Neural Network Kalman Smoother for Filling up Missing Probe Vehicle Data

Kam Cheong CHOY[1], Takashi NAKATSUJI[2]

[1]Student Member of JSCE, Master Student, Graduate School of Engineering, Hokkaido University
(Kita-13, Nishi-8, Kitaku, Sapporo 060-8628, Japan)
E-mail:choy@eng.hokudai.ac.jp
[2]Regular Member of JSCE, Professor, Graduate School of Engineering, Hokkaido University
(Kita-13, Nishi-8, Kitaku, Sapporo 060-8628, Japan)
E-mail:naka@eng.hokudai.ac.jp

Recently, taxi probes become very popular for obtaining traffic information so as to provide them to drivers. However, the percentage of taxi probes running in urban road network is not high enough. As a result, missing data causes some deficiency in traffic information. Therefore, filling up the missing data is necessary. In this study, a method called Kalman smoother is examined. Since travel time in urban road network during heavy traffic situation is non-linear, it is difficult to provide suitable formula. Therefore, artificial neural network (ANN) is introduced to Kalman smoother. Besides, a high amount of missing data occurs in Taxi probe so neural Kalman smoother is utilized for filling up the missing data. By assuming some available data to be missing, the accuracy of Kalman smoother can be found.

Key Words : *Artificial Nerual Network, Kalman Filter, Kalman Smoother, Probe Vehicle Data*

1. INTRODUCTION

In Sapporo city, taxis probe data system has been deployed in June, 2008[1)]. Nowadays, there are totally 1700 taxis probe which are running in this city[1)]. These taxis are used to collect the real time data which is so-called “Probe vehicle data”. Probe vehicle data can provide link travel time which is defined as the used time from one upstream intersection to downstream one and the corresponding longitudes and latitudes are provided.

However, since the percentage of taxis probe running in urban road network is not high enough, there must be a certain amount of data to be missing for each individual link in some time periods. Therefore, it probably causes some deficiency to the probe vehicle data set. Hence, it is considered that the estimation of the missing data is necessary.

Most of researchers has tried to fill up the missing data by combining some equations and Kalman smoother together [2), 3) and 4)]. In the study of Suzuki et al.[5)], travel time is estimated by integrating Neural network and Kalman filter together.

In this study, a new method called “Neural Kalman smoother” is proposed for filling up the missing data. This is featured by the combined usage of “Artificial neural network” and “Kalman smoother” together.

2. OBJECTIVE

This study aims at investigating how efficient the neural Kalman smoother is in filling up missing data:

(1) Develop a filling up scheme for missing data based on Kalman smoother whose equations are described by artificial neural network (ANN)
(2) Establish an efficient and stable training procedure by integrating Genetic Algorithm (GA) into the Back Propagation method for avoiding local minimums.
(3) Examine the accuracy of the proposed method through some numerical experiments by treating some actual measured data as filling data.

3. STUDY AREA

In this study, a general road which is next to Hokkaido University, Sapporo is picked up for the test site, as shown in **Fig.1**. The reason of choosing this road is that there are many vehicles running through it which the percentage of missing data is believed to be less than others. The total number of links in this route is 48 and only south bound direction is analyzed.

Fig.1 Studied road in Sapporo

4. METHODOLOGY

(1) Kalman Smoother

In this study, a method called neural Kalman smoother is newly developed for filling up missing data. The smoother can divided into two steps[6]. The first step is the Forward pass which is exactly defined as the state equation in Kalman filter;

$$\mathrm{x}(k+1) = f\big(x(k)\big) + v(k) \tag{1}$$

where *x(k)* is the travel time at time step k and *v(k)* is the system error, respectively. The function f describes how the *x(k)* changes over time. Kalman filter is an algorithm which uses a series of observed measurements over time containing noise to produce the unknown value. The second step is Backward pass which smoothes the values estimated by Kalman filter[6] defined as the measurement equation;

$$y(k+1) = g\big(x(k+1)\big) + w(k) \tag{2}$$

where *y(k)* is the measurement travel time at time step k and *w(k)* is the observation error, respectively. The function g defines the relationship between *x(k)* and *y(k)*.

(2) Neural Kalman Smoother (NKS)

NKS aims at estimating travel time for urban road network in future time step. Since there are a lot of traffic signals in urban road network so the travel time is influenced by many factors and fluctuated significally. It is very difficult to define the equations analytically. Therefore, Artificial Neural Network model (ANN) is introduced to define the equations in Kalman filter. The structure of the ANN model adopted in this study is depicted in **Fig.2**. The ANN for the state equation, consists of totally 4 layers. The input layer has 5 neurons for 5 input signals for different time steps and the output has 1 neuron for 1 signal for next time step. The general equation of ANN model is expressed as:

$$y_m = h\{\textstyle\sum_k w_{km} \cdot h\big[\sum_j w_{jk} \cdot h\big(\sum_i w_{ij} x_i\big)\big]\} \tag{3}$$

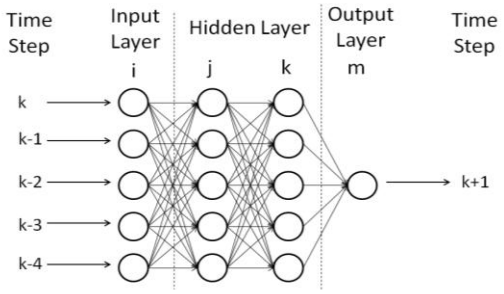

Fig.2 Structure of ANN model

where y_m: Output signal
x_i: Input signal
w: Weights between layer and layer
$h(x)$: Sigmoid function defined as:

$$h(x) = \frac{1}{1+e^{-x}} \tag{4}$$

Neural network can be considered as human brain network. To teach the ANN model by inputting the input signals and output signals from training data set so that the ANN model can act as a brain to solve problems. The process is called training process.

The function g is set to be identity in this study. Since the neural network function f is non-linear, the structures of Extended Kalman filter are applicable and also in NKS. That is, Eqs.(1) and (2) are approximately linearized as follows:

$$\mathrm{x}(k+1) = A_k x(k) + v(k) \tag{5}$$

$$y(k+1) = C_k x(k+1) + w(k) \tag{6}$$

where $Ak = \frac{\partial f}{\partial x_i}$ and $Ck = \frac{\partial g}{\partial x_i} = 1$

Since each link is defined an ANN model respectively, there are total 48 ANN models in this study.

(3) Training of artificial neural network

Since ANN model is used as the equations in Kalman smoother, the model is necessary to be trained by training data set in advance. Two training methods which are Genetic Algorithm (GA) and Back Propagation (BP) are used to train ANN model sequentially. GA is used for finding the optimal weights between layers without being entruppted into local minimums while BP is used for converging into the global optimum steadily by minimizing the error between input and output signals.

(4) Fill up missing data by NKS

As described before, NKS can basically divided sequentially into Forward pass and Backward pass .

a) Forward pass

Forward pass is exactly same as the Kalman filter algorithm. The calculation process is based on the following 4 steps.

1st Step: Initial setting:

Assume the covariance matries of *v(k)* and *w(k)*, V_k and W_k, and the initial values of $\hat{x}(0)$ and P_0

2^{nd} Step: Prediction:

$$\tilde{x}(k+1) = A_k \hat{x}(k) \quad (7)$$

$$\tilde{y}(k+1) = C_k \tilde{x}(k+1) \quad (8)$$

3^{rd} Step: Kalman gain

$$M_{k+1}^{xx} = A_k P_k A_k^T + V_k \quad (9)$$

$$M_{k+1}^{xy} = M_{k+1}^{xx} C_{k+1}^T = M_{k+1}^{xx} \quad (10)$$

$$M_{k+1}^{yy} = C_{k+1} M_{k+1}^{xx} C_{k+1}^T + W_{k+1} \quad (11)$$

$$K_{k+1} = M_{k+1}^{xy} \left(M_{k+1}^{yy}\right)^{-1} \quad (12)$$

4^{th} Step: Update after observing y(k)

$$\hat{\mathrm{x}}(\mathrm{k}+1) = \tilde{\mathrm{x}}(\mathrm{k}+1) + \mathrm{K}_{\mathrm{k}+1}\left[\mathrm{y}(\mathrm{k}+1)\text{-}\tilde{\mathrm{y}}(\mathrm{k}+1)\right] \quad (13)$$

$$P_{k+1} = M_{k+1}^{xx} - K_{k+1} C_{k+1} M_{k+1}^{xx} \quad (14)$$

Let k = k+1 and repeat 2^{nd} step to 4^{th} step until the end of study period. If there is missing in measured data in time step k+1, let $y(k+1) = \hat{x}(k)$ in Eq. (13).

b) Backward pass

Backward pass is run based on the values estimated by Kalman filter. The process is listed below.

$$L_k = P_{k-1} A_{k-1}^T M_k^{xx-1} \quad (15)$$

$$\hat{X}^s(k) = \hat{x}(k) + L_k(\hat{X}^s(k+1) - \tilde{x}(k+1)) \quad (16)$$

$$P_k^s = P_{k-1} + L_k(P_{k+1}^s - M_k^{xx}) L_k^T \quad (17)$$

Where $\hat{X}^s(k)$ is the updated state variable by Backward pass at time step k. Let k = k-1 and repeat Eqs.(15)-(17) above until the beginning of the study period.

5. EXPERIMENTAL DATA

In this study, two time periods are focused which are divided into two cases. Case 1 is morning peak hour, i.e., 07:00 – 10:00 while Case 2 is evening peak hour, i.e., 17:00 – 20:00. Before filling up missing data by NKS, the ANN model must be trained. For Case 1, the training data is selected from 07:00 – 10:00, 4^{th} to 6^{th}, Jan, 2010 and 2011. For Case 2, the training data is selected from 17:00 – 20:00, 4^{th} to 10^{th}, Jan, 2010 and 2011. Since there must be missing data, the training data set is used if and only if 6 data exist continuously in a row based on the assumption. By looking at the training data set, the percentage of missing data in evening peak is higher than that in morning peak so the number of days of training data is different for both cases. In the ANN model, since there are totally 55 weights, the number of training data set is at least double of the number of weights. However, since the pattern of missing data in each link is different, the number of training data set for each link is somewhat different but the number of training data set is about 110.

6. RESULT

In this part, by removing some existing data, the removed data is then filled up by NKS. The accuracy for both cases on 13^{th} Jan, 2011 are investigated:

Case 1: Morning peak hour, i.e., 07:00 – 10:00

Case 2: Evening peak hour, i.e., 17:00 – 20:00

For both cases, the R^2 of the smoothed values for each link are plotted in the following figure.

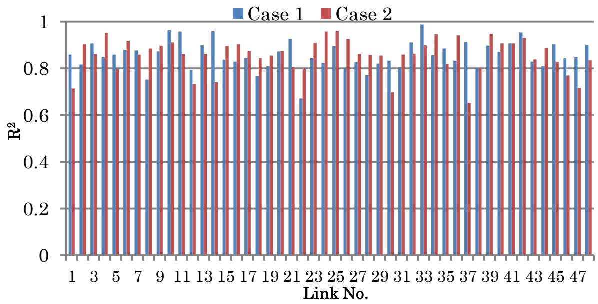

Fig.3 R^2 vs. link no. on 13^{th} Jan, 2011

From **Fig.3**, it can be noticed that the R^2 values for both cases in most of the links are higher than 0.8 and the lowest value is not less than 0.6. It is considered that NKS can be used in filling up missing data.

For Case 1, the travel time data at 8:30 and 9:30 in each individual link is removed for the entire route which is including totally 48 links.

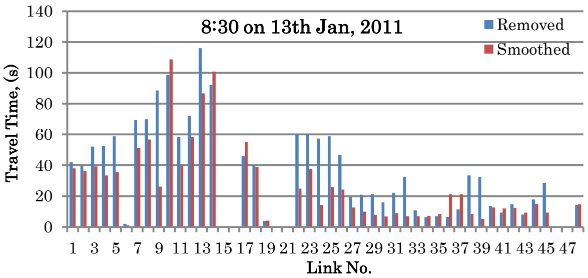

Fig.4 Travel time vs. link no. at 8:30

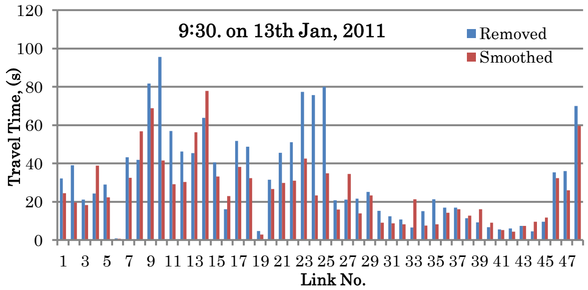

Fig.5 Travel time vs. link no. at 9:30

From **Fig.4** and **Fig.5**, it can be noticed that sometimes the removed and smoothed values are very closed to each other but sometimes it does not.

For Case 2, the travel time data at 18:00 and 19:30, 13th Jan, 2011 in each individual link is removed.

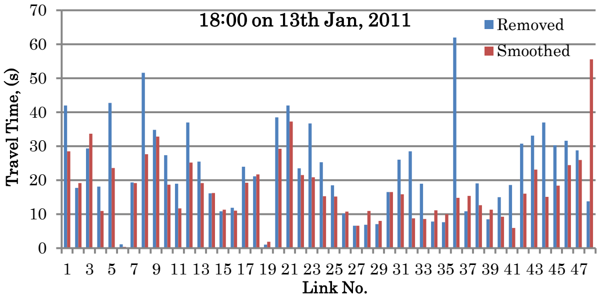

Fig.6 Travel time vs. link no. at 18:00

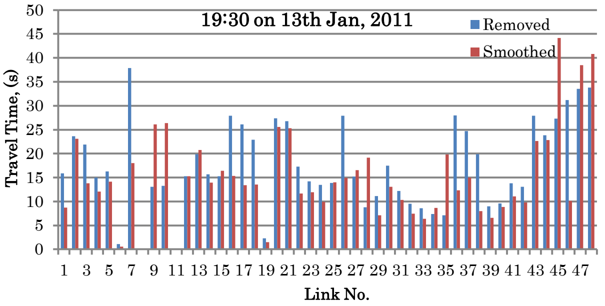

Fig.7 Travel time vs. link no. at 19:30

As shown in **Fig.6** and **Fig.7**, it can be stated that the removed and smoothed values are very similar to each other for some links but the removed and smoothed values are quite different from each other for some links.

In other to notice the profile of smoothed value for links, 2 links are chosen for better understanding by looking at their smoothed profiles. Link 14 of Case 2, and Link 25 of Case 1 are chosen for studying the smoothed profile.

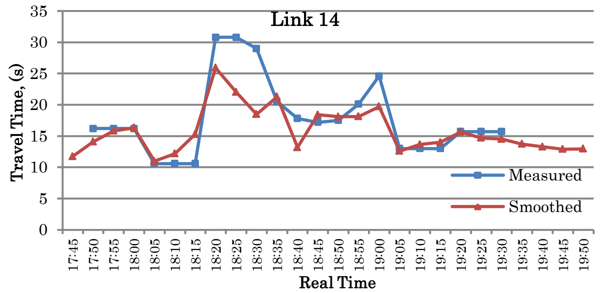

Fig.8 Travel time vs. real time for Link 14

From **Fig.6** and **Fig.7**, it can be noticed that the smoothed values are very closed to removed values for link 14. From **Fig.8**, it can be said that the actual measurement data at real time 18:00 and 19:30 is not dramatically changed comparing with the adjacent measurement data. It is believed that this is the main reason to cause the smoothed values to be very closed to the removed values.

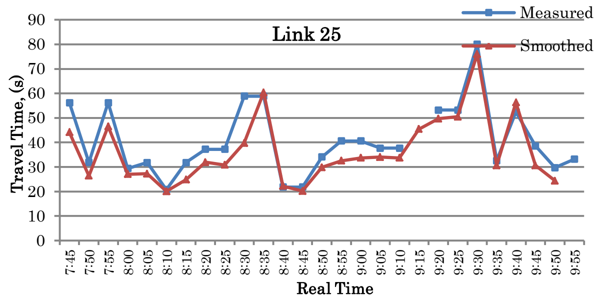

Fig.9 Travel time vs. real time for Link 25

Besides, from **Fig.4** and **Fig.5**, it can be noticed that the smoothed values are quite different from the removed values for link 25. From **Fig.9**, it can be stated that the actual measurement data at real time 8:30 and 9:30 is suddenly changed comparing with the adjacent measured data. It is considered that this is the main reason to cause the smoothed values not so accurate.

7. CONCLUSION

In this study, a method called Neural Kalman smoother is examined to fill up the missing probe vehicle data. From the results, it can be concluded that the NKS is very efficient and accurate while the slope of the adjacent data is consistent. Similarly, if the slope of adjacent data is so fluctuated, filling up missing data by Kalman smoother is not so accurate.

REFERENCES

1) *札幌圏商用車プローブ実用化検討会活動報告書.* Sapporo. (2011).
2) Fulton, J., Bitmead, R., & Williamson, R. (n.d.). Smoothing Approaches to Reconstruction of Missing Data in Array Processing.
3) Rahnama, S., & Arvan, M. Comparison of Extended and Unscented Kalman Smoother in Deriving Kinematic Characteristics of a High Maneuver Flying Target. *Proceedings of 2011 International Conference on Modelling, Identification, and Control*, (pp. 537-542).
4) Särkkä, S., Vehtari, A., & Lampinen, J. CATS Benchmark Time Series Prediction by Kalman Smoother with Cross-Validated Noise Density. *3rd International Conference on Development and Learning*, (pp. 2331-2341). (2007).
5) Suzuki, H., Nakatsuji, T., Tanaboriboon, Y., & Takahashi, K. A Neural-Kalman Filter for Dynamic Estimation of Origin-Destination (O-D) Travel Time and Flow on a Long Freeway Corridor. *Transportation Research Board*, (pp. 67-75). (2000).
6) Haykin, S. *Kalman Filtering and Neural Networks.* John Wiley & Sons, Inc. (2001).