World Academy of Science, Engineering and Technology
International Journal of Transport and Vehicle Engineering
Vol:12, No:1, 2018

Missing Link Data Estimation with Recurrent Neural Network: An Application Using Speed Data of Daegu Metropolitan Area

JaeHwan Yang, Da-Woon Jeong, Seung-Young Kho, Dong-Kyu Kim

***Abstract*—**In terms of ITS, information on link characteristic is an essential factor for plan or operation. But in practical cases, not every link has installed sensors on it. The link that does not have data on it is called "Missing Link". The purpose of this study is to impute data of these missing links. To get these data, this study applies the machine learning method. With the machine learning process, especially for the deep learning process, missing link data can be estimated from present link data. For deep learning process, this study uses "Recurrent Neural Network" to take time-series data of road. As input data, Dedicated Short-range Communications (DSRC) data of Dalgubul-daero of Daegu Metropolitan Area had been fed into the learning process. Neural Network structure has 17 links with present data as input, 2 hidden layers, for 1 missing link data. As a result, forecasted data of target link show about 94% of accuracy compared with actual data.

***Keywords*—**Data Estimation, link data, machine learning, road network.

I. Introduction

As the computing power of personal computers has been rapidly rising, artificial neural network algorithms, which were not used due to excessively long operation time, have been used by many kinds of scientific and engineering researchers for these days. These Artificial Neural Network (ANN) algorithms have been broadening its parts with Big Data and Machine Learning technologies.

Nowadays, ANN is used for development of artificial intelligence, system control, pattern recognition, medical diagnosis, regression, management, clustering, data handling. With the advance of data size and its technologies, ANN can be used for more parts of engineering and science.

Regarding transportation, ANN algorithms have been received for more and more fields of study. ANN algorithms have been used for data handling, regression models, image analysis, etc. So it can be used for traffic operation, planning, safety and human behavior.

ANN has many various models that are classified by functions and structures. For transportation engineering, Convolution Neural Network (CNN), Recurrent Neural Network (RNN), Fuzzy Neural Network (FNN) known as Neuro-fuzzy, Multilayer Perceptron Neural Network (MLPNN) are generally used.

The limitation of ANN is data abstracting. ANN should carry data through abstracting process. These processes are working as black-box, so relationship between input data and output data is hard to be proved.

As in [1], Neural Networks (NN), an extremely popular class of CI models, have been widely applied to various transportation problems. For operation, [2] uses Time delay neural network (TDNN), [3] uses FNN. MLPNN is also used for signal optimization in [4].

In planning, MLP has been widely used for modeling. References [5]-[7] studied MLP models for travel behavior analysis, mode choice modeling, trip distribution forecasting. Also, other ANN models are used actively. ANN models are also used for accident analysis. References [8], [9] used MLPNN and FNN to analyze accidents.

ANN models are also used for speed data estimation in many studies. Zhang and Qi tried simple NN algorithm for forecasting seasonal and trend time-series data [10]. In [11], Lee studied link speed estimation method with RNN and compared it with Kalman-filter and the naïve forecasting. As the result, RNN model for multi-links has higher accuracy than other methods. Vlahogianni studied MLPNN for forecasting [12]

As wave analysis method, Long Short-term Memory (LSTM) NN was used for speed estimation in [13]; RNN is also used for transportation mode detection using mobile devices. Reference [14] uses RNN-RBM (Restrict Boltzmann Machine) to predict large-scale transportation network congestion.

This study uses DSRC sensor data of Dalgubul-daero. For missing link data estimation, RNN is chosen. Because DSRC data are kinds of time-series data, typical Neural Network methods are not appropriate to handle DSRC data.

Dalgubul-daero is a central wide road of Daegu Metropolitan Area, which is southeast of South Korea. Dalgubul-daero has 8 or 10 lanes for 33.58 km length. There are about 37 DSRC sensors on it. DSRC speed data have 5 minute time interval, so number of data which are collected per a day is 288. This study used data from July 21 to 23 of 2016. From these days, 864 samples had been collected from each link. This study uses R 3.4.2 and Rstudio 1.1.383, library package rnn, readr, imputeTS

JaeHwan Yang is with the Civil and Environmental Engineering Department, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Rep. Korea, (phone: +82-2-880-9154; e-mail: iyanan05@snu.ac.kr).

Da-Woon Jeong, Seung-Young Kho, and Dong-Kyu Kim are with the Civil and Environmental Engineering Department, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Rep. Korea.

for programming.

II. Methodology

As it is well known, ANN are simulation models inspired by human brain. It consists with nodes called "neurons", and connectors between neurons. If an ANN system has hidden layers, the learning process of it is called "Deep Learning".

The structure of ANN is in Fig. 1. Each node has links to nodes of next step, and each link has its weights. In the learning process, weights of each link are calculated from data. Data of nodes on input layer are abstracted step by step and converged to the output layer.

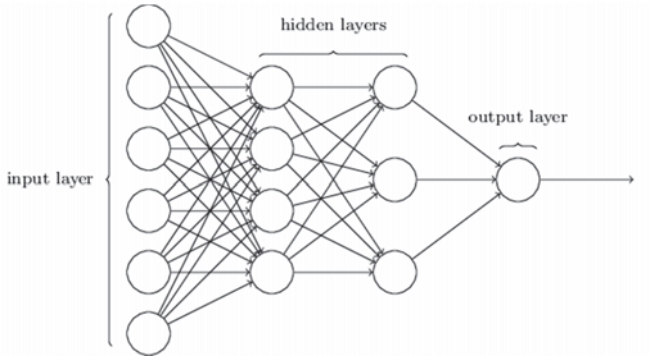

Fig. 1 Structure of ANN [15]

The process is used to optimize the structure same as follows [11]:

1. Give initiating random weight to all links; take summation of each value in the first hidden state.
2. Take sigmoid function, and repeat for next layer. i) and ii) steps are calling "Feed-Forward Propagation".
3. When feed-forward propagation steps are completed, difference between estimated values and observed values can be taken. And minimize this with "Back-propagation".

A RNN is an NN with self-connections whose delay is one. It is very efficient and powerful to handle time-series signals. In RNN models, a hidden layer with self-connections has the role of memory that accumulates information over time from an input sequence [16]. This network can be defined as stacked NN that is in sequence.

In RNN, the same task is applied to every factor of each step. Because of this, the output of each step might be influenced by the previous step. Fig. 2 shows the structure of RNN.

In [17], calculation within RNN process is as follows: When x_t refers input data of time step t, s_t is hidden state of time step t, o_t is the output data of time step t, o_t can be calculated from function $s_t = f(Ux_t + Ws_{t-1})$. So hidden state of step t-1 can influence next step t, and this process will be repeated. By this method, RNN can handle series data very appropriately. With this characteristic, RNN is mainly used for successive data handling like time series prediction, speech recognition, grammar learning, and so forth.

Normally, transportation data like speed, volume, density have characteristics of time-series data. If the characteristics of transportation infra of target area do not change, transportation data of each time step can be assumed to influence the following step. Therefore this RNN method can be suitable for transportation data.

In this paper, RNN is used to estimate speed data of missing link. Speed data of each link can be assumed to affect each other, and data of each time step can be also assumed to affect next steps, so RNN is expected to be appropriate for this problem. For this, data of every target links had been entered as input data of the learning process.

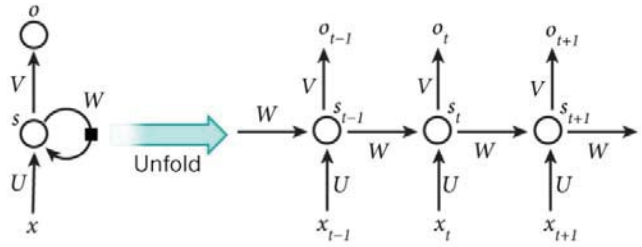

Fig. 2 Structure of RNN [17]

Fig. 3 Locations of Dalgubul-daero links (links with numbering)

III. Application

A. Using Data

As described above, this study uses DSRC speed data of Dalgubul-daero on Daegu Metropolitan Area. Locations of Dalgubul-daero links are marked in Fig. 3 (using QGIS ver. 2.18.7). Dalgubul-daero has a length of 33.5 km, 8 or 10 lanes with interrupted flows.

This study use speed data that were collected in July 21-23 2016. Its time interval is 5 minutes for each step. In this study, link no. 546 was target link to compare observed data and estimated data. So no. 546 and 18 links that are near to no. 546 were chosen for input data.

B. Data Imputation Process

DSRC sensor writes a datum when a vehicle which has the on-board device is passing it. Because of this, DSRC data inevitably have missing values where a vehicle which has no on-board device is passing by sensor location. When the learning process is complete, the missing data can be estimated by RNN estimation process. However, data is input to the learning process should not have missing data for simple RNN process. Accordingly, this study takes data imputation process.

Data imputation is carried out by the simple section average method. The width of the section is 5. This means that the data from 25 mins before the missing time to 25 mins after the missing time were used to take the average. The program package na.ma on R was used to achieve this.

C. RNN Structure

RNN used for this study should have 18 input nodes and one output node because of its purpose. The model is made to take speed estimation value of no. 546 from values of 18 links.

With the trial-and-error method, the structure of hidden state determined as 2 layers, that have 12 and 3 nodes each. In the same method, other properties of RNN structure were determined. Table I shows the model structure properties of RNN used in this study.

The learning process used data from earlier 2 days and 20 hours data as input data. After the learning process is finished, rest of 4 hours were used for the estimation process. Input and output data of earlier 68 hours are used to make the RNN structure. And data of rest 4 hours are used to verify the RNN structure. For verification of structure, observed data and estimated data of last 4 hours are compared. From this comparing, the accuracy of the structure would be calculated.

D. Comparing with Other Statistical Models

To check whether the accuracy of the RNN structure is high or low, we performed two other statistical estimations. First, the simple average method which is currently used for link data estimation in Daegu Metropolitan Area is performed. And the Generalized Least Square (GLS) method which is known as more efficient under heteroscedasticity and autocorrelation state to take linear regression model is performed.

The GLS model and the simple average model was built with speed data of 4 links, link no. 125, 442, 458, 422 which are nearest to link no. 546. Link no. 442 and 458 are just before and after links to link no.546 and link no. 125 and 422 are next.

TABLE I
RNN Model Properties

Property	explanation	Value
Layer Structure	Layer structure of RNN including input, hidden, output state	18-12-3-1
Batch Size	Number of input values which is entered at once	2
Learning rate	Size of interval which is changed at each process	0.01
Epoch	Number of the learning process iteration	5000

IV. Results

A. RNN Model

As a result, RNN model figures out estimated speed data of link no. 536, which have about 94.9% of accuracy. The result that comparing observed data and estimated data is in Fig. 4.

B. Statistical Models

With the same datum, the traditional simple average method shows 86.2% of accuracy. The GLS models came out as:

$$V_{546} = 0.0431V_{125} + 0.2804V_{442} + 0.1161V_{458} + 0.3858V_{422} + 5.7938$$

This model takes 88.9% of accuracy. Comparing result is on Table II.

TABLE II
Comparing RNN and Statistical Models

Method	Average Accuracy	Explain
Simple Average	94.9%	The traditional method
Generalized Least Square	88.9%	20% advanced accuracy than the traditional method
RNN	86.2%	60% advanced accuracy than the traditional method

V. Conclusion

This paper produced a model with RNN algorithms to estimate missing speed data. For this, DSRC data from Dalgubul-daero are used. As a result, the RNN model can estimate the data fast and with high accuracy and have higher accuracy than some statistical models.

The models from this paper could make more advanced result by some follow-up studies. First, the larger dataset can be input data for this model. The accuracy of machine learning system can be improved with a larger dataset which have shorter interval or many days. Second, the optimization of RNN structure should be performed for faster and more accurate models. For example, optimized epoch and layer structure can take more accurate result.

ANN models take high attention to many researchers for its accuracy and applicability. It is highly expected that ANN makes improvement of transportation engineering.

Acknowledgment

This research was financially supported by the Seoul R&D Program (PS160010) through the Research and Development for Regional Industry.

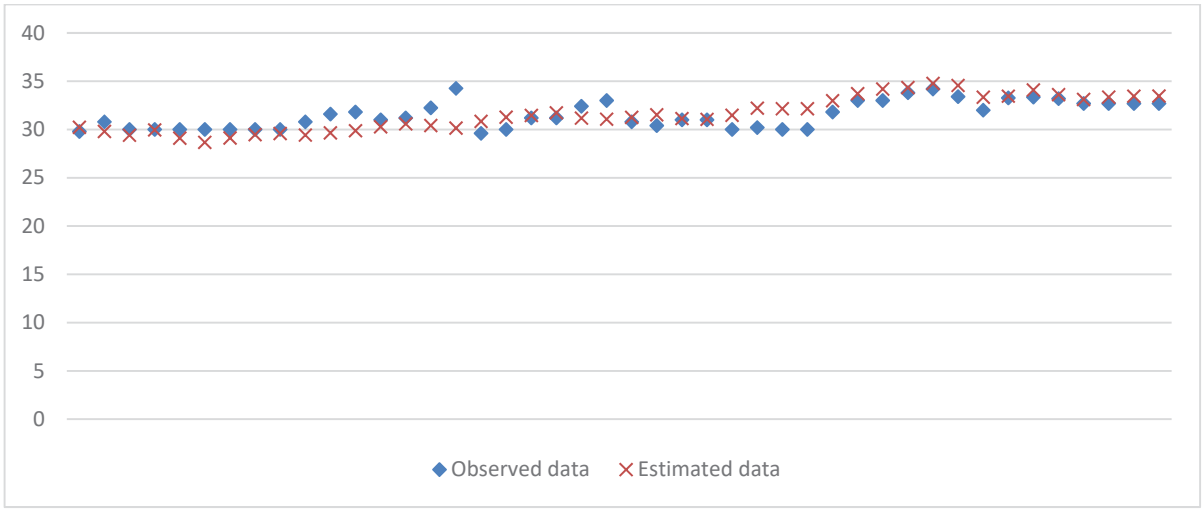

Fig. 4 Comparing Observed data and estimated data (5 mins interval)

REFERENCES

[1] M. G. Karlaftis and E. I. Vlahogianni, "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights," *Transp. Res. Part C Emerg. Technol.*, vol. 19, no. 3, 2011, pp. 387-399

[2] Zhong, M., Sharma, S., Lingras, P., "Genetically designed models for accurate imputation of missing traffic counts" *Transportation Research Record 1879*, 2004, pp. 71-79.

[3] Yin, H., Wong, S.C., Xu, J., "Urban traffic flow prediction using fuzzy neural approach", *Transportation Research Part C 10 (2)*, 2002, pp. 85-98.

[4] Teodorovic, D., Varadarajan, V., Jovan, P., Chinnaswamy, M., Sharath, R.," Dynamic programming neural network real-time traffic adaptive signal control algorithm", *Annals of Operation Research 143 (1),* 2006, pp. 123-131.

[5] Shmueli, D., Salomon, I., Shefer, D., 1996, "Neural network analysis of travel behavior: evaluating tools for prediction", *Transportation Research Part C: Emerging Technologies 4 (3)*, 1996, pp. 151-166.

[6] Nijkamp, P., Reggiani, A., Tritapepe, T., "Modelling inter-urban transport flows in Italy: a comparison between neural network approach and logit analysis", *Transportation Research Part C 4*, 1996, pp. 323-338.

[7] Mozolin, M., Thill, J.C., Lynn Usery, E.L., "Trip distribution forecasting with multiplayer perceptron neural networks: a critical evaluation", *Transportation Research Part B* 34 (1), 2000, pp. 53-73.

[8] Abdelwahab, H.T., Abdel-Aty, M.A., "Artificial neural networks and logit models for traffic safety analysis of toll plazas", *Transportation Research Record 1784*, 2002, pp. 115-125.

[9] Abdel-Aty, M.A., Abdelwahab, H.T., "Predicting injury severity levels in traffic crashes: a modeling comparison", *Journal of Transportation Engineering 130 (2),* 2004, 204-210

[10] G. P. Zhang and M. Qi, "Neural network forecasting for seasonal and trend time series," *Eur. J. Oper. Res.*, vol. 160, no. 2, 2005, pp. 501-514.

[11] M. Lee, "Forecasting short-term travel speed in a dense highway network considering both temporal and spatial relationship : using a deep-learning architecture," Chung-ang University, 2016

[12] Vlahogianni, E.I., Karlaftis, M.G., Golias, J.C., "Spatio-temporal urban traffic volume forecasting using genetically-optimized modular networks", *Computer-aided Civil and Infrastructure Engineering 22 (5)*, 2007, 317-325

[13] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," *Transp. Res. Part C Emerg. Technol.*, vol. 54, 2015, pp. 187-197.

[14] X. Ma, H. Yu, Y. Wang, and Y. Wang, "Large-scale transportation network congestion evolution prediction using deep learning theory," *PLoS One*, vol. 10, no. 3, 2015, p. e0119044.

[15] Taehoon Kim, https://carpedm20.github.io/2014/neural-net-translation/

[16] T. H. Vu and J.-C. Wang, "Transportation Mode Detection on Mobile Devices Using Recurrent Nets," *Proc. 2016 ACM Multimed. Conf. - MM '16*, pp. 392–396, 2016.

[17] Team AI Korea, "Recurrent Neural Network (RNN) Tutorial – Part1" http://aikorea.org/blog/rnn-tutorial-1/