITS), University of California Irvine, Irvine, CA 92697, USA*
[b]*Donald Bren School of Information and Computer Sciences (ICS), University of California Irvine, Irvine, CA 92697, USA*

Abstract

In transportation engineering, spatio-temporal data including traffic flow, speed, and occupancy are collected from different kinds of sensors and used by transportation engineers for analysis. However, the missing data influence the analysis and prediction results significantly. In this paper, Denoising Autoencoders are used to impute the missing traffic flow data. In our earlier research, we focused on a more general situation and used three kinds of Denoising Autoencoders: "Vanilla", CNN, and Bi-LSTM, to impute the data with a general missing rate of 30%. The Autoencoder models are used to train on data with a high missing rate of about 80% in this paper. We demonstrate that even under extreme loss conditions, and Autoencoder models are very robust. By observing the hyper-parameter tuning process, the changing prediction accuracy is shown and in most cases, the three models maintain the original accuracy even under the worst situations. Moreover, the error patterns and trends concerning different sensor stations and different hours on weekdays and weekends are also visualized and analyzed. Finally, based on these results, we separate the data into weekdays and weekends, train and test the model respectively, and improve the accuracy of the imputation result significantly.

Keywords: Transportation data analysis; Spatio-temporal problem; Denoising autoencoder; Missing data imputation

* Corresponding author. Tel.: 01-949-824-2611
E-mail address: aregan@uci.edu

10.1016/j.procs.2021.03.122

1. Introduction

Spatio (Spatial) refers to space. Temporal refers to time. Spatio-temporal or spatial-temporal is used to describe data collected over both space and time [1]. Spatio-temporal problems have long been studied in transportation engineering where the data are collected by a great number of sensors over a long period. However, due to sensor malfunction, communication failure, or measurement error part of the data is inevitably missing or corrupted [2]. Not only will the presence of missing data influence the final result of traditional transportation data analysis, but also the performance of machine learning models used for clustering and prediction will drop precipitously. To overcome this challenge, data must be processed before use in such models. The imputation of missing data is one way of improving the accuracy of final results. Recent advances in deep learning models provide more capability in exploring spatio-temporal problems [3]. Denoising autoencoders are the most common deep learning models used for missing data imputation. We consider missing data imputation in traffic flow data, which is complex spatio-temporal data, using denoising autoencoders. In [4], the researchers examine the performance of denoising autoencoders for missing data imputation, however, their analysis is not for spatio-temporal data. In [5], a layer-wise pre-training of fully connected layers is proposed to impute missing traffic flow data. In [2], the researchers use denoising stacked autoencoders for missing traffic flow data imputation. However, they only consider fully connected layers, and they do not consider convolutional and recurrent neural networks. Moreover, multiple imputations of denoising autoencoders are examined in [6]. In [7], they consider a deep convolutional neural network for missing data imputation. In [8], they show the improved performance of a recurrent neural network for missing data imputation in time-series data, and in [9], a convolutional-recurrent neural network is proposed for missing data imputation of spatio-temporal data. Many of these works focus on proposing fully-connected, convolutional, and recurrent denoising autoencoders with better performance for missing data imputation. In this paper, we examine the performance of fully connected, convolutional, and bi-directional LSTM denoising autoencoders for missing traffic flow imputation. We explore various missing data ratios. We also explore variations of missing data imputation error for spatial and temporal contexts, as missing data imputation error varies over the spatial and temporal domain. Such an analysis illustrates the capability of implemented autoencoders in imputing missing data under various missing data ratios and spatial and temporal domains.

This paper builds on our earlier more general work. In that paper, we focused on a more general situation and used three kinds of Denoising Autoencoders to implement the data with a general missing rate of 30%. In this paper, we verified the robustness of Autoencoder models, examined the error patterns and trends, and also improved the accuracy of the imputation results.

1.1. Dataset

The traffic flow data extracted from the PeMS (Caltrans Performance Measurement System) [10] are used in this study. In the original dataset, traffic flow data are gathered every 30 seconds and aggregated every 5 minutes by using the loop detectors, magnetometers, and radars. The middle 8 stations data of the Bay Area Region of US 101-South are used in this research. The selected sensors have more than 99% of the available data for this period.

1.2. Problem definition

Spatio-temporal data is represented by a matrix $X \in \mathbb{R}^{s \times t \times f}$, where s is the number of sensors, t is the number of time steps and f is the number of features. In this research, traffic flow data and 8 stations are considered, so the s is 8 and f is 1, and the matrix is $\in \mathbb{R}^{8 \times t \times 1}$. In the machine learning prediction problem, the error between the true value and prediction value is used to judge the accuracy of each model. We begin pre-processing work for our raw datasets. Firstly, the missing part of the original dataset is filled by using adjacent values, and a “perfect” dataset is formed. Then, since missing data may exist at individual points or for some periods, artificial missing data of random lengths are formed in the data. Finally, the training and testing process are implemented, and RMSE and MAE are used to calculate the accuracy of each model.

1.3. Sliding window

To apply the neural network model for time series imputation, the sliding window is needed to generate the data point for each time stamp and the parameter "look back" is used to adjust the shape of each data point. The dimension of each data point is $1 \times (s \times f \times l)$ where l is the value of look back. In our case, this can be simplified to $1 \times (8 \times 1 \times l)$ due to the number of sensors and features that have been fixed to 8 and 1. For example, if we consider the previous 4-time stamps for each data point that means the l is 4 and that each data point will be the matrix $\mathbb{R}^{1\times32}$.

1.4. Models

In this paper, 4 different methods are used for imputation. Weekly-Hourly Average is an industry-wide accepted method to predict traffic flow that uses a temporal average value to fill the missing gap. "Vanilla" (FCN) only has one hidden layer, and it is a simple ML model with a Fully Connected Network. CNN (Convolutional Neural Network) uses convolutional layers and the 2D representation of data is passed. In Bi-LSTM (Bi-directional Long Short Term Memory), the original data is fed to the learning algorithm twice, once from beginning to the end and once from end to beginning, and it learns to train the model using previous information.

1.5. Results of low missing data rate

In our earlier research, we focused on the general situation to implement the data with a general missing rate of 30% [6]. The "Vanilla" model did very well in the imputation of missing data. It almost had the same accuracy as the Bi-LSTM in RMSE and just fell slightly behind in MAE. Even though Bi-LSMT provided the least error rate, it also costs much more time than "Vanilla" in model training. Providing a 2D sliding window seemed to be useless for the CNN model in that research the CNN performance was the worst. So using a simpler model and optimizing it yielded much better results than anticipated. Even though Bi-LSTM has room for improvement, its complexity makes it a lot more difficult to optimize. Bi-LSTM is much more expensive computationally and yields only slightly better results than Vanilla.

We continue to research and explore the performance limit of these models, and also test their robustness by using high-missing rate data in this paper. Moreover, we go deeper into the error patterns and separate and analyze the errors at different stations. And the error trends over 24 hours for weekday and weekend data are also examined. And the data will also be separated according to the weekday and weekend, and we train and test these models with these data respectively.

2. Experimental results for high missing rate

The figure below shows the trend of RMSE and MAE of different training sets with missing rates from 10% to 90%), by applying "Vanilla" model, and the testing set keeps the missing rate at a constant 15%. These two values increase which means we will obtain a larger error with the increase in the missing rate.

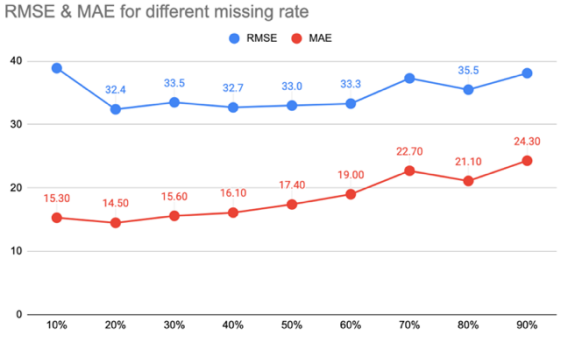

Fig. 1. RMSE and MAE of different missing rates

In this section, the missing rate of 80% is used for generating training data and 15% for testing (validation data) and this high missing rate dataset is used to verify the robustness of the Autoencoder model under the most extreme situations.

Even for the high missing rate dataset, the simplest model, “Vanilla”, also does very well in the imputation of missing data. It still almost has the same value as the Bi-LSTM in RMSE and falls just a bit behind in MAE. The Bi-LSTM also takes about 100 times much longer than “Vanilla” to compute the missing values although it has the best performance. Even though CNN improves and is better than W-H Avg in MAE, the RMSE is the worst of all the methods compared. As same as the low missing rate case, providing a 2D sliding window seems to have done nothing for the CNN model here. The optimal model seems to be the tried and true “Vanilla”.

Even though Bi-LSTM still has some room for improvement, its complexity makes it a lot more difficult to optimize. Bi-LSTM is a much more expensive computationally method and only has a slightly better result than “Vanilla”.

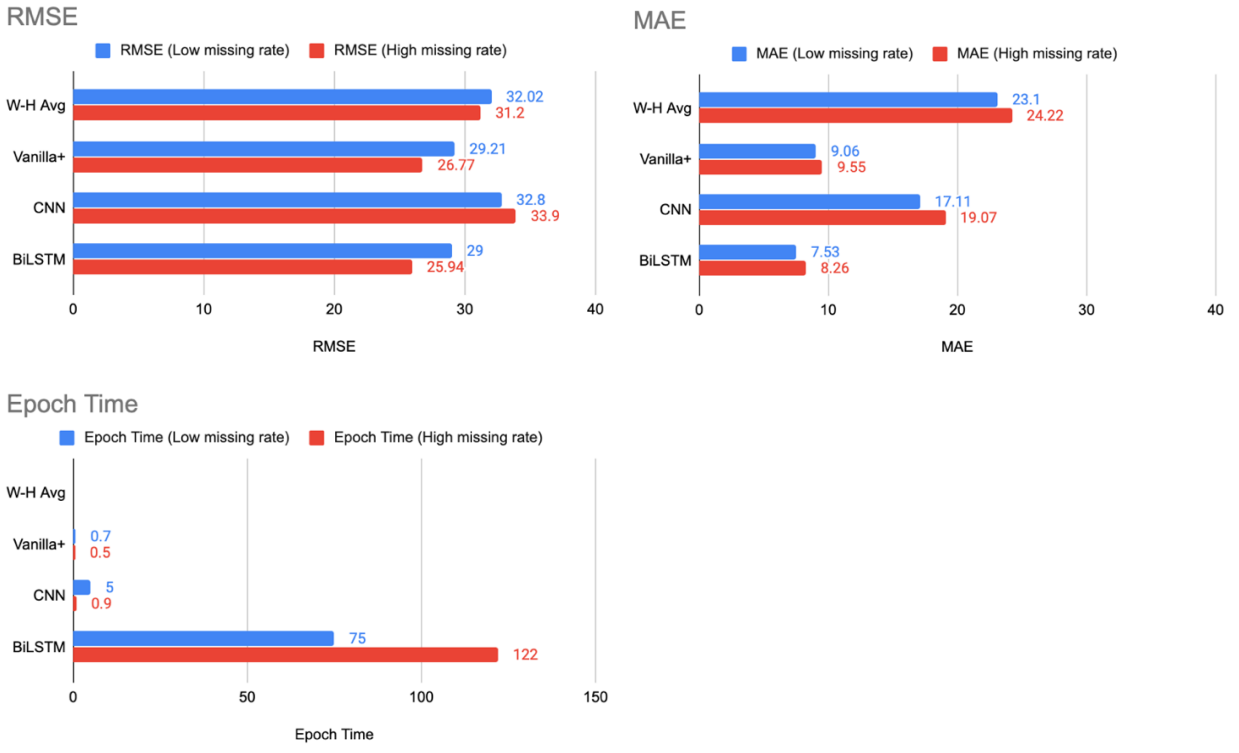

Fig. 2. Result comparison of high missing rate and low missing rate.

We also compare the value of RMSE, MAE, and epoch time for both high missing rate data and low missing rate data in Figure 2. And the robustness of these three Autoencoders can be verified because the difference of RMSE or MAE between low missing rate data and high missing rate data is small. And the increase of the missing rate doesn’t influence the final imputation result much. However, for Bi-LSTM, the high missing rate data cost much time to calculate and it has a higher computational expense. So we can say that these machine learning models can be used for traffic flow data imputation even though the gap within the data is large, for example, the extreme situation case: 80% missing rate.

3. Error patterns for different stations and hours

In this section, the whole dataset is considered while calculating the RMSE and MAE. And the data of different stations or different hours are separated and calculated independently for RMSE and MAE. Besides, the hyper-

parameters of the model in this section are selected generally and there is no tuning work in this section due to the main purpose being to find the error patterns instead of the most accurate rate of the model.

3.1. Error patterns for different stations

First, the data are separated according to the sensor stations and calculated respectively. The results are shown as follows.

From the figure below, all three models give almost error trends. The Bi-LSTM model has the best performance compared with the "Vanilla" and CNN, this mirrors the conclusion that Bi-LSTM has the lowest RMSE and MAE in the last section. Besides, two stations: 106 and 107 have the largest error value especially for 104 with the highest RMSE of about 63 no matter which model is used. Several reasons may cause this problem.

a) Facility error

The measurement results may be inaccurate due to malfunctions at the detected facility, for example, the loop detector or magnetometer could be broken. So the value at station 104 may be erratic and chaotic, and it's hard for a machine learning model to find out the certain pattern of traffic flow data.

b) Chaotic operational situation

Due to the unreasonable geometry design or extremely large traffic volume, the serious traffic congestion caused by accidents or large volumes may happen near station 104 frequently, and it may be hard to find a regular pattern especially in the case of accidents caused by bad road design.

Because of the lack of a sensor location distribution map, it is hard to track the exact problem. Instead, we give two possible causes of the high error problem. Now we can try to start exploring the problem origin from these two aspects. A future study could focus on the deeper details of it.

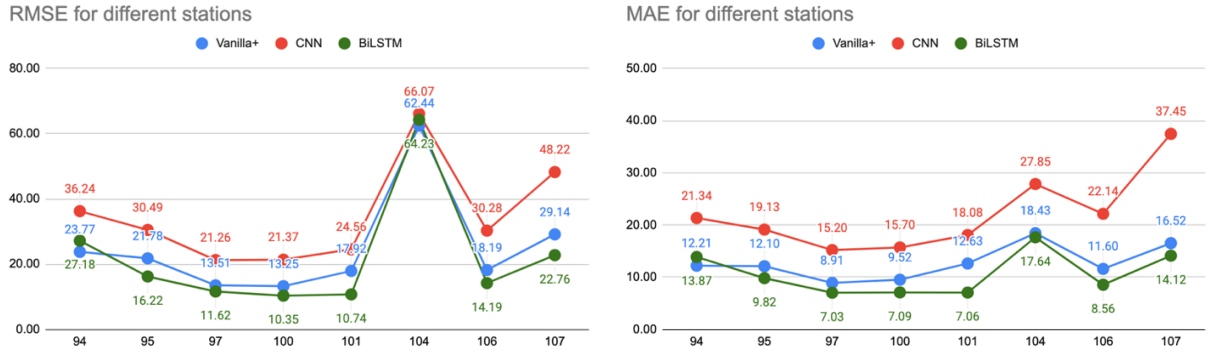

Fig. 3. RMSE and MAE for different stations.

3.2. Error pattern for different hours

In this part, the data of each hour is separated, and RMSE and MAE are calculated respectively as shown in Figures 4 and 5. We can see the error pattern for 24 hours on weekdays. Firstly, all three models show a similar error trend. And it can verify the conclusion in the previous section that the "Vanilla" and Bi-LSTM are better than CNN. Besides, two error peaks are concentrated on the morning peak hour (around 8:00) and afternoon peak hour (17:00) which are consistent with the daily flow peak hours.

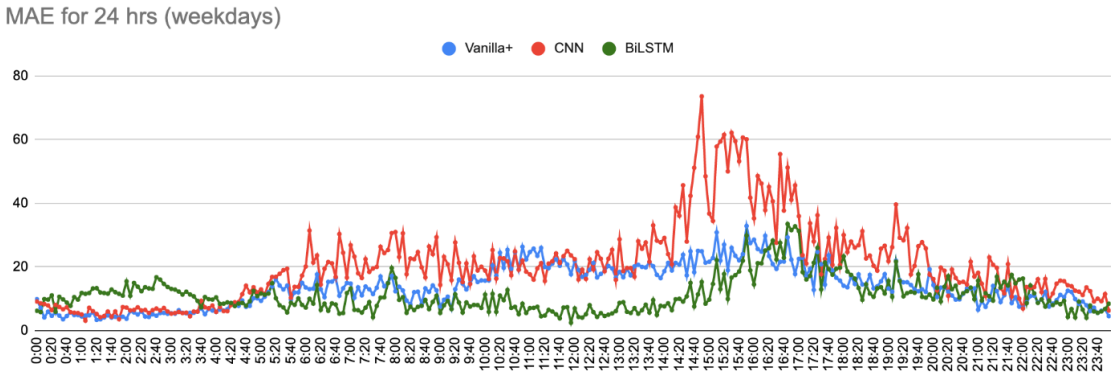

Fig. 4. MAE for 24 hours (weekdays).

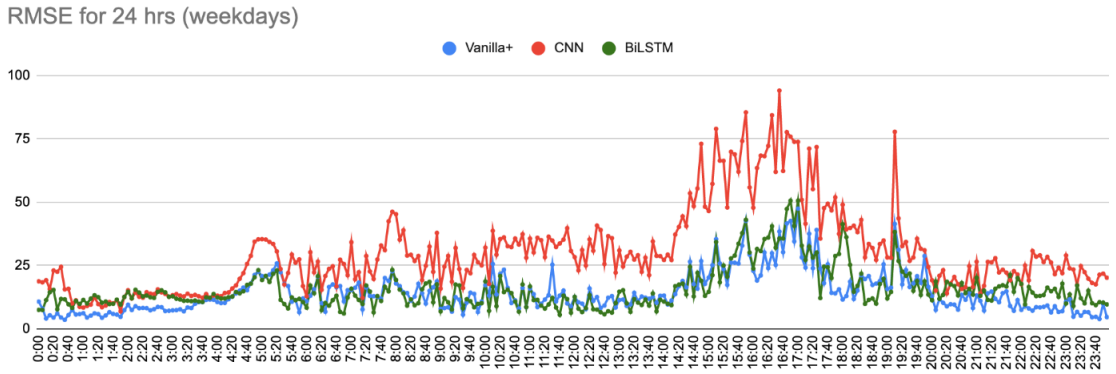

Fig. 5. RMSE for 24 hours (weekdays).

As shown in Figures 6 and 7, we can see the error pattern for 24 hours at weekends. Compared with weekdays data, the weekends' errors are much higher and chaotic. The highest error peak is around weekend afternoon peak hours (14:40). However, the high error lasts until midnight.

In short, it's easier to impute an accurate traffic flow value on weekdays compared with the weekend. And the higher traffic flow may cause a higher error.

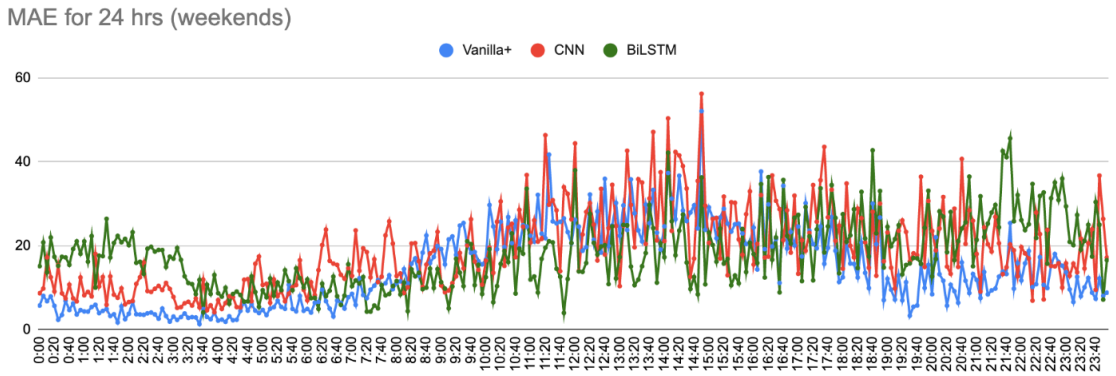

Fig. 6. MAE for 24 hours (weekends).

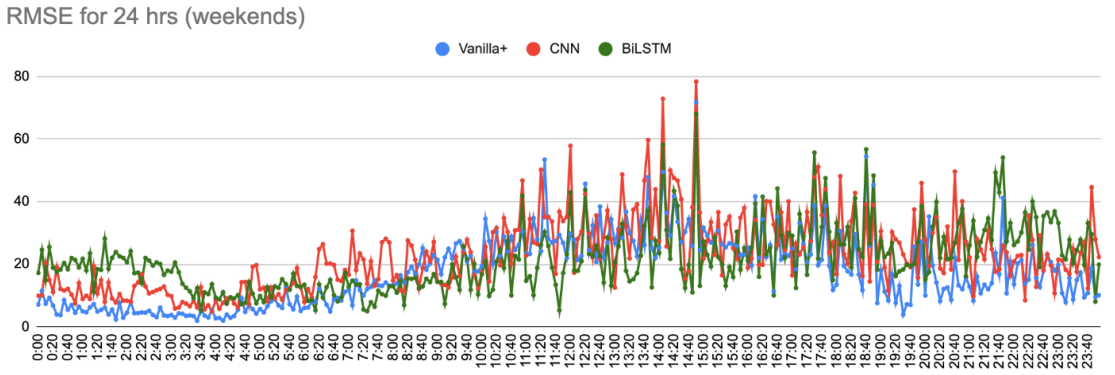

Fig. 7. RMSE for 24 hours (weekends).

4. Experimental results after data separation

Section 3 shows the different error patterns for weekdays and weekends, and it shows that the error for weekends is much higher than on weekdays. In this section, the dataset is separated into weekdays and weekends, and these are trained, tested, and errors are calculated respectively. To compare the effect of separation, all three groups still use the same hyper-parameter without much tuning. The volume of training is 7200 and 2592 for testing. To obtain the same volume of data, different day periods are selected. The details about the period are shown as follows.

	Training period	Training set volume	Testing period	Testing set volume
Total	Feb 1 – Feb 26	7200	Mar 1 – Mar 10	2592
Split-weekdays	Feb 1 – Mar 5	7200	Apr 1 – Apr 14	2592
Split-weekends	Feb 1 - May 1	7200	May 1 – Jun 1	2592

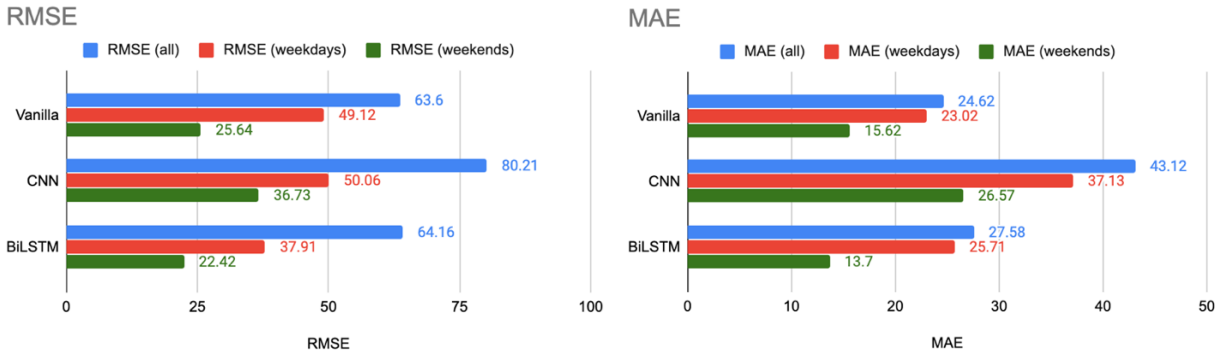

Fig. 8. MAE and RMSE after data separation.

The testing results are shown in Figure 8. First, the previous conclusion is verified again that the Bi-LSTM is better than “Vanilla” and “Vanilla” is better than CNN. It can seem clear that the training error after splitting the data is much lower than with the whole dataset, especially for the MAE value of the weekend, no matter which model is applied. Therefore, the separation could be a good idea to train the model and test the prediction value, and it can be used in practice in traffic data prediction to obtain more accurate imputation results.

5. Conclusion

In this paper, Denoising Autoencoders are used to process and generate the missing traffic flow data. RMSE, MAE, and epoch times for both high and low missing data rates are shown and compared and are used to evaluate the performance of the models, and three models: “Vanilla”, CNN, and Bi-LSTM are also compared to the industry-widely accepted method of using Weekly-Hourly Average of predicting traffic flow.

The training and testing result shows that Bi-LSTM is better than “Vanilla” and “Vanilla” is better than CNN. The simplest model “Vanilla” does very well in the imputation of missing data even if the missing rate is high. CNN is not suitable for data imputation instead it is good at image processing. And Bi-LSTM has room for improvement although it’s too complex to finish optimization in a short time. Bi-LSTM is much more expensive computationally and yields only slightly better results than Vanilla.

Second, the robustness of these three Autoencoders can be verified because the difference of RMSE or MAE between low missing rate data and high missing rate data is small. And the increase of the missing rate doesn’t influence the final imputation results much. However, for Bi-LSTM, the high missing rate data has higher computational time and expense. So we can say that these machine learning models can be used for traffic flow data imputation even though the gap within the data is large.

Besides two possible reasons for the high errors of some stations are proposed and analyzed and can be studied further after the providing of more detailed corresponding information.

The error patterns for different hours on weekdays and weekends are also shown. For weekdays, the error concentrates at the two daily peak hours in the morning and the afternoon. And the error of weekends is much higher and concentrated from afternoon to midnight.

In the end, the data are separated into weekdays and weekends and trained and tested respectively. And we can see that the method of separation can lower the error significantly, especially for the weekend, and improve the accuracy of imputation. Therefore, this separation method can be used for accuracy improvement.

Acknowledgments

The authors acknowledge helpful discussions with Professors R. Jayakrisnan and Michael Dillencourt at the University of California, Irvine.

References

[1] Atluri, Gowtham, Anuj Karpatne, and Vipin Kumar. (2018) "Spatio-temporal data mining: A survey of problems and methods." ACM Computing Surveys (CSUR) 51.4: 1-41.

[2] Duan, Yanjie, Yisheng Lv, Yu-Liang Liu, and Fei-Yue Wang. (2016) "An efficient realization of deep learning for traffic data imputation." Transportation research part C: emerging technologies 72: 168-181.

[3] Asadi, Reza. (2020) “Deep Learning Models for Spatio-Temporal Forecasting and Analysis.” DISSERTATION. Diss. UNIVERSITY OF CALIFORNIA, IRVINE.

[4] Costa, Adriana Fonseca, Miriam Seoane Santos, Jastin Pompeu Soares, and Pedro Henriques Abreu. (2018) "Missing data imputation via denoising autoencoders: the untold story." International Symposium on Intelligent Data Analysis. Springer, Cham.

[5] Duan, Yanjie, Yisheng Lv, Wenwen Kang, and Yifei Zhao. (2014) "A deep learning based approach for traffic data imputation." 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). IEEE.

[6] Gondara, Lovedeep, and Ke Wang. (2018) "Mida: Multiple imputation using denoising autoencoders." Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, Cham.

[7] Zhuang, Yifan, Ruimin Ke, and Yinhai Wang. (2018) "Innovative method for traffic data imputation based on convolutional neural network." IET Intelligent Transport Systems 13.4: 605-613.

[8] Zhang, Jianye, and Peng Yin. (2019) "Multivariate Time Series Missing Data Imputation Using Recurrent Denoising Autoencoder." 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE.

[9] Asadi, Reza, and Amelia Regan. (2019) “A convolutional recurrent autoencoder for spatio-temporal missing data imputation.” Proceedings of 2019 International Conference on Artificial Intelligence (ICAI’19), pp. 206-212.

[10] “California. pems, http://pems.dot.ca.gov/, 2017.”