DEVELOPMENT OF A DATA PRE-PROCESSING ALGORITHM FOR HI-PASS TRAFFIC INFORMATION SYSTEM

Sangsoo Lee, Keechoo Choi, Sangwoo Shim
Professor, Ajou University, Korea
sslee@ajou.ac.kr, keechoo@ajou.ac.kr, artmania@ajou.ac.kr
Seong J. Namkoong
Director, Korea Expressway Corporation, Korea
jake@ex.co.kr
Kangwon Shin
Instructor, Kyungsung University, Korea
kangwon@ks.ac.kr

ABSTRACT

Since 2009, Korea Expressway Corporation (KEC) in Korea installed the 'Hi-pass traffic information system' on expressways of Korea to provide with reliable and accurate travel information using the 'dedicated short-range communication' (DSRC) technology. Currently total 305 roadside equipments (RSE) are installed on 1,050km of expressway. This paper investigated the pattern and outlier characteristic of traveling data and proposed an enhanced data pre-processing algorithm for the Hi-pass traffic information system. Initial investigation results showed that the pattern and magnitude of travel time between passenger cars and buses are significantly different due to the median bus lane operation. In addition, it was identified that many variables affect the quality of raw data including the existence of service area and alternative route, number of RSE, pricing information, and the type of vehicles. The structure of the enhanced algorithm is described and comparison of the results is also discussed in this paper.

KEYWORDS

Pre-processing, DSRC, Hi-pass Traffic information System, Outlier, RSE

1. INTRODUCTION

Korea Expressway Corporation (KEC) had started the building of Freeway Traffic Management System (FTMS) since 1993. The system is now operating on all nationwide expressways. The data collection methods being used are Vehicle Detection System (VDS) and Toll Collection System (TCS).

However, the accuracy of VDS needs improvement because it estimates the travel time by converting the 'spot speed', which is estimated by the detection-time-difference between two detectors installed in one location, to 'spatial average speed'.
The TCS also has an issue of relatively big 'time lag'; because the data for TCS can be collected only after the vehicle would arrive at the next operation point from previous operation point while the distance between the two operation points is too far.

In order to resolve these issues in FTMS, KEC began to build ‘Hi-pass Traffic Information System’ as a part of FTMS. The Hi-pass Traffic Information System is based on dedicated short-range communication (DSRC) and it uses Hi-pass On-board equipment (OBE) as the probe. The system has 305 Roadside equipment (RSE) units at every 3km of 1,050km length covering 15 expressways. The system installation is still ongoing. Meanwhile, more than 3 million units of Hi-pass OBE are being used now.

However, there are no enough studies on the characteristic of data to be collected in accordance with the introduction of the new system. Also, there is no proper pre-processing algorithm for the new system. Accordingly, this study will develop an efficient pre-processing algorithm by way of analyzing the characteristic of collected data.

2. LITERATURE RIVIEW

The statistical method to remove the outliers is removing the values that exceed the upper limit and lower limit after estimating the average traveling time of each vehicle. Methods to estimate such range are ‘median absolute deviation method’ and ‘confidence interval method’.

The median absolute deviation method (MAD) uses ‘median absolute deviation’ instead of ‘standard deviation’ to identify the data distribution in the removal of outliers and the estimation of traveling time. The method is often considered in the robust estimation. It is an outlier method which does not have to assume the data distribution (Xuegang et al., 2010; Barnett and Lewis, 1984; Hoaglin et al., 1983).

The confidence interval method uses the confidence interval estimated by average and standard deviation. Kang et al.(2002) removed the data that exceed 100% of design speed or data that are less than 10km/h. Then they assumed that the values exceeding 68% of confidence interval as the outliers and removed those.

Do et al.(2008) applied ‘median absolute deviation method’ and ‘confidence interval method’ to TCS data. In case of ‘median absolute deviation method’, the representative values were underestimated because of the data with short traveling time. The segment distance also increased and the representative values were more distorted when the dispersion was big. In case of ‘confidence interval method’, the representative values were overestimated because relatively bigger range data were handled as valid data. This was caused by the issue of normality assumption of the samples.

‘TransGuide’ is the algorithm being used by San Antonio Expressway Traffic Management System in U.S. The traveling time estimate using AVI data is a moving-average algorithm that automatically removes the traveling time values that exceed the range determined by the user within the relevant collection period (Dion and Rakha, 2003).

‘Transmit’ is the algorithm being used in New York and New Jersey. It uses a data-smoothing method during 15 minutes collection period. This method first takes the average of traveling time of relevant collection period. Then the data are smoothed by the data of same time band and same weekday in the past so that updated average traveling time can be obtained. The method gives updated average traveling time as above and the average traveling time that was smoothed during the previous collection period.

The existing studies on outlier removal are mostly based on statistical methods or they use smoothing method. However, the statistical methods assume normal distribution. If the distribution of collected data is not normal, there can be a problem. Therefore, development of a outlier removal method that does not have to consider the data distribution is required.

3. DATA ANAYSIS

Hi-pass Traffic Information System can estimate traveling time by the difference in collection times of unique vehicle ID collected through the RSE. In addition, it can classify the vehicle types; therefore, if the distribution of traveling time would be different dependent on vehicle type, the system can classify them in advance.

The traveling time distribution of passenger cars in general lanes and buses in High Occupancy Vehicle (HOV) lane were analyzed. When the traffic is smooth, there was no difference; however, when there was traffic congestion, differences in the traveling time distribution occurred as shown in Figure 1. When the statistical outlier removal method was applied on the two vehicle types, the dispersion was too big and even normal values were removed as outliers. Therefore, the bus data had to be removed in advance.

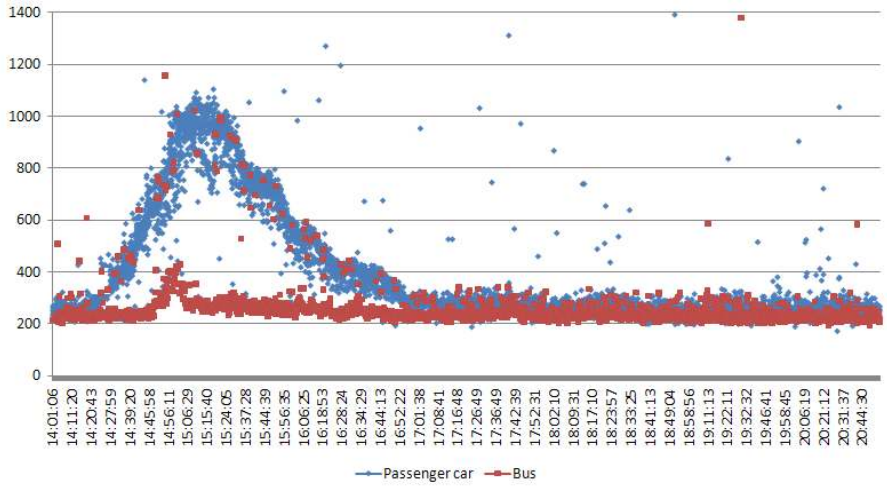

Figure 1 - Plotting of passenger cars and buses traveling time

When the normality test was done on the passenger cars, there were differences dependent on the segment; however, there was no normality in the time band of 30 to 55 %. Therefore, it is believed that a statistical method that assumes the normal distribution cannot be applied.

Table 1 - Number of samples collected by segment and the ratio of normal distribution

Segment	Start RSE ID	End RSE ID	Length (km)	# of Sample (Vehicles/5min)	Ratio of Normal Distribution (%)
Seoul~Suwon	4043	3936	10.7	56	67
Suwon~Giheung(old)	3936	3877	5.9	57	55
Giheung(old)~Shin-giheung	3877	3855	2.2	47	55

Shin-giheung~Osan	3855	3789	6.6	42	44
Osan~Anseong	3789	3619	17	32	60
Anseong~Cheonan	3619	3423	19.6	42	45
Cheonan~Mokcheon	3423	3319	10.4	23	65
Mokcheon~Cheongju	3319	3068	25.1	22	39
Cheongju~Cheongwon	3068	2952	11.6	17	71
Cheongwon~Shintanjin	2952	2852	10	29	68
Shintanjin~Daejeon	2852	2744	10.8	16	67

These results indicate that the outliers should be removed after removing the bus data first and another method is required for the outlier removal, other than the statistical method assuming normal distribution.

4. DATA PRE-PROCESSING ALGORITHM

The data analysis results indicated that buses using HOV lane show different traveling characteristic from passenger cars when there is traffic congestion. And it was found that the outlier removal method assuming normality cannot be applied; because there are too many data without normality. The pre-processing algorithm was developed considering these issues. The pre-processing algorithm consists of; error data handling, outlier removal and summation.

The Hi-pass Traffic Information System can track the vehicle type and its route. Using this characteristic, buses and passenger cars using service area and alternative roads were judged as error data and removed. In case of buses, vehicles detected by roadside equipments (RSE) at service areas were judged as error data and removed. In case of passenger cars using alternative roads, if the traveling sequence of roadside equipments (RSE) is not in sequence, they were judged as error data and removed.

The outlier removal is done in two stages.
First, since the driving styles of drivers are different from each other, they cannot meet the principle of FIFO (First In, First Out). However, if they would travel in normal way, the difference in traveling time will not be big. Therefore, the first outlier removal algorithm removes the vehicles that have big difference from other vehicles due to abnormal traveling. For this, the traveling times of the vehicles were lined-up in sequence based on their departure times. Then the traveling times of vehicles for judgment were compared with vehicles before and after the vehicles for judgment. Data that exceeds the set up range were judged as outliers and removed. The formulas are as following.

IF ($TT_t > \alpha \times TT_{t-1}$ **and** $TT_{t-1} + TT_{t+1} < TT_t$**),** TT_t **is an outlier**

IF ($TT_t > \frac{TT_{t-1}}{\alpha}$ **and** $TT_{t-1} < TT_{t+1} + TT_t$**),** TT_t **is an outlier**

Where,
TT : traveling time of each vehicle
t :sequence of vehicle lined-up based on their departure times
α : a parameter to test outliers (Default=2.0)

Since the data with big impact had been removed in the first outlier removal, the average does not get big impact. Therefore, a variation coefficient was applied on the average traveling speed in the second outlier removal. The variation coefficient had been estimated based on the analysis data. It was found to be average 0.3. The formula for the second outlier removal is as following.

IF ($TV_i < \overline{TV_t} \times (1-\beta)$ **or** $TV_i > \overline{TV_t} \times (1+\beta)$ **),** TV_i **is an outlier**

Where,

TV_i : traveling time of each vehicle

$\overline{TV_t}$: average traveling speed of t summation period

β : the variation coefficient (0.3)

T-test with significance level 0.05 was done in order to compare the error-data handling and outlier-removal method developed in this study with the 'median absolute deviation method' used in previous studies. The P-value was more than 0.86, which was bigger than significant probability 0.05. Therefore, there is no difference in traveling time between the 'median absolute deviation method' and the proposed method in this study.

When the distribution was unbalanced to one side, the result was similar; however, when the distribution was 'bimodal', the dispersion of proposed method was smaller than the dispersion of 'median absolute deviation method'.

Based on these results, it is believed that the proposed method will be quite useful because it can obtain similar results with 'median absolute deviation method' while it does not assume the normality.

Table 2 - The application results of 'median absolute deviation method' and proposed method.

Direction	Start RSE	End RSE	Length (km)	Number of samples (per day)	Median absolute deviation method		Proposed method		T-test result
					Normal value	Outlier	Normal value	Outlier	P-value
NB	3619	3669	5	14,742	14,353	389	14,518	224	0.946
	3936	3974	3.8	16,275	15,541	734	15,614	661	0.998
	3841	3855	1.4	17,574	17,251	323	17,290	284	0.938
SB	3669	3619	5	12,966	12,053	913	12,070	896	0.870
	3974	3936	3.8	16,741	16,015	726	16,097	644	0.857
	3855	3841	1.4	15,011	14,383	628	13,973	1,038	0.962

5. CONCLUSION

The Hi-pass Traffic Information System based on DSRC is being built to resolve the issues in existing FTMS. However, there is no proper method for the analysis and pre-processing of traveling data. Accordingly, this study suggested a pre-processing algorithm based on data analysis.

According to the results of data analysis, buses using HOV lane show different traveling characteristic from passenger cars when there is traffic congestion. It was found that the outlier removal method assuming normality cannot be applied because there are too many data without normality.
A pre-processing algorithm was developed considering these issues. When the result of proposed algorithm and the result of 'median absolute deviation method' were compared, they were similar.
The proposed method in this study is expected to be quite useful because it does not assume normality, different from the 'median absolute deviation method'.
Meanwhile, since the analysis of this study was done in the limited segments, further examination would be required with more segments.

ACKNOWLEDGEMENTS

"This work was partially supported by the National Research Foundation of Korea grant funded by the Korea government(MEST) (NRF-2010- 0028693)."

REFERENCE

1. Barnett V., and Lewis T., Outliers in Statistical Data, New York: Wiley, 1984
2. Dion F., and Rakha H., Estimation Spacial Travel Time using Automatic Vehicle Identification Data, Transportation Research Board, 2003
3. Hoaglin D. C., Mosteller F., and Tukey J. W., Understanding Robust and Exploratory Data Analysis, New York; Wiley, 1983
4. Jin-Kee Kang, Youngtae Son, Yeo-Hwan Yoon, and Sangchul Byun, Regional Traffic Information Acquisition by Non-intrusive Automatic Vehicle Identification, Journal of Korea Society of ITS, 1(1), 2002
5. Myungsik Do, Hyangmi Lee, and Jake Namkoong, Outlier Filtering and Missing Data Imputation Algorithm using TCS Data, Journal of Korea Society of Transportation, 26(4), pp. 241-250
6. Xuegang J. B., Yuwei L., Alexander S., and Margulici J. D., Performance Evaluation of Travel-Time Estimation Methods for Real-Time Traffic Applications, Journal of Intelligent Transportation Systems, 14(2), pp. 54-67, 2010