Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

Truck identification on freeways using Bluetooth data analysis

Ali Nadi[a]*, Maaike Snelder[a,b] ,Lori Tavasszy[a],Salil Sharma[a],Hans Van Lint[a]

[a]Delft University of Technology, Civil Engineering and Geosciences, Stevinweg 1, 2628 CN Delft, The Netherlands
[b] TNO, Anna van Buerenplein 1, 2595 DA Den Haag, The Netherlands

Abstract

Bluetooth technology is receiving more and more attention to support travel time measurement for intelligent transportation systems (ITS) applications. Bluetooth receivers are used to time-stamp passing identical vehicles at different locations based on their unique MAC addresses. This information is useful to predict travel times and estimate origin-destination flows on freeways. However, there is more valuable information in this big data source than has been explored to date. The main objective of this paper is to show vehicle type as a new feature that can be extracted from Bluetooth data, presenting a semi-supervised learning methodology which can be used to identify trucks on freeways. In this paper we also address how to deal with outliers in the Bluetooth data using an unsupervised machine learning technique to make vehicle identification and other data analysis more reliable. The predominant application for this vehicle identification is to predict travel time and estimate origin-destination specifically for freight transport. We use the A15 freeway in the Netherlands as a testbed. This corridor connects the port of Rotterdam to its hinterland and is one of the important freeways for logistic trip planning. The results show that the proposed method can identify trucks next to passenger cars with acceptable certainty and improved accuracy.

*Keywords:*Truck Travel Time; freight transport; Bluetooth Technology; Big data;

1. Introduction

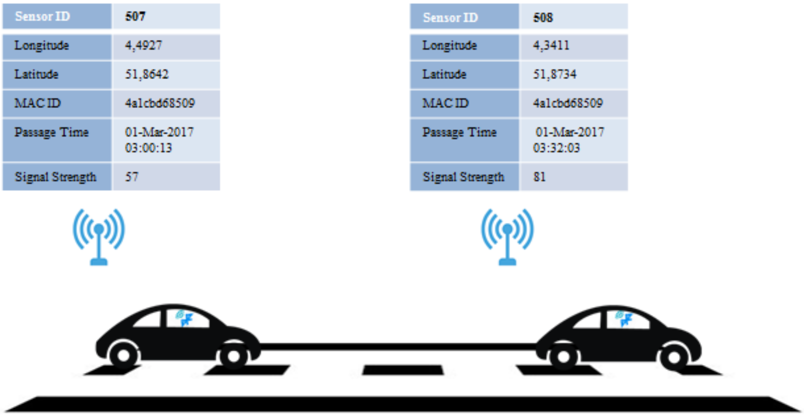
Predicting the travel times of trucks is vital mostly for freight carriers and third party logistics service providers who are responsible for trip planning for freight transport. For transport planners, knowledge of aggregate truck flows is important for studying the relations between logistics operations and traffic. Therefore, travel time prediction and origin-destination estimation for trucks has long been studied, using different sources of data. Weigh-in-motion, inductive loop detectors and GPS are the most popular data sources that have been used to estimate travel time. Some of these sources like weigh-in-motion need some sort of preprocessing to re-identify vehicles (i.e. to

* Corresponding author. Tel.: +31-626469382.
E-mail address: a.nadinajafabadi@tudelft.nl

classify vehicles based on the recorded signal). For example, Cetin and Nichols (2009) presented a two-stage methodology for vehicle re-identification and classification based on data collected by weigh-in-motion sensors. They used a Bayesian method to match vehicles between different locations in the first stage and solved a one-to-one assignment optimization problem in the second stage to make sure every vehicle is assigned only once. The results showed 99% accuracy for matching vehicles based on weigh-in-motion data. Ndoye et al. (2011) used the maximum a posteriori probability method for matching the vehicle signature detected downstream by inductive loop detectors with vehicles signatures detected upstream. Although this method showed accurate results in re-identification of vehicles using inductive loop detectors, it cannot classify vehicles into types. There are a few methodologies in literature that have been developed to improve the functionality of inductive loop detector devices in order to classify vehicles (Jeng et al. (2013), Chaudhuri et al. (2011), Ki and Baik (2006), Zhang et al. (2008), Keawkamnerd et al. (2008), Meta and Cinsdikici (2010)). Even though specific types of inductive loop detectors can classify vehicles, there are not enough installed devices with this option yet that can cover transportation networks. Therefore, most of the researchers focus on travel time estimation based on inductive loop detectors without considering any specific class of vehicle (Vanajakshi et al. (2009), Van Lint and Van der Zijpp (2003), Van Lint et al. (2005)).

GPS is another source of traffic data which can be used for more class specific travel time prediction and origin-destination estimation. Wang et al. (2016) described the speed distribution coefficient of variation to measure travel time reliability of trucks using probe data collected by GPS. Figliozzi et al. (2011) also used GPS data to calculate truck travel time and reliability for freight movements and also to assess the impact of congestion on freight vehicles. Another example of applications for class specific travel time prediction using GPS data is tracking real-time information of buses, aiming to reduce waiting times at bus stops (Vanajakshi, Subramanian et al. (2009), Lin and Zeng (1999)). The major challenges to use GPS data are map matching which requires extensive processing and privacy considerations, which limit the access to data.

The growing number of mobile devices has introduced another type of sensor for data collection. These sensors integrate the wireless communications technology (Wi-Fi spectra) and Bluetooth technology to connect sensors and mobile devices to each other. Bluetooth sensors record the unique MAC addresses of bypassing devices. To ensure privacy issues, providers hash this MAC address to unique IDs which are not trackable. This MAC addresses are time-stamped once they are detected by a sensor. The time difference between matching MAC addresses at different locations gives the travel time of different devices between different locations (see Fig. 1).



Sensor ID	507
Longitude	4,4927
Latitude	51,8642
MAC ID	4a1cbd68509
Passage Time	01-Mar-2017 03:00:13
Signal Strength	57

Sensor ID	508
Longitude	4,3411
Latitude	51,8734
MAC ID	4a1cbd68509
Passage Time	01-Mar-2017 03:32:03
Signal Strength	81

Fig. 1. Representation of how Bluetooth devices can capture vehicles.

This approach is becoming very popular because it is cost-effective, easy to use, and without any privacy issues compared to the three other methods used in the travel time data collection. For instance, Haghani et al. (2010) discussed about data processing algorithm for collecting ground truth travel times from Bluetooth technology. Martchouk et al. (2010) used Bluetooth data to study on travel time variability in freeways. Beside the advantages of this data collection technology, the presence of outliers may significantly affect the accuracy and reliability of travel time estimation obtained based on Bluetooth sensors (Araghi et al. (2015)). Therefore, Díaz et al. (2016) studied the reliability of the measurements, the representativeness of the travel time estimates and the issues regarding data filtering and outliers detection in Bluetooth data. Barceló et al. (2010) also applied Kalman filtering on the data obtained from Bluetooth sensors for short-term travel time prediction on freeways and to identify time-dependent origin-destination flow volumes. All these studies proved the quality of Bluetooth data for the travel time prediction and dynamic origin-destination flow estimation. However, we believe that there is more valuable information in this big data, which could be exploited. The main contribution of our paper is to show that beside ease of development, straightforward data processing, privacy friendliness and cost-effectiveness, Bluetooth data can be used to classify vehicle types on freeways as well. In this paper, we present a two-stage methodology using semi-supervised techniques to identify truck movements from Bluetooth data. In the first stage we use an unsupervised clustering approach using a Gaussian mixture model to identify truck movements within a series of travel time observations. In a second stage, we use support vector machine as a supervised classification method to improve the certainty of the truck identification using other spatial-temporal features which are driven from vehicle trajectory data. Fig.2 shows the process of the truck identification model based on Bluetooth data.

This paper is organized as follow: In section 2, the travel time visualization, filtering data and the outlier detection are discussed. Details about the proposed methodology for truck identification and the experimental results are given in section 3. We conclude in section 4 and discuss possible future research topics.

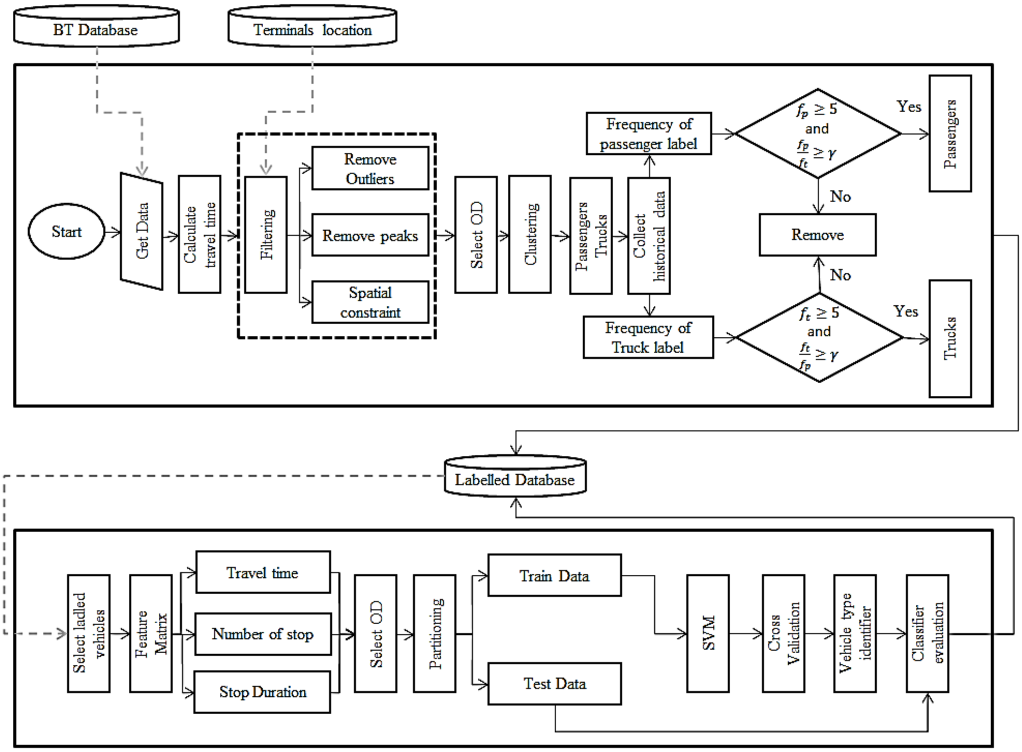


Fig. 2. Process of proposed methodology .

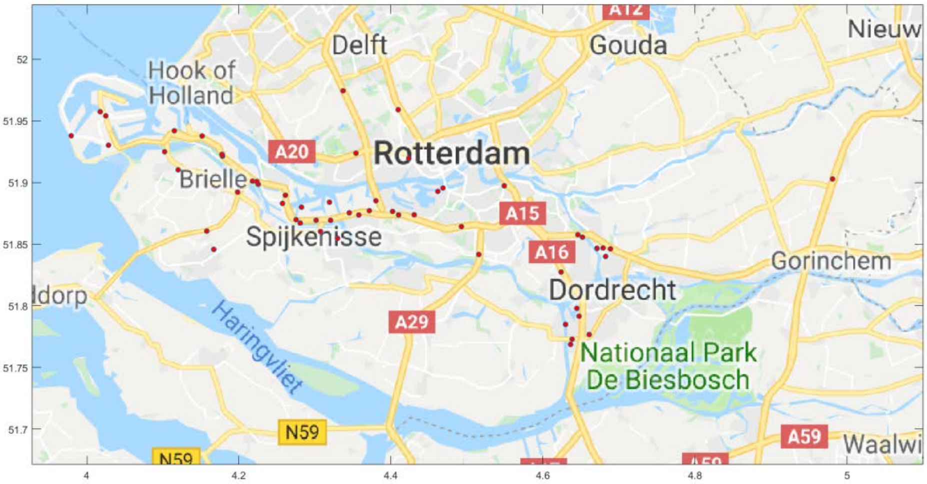


Fig. 3. location of Bluetooth devices along A15 motorway

2. Getting Data and pre-processing

There are 71 Bluetooth sensors (red dots in Fig. 3) located around A15 corridor and its connected links. This freeway connects port of Rotterdam to hinterland and is the most important transit motorway in Netherlands (see Fig.3). The data collected from all these sensors consist of 4 to 7 million of detections along one day. We collected all the detections from all devices for the 365 days in the year 2017.

Working with such a big data needs some sort of pre-processing and data filtering which will be discussed in this section. This data consists of Longitude of sensors ,Latitude of sensors, Devices MAC address, sensors ID, Passage Time, and signal strength (see Table 1).

Table 1. An example of raw data collected from Bluetooth sensors along A15 motorway.

Hashed MAC ID	Sensor ID	Longitude	Latitude	Passage time	Signal strength
"x4a1cbd68509"	526	5,607824	51,419942	01-Mar-2017 04:16:50	71
"x4a1cbd68509"	526	5,607824	51,419942	01-Mar-2017 04:16:52	-84
"x4a1cbd68509"	507	4,492719	51,864233	01-Mar-2017 11:31:28	86
"bfe4bad7d45 "	514	5,310883	51,640003	01-Mar-2017 05:48:17	76
"bfe4bad7d45 "	514	5,310883	51,640003	01-Mar-2017 05:48:17	69
"bfe4bad7d45 "	514	5,310883	51,640003	01-Mar-2017 05:48:18	-71
"bfe4bad7d45 "	514	5,310883	51,640003	01-Mar-2017 15:24:05	73
"x0d3c05563a2"	1580339	4,338537	51,87211	01-Mar-2017 13:33:47	89
"x0d3c05563a2"	1580335	4,32062	51,86935	01-Mar-2017 13:33:33	51
"x0d3c05563a2"	1580335	4,32062	51,86935	01-Mar-2017 13:33:36	71

One device might be detected by a sensor frequently and within some seconds with different signal strength (see Table 1). Because the resolution in this study is one second, we considered two times for each detection: one for arrival time of device to the sensor and the other for departure time of device from that sensor.

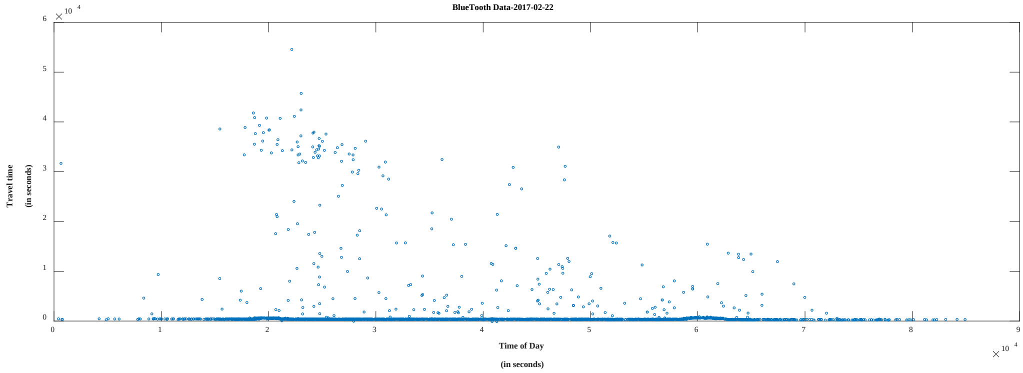


Fig. 4. travel time visualization for one day using Bluetooth data

2.1. Calculate travel time

To visualize this data, the time difference between arrival time of one (hashed) MAC ID to the one particular sensor and the departure time of the same MAC ID from the previous sensor is calculated. The result is the travel time of the vehicle between that sensor pair. For example, the travel time is illustrated in Fig. 4 for every vehicles passing sensor IDs “507” and “508” which are located along A15 Motorway from East to west direction toward Port of Rotterdam. The x-axis in this figure is time of day in seconds and the y-axis is travel time in seconds.

3. filtering

We can see from Fig. 2 that there exist a lot of outliers through travel time data. The reason for this outliers is that some vehicles passed sensor “507” but remained for some time in between sensors “507” and “508” for some reason (e.g. stop for gas station, break, loading or unloading, etc.) and then passed “508”. In this case, an abnormal travel time can be seen. In the following we describe how one can detect these outliers and remove them from the data set.

3.1. Outlier detection

One method used in literature to remove outliers is to limit data to those with travel time between a defined lower-bound and upper-bound (Barcelo et al. (2010)). In this method, the probability distribution of observed travel time is formed for a past period of time. Then, a moving average of the travel time frequencies is calculated which can be used to define the lower and upper bounds. Observations that represent travel time beyond these lower and upper limits are considered as outlier and removed from data. However, defining the lower and upper cut-off line for travel time cannot accurately detect all outliers; especially in congestion periods, where travel time is abnormally higher than in normal conditions, and also when two or more patterns of frequent travel time appear in the data. Therefore, we propose a density based clustering algorithm which can detect outliers based on their density of occurrence. This method is not based on defining lower and upper bound for travel time, instead, the approach is based on how travel time of one vehicle is far enough from other travel times so as to be clustered as noise. This method is based on the DBSCAN algorithm developed by Ester et al. (1996). Considering the set of travel times in (day time)×(travel time) space, points are classified as (1) core points, (2) reachable points and (3) outliers. A point p is a core point if at least a minimum number of points are within distance ε of it. A point q is directly reachable from p if point q is within distance ε from point p where p must be a core point. All those points which are not reachable from any other point are outliers. Considering 4 as the minimum number of points and 250 as the epsilon, the cleaned data is illustrated in fig. 5.

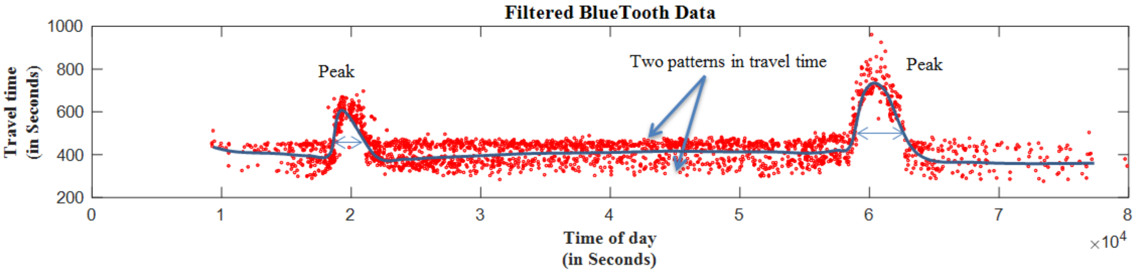


Fig. 5. Detection of peak period in travel time .

3.2. Removing peak periods

It is clear from Fig. 5 that there are two frequent patterns in travel times captured from those two Bluetooth sensors. This two patterns can only be explained under two conditions. The first reason is that there might be two different routes with two different travel time between sensors “507” and “508”. And the other reason is that there might be two class of vehicles with different speed limits. The first reason is not true due to this fact that there is only A15 corridor between these two sensors and vehicles only have to drive through A 15 to reach sensor “508” after passing “507”. In addition, we know that because of Port of Rotterdam, there are two class of vehicles, one trucks and one passenger cars, with different speed limits of 80 km\h and 120 km\h, respectively. Therefore we can infer that the vehicles with higher travel time are likely to be trucks. However, it can be seen in Fig. 3 that these two patterns convolved during the congestion period and thus make it impossible to see the clusters. This is because all types of vehicles drive at the same speed while they are in congestion. Therefore, the peak periods in the travel times must be detected and removed from data. To detect the peak periods, A general approach is to smooth the signal and then find peaks by comparing the local maximums of the fitted function. The same approach is used here using signal processing toolbox of Matlab to find local maxima and the peaks width in the travel time signal. By removing peaks width from travel time data, we have off-peak period of travel time data represented in Fig. 6

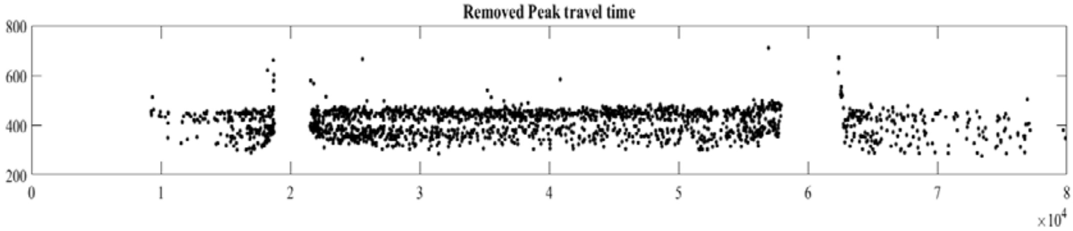


Fig. 6. (a) Filtered travel time

The values for Gaussian functions fitted on travel time is shown in Table 2 and the filtered travel time is presented in Fig. 6.

Table 2. the parameters of travel time signal in peak periods

Number of peaks	peak	location (× 10^4)	width	Prominence
1	654	2.12	8005	340.3904
2	742	6.27	7406	420.1834

3.3. Spatial constraint

In addition to the outlier and peak period filtering, we used the terminal location dataset to put spatial constraint on Bluetooth data for clustering vehicles (in the first stage) and training classifier (in the second stage). It means that we filtered the dataset for those vehicles that have at least one of the terminal locations around port of Rotterdam in their trajectories. We applied this spatial filtering to increase the certainty of the clustering due to this fact that there are less passenger cars and more trucks that pass through A15 motorway and meet terminals. However, later we use the whole dataset and trained classifier to assign other vehicles to one of these two classes.

4. Methodology for truck identification

To identify trucks using travel times captured by the Bluetooth sensors, a two-stage semi-supervised learning model is proposed here. This model, in its first stage, solves an unsupervised clustering problem to detect two classes of vehicles and then, in the second stage, identifies vehicles that are likely to be truck using supervised classification.

4.1. Clustering travel time

Looking at the distribution of travel time (see Fig. 7), the data looks multimodal: there are two peaks in the distribution of travel times (TT). A mixture of many unimodal Gaussian distributions can be used to model such data. The Gaussian mixture model is a parametrized kernel function with three values, the mixture weights, means and variances. Having a univariate Gaussian mixture model with *K* kernels for travel time data, the *i*th kernel has a mean of μ_i and variance of σ_i . the weight for kernel *i* is also defined as Θ_i.

$$f(TT) = \sum_{i=1}^{K} \theta_i N(TT \mid \mu_i, \sigma_i) \tag{4}$$

$$N(TT \mid \mu_i, \sigma_i) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp(-\frac{(TT-\mu_i)^2}{2\sigma_i^2}) \tag{5}$$

$$\sum_{i=1}^{k} \theta_i = 1 \tag{6}$$

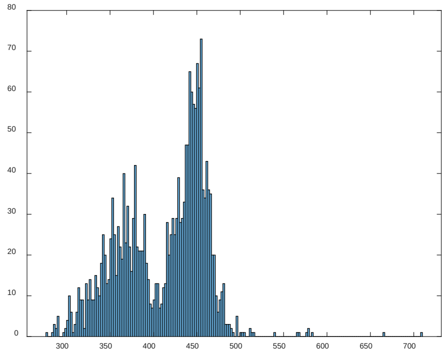


Fig. 7. Distribution of travel time

The equation 6 normalizes the probability distribution. Given a univariate model's parameters, the probability that a point in data belongs to a cluster C$_i$ is calculated using Bayes' theorem as shown in equation 7.

$$P(C_i \mid TT) = \frac{P(TT, C_i)}{P(TT)} = \frac{P(C_i)P(TT \mid C_i)}{\sum_{j=1}^{K} P(C_j)P(TT \mid C_j)} = \frac{\theta_i N(TT \mid \mu_i, \sigma_i)}{\sum_{j=1}^{K} \theta_j N(TT \mid \mu_j, \sigma_j)} \tag{7}$$

The a-posteriori estimates of the component probabilities are typically trained by using maximum likelihood estimation techniques, which maximize the similarity, or likelihood, of the observed data given the model parameters. The expectation maximization (EM) is the most popular numerical technique which is used to estimate maximum likelihood. As it is mentioned, there are two clusters in the travel time distribution; a two-kernel mixture is needed to cluster travel times. Fig.8 shows the result of clustering after the parameters of the Gaussian mixture model are obtained (see Table 3).

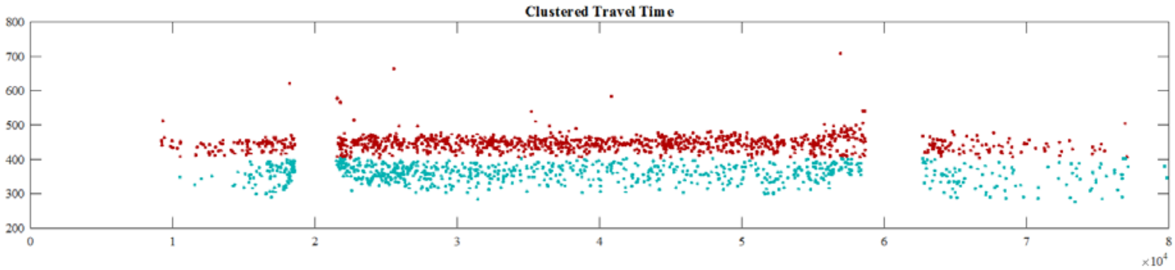


Fig. 8. Clustered travel time

The estimated parameters of two components Gaussian mixture are in Table 3. The Akiak’s Information Criterion (AIC) and Bayesian Information Criterion (BIC) are minimized to estimate the number of Gaussian mixture components.

Table 3. Estimated parameters of the two components Gaussian mixture

Number of components		μ	σ
1		654	2.12
2		742	6.27
AIC:	19532,16		
BIC:	19559.76		
Log-likelihood	9.761e+03		

The center of clusters μ_1 and μ_2 are 446,9622s and 359,7014s respectively. This means there are two average travel time between two selected sensors “507” and “508”. The cluster with higher average travel times are more likely to include trucks. However, we cannot label them truck by certainty because, one passenger car may drive with the speed of a truck; in this case many passenger cars that drive slowly could be included in the cluster with higher average travel time.

To increase the certainty of truck identification, more features should be used as indicators. The proposed approach in this paper is to do the same process as mentioned above for multiple locations and for a duration of one month (the October 2017) to create a historical data set for clustered travel times. The historical data set consists of 28213 Vehicle IDs that was labeled as passengers and 101275 vehicle IDs that was labeled as Truck. Some of these vehicles belong to one category and some of them belong to both categories but with different frequency. For example one vehicle might be labeled as truck at some certain times and locations but also be labeled as passenger at different times and locations. The frequency of vehicles being labeled as truck (f_t) and passenger (f_p) for each class is presented in fig. 9.

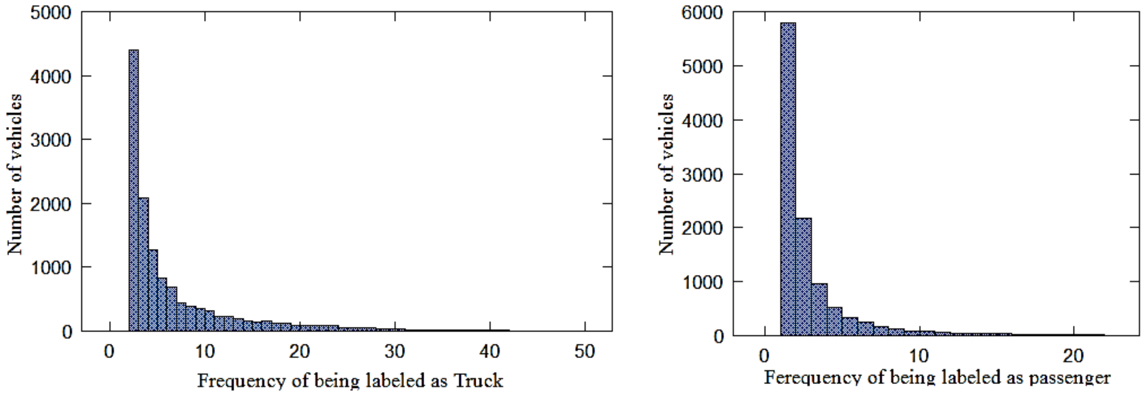


Fig. 9 the frequency of labels in each class for every vehicle

To increase the certainty, We only accept the labels with high frequency (let say more than 5 depending on the level of certainty needed) in each class. For those vehicles which have high frequency and they are in both classes we only accept the vehicles which their frequency proportion are higher than a certain level of certainty (γ). In this research we set the value of $\gamma = 2$ which means the vehicles should have been labeled as truck at least twice more than as passenger in order to be accepted as truck and the other way around. Of course the more level of certainty we choose, the fewer labeled data we have. It is because we remove many of those vehicles we are not certain about considering the rules below.

$$f_t \geq 5 \text{ and } \frac{f_t}{f_p} \geq \gamma \ \rightarrow Truck \tag{8}$$

$$f_p \geq 5 \text{ and } \frac{f_p}{f_t} \geq \gamma \rightarrow Passenger \tag{9}$$

Given the labeled data set, we can train a classifier to predict the label of other vehicles which we could not labeled through the previous process. To do so, we use a support vector machine to model the vehicle type identification.

4.2. Vehicle type Identification

In order to identify vehicle types even in fuzzy location and time (e.g. peak periods), we have to train a classifier. Therefore, the labeled vehicles (whether Trucks or Passengers) data are selected for a specific day (25th of October 2017) matching BT database and Labelled database. To train a classifier, we need to have more features beside travel time that can distinguishes between truck and passenger cars.

4.3. Feature matrix

Having trajectory data of labelled data, we consider number of activities and activity duration next to the travel time as predictor features. Because the distribution of number of activities and average activity duration for truck and passenger cars can be quite different. The process of calculating travel time from BT dataset has been described in section 2.1. However, the process of calculating number of activities and average activity duration is a bit more complicated because prior to that we have to understand when an activity happens.

4.3.1. Number of activity

In order to find out when an activity happens, we need to track every vehicles' trajectory. The activity is assumed that has happened when a vehicle passed two sequential sensors with a large delay in between. There could be two reasons for such delays. (1) the vehicle stuck in the congestion or (2) the vehicle stopped for doing an activity (e.g. gas station, rest , loading and unloading, work, etc.). The last one can be considered as an actual activity. we use the Inter Quartile Range (IQR) of travel times to detect activities in every vehicle trajectories. The IQR is a measure of variability, based on dividing distribution of a data set into quartiles. The values that separate each part are called the first (Q1), second (Q2), and third (Q3) quartiles. every vehicle's trajectory consists of at least one pair of sequential O/D. for all O/Ds in all trajectories, activity for a vehicle happens if the travel time of that vehicle exceeds Tukey upper bound in box-and-whisker plot (Q3+1.5×IQR). As peak periods in travel time may shift the IQR inappropriately, we divided the time of day into 5 peak and off peak periods (00:00 to 6:00, 6:01 to 9:00, 9:01 to 16:00, 16:01 to 19:00, 19:01 to 24:00).

4.3.2. Average activity duration

The activity duration is the amount of delay which happens between two sequential O/D. this delay can be calculated by subtracting passage time of vehicles doing some activity from upper bound of expected travel time. The upper bound of travel time is the Tukey upper bound in box-and-whisker plot (Q3+1.5×IQR) of travel time. The average activity duration is the mean of activity durations through every vehicle's trajectory.

Having the travel time, number of intermediate activities, and average activity duration for every vehicle and for every O/D, we can form the feature matrix of vehicle identification model (see Table 4).

Table 4 a sample of feature matrix including labels and Users ID for 506 to 507 sensor ID

Users	Label	Travel Time	Number of activities	Average activity Duration
11	0	373	0	0
29	1	516	0	0
37	1	473	2	13106.75
39	0	415	2	15776.48
67	1	530	2	312.57
76	1	522	0	0
83	1	515	1	8033.63
85	0	446	1	44
90	1	480	0	0

To see if there is a difference in means of average activity duration and Number of activities between Truck and Passenger in the population, we perform independent samples t-test for these variables. The null hypothesis for average activity duration is that there is no difference in means of average activity duration of Passengers and Trucks and the null hypothesis for number of activities is that there is no difference in means of number activity duration of Passengers and Trucks.

Table 5 independent samples t-test for number of stop and average stop duration

	t-test for Equality of Means						
	t	Degree of freedom	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
Average activity duration	9.26	1360	0.0	4163.56	449.68	3281.42	5045.70
Number of activities	-2.40	1360	0.02	-0.25	0.10	-0.454	-0.46

The t-test result reject the null hypothesis and proves that Passenger's average activity duration is significantly longer than Trucks with a confidence of 95% in the population. The t-test results also reject the null hypothesis and proves that Truck's number of activities is significantly more than passengers with a confidence of 95% in the population. It make sense because passengers usually have direct trip to work and home with limited intermediate activities. The duration of activities for passengers is also long as they have specific working duration of 8-9 hours. On the other hand, trucks usually have more frequent activities like (un)loading and their activity duration is limited compared to that of the passenger's activities.

4.4. Support vector machine

Give a set of training dataset I=$\{x_i, y_i\}$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1,1\}$, the support vector machine is to find a classifier function as follow.

$$w^T x_i + b = 0 \tag{10}$$

As our data is not linear, the SVM should find a nonlinear function to classify data by mapping x_i to a higher space using a transfer function $\emptyset(x)$.

$$w^T \emptyset(x_i) + b = 0 \tag{11}$$

The SVM considers a margin around this hyperplane to find the best classifier with least operational risk. Therefore the aim is to label y_i as:

$$y_i = 1 \Longrightarrow w^T \emptyset(x_i) + b > 1 \qquad 12$$

$$y_i = -1 \Longrightarrow w^T \emptyset(x_i) + b < -1 \qquad 13$$

Therefore the objective function to find the best classifier in SVM is to maximize the distance between two above mentioned hyperplane. Considering the constraints and relaxing the problem by Lagrange multiplier, The problem that SVM tries to solve is an optimization problem as below:

$$min \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j \, K(x_i, y_j) - \sum_i \alpha_i \qquad 14$$

S.t.

$$\sum_i \alpha_i y_i = 0 \qquad 14$$

$$0 \leq \alpha_i \leq C \qquad 15$$

$$K(x_i, x_j) = \emptyset(x_i)^T \emptyset(x_j) \qquad 16$$

Where α is Lagrange Multiplier , C is the penalty factor which is the upper bound of the box constraint and $K(x_i, x_j)$ is the kernel function. This optimization problem can be re-written as a quadratic programming problem :

$$\min \frac{1}{2} \alpha^T H \alpha + f^T \alpha \qquad 17$$

$$\sum_i \alpha_i y_i = 0 \qquad 18$$

$$0 \leq \alpha_i \leq C \qquad 19$$

$$h_{ij} = y_i y_j K(x_i, x_j) \qquad 20$$

$$H = [h_{ij}]_{n \times n} \qquad 21$$

$$f = [-1]_{n \times 1} \qquad 22$$

Solving this quadratic programing will give us a set of solution α which can classify our data as follow:

$$Sv = \{i | 0 \leq \alpha_i \leq C\} \qquad 23$$

$$b = \frac{1}{|Sv|} \sum_{i \in Sv} (y_i - \sum_j \alpha_j y_j K(x_j, x_i)) \qquad 24$$

$$y = sign(\sum_i \alpha_i y_i \, K(x_i, x) + b) \qquad 25$$

Where Sv is the set of support vectors and b is the bias term. There are various type of kernels that can be used for SVM binary classification. the most famous one for nonlinear space is Gaussian kernel which is also an radial based function (RBF).

$$K(x_i, x) = \exp(-\frac{1}{2\sigma^2}\left\|x_i - x_j\right\|^2) \qquad 26$$

Where σ is standard deviation of kernel function which defines the kernel scale. To train the SVM classifier for truck identification, we select the sensors 506 and 507 as a candidate OD. There are 1362 labelled vehicles passing these two sensor during 25th of October 2017. We partitioned data into two train and test datasets. The test datasets consists of a sample of 100 vehicles which are randomly drawn from labelled dataset and the rest of the data (1262 vehicles) are kept in train dataset. The train data set will be used to train SVM and the test dataset will be used to evaluate classifier after being trained. We centre and scale each predictor variable by the corresponding weighted column mean and standard deviation.

4.4.1. SVM hyper parameter

As we can see C and kernel scale and kernel type are hyper parameters of SVM which should be set for binary classification. These hyper parameters should be optimized to get the best classification results. We used Bayesian optimization for this purpose.

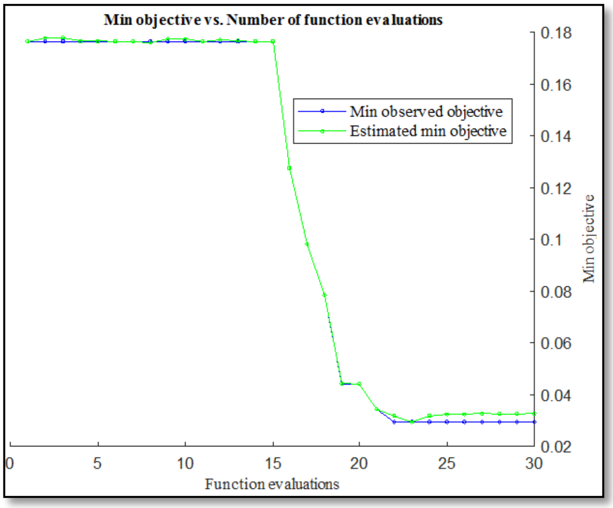


Fig. 10 objective function value per iteration during hyper parameter optimization

The results shown in table 6 indicates the best box constraint , kernel scale and kernel type. The Fig. 10 also shows that the optimization process minimized after evaluating the objective function for 30 times.

Table 6. Best estimated feasible point

Box Constraint	Kernel Scale	Kernel function	Observed objective	Estimated objective
2.1928	1.3411	RBF	0.029412	0.032597

4.4.2. Cross validation

To have an insight on how this model will generalize to an unlabelled dataset, we use 10-fold cross validation during training process. In 10-fold cross-validation, the original dataset is randomly partitioned into 10 subsamples with equal size. One of these 10 subsamples is used as the validation data for testing the model, and the remaining 9 subsamples are used as training data. The cross-validation process is then repeated 10 times in a way that each of the 10 subsamples used exactly once as the validation data. The average of 10 results can then be used to produce a single estimation. In our model, we hold out 10% of data for validation and the remaining 90 % for training process. The result of the cross validation show that the out of sample misclassification rate is 0.0423. In other words the generalization rate of this model is approximately 96 %.

4.4.3. Vehicle type identifier

The optimized SVM has been trained with the train data set using above hyper parameters. The parameters of the trained vehicle type identifier are in the table 7.

Table 7 shows the parameters of the trained vehicle type identification

Variables		σ	µ
Travel Time		73.7059	494.6268
Number of Stop		1.6062	1.5593
Average Stop duration		6661.3	4371.8
Bias		-0.1438	
Misclassification rate		0.0423	
C		2.1928	
Kernel Scale		1.3411	
MSE		0.01	

Score transform	Sigmoid	A=-2.145
		B=0.236	

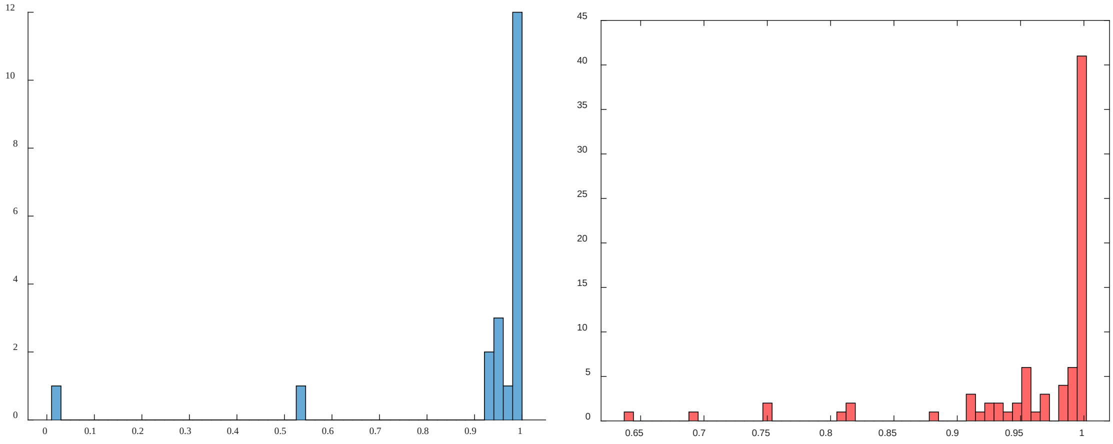
where the parameters A and B in score transform function are the slope and intercept parameters, respectively.

4.4.4. Classifier evaluation

To evaluate the accuracy of the model, we used the test data to predict their labels using vehicle type identifier model. In this section we compare the true label of the test data with the predicted labels using mean square error.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} ||t - l||^2 \qquad 27$$

Where, N is the number of observation, *t* is the true labels of input test data and the *l* is the predicted labels. As the number of observation in the test data is 100 and as MSE value is 0.01, it means that the model only misclassified 1 out of 100 samples which is a high accuracy for the model. For the test data the distribution of estimated posterior probability for both Truck and Passenger class is shown in Fig 11. To have a better insight the table 8 shows the first 10 row of the tests data as an example



(a) Posterior probability for Passenger class

(b) Posterior probability for the Truck class

Fig. 11 (a) the distribution of posterior probability for passenger class (b) the posterior probability distribution for truck class

The results show that the average posterior probability for Truck class is 96% and the average posterior probability for passenger class is 91%. It means that the average probability of a truck being labelled as truck is 96 % and the average probability of a passenger car being labelled as passenger car is 91% which it proves the high accuracy of vehicle identifier model.

Table 8 shows the predicted labels verses true labels and their corresponding class score and posterior probability

True Labels	Predicted labels	scores	Passenger class posterior probability	Truck class posterior probability
0	0	0.027	0.54	0.455
1	1	3.24	0.001	0.998
0	0	-1.70	0.98	0.019
1	1	1.48	0.049	0.950
1	1	1.31	0.070	0.929
1	1	3.11	0.001	0.998
1	1	3.33	0.001	0.998
1	1	1.50	0.047	0.952
1	1	2.81	0.003	0.996
0	1	2.02	0.016	0.983
Number of observation:		100		
Average posterior probability for Truck class:		0.96		
Average posterior probability for Passenger class:		0.91		

5. **Conclusions and recommendations**

Our paper presents a robust method for vehicle classification based on BT data. It allows to separate trucks and passenger cars with high accuracy. This is especially important in cases where (1) flows are heterogeneous, such as around industrialized areas and (2) where predictions are needed that are customized towards a specific vehicle class. Our paper adds to recent work that uses BT data for other purposes, such as travel time estimation and O/D estimation. Next steps for research may include

- Testing the effect of classification on travel time predictions and O/D matrix estimation.
- Testing the effect of classification on estimation of freight travel time variability.
- Predicting the activity-travel sequences for trucks

Acknowledgements

This work is part of the research project ToGRIP which is supported by NWO, TKI Dinalog, Commit2data, Port of Rotterdam, SmartPort, Portbase, TLN, Deltalinqs, Rijkswaterstaat and TNO. We would like to thank Taoufik Bakri and Peter Knoppers for their work on Bluetooth data.

References

Araghi, B. N., K. S. Pedersen, L. T. Christensen, R. Krishnan and H. Lahrmann (2015). "Accuracy of travel time estimation using Bluetooth technology: Case study Limfjord tunnel Aalborg." International Journal of Intelligent Transportation Systems Research **13**(3): 166-191.

Barceló, J., L. Montero, L. Marqués and C. Carmona (2010). "Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring." Transportation Research Record: Journal of the Transportation Research Board(2175): 19-27.

Cetin, M. and A. Nichols (2009). "Improving the accuracy of vehicle reidentification algorithms by solving the assignment problem." Transportation Research Record: Journal of the Transportation Research Board(2129): 1-8.

Chaudhuri, P., P. T. Martin, A. Z. Stevanovic and C. Zhu (2011). "The effects of detector spacing on travel time prediction on freeways." Int J Eng Appl Sci **7**(1): 1-10.

Díaz, J. J. V., A. B. R. González and M. R. Wilby (2016). "Bluetooth traffic monitoring systems for travel time estimation on freeways." IEEE Transactions on Intelligent Transportation Systems **17**(1): 123-132.

Ester, M., H.-P. Kriegel, J. Sander and X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd.

Figliozzi, M., N. Wheeler, E. Albright, L. Walker, S. Sarkar and D. Rice (2011). "Algorithms for Studying the Impact of Travel Time Reliability Along Multisegment Trucking Freight Corridors." Transportation Research Record: Journal of the Transportation Research Board(2224): 26-34.

Haghani, A., M. Hamedi, K. Sadabadi, S. Young and P. Tarnoff (2010). "Data collection of freeway travel time ground truth with bluetooth sensors." Transportation Research Record: Journal of the Transportation Research Board(2160): 60-68.

Jeng, S.-T., L. Chu and S. Hernandez (2013). "Wavelet-k nearest neighbor vehicle classification approach with inductive loop signatures." Transportation Research Record: Journal of the Transportation Research Board(2380): 72-80.

Keawkamnerd, S., J. Chinrungrueng and C. Jaruchart (2008). Vehicle classification with low computation magnetic sensor. ITS Telecommunications, 2008. ITST 2008. 8th International Conference on, IEEE.

Ki, Y.-K. and D.-K. Baik (2006). "Vehicle-classification algorithm for single-loop detectors using neural networks." IEEE Transactions on Vehicular Technology **55**(6): 1704-1711.

Lin, W.-H. and J. Zeng (1999). "Experimental study of real-time bus arrival time prediction with GPS data." Transportation Research Record: Journal of the Transportation Research Board(1666): 101-109.

Martchouk, M., F. Mannering and D. Bullock (2010). "Analysis of freeway travel time variability using Bluetooth detection." Journal of Transportation Engineering **137**(10): 697-704.

Meta, S. and M. G. Cinsdikici (2010). "Vehicle-classification algorithm based on component analysis for single-loop inductive detector." IEEE Transactions on Vehicular Technology **59**(6): 2795-2805.

Ndoye, M., V. F. Totten, J. V. Krogmeier and D. M. Bullock (2011). "Sensing and signal processing for vehicle reidentification and travel time estimation." IEEE Transactions on Intelligent Transportation Systems **12**(1): 119-131.

Van Lint, J., S. Hoogendoorn and H. J. van Zuylen (2005). "Accurate freeway travel time prediction with state-space neural networks under missing data." Transportation Research Part C: Emerging Technologies **13**(5-6): 347-369.

Van Lint, J. and N. Van der Zijpp (2003). "Improving a travel-time estimation algorithm by using dual loop detectors." Transportation Research Record: Journal of the Transportation Research Board(1855): 41-48.

Vanajakshi, L., S. C. Subramanian and R. Sivanandan (2009). "Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses." IET intelligent transport systems **3**(1): 1-9.

Wang, Z., A. Goodchild and E. McCormack (2016). "Measuring truck travel time reliability using truck probe GPS data." Journal of Intelligent Transportation Systems **20**(2): 103-112.

Zhang, W., G. Tan, N. Ding, Y. Shang and M. Lin (2008). Vehicle classification algorithm based on binary proximity magnetic sensors and neural network. Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on, IEEE.