Study on Enhancement of TRANSGUIDE Outlier Filter Method under Unstable Traffic Flow for Reliable Travel Time Estimation
-Focus on Dedicated Short Range Communications Probes-

Moataz Bellah Ben Khedher[1], Duk Geun Yun[2*]

[1]Dep. of ITS Engineering, University of Science and Technology

[2]Highway & Transportation Research Institute, Korea Institute of Civil Engineering and Building Technology / Dep. of ITS Engineering, University of Science and Technology

불안정한 교통류상태에서 TRANSGUIDE 이상치 제거 기법 개선을 통한 교통 통행시간 예측 향상 연구
-DSRC 수집정보를 중심으로-

모타즈 케들러[1], 윤덕근[2*]
[1]과학기술연합대학원대학교 교통물류 및 ITS 공학과
[2]한국건설기술연구원 도로연구소 /과학기술연합대학원대학교 교통물류 및 ITS 공학과

Abstract Filtering the data for travel time records obtained from DSRC probes is essential for a better estimation of the link travel time. This study addresses the major deficiency in the performance of TRANSGUIDE in removing anomalous data. This algorithm is unable to handle unstable traffic flow conditions for certain time intervals, where fluctuations are observed. In this regard, this study proposes an algorithm that is capable of overcoming the weaknesses of TRANSGUIDE. If TRANSGUIDE fails to validate sufficient number of observations inside one time interval, another process specifies a new validity range based on the median absolute deviation (MAD), a common statistical approach. The proposed algorithm suggests the parameters, a and β, to consider the maximum allowed outlier within a one-time interval to respond to certain traffic flow conditions. The parameter estimation relies on historical data because it needs to be updated frequently. To test the proposed algorithm, the DSRC probe travel time data were collected from a multilane highway road section. Calibration of the model was performed by statistical data analysis through using cumulative relative frequency. The qualitative evaluation shows satisfactory performance. The proposed model overcomes the deficiency associated with the rapid change in travel time.

요 약 신뢰성 있는 통행시간 예측을 위해 DSRC로부터 수집된 통행시간에서의 이상치(outlier) 필터링은 필수이다. 통행시간 예측을 위해 사용되는 보편적 기법인 TRANSGUIDE는 특정 분석 시간동안 통행시간의 변동이 크게 발생하는 조건에서 수집데이터의 이상치 제거를 효율적으로 처리하지 못하는 문제점이 존재한다. 이에 본 연구에서는 TRANSGUIDE의 한계점을 보완할 수 있는 알고리즘을 제안하고자 한다. TRANSGUIDE가 특정 분석 시간대 충분한 데이터 관측이 어려울 경우 Median Absolute Deviation(MAD)를 이용하여 이상치 제거를 위한 새로운 유효 분석 영역을 설정하였다. 새로운 분석 영역 설정 후 특정 시간대 교통 조건하에서 최대 허용 가능한 이상치를 고려한 변수 α, β를 제안하였다. 변수 α, β를 추정하기 위해 과거 데이터와 도로 구간의 특성을 반영하였다. 개발된 알고리즘은 수도권 일반국도 3호선, 2013년 1월 1달간 DSRC 데이터가 존재하는 다차로 일반국도에 적용하였다. 누적상대도수를 이용하여 모형의 정산 수행 후 성능에 대해 정량적 평가를 수행하였다. 개발된 알고리즘은 기존의 TRANSGUIDE가 특정 조건, 즉 일정 분석 시간동안 교통 조건이 급하게 변동되는 구간에서 이상치 제거에 취약한 점을 보완하는 것으로 판단되었다. TRANSGUDIDE가 특정 조건에서 통행시간 예측이 어려울 경우 본 개발 알고리즘은 활용될 것으로 판단한다.

1. Introduction

Travel time information is necessary to provide travelers and road transportation agencies with accurate data that may be used to make decisions on their current trips namely roadways that show high variability in traffic flow. Travel time is the best indicator of the level of service in a road section as it can be easily understood by all road users [1], offering them the ability to pick the least congested route and fastest path toward their destination before starting the trip based on Advanced Traveler Information Systems (ATIS). The travel time data collection is made through various systems that take advantage of the current advancement in technologies and forward it to ATIS which processes data and provide estimation results to public. Among these technologies, loop detectors, magnetic signature re-identification, automatic license plate recognition, closed circuit televisions (CCTV) and others have been widely used. Also, Dedicated Short Range Communication (DSRC) probes are attracting more interest. Originally, DSRC is used for electronic toll collection system by installing on-board units in vehicles. By the end of 2014, half of registered cars in South Korea are equipped with DSRC on-board units [6]. DSRC is a direct measurement method for travel time estimation that provides time stamps at which the vehicle passes two successive control points. DSRC makes use of a unique identification of on-board units placed on vehicles passing through two consecutive road‐side units (RSUs) and the travel time is acquired by pairing both records.

Fig. 1 shows a simple illustration of the problem. A vehicle i is detected the first time by the road-side unit A at time t_A where both t_A and the unique identification of corresponding vehicle are recorded. When vehicle i arrives at RSU B, the vehicle is re-identified and the arrival time t_B is recorded. Suppose N vehicles are detected during a given interval of time t, the problem can be formally expressed as follow:

$$T_{AB}(t) = \frac{\sum_{i}^{N}(t_{B_i} - t_{A_i})}{N} \qquad (1)$$

where $T_{AB}(t)$ is the link travel time, N is the total number of detected vehicles, t_{B_i} is arrival time of vehicle i and t_{A_i} is entry time of vehicle i. By matching DSRC probe data and measuring the time interval needed by a car to pass through a road segment, an individual vehicle travel time is obtained. However, in some cases, the reported result does not show a representative value of travel time. Methods based on direct measurement may be inaccurate due to the presence of outliers in the travel time observations. An outlier is an observation that its corresponding time stamp is numerically distant from the other observations, and consequently, does not comply with general pattern of the data. The presence of outliers is explained by many reasons including vehicle stop between RSUs, vehicle entering/exiting between measuring stations, abnormal driving behavior and also possible measurement error. Also, unlike license plate recognition systems, DSRC collects probe data bi-directionally and because of that, this situation takes place many times mainly in urban and rural highways which have many intersections. The observation that indicates a very long travel time compared to other observations, is considered as an outlier. Also, abnormal short travel time is considered as outlier since it can be due excessive driving speed and because of that it does not represent the overall link travel time. Thus, outliers have to be identified in the dataset and eliminated. Therefore, many researchers have proposed outlier filtering algorithms in order to tackle this problem, but, no optimal outlier treatment method have been found. TRANSGUIDE is one of the earliest developed algorithms proposed by the Southwest Research Institute (SWRI) in 1998 and it is based on moving average [2]. Previous research emphasized limitations of this algorithm and mainly the poor response to rapid change in travel time in certain aggregation time. Our study intends to enhance the

performance of TRANSGUIDE by incorporating an algorithm that identifies data based on median absolute deviation at time intervals where TRANSGUIDE fails to perform and, instead, identifies valid observations as outliers.
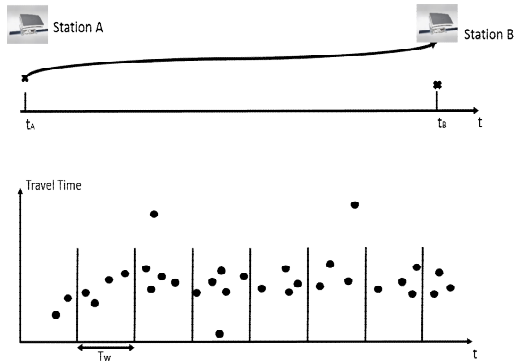

Fig. 1. Illustration of travel time estimation problem

2. Literature review

The existence of outlying observations in collected travel times influences the accuracy of travel time estimation. In this regard, various outlier filtering algorithms were proposed for Automatic Vehicle Identification (AVI) based technologies. In this section, we introduce some previous studies that investigated this issue. Numerous authors have been dealing with outlier elimination for the measured travel times that are not related to traffic conditions on road segments but are the outcome of individual vehicle behavior. TRANSTAR Algorithm was proposed by SWRI [3]. This algorithm identifies travel time by calculating the average of current observations that are within a user-defined threshold based on the previous average. The travel time is recalculated every time there is a new observation detected. Say for example the threshold is 20%, if the new travel time is within 20% from the previously estimated link travel time, this observation is considered as valid and average link travel time is updated, otherwise, the observation is an outlier and link travel time remains unchanged.

TRANSMIT is another outlier filtering algorithm. It is relatively similar to TRANSTAR but the link travel time is updated at fixed time intervals which is 15-minute observation intervals with no threshold restrictions [4]. Instead, the average of observations goes through a smoothing process using a smoothing factor that is calculated based on historical data. Dion and al. proposed a low-filtering algorithm based on the assumption of log-normal distribution of probe travel times. This algorithm is based on several factors; the expected average travel time and variability in the upcoming time interval, number of successive intervals with no readings since the last obtained record, number of successive outliers and the travel time variability inside an analysis interval. The data validity window is dynamically varying according to the number of observations inside one sampling interval [3]. Ma and al. proposed an approach based on median filter as a measure of location [5].The authors evaluated the statistical median filtering approach and suggested modifications for automatic vehicle identification data estimation. Jang proposed an outlier filtering algorithm for DSRC probe travel times on signalized rural arterials under low size sample situations [6]. The proposed method handles two situations. For the low size sample intervals, the algorithm allocates a validity range based on previous interval value, otherwise, the algorithm implicates a modified z-score to specify a modified median filter in order to determine a new validity range. TRANSGUIDE algorithm, is one the earliest outlier filtering algorithms. It was proposed by the SWRI and it is based on moving average [2]. In this paper, a major deficiency in TRANSGUIDE algorithm performance will be outlined and a framework will be proposed to overcome the examined problem.

3. TRANSGUIDE filtering algorithm

3.1 TRANSGUIDE framework

In the TRANSGUIDE system, link travel time between successive AVI readers is obtained by

automatically filtering out all records of travel times that are beyond a user pre-defined link threshold travel time using a rolling average algorithm. The algorithm is defined by the following Equations:

$$S_{AB}(t) = \left\{ \begin{array}{l} t_{B_i} - t_{A_i} \mid t - t_w \le t_{B_i} \le t \cap \\ T'_{AB}(t)(1-th) \le t_{B_i} - t_{A_i} \le T'_{AB}(t)(1+th) \end{array} \right\} \quad (2)$$

$$T_{AB}(t) = \frac{\sum_{i}^{|S_{AB}(t)|} (t_{B_i} - t_{A_i})}{|S_{AB}(t)|} \quad (3)$$

where $S_{AB}(t)$ is the set of valid recorded travel times from reader A to reader B at time t, t_{A_i} is the detection of vehicle i at reader A (seconds), t is time at which travel time estimation takes place (seconds), t_w is Rolling average window (seconds), th is user pre-defined link threshold travel time parameter, $T_{AB}(t)$ is the average travel time from reader A to reader B that is estimated at time t (seconds), and $T'_{AB}(t)$ is the previously estimated average travel time from reader A to reader B.

Eq. (2) defines the set of valid recorded travel times that are used at each evaluation time to estimate the current average travel time between two AVI readers. Eq. (3) is the calculation of the new estimated average travel time from the valid observations. The leading parameters of TRANSGUIDE algorithm are the rolling average window t_w and the link threshold travel time th. The rolling average window specifies the interval of time (aggregation time) that is required to be considered when estimating the current average travel time. For example, if aggregation time t_w is set to 5 min and a threshold of 20%, then any observation from travel time that deviates by 20% from the previously estimated travel time will be considered as outlier and thus, eliminated from the estimation of travel time in that interval.

In the following sections, a major deficiency in TRANSGUIDE algorithm performance will be outlined and a framework will be proposed to overcome the examined problem.

3.2 Performance of TRANSGUIDE

The TRANSGUIDE outlier filtering algorithm is based on moving average. The choice of parameters t_w and th defined by the user and thus requiring historical data examination and an engineering judgment for final decision. The original report of TRANSGUIDE recommends 2 minutes interval for the rolling average window and a threshold of 20%. The performance of TRANSGUIDE under different values of the parameters was investigated in previous research [3,5]. The choice of aggregation time is important since two major objectives have to be reached: the rolling average window has to be large enough so that it can be capable of acquiring suitable number of observations for more reliable travel time estimation, at the same time, t_w has to be the shortest possible in order to provide updates more frequently and keep track of travel time progression without smoothing the rapidly changing conditions [1].

Previous studies[3], show that TRANSGUIDE performs poorly in case of sudden change in travel time. Fig. 2 shows an example for the result of application of TRANSGUIDE on 24 hours block collected travel time data in January, 10. In this example, we used a rolling average window of 5 minutes and link threshold of 20%. In time interval between 7AM and 9AM, the algorithm labeled all travel time observations as outliers. The validity window between the upper bound and lower bound could not track the rapid change in travel time. For the other time intervals, the algorithm shows satisfactory results. A similar poor performance was examined when testing the algorithm against data from other week days.
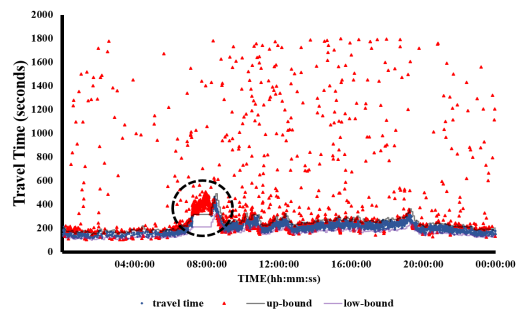

Fig. 2. Performance of TRANSGUIDE in January, 10.

4. Proposed model

Based on the aforementioned deficiency, TRANSGUIDE proved the failure of dealing with sudden change in travel time. Therefore, we propose a model that can detect and respond to the rapid change in travel time. First, we input the raw data of matched travel times. Next, an initialization step is required to specify the rolling average window and link threshold. Then, travel time observations, in the current time interval, that fall outside the validity window specified by Eq. (2) are considered as candidate outliers. Up to this stage, all performed operations are adopted from TRANSGUIDE. In order to acknowledge the previous results, our method verifies if total number of candidate outliers and the relative fraction of outliers in the current time interval fall outside two maximum allowed limits. This is realized through the following condition:

$$\text{if } (n_c > \alpha) \text{ and } \frac{n_c}{n_o} > \beta) \qquad (4)$$

Where, n_c is the number of candidate outliers in the current time interval and n_o is the total number of observations inside the current time interval. The parameter α is the maximum allowed number of outliers for one time interval and β is the maximum allowed relative fraction of number of outliers over total number of travel time observations per time interval. Parameters α and β are related to the characteristics of road section from where the travel time observations are obtained. They rely on the traffic and road conditions of a specific road segment including traffic flow, road geometry and road type (e.g. rural or urban). Accordingly, an estimation of both parameters can be obtained off-line using historical data and hence, updated on a regular basis. An example of parameter estimation is shown in the next section.

In case the number of candidate outliers and the relative fraction of outliers are within the specified conditions, our method approve the obtained results and thus, the outliers are eliminated and the output carries only the valid travel observations that will be included in the new estimation of travel time using Eq. (3). However, if the conditions are not meet, the algorithm rejects the previous result and initializes another process to identify outliers. A new validity window is established based on Median Absolute Deviation (MAD). The absolute deviation from the median was re-exposed by Hampel who accredits the idea to Carl Friedrich Guass [7]. It is also known as the median absolute deviation about the median. It is a very robust scale estimator. The Median is, similar to the mean, a measure of central tendency. The advantage of median over the mean is that the median is less sensitive to the presence of outliers [8]. Therefore, MAD is straightforward since it implicates finding the median of absolute deviations from the median. The Median Absolute deviation is defined as follow:

$$MAD = c \times median\left(\left|(t_{B_i} - t_{A_i}) - median(n_o)\right|\right) \quad (5)$$

where, $c = 1.4826$, is a constant associated with the assumption of normality of data. This constant is needed to ensure the consistency of estimator for the parameter of interest. Once MAD is calculated, a new validity window has to be defined. We must define a rejection criteria for observations. Outlier rejection based on MAD has been applied in many other science fields. Miller suggests the values 3 (very conservative), 2.5 (moderately conservative) or 2 (poorly conservative) [9].The selection of the value is relatively subjective. In our algorithm, we decide to acquire very conservative value which means choosing value 3 as a rejection criterion. So, the new set of valid travel time records is defined as follow:

$$S_{AB}(t) = \left\{ \begin{matrix} t_{B_i} - t_{A_i} | t - t_w \le t_{B_i} \le t \cap \\ M - 3MAD \le t_{B_i} - t_{A_i} \le M + 3MAD \end{matrix} \right\} (6)$$

where M is the median value of all travel time observations within an interval of time. At this point,

the filtering process is accomplished and valid travel time observations are used to estimate the new average travel time using Eq. (3). The flowchart of the proposed filtering mechanism is shown in Fig. 3.
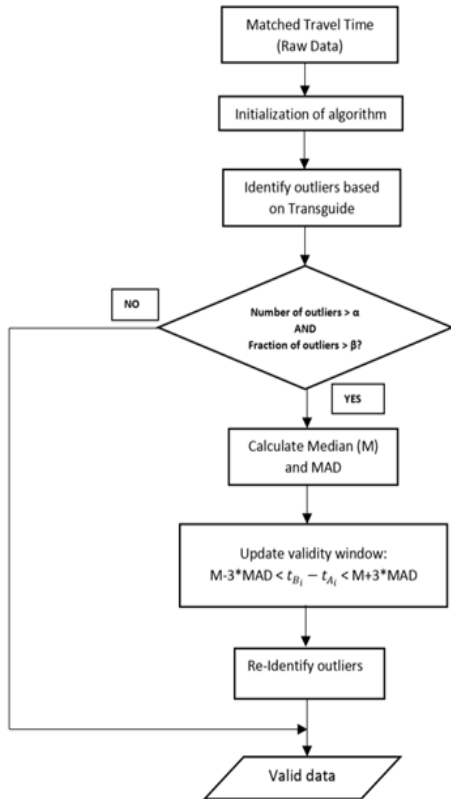

Fig. 3. Flowchart of the proposed model.

Fig. 4. Collection site.

In order to estimate the values of α and β, we do a statistical analysis based on historical data. To achieve this end, we select two data samples from historical data where TRANSGUIDE shows a satisfactory performance. The first sample contains values of number of outliers that were detected in each time interval of five minutes. The second data sample contains fractions of the number of labelled as outliers inside each 5-minute interval. Both datasets are selected from five weekdays where no traffic fluctuations was observed. A data summary of travel times during the chosen 5 weekdays is presented in Table 1 after application of TRANSGUIDE filter.

Table 1. Summary of data used for parameter estimation

	Day 1	Day 2	Day 3	Day 4	Day 5
Sample size	5008	6155	5990	6896	6118
Average	263.6	234.5	229.6	233.0	238.7
median	205	230	224	230	219
Standard Deviation	246.4	50.1	51.7	45.9	76.4
85th percentile	242	285	280	279	287
Number of 5-minutes Intervals	145	288	288	288	287
Average observations per 5-minutes	34.5	21.4	20.8	23.9	21.3
Average number of outliers per 5-minutes	4.4	3.9	3.4	3.6	3.2
Average ratio of outliers per 5minutes	0.1	0.2	0.2	0.2	0.2

5. Application of the proposed model:

After developing the model, parameter estimation is required to calibrate the algorithm followed by an application of the model to verify the performance. For this reason, probe travel times were collected using DSRC from a multilane highway near Seoul Area on National Highway No. 3. The suburban roadway section spans around 3km with one intersection and one interchange as shown in Fig. 4. We obtained probe travel time records for 30 days during the month of January 2013.

Next, we calculate the cumulative relative frequency that shows the proportion of data items. Results shown in Fig. 5 reveal that for a rolling average window of five minutes, less than 1% of travel time intervals had more than 20 observations are classified as outliers.

Similarly, as shown in Fig 6, less than 1% of intervals had 90% of travel time observations labelled as outliers. We assume that there is a very low probability that number of outliers and ratio of travel time observations labelled as outliers will exceed the 99th percentile corresponding value from historical data in 5-minute time interval for this roadway section. Accordingly, the selected value are $\alpha = 20$ and $\beta = 0.9$.

Once parameters are obtained, we test our proposed model using the previously estimated values for parameters. Also, the link threshold and rolling average are set to 20% and 5 minutes, respectively. Among the datasets that were obtained from collection, four of them TRANSGUIDE shows a poor response for the sudden change in travel time. Therefore, we show the result of application of our model for filtering outliers.
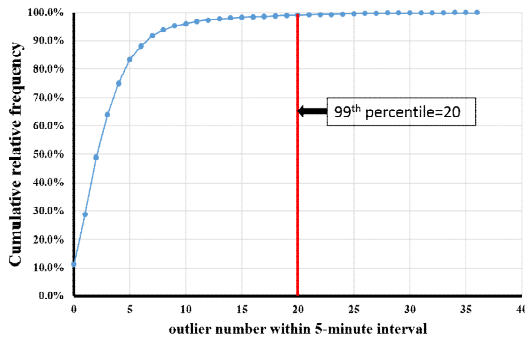

Fig. 5. Number of outliers inside 5-minute interval.
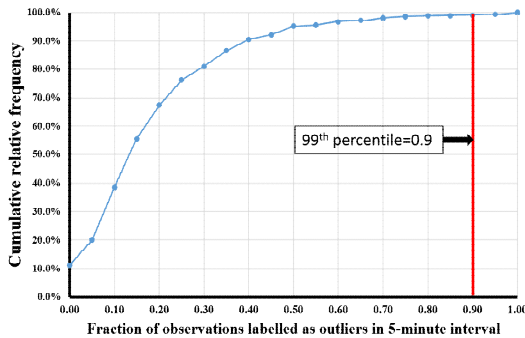

Fig. 6. Fraction of outliers inside 5-minute interval.
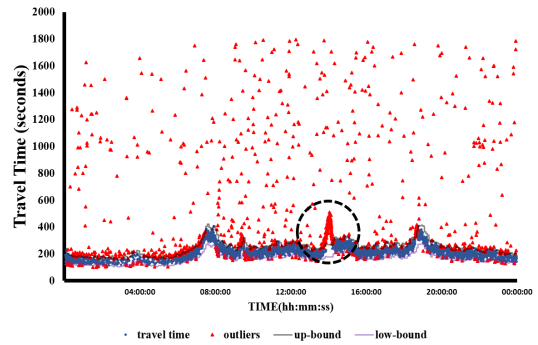

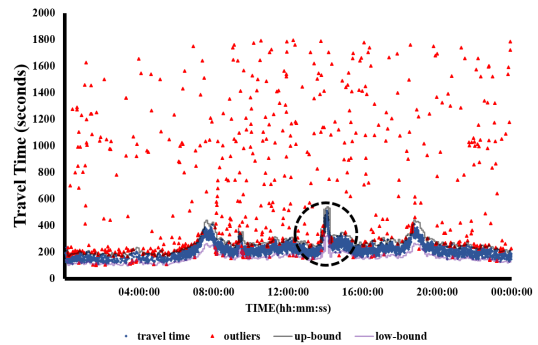Fig. 7. TRANSGUIDE Results (January, 17).

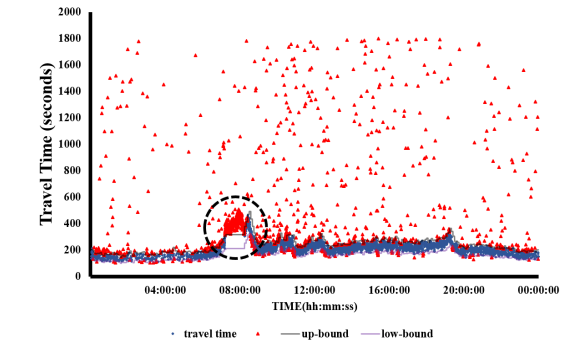Fig. 8. The proposed model results (January, 17).
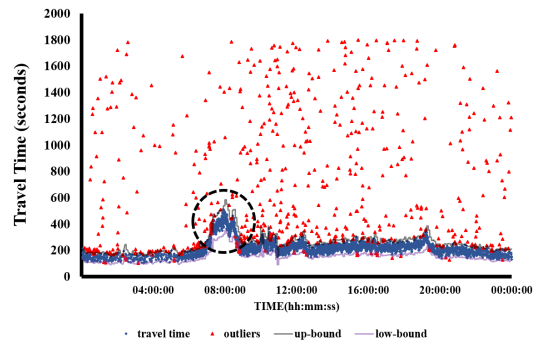

Fig. 9. TRANSGUIDE results (January, 10).

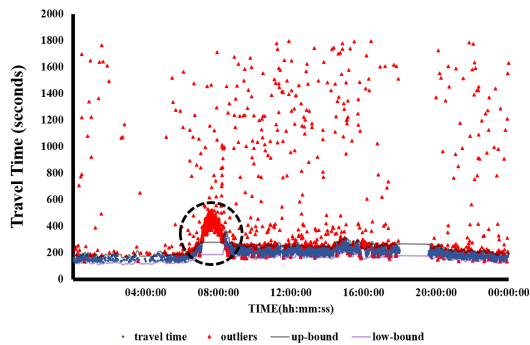Fig. 10. The proposed model results (January, 10).

Fig. 11. TRANSGUIDE results (January, 28).

As shown in Fig.7, original TRANSGUIDE algorithm failed to identify valid travel times in the time period between 1PM and 2PM, where all observations are labeled as outliers and because of that, no travel time estimation is performed for that specific duration of time.
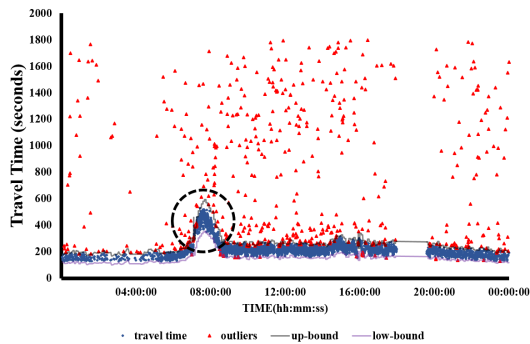

Fig. 12. The proposed model results (January, 28).

However, our model successfully overcome the deficiency associated with the rapid change in travel time as shown in Fig. 8. When TRANSGUIDE was unable to produce an estimation for travel time, our method is executed to determine the new validity range. Due to the absence of out-of-sample data for validation, approval of the outlier filtering result is qualitative and established on whether travel profile capture the trend illustrated by the majority of collected data. More results are shown in figures from Fig.9 to Fig.12. The results are associated with previously calibrated parameters where satisfactory results are

obtained. The figures unveil the ability of the proposed filtering algorithm to effectively respond to abrupt increases in travel times. Further research is required to evaluate the overall performance of the algorithm. Note that for Fig. 11 and Fig. 12, the gap between 6PM and 8PM is due to missing raw data originated from a system breakdown during that interval of time.

6. Conclusions and further study

Filtering the data for travel time records obtained from DSRC probes is necessary for better estimation of link travel time. In this study, a major deficiency in the performance of TRANSGUIDE in removing anomalous data was outlined. This algorithm is unable of handling unstable traffic flow conditions for certain time intervals where fluctuations are observed. In this regard, we proposed an algorithm that is capable of overcoming the weakness of TRANSGUIDE. In case TRANSGUIDE fails to validate sufficient number of observations inside one time interval, another process specifies a new validity range based on median absolute deviation, a common statistical approach. Parameter estimation of the model rely on historical data, as it has to be updated frequently and depending on the road segment characteristics. To test the proposed algorithm, DSRC probe travel time data were collected from multilane highway road section. Calibration of the model was made using statistical data analysis with the use of cumulative relative frequency. The results show a satisfactory performance for specific cases where rapid change in travel time observed. Even though the application of the proposed algorithm successfully handled the aforementioned failure, the evaluation was based only on qualitative validation. A further research is required to validate the findings of the current study and evaluate the performance travel time probes from other road types (freeways). In addition, the rejection criterion need to be investigated in future to justify the choice of the

value. The proposed extension will be tested with other outlier filtering algorithms that suffer from similar deficiencies such as Dion and Rakha method. The algorithm will be applied to travel time probes collected from smart phones which is a widely used practice in South Korea.

REFERENCES

[1] Soriguera, F., Rosas, D., Robusté, F., Travel time measurement in closed toll highways, *Transportation Research Part B*, vol. 44, no. 10, pp. 1242-1267, 2010. DOI: https://doi.org/10.1016/j.trb.2010.02.010

[2] Automated Vehicle Identification Model Deployment Initiative System Design Document, *Texas Department of Transportation*, San Antonio, 1998.

[3] Dion, F., and H. Rakha. Estimating Dynamic Roadway Travel Times Using Automatic Vehicle Identification Data, *Transportation Research Part B*, Vol. 40, No. 9, pp. 745‒766, 2006. DOI: https://doi.org/10.1016/j.trb.2005.10.002

[4] Mouskos, K.C., E. Niver, L.J. Pignataro, and S. Lee. TRANSMIT System Evaluation, Final Report, *Institute for Transportation, New Jersey Institute of Technology*, Newark, NJ, 1998.

[5] Ma, X., and H. Koutsopoulos. Estimation of the Automatic Vehicle Identification Based Spatial Travel Time Information Collected in Stockholm, *IET Intelligent Transport Systems*, Vol. 4, No. 4, 2010. DOI: https://doi.org/10.1049/iet-its.2009.0149

[6] J. Jang. Data-Cleaning Technique for Reliable Real-Life Travel Time Estimation. Transportation Research Record: *Journal of the Transportation Research Board, No. 2593, Transportation Research Board*, Washington, D. C., pp. 85‒93, 2016.

[7] Hampel, F. R. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 383‒393, 1974. DOI: https://doi.org/10.1080/01621459.1974.10482962

[8] Huber, P. J. Robust statistics, *New York: John Wiley*, 1981. DOI: https://doi.org/10.1002/0471725250

[9] Miller, J. Reaction time analysis with outlier exclusion: Bias varies with sample size, *The Quarterly Journal of Experimental Psychology*, vol. 43, no. 4, pp. 907‒912, 1991. DOI: https://doi.org/10.1080/14640749108400962

Moataz Bellah Ben Khedher   [Regular member]

- Feb. 2015 : KAIST, Bachelor, Dep. of Electrical Engineering
- March. 2015 ~ current : University of Science and Technology, MS, ITS Engineering

<Research Interests>
ITS, Travel time estimation, Outlier filtering

Duk Geun Yun   [Regular member]

- Feb. 2009 : Myongji Univ. Ph. D.
- March. 2001.1 ~ current : Senior Researcher of KICT
- March. 2011.1 ~ current : Associate Professor of University of Science and Technology

<Research Interests>
Road Safety, Road Maintenance, Road Design, Autonomous Vehicle