# Urban travel time data cleaning and analysis for Automatic Number Plate Recognition

Li, Jie; Van Zuylen, Henk; Deng, Yuansheng; Zhou, Yun

**DOI**
10.1016/j.trpro.2020.03.151

**Publication date**
2020

**Document Version**
Final published version

**Published in**
Transportation Research Procedia

Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 47 (2020) 712–719

Transportation Research Procedia

www.elsevier.com/locate/procedia

22nd EURO Working Group on Transportation Meeting, EWGT 2019, 18-20 September 2019, Barcelona, Spain

Barcelona, Spain

# Urban travel time data cleaning and analysis for Automatic Number Plate Recognition

Jie Li [1,4*], Henk van Zuylen [1,2,4,5], Yuansheng Deng[3], Yun Zhou [1,4]

[1]*Key Laboratory of Building Safety and Energy Efficiency of the Ministry of Education, Hunan University, Changsha 410082, P.R. China*
[2]*Transport and Planning Department, Delft University of Technology, P. O. Box 5048, Delft 2600 GA, the Netherlands*
[3]*Traffic Police Detachment, Changsha Public Security Bureau, Fenglin Road 2, Changsha 410006, P.R. China*
[4]*Civil Engineering College, Hunan University, Lushan South Road 1, Changsha 410082, P.R. China*
[5]*School of Transportation and Logistics, Southwest Jiaotong University, Western Hi-tech Zone, Chengdu 611756, P.R.China*

**Abstract**

Data recorded by Automated Number Plate Recognition (ANPR) cameras can be used to determine several important traffic characteristics, such as real time travel time, travel time statistics, travel time reliability and OD matrices. In this paper ANPR data collected in Chinese city Changsha have been validated. Travel time extracted from ANPR data includes some outliers which are often caused by drivers who have an intermediate stop between two observation points or deviate from the straight route. Exceptional travel time reduces the validity of the estimation of the travel time and reliability. Firstly, the Rapid-Moving Window method is introduced to identify outliers. Afterwards, another method based on wavelet analysis is put forward to identify and remove the outliers in the travel time series. The wavelet analysis method is compared with the Rapid-Moving Window method and shows to be more accurate in outlier identification. The method for eliminating outliers in travel times can be implemented in real time to enhance the data quality for traffic network monitoring and management. After the removal of the outliers, the resulting travel times are used for the analysis of the relation between average travel time and standard deviation/skewness.



*Keywords:* Automated Number Plate Recognition data; Data clearning; Travel time reliability; Wavelet analysis

*Corresponding author. Tel.: +86-731-88821491; fax: +86-731-88823579
*E-mail address:* ljlj369@msn.com


10.1016/j.trpro.2020.03.151

## 1. Introduction

Urban traffic is very complicated due to the variety of interacting traffic participants, the road network, and the traffic control. Characterizing traffic states is more difficult for urban road networks than for freeways. One of the reasons is that empirical traffic data of urban road networks are often incomplete, out-of-date or unreliable, as demonstrated by Li et al. (2011). For instance, Li et al. (2014) found loop detectors can count traffic flow, but the reliability is impacted by internal and external conditions including hardware failures. The upcoming of probe vehicles installed with Global Positioning System (GPS) gives more possibilities to determine travel times for a sample of vehicles (Woodard et al. 2017). Bluetooth scanners have been used to measure travel times of freeways in many countries(Aliari and Haghani 2012; Diaz et al. 2016; Martchouk et al. 2011). These scanners receive the signals from Bluetooth devices in a range up to about 100 m radius, which only has limited influence on the travel time estimation on freeways, but can results in obvious deviation for travel time estimation on urban streets. A more serious problem derives from the fact that Bluetooth devices can be carried by vehicle passengers, pedestrians or cyclists in urban areas. The last two carriers have nothing to do with the traffic situation evaluation, but they are difficult to be distinguished. Some researchers proposed to fuse different types of data that come from different sources (inductive loop detectors and toll tickets) to predict highway travel times (Pirc et al. 2016; Soriguera and Robuste 2011). The validity of these algorithms has been verified on freeways, but traffic environments in urban areas are more complex and challenging. Moghaddam and Hellinga (2013) found it isn't feasible to use Bluetooth field data to examine the measurement errors because the errors are inherent within the observations and cannot be separated.

Automated Number Plate Recognition (ANPR) cameras register the number plates and the passing moments of vehicles driving on a certain lane. Chow et al. (2014) determined the total time spent (travel time weighted with the flow rate) as a one-dimensional characteristic of the traffic state. They used this characteristic to identify the emerging of congestion. Besides travel time, ANPR cameras data are also used to deduce some other relevant traffic characteristics, such as traffic flow, vehicle headways (Gunay 2012), Origin Destination tables (Sbaï et al. 2017), etc.

On the other hand, ANPR data are not error free and some traffic characteristics derived from ANPR data are less reliable than what some researchers expected. Actually, ANPR camera registration is possibly sensitive to lighting conditions and the way number plates are fixed (Rhead et al. 2012). Furthermore, travel times estimated from ANPR data may include outliers because drivers may stop between two camera sites, e.g. for shopping, working, delivering goods etc., or deviating to a side road and coming back to the route after some time. Models of the traffic process facilitate the identification and removal of outliers (Zhu et al. 2016). The traffic data extracted from ANPR records have to be cleaned in order to remove invalid, unnecessary and irrelevant travel times before they can be used for travel time information and statistics. Valid and reliable ANPR data can be used to monitor the traffic situation in terms of network capacity and travel time reliability. The main research questions in this paper are:

- How is the quality of the data recorded by ANPR cameras?
- How to eliminate the outliers in travel times deduced from ANPR data?
- How do the outliers influence the analysis of travel time reliability?

The remaining part of this paper is structured as follows: the next section introduces ANPR data collection and a general analysis is made for the collected data. Afterwards, the outliers in travel time series are analyzed and are eliminated with two methods: the Rapid-Moving Window method and the Wavelet Analysis Method. The performance of these two methods are compared. After removing the outliers, the remaining travel times are used for the analysis of the relation between average travel time and standard deviation/skewness. Finally, some conclusions are drawn and some potential future research is described.

## 2. Data collection and description

In Changsha, the capital of Hunan province in China, the traffic monitoring system includes 1256 Automated Number Plate Recognition (ANPR) cameras installed on 112 intersections, one camera per lane. The data processed in this study were collected from April 20 to 22, 2015. The ANPR data comprise five pieces of information of every passing vehicle: the name of the intersection, the number plate, the passing approach, the passing lane and the passing moment with one second accuracy. After checking the ANPR data, multiple recording and fail-recognition are revealed.

### *2.1. Multiple recording*

ANPR data can be used to estimate traffic flow rates. However, ANPR cameras may record the number plate of one vehicle more than once in a short time, especially during peak hours. The double counting can be identified by comparing the recording time of the same number plate at the same location.

Multi-recording nearly never occurred on some intersections and rather frequently (10% to 30%) on the other intersections. The most probable reason for double counting is misalignment of some cameras which register vehicles on more than one lane. The fact that some vehicles do not always drive over one single lane, increases the chance of double counting. This happens more often in peak hours due to lane changing. In this study, the multiple ANPR records have been identified and removed.

### *2.2. Fail-recognition*

ANPR cameras can't always recognize vehicle number plates successfully (Rhead et al. 2012). The unrecognized number plates are recorded as '0000000' in the log, which makes it possible to use them still for traffic volume counting. The percentage of recognition failure varies from 5% to 88 % in different periods and on different intersections. At night, before 6 AM and after 6 PM, the percentage of failed recognitions is higher than during daytime. That might be due to the light conditions which have an influence on the performance of ANPR cameras (Lubkowski and Laskowski 2017). The missing of number plate information makes it impossible to estimate the travel time on a link for these vehicles. To estimate travel time, multiple and fail-recognition recordings should be removed from the ANPR data. Even if the travel times extracted from ANPR are not fully complete, the urban traffic state still can be characterized if the remaining observed travel times are adequate and valid.

## 3. Travel time outliers identification

In urban road networks, outliers are rather common in travel times, which can be attributed to the difference in drivers' activities. One group of drivers continually drive along their route without any other activity and the other group of drivers have some intermediate activities. For a traffic state study, travel time statistics analysis should be based on the first kind of drivers. Therefore, the travel times of the second group of drivers can be considered as outliers and should be excluded in the determination of travel time distributions. Kazagli and Koutsopoulos (2013) developed two lognormal distributions to describe the travel times of these two groups of drivers. However, if the analysis period comprises of peak hours and off-peak hours, the distributions of travel times of straight driving vehicles and interrupted trips are overlapping due to signal control and congestion (Zheng and Van Zuylen 2011). Fig. 1(a) describes the raw travel times on a 0.560 km link, part of an arterial in Changsha city down town. The speed limit of this arterial road is 60km/h. On April 20, 2015, the minimum and maximum travel times are 27s and 84893s respectively. To be seeable, the raw data are truncated at 1800s in Fig. 1(a). It is not appropriate to identify outliers by setting a fixed threshold or a specific distribution for the whole travel times set. In this study, the Rapid-moving windows and Wavelet analysis methods are applied to identify outliers in travel times.

### *3.1. Rapid-Moving Window method*

To identifying outliers of travel times, the Moving Window is a commonly used method (Dion and Rakha (2006). The Moving Window gives maximum and minimum values for certain time intervals during the day, travel times outside the window are considered to be outliers. Robinson & Polak (2006) developed a similar method to clean travel times from ANPR data by identifying overtaking vehicles. In this study, every three contiguous vehicles passing the same link comprise a Window and two adjacent Windows overlap with two vehicles. The outliers in travel times can be distinguished for every Window. One simple method to distinguish the outliers is to compare the travel time of one vehicle ($t_i$) with those of the preceding and the following vehicle ($t_{i-1}$, $t_{i+1}$). If the travel time is $C$ seconds longer than the average travel time of the other two vehicles in the same Window, it is identified as an outlier, as described in Function 1:

$$\begin{cases} t_i > (t_{i-1} + t_{i+1})/2 + C, t_i \in outliers \\ t_i \le (t_{i-1} + t_{i+1})/2 + C, t_i \in regular \end{cases} \quad (1)$$

In Formula 1, $C$ is an important parameter to distinguish outliers. Considering the influence of traffic control, $C$ is determined as one cycle time of the corresponding signalized intersection in every link. To distinguish this method from the Moving Window Methods as discussed by Dion and Rakha (2006), we call this method the Rapid-Moving Window method because of the small Window size and the short moving step.

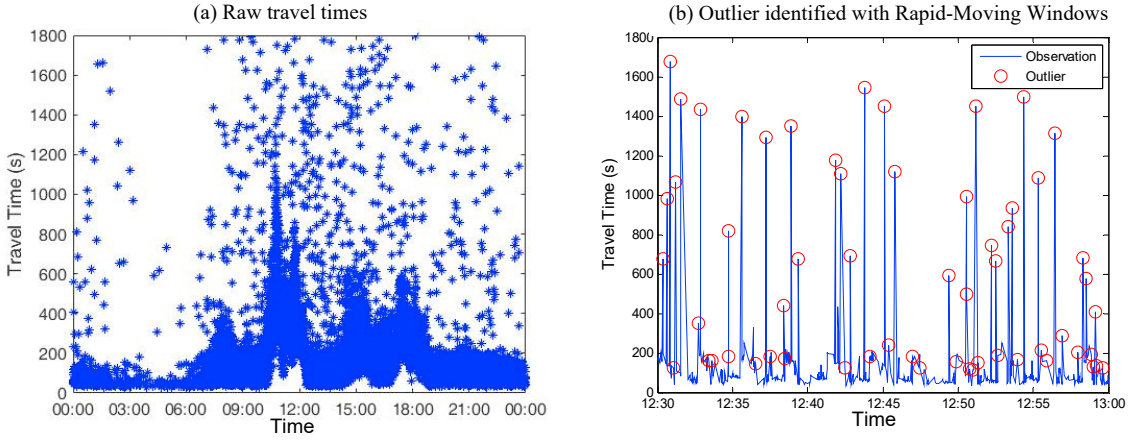


Fig. 1. (a) Raw travel times; (b) Outliers identified with Rapid-Moving Window method for Link 1225-1227 on Apr. 20, 2015.

By using the Rapid-Moving Window method, lots of travel times are identified as outliers, as shown in Fig. 1(b). In the period of 12:30~13:00, 581 raw travel times are estimated from ANPR data on Link (1225-1227). Totally, 59 vehicles' travel times (about 10%) are identified as outliers based on the rule described in function (1). However, this method seems a little too strict in data cleaning. Often, the travel times of the first vehicle or the last vehicle in a queue during red time are identified as outliers. The Rapid-Moving Window method can also be used by involving 5 or 7 vehicles. A test shows that the wider Windows still result in over-cleaning.

### *3.2. Wavelet Analysis Method*

Another more systematic way to eliminate outliers is to analyse the time series of travel time with the wavelet analysis method (Jiang and Adeli 2005). The wavelet analysis follows the general trend of the travel time as a function of the time but ignores the outliers. The wavelet analysis method developed in this study comprises 4 steps:

*Step 1: Wavelet Decomposition*

Travel time series are decomposed at level 4 using Daubechies wavelets. The output decomposition structure contains a wavelet decomposition vector and a bookkeeping vector. Daubechies wavelets are compactly supported wavelets with extremal phase and highest number of vanishing moments for a given support width. Associated scaling filters are minimum-phase filters (Wikipedia 2017). The wavelet decomposition vector $C$ is presented in Fig. 2(a).

*Step 2: Compute Approximation Coefficients and Detail Coefficients*

The approximation coefficients of travel time series are extracted at level 4 from the wavelet decomposition structure, as shown in Fig. 2(b). Afterwards, detail coefficients at levels 1, 2, 3 and 4 are calculated from the wavelet decomposition structure, as shown in Fig. 2(c~f).

*Step 3: Reconstruct Travel Time Series*

Travel time series is reconstructed at level 4 by summing up the approximation (a4) and detail coefficients (d4), which is shown as the red dashed line in Fig. 3(a). The reconstructed travel time series can't directly replace the raw series because it includes some very short travel times, as short as 4 seconds. A more important issue is that the mean value of the reconstructed travel time series is almost equal to that of the original series, which means that wavelet reconstruction only suppresses the outliers rather than removes them.

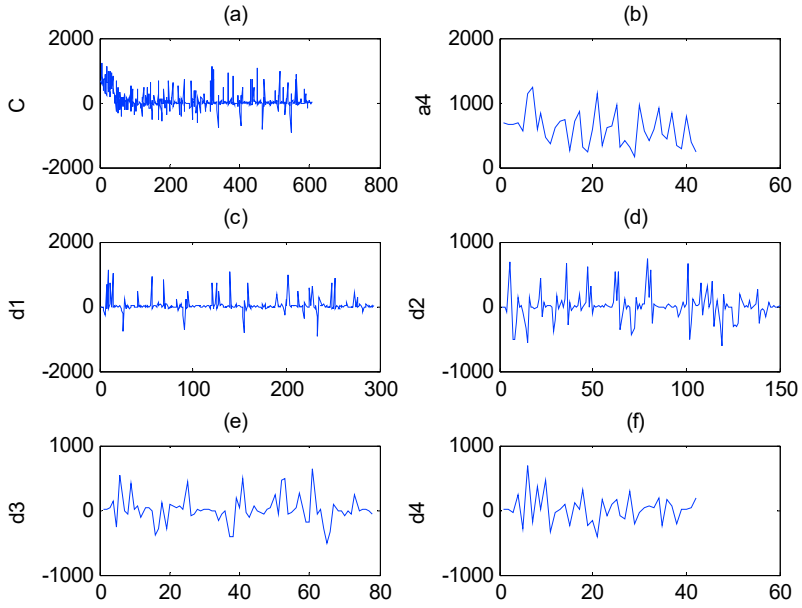


Fig. 2. Coefficients for approximation at level 4 and for details at levels 4, 3, 2 and 1.

*Step 4: Outlier Identification and Elimination*

The reconstructed travel time series is compared with the original one, as shown in Fig. 3(b). The difference between the original travel times and the reconstructed travel times are estimated as $\Delta t_i = t_{ori.,i} - t_{rec.,i}$, which is used to distinguish outliers. Since the outliers are always larger than the regular travel times, a boundary condition to identify outliers is described in Formula 2.

$$\begin{cases} \Delta t_i > 2 \times std(\Delta t_i), t_i \in outliers \\ \Delta t_i \le 2 \times std(\Delta t_i), t_i \in regular \end{cases} \tag{2}$$

According to Formula 2, if the difference between the reconstructed travel time and the original one is larger than twice the standard deviation, the raw travel time is identified as an outlier and should be removed from the travel time series. The outliers in travel times on Link 1225-1227 are identified and shown in Fig. 4(a). The travel times of the first vehicle or the last vehicle in a queue during red time are identified as outliers by Rapid-Moving Windows method, but still can be preserved by Wavelet method. In Changsha, the maximum cycle time is 220s. Thus, it is reasonable to have travel times between 30s to 200s in a short time interval. By comparing with Rapid-Moving Window method, the Wavelet Analysis Method distinguishes outliers more accurately.

A statistical analysis is made for Link 1225-1227 on Apr 20, 2015. On the south approach of Intersection 1227, 33615 vehicles have been registered, which include 374 multiple records and 6461 records with number plate ‘0000000’. 21338 vehicles were firstly registered on Intersection 1225 and then moved to Intersection 1127. There are 3736 records (about 17.51%) identified as outliers. The distribution of outliers is shown in Fig. 4(b). Outliers occur most frequently in the morning and decrease after lunch time. This pattern indicates that in the morning drivers have more activities on their trip than on the trips later in the day.

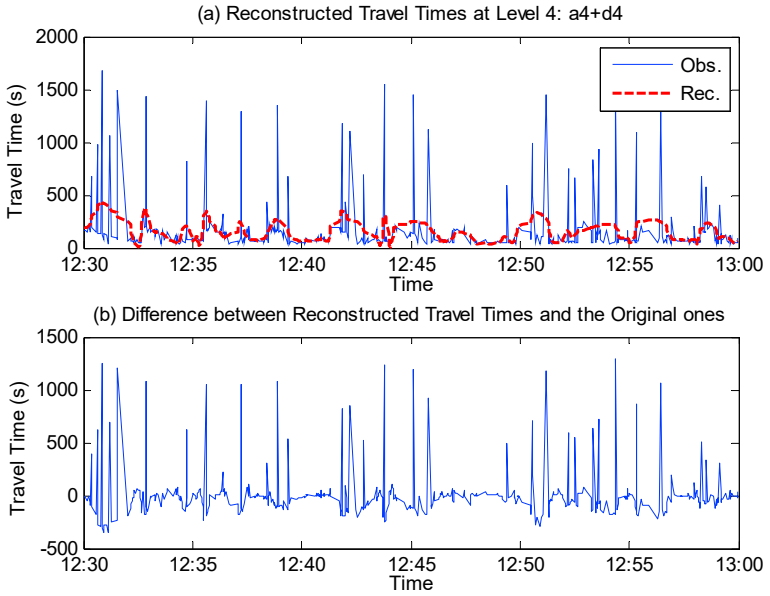


Fig. 3. (a) Reconstructed travel time series; (b) Difference between reconstructed travel times and the original ones.

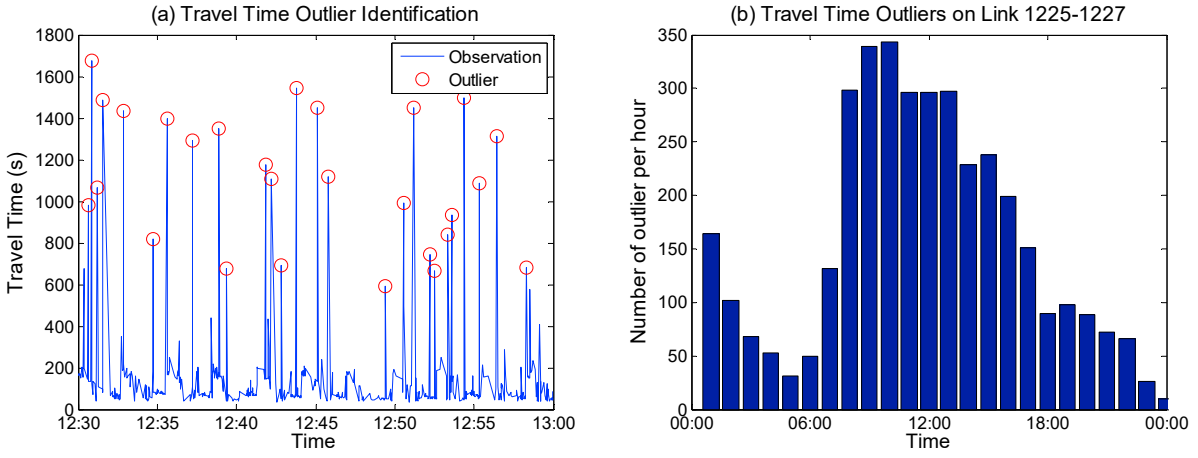


Fig. 4. (a) Outliers identified with wavelet analysis method; (b)Travel time outlier distribution of Link 1225-1227 on Apr. 20 2015.

It is obvious that after removing the outliers, the mean travel time, the standard deviation and the skewness all decrease significantly. The travel time data cleaned in this way can be used in the further study of traffic state, such as the analysis of travel time reliability.

## 4. Relation between mean travel time and standard deviation /skewness

Standard deviation of travel time is often used as a measure for travel time reliability. The skewness is also an important factor which has influence on route planning and departure time choice of travelers (Bogers 2009). In this section, we take Link 1225-1227 as an example to reveal the relation between mean travel time and standard deviation/skewness. The raw data include lots of noise as shown in Fig. 1. After removing the outliers from the raw travel times with Wavelet Method, the maximum travel time on Link 1225-1227 is identified as 900 seconds during 07:00~23:00 on April 20, 2015. To be comparable, travel times longer than 900 seconds are removed from the raw

data. Afterwards, the remaining raw data (i.e. truncated raw data) are compared with the filtered data, as shown in Fig. 5.

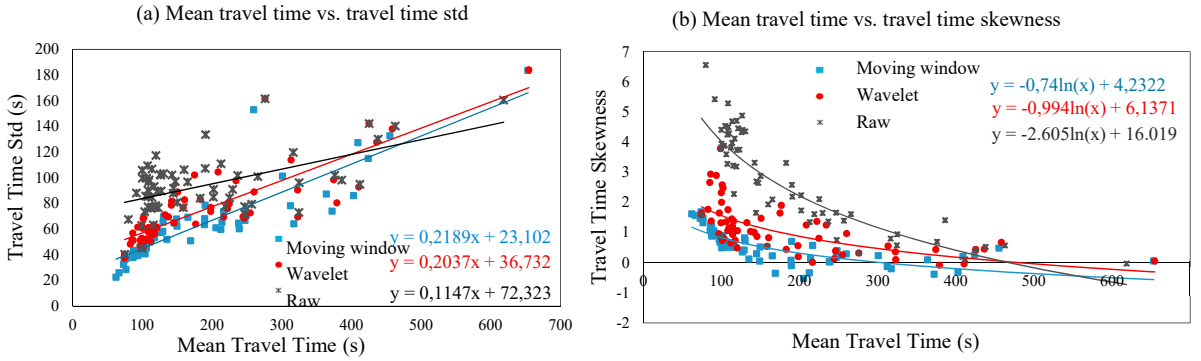


Fig. 5. Relation between mean travel time and standard deviation/skewness of Link 1225-1227 for different filtering results.

After removing the outliers, the mean travel time, the standard deviation and the skewness all decrease significantly, as shown in Fig. 5. The relation between mean travel time and standard deviation/skewness depend on the filtering results: the raw data have a larger mean travel time, standard deviation and skewness. The standard deviation of travel time increases linearly with the increase of the mean travel time. The truncated raw data still present much larger standard deviations and skewness values than the filtered data processed by the Wavelet Method and Rapid-Moving Window Method. Concerning the relation between mean travel time and skewness, the difference among three data sets is rather obvious, which is consistent with the above-mentioned finding: Rapid-Moving Window Method overly cleans the outliers. The quality of the outliers cleaning is important for the analysis of travel time reliability.

## 5. Discussion and conclusions

Automated Number Plate Recognition (ANPR) data contain relevant information about the traffic situation and make it possible to study both microscopic and macroscopic properties of traffic. Recently, much research on the application of ANPR data has been carried out and been published. However, limited attention has been given to the ANPR data quality. This study firstly checks ANPR records, and then applies two kinds of data cleaning methods (Rapid-Moving Window Method and Wavelet Analysis Method) to identify and remove the outliers in travel time series. The remaining valid travel times are analyzed to show the relation between mean travel time and standard deviation/skewness.

ANPR cameras mainly have two kinds of registration errors: multi-recording and fail-recognition. Multi-recording nearly never occurred on some intersections and rather frequently (10% to 30%) on other intersections and happened more often in peak hours. The percentage of failed recognition varies from 5% to 88 % on different intersections and increases remarkably at night. Light condition is a significant factor for the number plate recognition. ANPR records must be checked, so that invalid registrations can be identified and removed.

A method based on wavelet analysis is proposed to identify the outliers in travel times. Lots of outliers exist in the travel times due to stops or detours between two observations. The Rapid-Moving Window method, which compares the travel times of nearby vehicles, removes too many valid travel times, especially travel times of vehicles that are delayed by the beginning of a red phase. Wavelet reconstruction is often used in outlier identification in different application domains, but rarely applied for the elimination of travel time outliers. A comparison between reconstructed travel times and original ones can effectively identify the outliers. In comparison with the Rapid-Moving Window method, the Wavelet Analysis Method is more accurate for outlier identification.

Further research on outlier identification for travel time is expected to be executed with different wavelet decomposition functions at different levels. ANPR data can be fused with loop detector data and taxi GPS data to develop a comprehensive traffic prediction model for traffic management of a road network. After the identification

of outliers, trajectories of vehicles as identified by the ANPR cameras can be analyzed to get the origin destination matrix of the monitored network.

## Acknowledgement

This research has been made possible by grants for Changsha Science and Technology Commission under project kq1801010, Department of Communications of Guangdong Province under number 2016-03-013, the National Science Foundation of China (NSFC) under project 51878264, National Key Technology Support Program under number 2015BAJ03B01.

## References

Aliari, Y., Haghani, A., 2012. Bluetooth Sensor Data and Ground Truth Testing of Reported Travel Times. Transportation Research Record. 2308, 167-172.

Bogers, E.A.I., 2009. Traffic Information and Learning in Day-to-Day Route Choice. Delft University of Technology, Delft, the Netherlands.

Chow, A.H.F., Santacreu, A., Tsapakis, I., Tanasaranond, G., Cheng, T., 2014. Empirical assessment of urban traffic congestion. Journal of Advanced Transportation. 48.8, 1000-1016.

Diaz, J.J.V., Gonzalez, A.B.R., Wilby, M.R., 2016. Bluetooth Traffic Monitoring Systems for Travel Time Estimation on Freeways. Ieee Transactions on Intelligent Transportation Systems. 17.1, 123-132.

Dion, F., Rakha, H., 2006. Estimating dynamic roadway travel times using automatic vehicle identification data for low sampling rates. Transportation Research Part B -Methodological. 40.9, 745-766.

Gunay, B., 2012. Using automatic number plate recognition technology to observe drivers' headway preferences. Journal of Advanced Transportation. 46.4, 305-317.

Jiang, X.M., Adeli, H., 2005. Dynamic wavelet neural network model for traffic flow forecasting. Journal of Transportation Engineering-Asce. 131.10, 771-779.

Kazagli, E., Koutsopoulos, H.N., 2013. Estimation of Arterial Travel Time from Automatic Number Plate Recognition Data. Transportation Research Record: Journal of the Transportation Research Board. 2391, 22-31.

Li, J., van Zuylen, H., Liu, C.H., Lu, S.F., 2011. Monitoring Travel Times in an Urban Network Using Video, GPS and Bluetooth, 14th Meeting of the Euro Working Group on Transportation: Procedia - Social and Behavioral Sciences. Poznan, Poland, paper 990, 20. 630-637.

Li, J., van Zuylen, H.J., Wei, G.R., 2014. Diagnosing and Interpolating Loop Detector Data Errors with Probe Vehicle Data. Transportation Research Record: Journal of the Transportation Research Board. 2423, 61-67.

Lubkowski, P., Laskowski, D., 2017. Assessment of Quality of Identification of Data in Systems of Automatic Licence Plate Recognition, in "Smart Solutions in Today's Transport". In: Mikulski, J. (Ed.). Springer-Verlag Berlin, Berlin, pp. 482-493.

Martchouk, M., Mannering, F., Bullock, D., 2011. Analysis of Freeway Travel Time Variability Using Bluetooth Detection. Journal of Transportation Engineering. 137.10, 697-704.

Moghaddam, S.S., Hellinga, B., 2013. Quantifying Measurement Error in Arterial Travel Times Measured by Bluetooth Detectors. Transportation Research Record. 2395, 111-122.

Pirc, J., Turk, G., Zura, M., 2016. Highway travel time estimation using multiple data sources. Iet Intelligent Transport Systems. 10.10, 649-657.

Rhead, M., Gurney, R., Ramalingam, S., Cohen, N., 2012. Accuracy of Automatic Number Plate Recognition (ANPR) and Real World UK Number Plate Problems, in "46th Annual 2012 Ieee International Carnahan Conference on Security Technology". In: Pritchard, D.A. (Ed.), pp. 286-291.

Sbaï, A., van Zuylen, H.J., Li, J., Zheng, F., Ghadi, F., 2017. Estimation of an Urban OD Matrix Using Different Information Sources, 17th International Conference on Computational Science and Its Applications. Trieste, Italy, paper 1433.

Soriguera, F., Robuste, F., 2011. Highway travel time accurate measurement and short-term prediction using multiple data sources. Transportmetrica. 7.1, 85-109.

Wikipedia. Daubechies wavelet, https://en.wikipedia.org/wiki/Daubechies_wavelet. Accessed July, 2017.

Woodard, D., Nogin, G., Koch, P., Racz, D., Goldszmidt, M., Horvitz, E., 2017. Predicting travel time reliability using mobile phone GPS data. Transportation Research Part C-Emerging Technologies. 75. 30-44.

Zheng, F., Van Zuylen, H., 2011. Modeling Variability of Urban Travel Times by Analyzing Delay Distribution for Multiple Signalized Intersections. Transportation Research Record: Journal of the Transportation Research Board. 2259, 80-95.

Zhu, G.-Y., Du, C., Zhang, P., 2016. A Travel Time Forecasting Model Based on Baseline Drift Correction. Journal of South China University of Technology (Natural Science Edition). 44.8, 131-138.