Journal of the Korea Academia-Industrial cooperation Society
Vol. 22, No. 10 pp. 1-10, 2021

https://doi.org/10.5762/KAIS.2021.22.10.1
ISSN 1975-4701 / eISSN 2288-4688

Statistical Method for Real-Time Detection of Travel Time Outliers

Moataz Bellah Ben Khedher[1], Duk Geun Yun[2*], Jung Gon Sung[2]
[1]Department of Civil & Environmental Engineering, KICT School, University of Science and Technology
[2]Department of Highway & Transportation Research, Korea Institute of Civil Engineering and Building Technology

실시간 통행시간 이상치 판별을 위한 통계적 방법

모타즈 케들러[1], 윤덕근[2*], 성정곤[2]
[1]과학기술연합대학원대학교, [2]한국건설기술연구원 도로연구소

Abstract Travel time information plays a crucial role in supporting road users and transportation agencies with precise data that can be used for traffic analysis and route choosing. However, the acquired data does not always show a representative value of the travel time due to the presence of outliers. The outlier must hence be removed, and various techniques of outlier filtering rely on parametric models that require extensive historical data for parametric estimation. In this study, we propose a non-parametric outlier filtering method for a reliable travel time estimation. As part of the study, an Outskewer method was adopted and enhanced to overcome its deficiency and to be able to detect outliers in online mode. Dedicated short-range communication probe data, collected from a multi-lane highway, was used for the quality assessment. In lack of ground truth records, the performance of the proposed algorithm was evaluated qualitatively, based on graphs and quantitatively using the confidence interval method. Visual inspection of graphical results shows the satisfactory performance of the present method. The hypothesis test for travel time data quality indicates a maximum relative error of less than 20% and a percentage correct classification of higher than 90%. Hence, the proposed methodology can serve as an alternative to other parametric methods used for travel time outlier treatments and can support the delivery of more accurate information for road travelers and transportation practitioners.

요 약 도로 이용자나 도로관리자에게 교통 분석이나 경로 선택에 있어 통행시간은 중요한 정보이다. 그러나 수집된 개별 통행시간에 이상치(outlier)가 존재함으로 인해서 수집된 통행시간 정보는 항상 그 값을 대표한다고 보기에 한계가 존재한다. 통행시간 추정을 위한 기존의 방법은 방대한 과거 데이터를 사용하여 계수 추정(parametric models)을 위한 다양한 모형을 사용하였다. 기존 전통적인 방법에서 Outlier를 검지하고, 처리하여 신뢰성 있는 통행시간을 추정하기 위해 본 연구에서는 Non-parametric outlier filtering 방법인 Outskewer 방법을 제안하였다. Outskewer 적용시 기존에 사용하던 Sliding Window 방법 대신 분석할 데이터양이 충분하지 못한 구간에서는 알고리즘이 충분히 분석할 수 있는 데이터양을 확보할 수 있도록 직전의 데이터를 더 추가할 수 있는 Extendable Moving Window 기법을 적용하였다. 제안된 모형을 평가하기 위하여 다차로 도로에서 수집된 DSRC(Dedicated Short Range Communication)를 사용하였다. 모형 평가시 통행시간의 참 값(ground truth)의 부재로 인하여 그래프와 통계의 유의성 검증을 통하여 모형을 평가하였다. 통행시간에 대한 통계적 검증결과 최대 상대 오차(maximum relative error)는 20%보다 적게 나타났고, Correct Classification Rate는 90% 이상으로 나타났다. 결과를 고려할 때 제안된 방법론은 통행시간 추정 시 Outlier를 처리하는데 있어서 기존의 방법론보다 더 신뢰성 있게 나타나 도로이용자와 관리자에게 더 정확한 통행시간을 제공할 수 있을 것으로 기대하였다.

Keywords : Data Mining, Outlier Filtering, Travel Time, DSRC, ITS

This work was supported by a grant (20SBRT-C158062-01) from Development of S-BRT Priority Signal and Safety Management Technology funded by Ministry of Land Infrastructure and Transport (MOLIT) of the Korea government.
*Corresponding Author : Duk Geun Yun(Korea Institute of Civil Engineering and Building Technology)
email: dkyun@kict.re.kr
Received June 23, 2021 Revised July 30, 2021
Accepted October 1, 2021 Published October 31, 2021

1. Introduction

Travel time information plays a crucial role to support road users and transportation agencies with precise data that can be used for traffic analysis and route choice. Road travelers often rely on Advanced Traveler Information System (ATIS) to make decision regarding their current trips namely the least congested road and the fastest path toward their destination. Current advancements in technologies have enabled more sophisticated techniques to collect travel time data. The acquired data is transmitted through ATIS for processing and later link travel time estimation will be provided for the public in real time. Loop detectors, magnetic signature re-identification, closed circuit television (CCTV), and Bluetooth detectors are among technologies utilized to acquire probe travel time records. Likewise, Dedicated Short Range Communication (DSRC) probes are attracting more interest. Initially, DSRC was used in an electronic toll collection system operated by installing a vehicle-mounted device. By the end of 2012, half of registered cars in South Korea were equipped with DSRC units.[1] DSRC is a direct measurement method for travel time estimation, where time stamps of vehicle passing through two consecutive control points are recorded. DSRC makes use of a unique identification device mounted on a vehicle that passes through two successive road-side units (RSUs) and pairs the two records to obtain the travel time record. DSRC measures the time period that a vehicle needs to cross a road section by matching probe data recorded at both RSUs. However, reported results do not always show a representative value of travel time. Methods based on direct measurements can be unreliable due to the presence of anomalies in travel time observations, often called outliers. There is no explicit definition of "outliers", in the field of transportation, but an outlier can be characterized as an observation whose time stamp is numerically separated from other observations, and consequently does not follow the general pattern of data. The existence of outliers can be explained by various reasons, such as a vehicle stopping between RSUs, a vehicle entering between measurement stations, abnormal driving behavior, and measurement errors. Also, unlike license plate recognition systems, DSRC collects probe data bi-directionally, so this situation often occurs multiple times on urban and rural highways with many intersections. Observations that represent very long travel times compared to other observations are considered outliers. Abnormal short travel times can also be attributed to excessive vehicle speeds, which are considered outliers as well since they do not represent the overall link travel time. Therefore, outliers have to be identified and removed from the dataset.

2. Literature review

Numerous authors have addressed the elimination of anomalies in measured travel times that are not related to traffic on the road segment, but are the result of individual vehicle behavior. TRANSGUIDE is one of the earliest developed algorithms and was proposed by the Southwest Research Institute (SWRI) in 1998. It is based on moving average.[2] This algorithm estimates travel time by calculating the average of current observations that are within a user-defined threshold based on the previous average. Previous studies highlighted the limitations of this algorithm and the adverse effects of rapid changes in the travel time during certain aggregation times.[1] Dion and Rakha proposed an algorithm based on the assumption that travel time probes follow a log normal distribution.[3] To overcome the shortcomings of the TRANSGUIDE algorithm, Dion and Rakha's algorithm takes into account several considerations; the expected average travel time and the variability of the upcoming time interval, the number of consecutive intervals with no readings since the last acquired record, the number of consecutive outlier values, and the time variability within the analysis interval. The determining parameters of this algorithm are delicate due to the fact that parameters typically have a combined nonlinear effect. A thorough sensitivity test is required to determine coefficients. Ma et al. proposed an approach based on the median filter as a measure of location.[4] The authors evaluated the statistical median filtering approach and proposed modifications to the AVI travel time estimation. Moghaddam and Hellinga proposed an outlier filtering algorithm for Bluetooth travel time data.[5] The suggested method is based on a proactive adaptive filtering approach that combines the recent available travel time estimates with quasi predictions obtained from a k-nearest neigbor model to establish a suitable validity window. However, this algorithm requires a model calibration and an estimation of a parameter adopted from Dion and Rakha's algorithm through a sensitivity analysis. Jang proposed an outlier filtering algorithm for DSRC probe travel times with a low sample size on signalized rural arterials. The proposed method handles two situations; for low sample size intervals, the algorithm allocates an effective range based on the previous interval value. Otherwise, the algorithm implies a modified z score to designate a modified median filter to determine the new validity range.[1] Similarly, Jang's algorithm requires an estimation of three parameters used to define a validity range. The common feature of the outlier filtering methods presented above is the necessity to perform a parameter estimation, which relies on historical data. In this paper, a statistical outlier filtering method called Outskewer is adopted and modified to perform in real-time with fixed time intervals. The adopted method detects outliers based on skewness distributions with no prior hypothesis on data.

3. Outskewer method

3.1 Overview

Outskewer is an outlier filtering algorithm proposed by Heymann and originally it was developed to observe events in the dynamics of internet topology and search engine queries.[6] This method defines outliers as extreme values, which skew a distribution of data. Thus, outliers are detected based on skewness, the measure of symmetry of a distribution. The skewness coefficient is given by

$$\gamma(X) = \frac{n}{(n-1)(n-2)} \sum_{x \in X} \left(\frac{x - \bar{x}}{\sigma} \right)^3 \tag{1}$$

where, γ denotes skewness coefficient, X denotes multiset of values, n represents the sample size, $\bar{x}$ denotes the mean, and σ is the standard deviation. The removal of extreme values on a one by one basis is expected to lower the skewness of the distribution since skewness is sensitive to extreme values. In case the distribution is symmetrical, the concept of outliers becomes irrelevant according to the suggested definition of outliers. The authors introduced two main notions within the Outskewer method; skewness signature and p-stability. The skewness signature tracks the change of the skewness coefficient, i.e., every time an extreme value is removed, the new value of γ is recorded. The p-stability concept is used to check the distance relativity to the proportion of extreme values that have been removed, where p is the fraction of extreme values removed. A skewness signature is considered as p-stable if the following condition is satisfied;

For any $p \in [0,0.5], |s(p',X)| \le 0.5-p$, for all $p' \in [0,0.5]$ where $s(p',X)$ is the function of the skewness signature. The value of p is limited to 0.5 which corresponds to half of the sample size. If dataset is p-stable, the algorithm classifies values as *outliers*, *potential outliers* and *not outliers*. Otherwise, all values are classified as *unknown* since the notion of outliers becomes irrelevant in the considered sample. Further details about how this algorithm labels classes can be found in literature [6].

3.2 Performance of Outskewer

Outskewer is able to detect outliers in both static and temporal data with no prior knowledge of the considered data. For time series data, multiple outliers can be detected dynamically, through a sliding window. The width of the sliding window is a fixed number of values that has to be given through a user input parameter. Fig. 2*a.* illustrates the application of the Outskewer algorithm on probe travel times collected from a multilane highway. The width of sliding window is set to as recommended in previous research.[6] The outcome shows that the algorithm successfully classifies probe travel times and the *unknown* case was not observed.

The algorithm can be applied for online and offline data. However, the sliding window is based on size of sample, rather than time intervals. For the online scenario, the window remains idle until a required number of data is acquired. For example, when window size is set to, the system does not update the new travel time until 100 data points are collected. When the penetration rate is low, updating travel time takes longer time than recommended for traffic analysis. This situation raises a challenge related to providing real-time travel time for consumers in fixed time intervals. Travel time has to be updated frequently with fixed time intervals in order to reflect the instantaneous condition of a road segment.

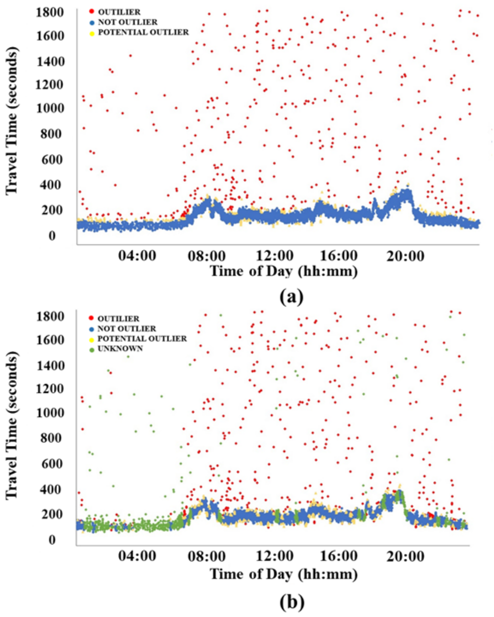

Fig. 1. Performance of Outskewer: (a) with sliding window and (b) with moving window

In order to solve this problem, a moving window is implemented instead of a sliding window. A moving window is based on time interval rather than sample size. The system updates new travel times at fixed time intervals independently from the number of time records. For every time interval, Outskewer is applied to the current collected data. Fig. 1*b.* shows results after the application of Outskewer with moving window using same data from Fig. 1*a.* The time interval was set to 10 minutes. It is clear that the system fails to perform the classification of values at certain time intervals and results are reported as *unknown*. During those time intervals, the skewness signatures of data samples are never p-stable and hence, the Outskewer method is not applicable in such situations. Also, another problem related to sample size was observed when we used 5 minutes intervals. The method fails to perform during time intervals, in which the sample size is below 3. This situation can be explained by the minimum sample size required to calculate skewness. From equation (1), the sample size n has to be at least 3 in order to calculate skewness and more than 3 to track the skewness signature. In the next section, we propose a new approach to solve the presented problem.

4. Outskewer method

Outskewer can perform better with large data samples. However, collecting sufficient travel time records within fixed time intervals is not guaranteed. At this point, two conditions have to be satisfied; the size of travel time records has to be large enough to perform Outskewer, and at the same time, link travel time has to be reported at fixed time intervals. In this regard, we propose the concept of an extendable moving window. Consider $X(t)$ a set of n travel time records at a current time interval such as

$$X(t) \equiv \{x_1, x_2, .., x_n\} \tag{2}$$

where $X(t)$ denotes the set that includes all travel times recorded during interval t, and n denotes the size of the set. Similarly, we consider a set $V(t)$ that includes all previous valid travel time records obtained before the current time interval ordered with respect to its

corresponding time such as

$$V(t) \equiv \{v_1, v_2, ..., v_m\} \tag{3}$$

where $V(t)$ denotes the set that includes all previous valid travel time records obtained until time interval $t-1$ (ordered with respect to collection time where v_m being the last valid time record), and m represents the order of the nearest previous valid travel time record. Now, consider a set $M(t)$ which will include travel time records over which OUTSKWER will be performed. At time interval, t, the algorithm verifies that the number of collected probe travel times meets the minimum required sample that corresponds to $n = 3$. In case that condition is satisfied, Outskewer is performed over the current dataset $M(t)$ such that $M(t) = X(t)$. Otherwise, the nearest previous valid value is included in the current dataset until the minimum sample size is acquired. When there are only two data points in the current interval, the nearest previous valid value is added to a new dataset that includes the current values and the new added value as follow:

$$M(t) \equiv X(t) \cup \{v_m\} \tag{4}$$

where, $M(t)$ is the set that includes travel time records over which OUTSKWER will be performed at time interval t, $X(t)$ is the set that includes all travel times recorded during time interval t, and v_m denotes the nearest previous valid travel time record. Similarly, in case of one single data in the current time interval, the two nearest previous valid values are added to a new dataset, along with the data from the current time interval such as:

$$M(t) \equiv X(t) \cup \{v_{i-1}, v_i\} \tag{5}$$

where $M(t)$ denotes the set that includes travel time records over which OUTSKWER will be performed at time interval t, $X(t)$ represents te set that includes all travel times recorded during time interval t, and v_i nearest previous valid travel time records with order with order i

Once the minimum sample size condition fulfilled, Outskewer is performed over the new dataset $M(t)$. At this level, if the result is reported as *unknown*, the new proposed approach is executed once again. The nearest previous valid travel time record is added to $M(t)$. Likewise, Outskewer is applied over the new dataset. This procedure is executed repeatedly until the outcome from application of Outskewer over $M(t)$ is different from *unknown*. Once the classification results are obtained, the average travel time is estimated. All data points that were added from $V(t)$ are excluded from calculations since $V(t)$ is used only to facilitate outlier detection at time intervals where result is reported as *unknown*. Link travel time is obtained as follow:

$$T_{AB}(t) = \frac{\sum_{i=1}^{|S_{AB}(t)|} (t_{B_i} - t_{A_i})}{|S_{AB}(t)|} \tag{6}$$

where $T_{AB}(t)$ denotes the average travel time from road side unit reader A to reader B that is estimated at time t (seconds), $S_{AB}(t)$ represents Set of valid recorded travel times from reader A to reader B at time t from $M(t)$ and excluding $V(t)$ values, t_{B_i} is the detection time of vehicle i at reader B (seconds), t_{A_i} is the detection time of vehicle i at reader A (seconds), and t is the time at which travel time estimation takes place (seconds).

Initialization is another essential consideration when the system is used for the first time. In case OUSTKEWER reports *unknown* results or a low sample size in the first time interval, the algorithm may fail to continue. In this regard, we propose to provide an initialization to the algorithm by manually providing valid travel time records. This procedure was adopted by various previous outlier filtering algorithms such as TRANSGUIDE and Dion and Rakha' algorithms.

5. Assesment of the proposed algotrithm

5.1 Field data

With the purpose of evaluating the proposed travel time estimation model, travel time probes were collected using DSRC detectors in a multilane highway in Seoul, South Korea (National Highway No. 3). The DSRC detectors were deployed along 3 km of the suburban roadway. The considered section includes one intersection and one interchange. Approximately 30 days of probe travel time records were obtained during the year 2015. Among those days, 4 days of data were randomly selected for evaluation. The chosen data samples include two 24-hour records from weekdays and two 24-hour records from weekends. The proposed algorithm has to be tested in such a way that the data is being collected in real time. To achieve this goal, an off-line software program was developed to establish a real time analysis framework.

5.2 Results

The proposed model was used for cleaning DSRC probe travel times. The sampling interval was selected as 5 minutes. Fig. 2*a.* and Fig. 2*b.* show the performance of the outlier filtering procedure during two weekdays. On the other hand, performance during weekend is shown in Fig. 2*c.* and Fig. 5*d.* The algorithm successfully classifies data as *outlier, not outlier* and *potential outlier*. By observing the behavior of the proposed algorithm, it seems clearly that the algorithm is following the general pattern of data when detecting outliers. Thus, the proposed methodology can be deemed as efficient from a qualitative point of view. However, a quantitative evaluation of the performance is required to validate the findings.

5.3 Model evaluation

The evaluation of outlier filtering algorithms is a challenging task. In some previous studies, the evaluation was performed using a statistical comparison between estimation results and ground truth observations obtained from global positioning system-equipped floating probe vehicles. However, this technique can be unreliable for real-time evaluation, since the sample of collected travel time observations are limited to a few values and in many cases to only one observation per link and time interval. The

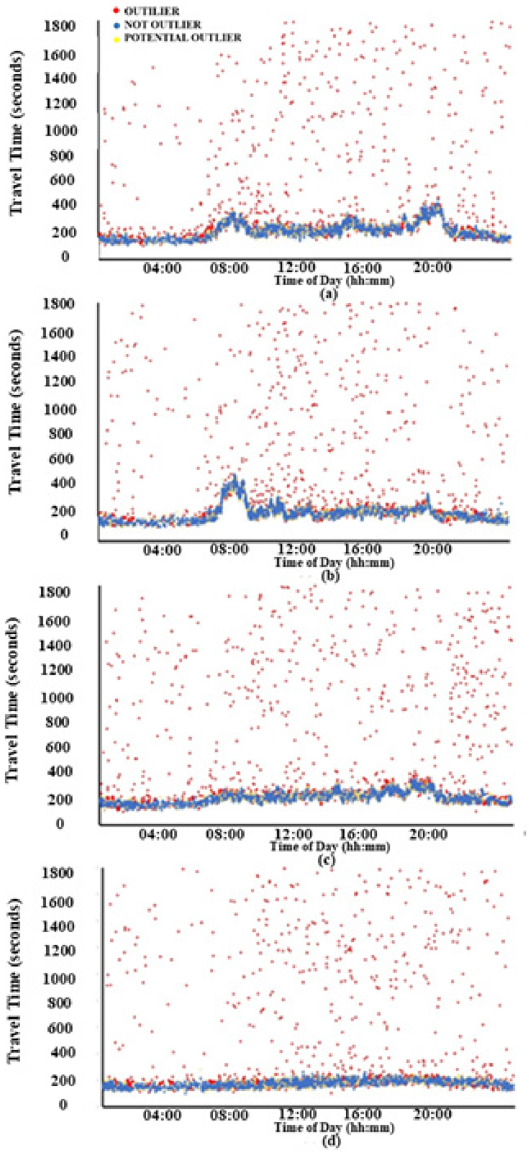

Fig. 2. Performance of Performance of the proposed algorithm: (a) a weekday January, 9, (b) a weekday January, 10, (c) a weekend January, 12, and (d) a weekend January, 13

evaluation of travel time estimation methods based on the floating car is built upon assumption that floating car data can represent an approximation of average travel times in the selected link. Data simulation is another appealing approach to evaluate travel time estimation methods. Through this method, true travel time can be obtained and later compared to results acquired by the travel time estimation method under evaluation. Nevertheless, the reliability of this method is undermined, since it requires introducing the sources of errors, which take place in field data, to the simulated data.

In this study, a statistical analysis method is used to evaluate the outcome results of the proposed outlier filtering method. Richardson et al. proposed a confidence interval method, based on Maximum Relative Error (MRE) for travel time data quality assessment.[7] The confidence interval estimate of population mean is obtained by [7]

$$\overline{X} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \overline{X} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \qquad (6)$$

The maximum relative error $e_{\max}$ is given by [7]

$$e_{\max} = \frac{2t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}}{\overline{X} - t_{t_{n-1,\alpha/2}}\frac{s}{\sqrt{n}}} \qquad (7)$$

When $e_{\max}$ is known, the minimum required sample of observation is given by [7]

$$n_{\min} = \left(\frac{s t_{n-1,\alpha/2}(2 + e_{\max})}{e_{\max}}\right)^2 \qquad (8)$$

Where, μ denotes population mean, $\overline{X}$ is sample mean, s represents sample standard deviation, $n_{\min}$ stands for sample size, $t_{n-1,\alpha/2}$ denotes student t-statistic, and $e_{\max}$ is the maximum relative error.

Richardson et al. suggested 20% as the

maximum relative error $e_{\max}$ for this method. Therefore, in order for an estimation value to be classified as accurate, it has to be within 20% of the population mean. Thus, the confidence interval, assuming the minimum number of samples obtained, can expressed as follow [7]

$$\overline{X}-\frac{e_{\max}\overline{X}}{2+e_{\max}}\leq T\leq \overline{X}+\frac{e_{\max}\overline{X}}{2+e_{\max}} \qquad (9)$$

where T denotes the average travel time from reader A to reader B that is estimated at time t (seconds), $\overline{X}$ is the sample mean, and $e_{\max}$ represents the maximum rlative error.

Once the final confidence interval is established, travel time estimation results are classified as accurate or inaccurate based on equation (9) for each interval of time. After obtaining accuracy classifications, the percentage of correct classifications (PCC) is determined. Since the results of classifications are either accurate or inaccurate, the system can be thought of as a sum of a series of independent Bernoulli random variables, resulting in a binominal distribution.[7] A travel time estimation system that has a high PCC and low MRE is presumed to estimate travel time accurately with a high probability. Let p be the probability of correct classification and c be the critical quality threshold desired. The probability p is not recognized since the actual performance of the travel estimation system is not known. Therefore, a hypothesis test is constructed to test that the parameter of the binominal distribution is less than or equal to the PCC value.

The hypothesis test follows:

$H_0 : p \leq c$

$H_a : p > c$

The system is accurately estimating travel time at the desired PCC if the null hypothesis can be rejected at the desired alpha level, otherwise, the system may be operating below standards. The aforementioned method was used to evaluate the performance of the traveler information systems (TISs) using Bluetooth identification (BTR) data as a benchmark. In this paper, this method is adopted for the evaluation of the proposed method. DSCR data, similar to BTR, are based on a re-identification process. Therefore, in this paper, DSRC matched records are considered as the benchmark and the proposed method is considered as the TIS.

For a quantitative evaluation, the confidence interval method, previously described, is used for the assessment. For convenience and based on observations from Fig.2., data points labelled as *potential outlier* are considered as valid data when estimating travel time in each time interval. The maximum relative error $e_{\max}$ is set to 20%; the desired PCC was 90% as recommended in a previous study.[7] Sample mean, standard deviation, sample size, and estimated travel time are recorded. For data obtained on January, 9 which is a weekday, 285 intervals of 5-min blocks are obtained. Only 76 intervals are used for evaluation, and the remaining are excluded from evaluation since they do not meet the minimum required simple size $n_{\min}$. The number of accurate classification is 76 with no recorded inaccurate estimations.

The hypothesis test of true parameter p of the binominal distribution is constructed as

$H_0 : p \leq 0.9$

$H_a : p \geq 0.9$

$p(K \geq 76 | n = 76, p = 0.9) = 0.000333$

The null hypothesis is rejected since the probability of seeing more than 76 of 76 correct classification in a binominal distribution with $p = 0.9$ less than 5% is 0,033%. Therefore, it can be concluded that the travel time estimation performed by the proposed model is accurate. Similarly, the null hypothesis was tested for the other weekday data obtained from January, 10 as well as data from the weekend days. Finally, the hypothesis test was performed for 4 days in 2015 combined to check the overall performance of the algorithm. The overall probability of a correct classifications is 0.000012%. Therefore, the proposed algorithm can be considered as accurate. Table 1. shows a summary of obtained results.

Table 1. Summary of quantitative evaluation results

Data Type	Weekdays		Weekend	
Day	January, 9	January, 10	January, 12	January, 13
Total obtained intervals	285	281	286	282
Accurate classifications	76	65	59	44
Inaccurate classifications	0	1	2	1
Evaluated intervals	76	66	61	45
Probability (per day)	0.000333	0.0079583	0.0491182	0.05
Probability (per data type)	0.00000534		0.00488574	
Overall probability	0.00000012			

6. Conclusion

In this paper, a new outlier filtering algorithm for travel time estimation was proposed. The literature review outlined that previous outlier treatment methods relied on prior assumptions on the distribution of datasets. Also, previous work showed that a sensitive analysis is required for parameter estimation in order to calibrate the proposed models. In this regard, the Outskewer method was adopted and tested. Outskewer is a statistical outlier filtering method based on skewness distribution of dataset and requires no prior knowledge of the data. The adopted method is able to detect outliers in real time, using a sliding window with a fixed width size. However, the online performance of Outskewer is not practical in real world situations since travel time has to be reported at fixed time intervals. A sliding window moves by discarding old values when new values arrive. This approach may be unacceptable when new values arrive at low rate and travel time information take longer time periods to be updated. In this regard, an extendable moving window was introduced to solve the problem. When Outskewer fails to detect outliers at certain time intervals, the nearest previous valid time records are included progressively in the current dataset, until the algorithm performs classification successfully. The values are classified as outliers, potential outliers or not outliers. The performance of the proposed method was evaluated using a confidence interval method with a hypothesis test. The algorithm was tested in such a way that the data was being collected in real time. To achieve this goal, an off-line software program was developed to establish a real time analysis framework. Data from weekdays and weekends was used for the quality assessment. Results showed a satisfactory performance. The next step of study would be to investigate the performance of the proposed outlier removal method using probe data from other road types such as freeways and suburban arterials. Also, data labeled as potential outlier was considered as valid data during the evaluation and when estimating travel time. A further investigation is required to justify this choice. A major limitation of our proposed approach is the lack of ground truth data. The proposed method has to be further tested with larger data under different conditions with more reliable validation technique to quantify the errors by comparing the outcome results to ground truth data records. Future research will be carried to validate the findings using simulated data.

References

[1] J. Jang, "Data-Cleaning Technique for Reliable Real-Life Travel Time Estimation: Use of Dedicated Short-Range Communications Probes on Rural Highways", *Transportation Research Record*, Vol. 2593, No.1, pp. 85-93, Jan.2016.
DOI:https://doi.org/10.3141/2593-09

[2] Southwest Research Institute, Automatic vehicle identification model deployment initiative—system design document. Report prepared for TransGuide Texas Department of Transportation, 1998, Available From:
https://rosap.ntl.bts.gov/view/dot/48800/dot_48800_DS1.pdf (accessed Jun. 09, 2021)

[3] F. Dion, H. Rakha, 2006. "Estimating dynamic roadway travel times using automatic vehicle identification data for low sampling rates", *Transportation Research Part B: Methodological,* Vol.40, No.9, pp.745-766, Nov.2006.
DOI: https://doi.org/10.1016/j.trb.2005.10.002

[4] X. Ma, H. Koutsopoulos, "Estimation of the automatic vehicle identification based spatial travel time information collected in Stockholm", *IET intelligent transport systems,* Vol.4, No.4, pp.298-306, Jan 2011.
DOI: https://doi.org/10.1049/iet-its.2009.0149

[5] S.S. Moghaddam, B. Hellinga, "Algorithm for detecting outliers in Bluetooth data in real time", *Transportation Research Record,* Vol.2442, No.1, pp.129-139, Jan.2014
DOI: https://doi.org/10.3141/2442-14

[6] S. Heymann, *Exploratory link stream analysis for event detection,* Ph.D dissertation, Pierre and Marie Curie University, Paris, France, 2013.

[7] J.K. Richardson, B.L. Smith, 2012. "Development of hypothesis test for travel time data quality". *Transportation Research Record,* Vol.2308, No.1, pp.103-109, Jan.2012
DOI:https://doi.org/10.3141/2308-11

Moataz Bellah Ben Khedher [Associate member]

- Feb. 2015 : KAIST, Bachelor, Dep. of Electrical Engineering
- Feb, 2018 : University of Science and Technology, MS
- Sep, 2018 ~ current : University of Science and Technology, PhD

〈Research Interests〉
Smart Mobility, Machine Learning, Road Safety, Autonomous Vehicle

Duk Geun Yun [Regular member]

- Feb. 2009 : Myongji Univ. Ph. D.
- March. 2001 ~ current : Research Fellow of KICT
- March. 2011 ~ current : Professor of University of Science and Technology

〈Research Interests〉
Road Safety, Road Maintenance, Road Design, Autonomous Vehicle

Jung Gon Sung [Regular member]

- Feb. 1997 : Univ. of Wisconsin-Madison. Ph. D.
- Dec. 1997 ~ current : Senior Research Fellow at KICT

〈Research Interests〉
Road Safety, Road Maintenance, Road Design, Autonomous Vehicle