Accepted for publication in Journal of Intelligent Transportation Systems

# Gaussian Mixture Model-Based Speed Estimation and Vehicle Classification Using Single Loop Measurements

Yunteng Lao[1]; Guohui Zhang[2]; Jonathan Corey[1]; Yinhai Wang[1*]

[1]Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195, USA

[2]Department of Civil Engineering, University of New Mexico, Albuquerque, NM 87106 USA

* Corresponding author

**Abstract:**

Traffic speed and length-based vehicle classification data are critical inputs for traffic operations, pavement design and maintenance, and transportation planning. However, they cannot be measured directly by single-loop detectors, the most widely deployed type of traffic sensor in the existing roadway infrastructure. In this study, a Gaussian Mixture Model (GMM)-based approach is developed to estimate more accurate traffic speeds and classified vehicle volumes using single-loop outputs. The estimation procedure consists of multiple iterations of parameter correction and validation. After the GMM is established to empirically model vehicle on-times measured by single-loop detectors, the optimal solution can be initially sought to separate length-based vehicle volume data. Based on the on-time of the separated short vehicles from the GMM, an iterative process will be conducted to improve traffic speed and classified volume estimation until the estimation results become statistically stable and converge. This method is straightforward and computationally efficient. The effectiveness of the proposed approach was examined using data collected from several loop stations on Interstate-90 in the Seattle area. The traffic volume data for three vehicle classes are categorized based on the proposed method. The test results show the proposed GMM approach outperforms the previous models, including conventional constant g-factor method, sequence method, and moving median method, and produces more reliable, accurate estimates of traffic speeds and classified vehicle volumes under various traffic conditions.

*Key words*: Gaussian Mixture Model, Single-loop detectors, Speed estimation, Vehicle length estimation, Vehicle classification

## 1. Introduction

Traffic speed and vehicle classification data are important inputs for traffic operations, pavement design and maintenance, and transportation planning (Zhang et al. 2006). Long vehicles, such as, buses, large trucks, and recreational vehicles are associated with slow acceleration, inferior braking, and large turning radii. Due to the heavy weight of these vehicles when they are fully loaded, they significantly impact pavement life time and have higher relative rates of fatal crashes (NCSA, 2005; Zhang et al., 2006). Among existing traffic detectors, inductive loop detectors (ILDs), especially single-loop detectors, are the most commonly used on highway systems (Coifman and Kim, 2009). However, they cannot measure both traffic speed and vehicle classification data directly (Dailey, 1999). Dual-loop detectors are capable of measuring classified vehicle volumes and speeds, but they are not widely available in the existing roadway infrastructure (Coifman and Kim, 2009).

Many previous studies have been conducted to estimate classified vehicle volumes and speeds from single-loop measurements in the past (Wang and Nihan, 2004). For example, the fundamental speed, volume, and density formulation was used to estimate space-mean speed (Athol, 1965; Mikhalkin et al., 1972; Wang and Nihan, 2000; Hellinga, 2002). In this conventional method, a constant, $g$, related to the average effective vehicle length is used for speed estimation. The limitation of this method is that the $g$ value may not be constant, but instead vary with vehicle lengths (Hall and Persaud, 1989; Pushkar et al., 1994).

In the past decades, a series of studies have been conducted to advance the constant g-factor method, such as the cusp catastrophe theory model (Pushkar et al., 1994), the linear model using slew rates of single loop inductive waveforms (Sun and Ritchie, 1999), moving median method (Coifman et al., 2003), the Kalman filter method (Dailey,1999; Ye et al., 2006), the artificial neural network algorithm (Zhang et al., 2006), the sequence method (Coifman and Kim, 2009), and the Bayesian method (Li, 2009; Jin et al., 2010). These methods have improved the accuracy and consistency of speed estimation in most situations. However, their accuracy may be constrained by the percentage of long vehicles or by training data. For example, the moving median method is sensitive to the percentage of long vehicles (Coifman and Kim, 2009) and the artificial neural network approach is constrained by its training data set, and its transferability needs further testing. Coifman and Kim (2009) developed a distribution method that considered the on-time distribution of current data. Although this distribution method outperformed the previous methods, it requires separating on-times into four groups by some specific thresholds. The setting of thresholds can considerably affect the estimation results.

The previous studies indicate that the regular traffic flow composition consists of a number of vehicle classes (May et al., 2004; Coifman and Kim, 2009). Based on this characteristic, the distributions of vehicle lengths and on-times follow a mixture distribution (McLachlan and Basford, 1988; McLachlan and Peel 2000) as observed empirically. A mixture distribution is the probability distribution of a random variable whose value can be interpreted as a set of other random variables. Based on the central limit theorem, the mean of vehicle lengths is approximately normally distributed when large amounts of data are analyzed. The Gaussian Mixture Model (GMM) (McLachlan and Basford, 1988; Reynolds and Rose, 1995) is an analytical tool to model the data characterized by a mixture of normal distributions. Due to its adaptability and applicability, the GMM has been successfully applied across a wide variety of fields, such as speech recognition (Reynolds and Rose, 1995), robotics (Ju et al., 2008), clustering operations (Hasan and Gan, 2009), and data quality control (Wang et al., 2009; Corey et al., 2011). Thus, this research aims at developing a GMM-based approach to extract traffic speeds and classified vehicle volumes using single-loop measurements. The proposed GMM approach requires only single loop data input and can produce more accurate estimates of traffic speeds and classified vehicle volumes on the test locations with the whole day data.

In the remaining parts of this paper, the relationship among speed, vehicle length and on-time is introduced. After that, three previous speed estimation methods: a conventional constant g-factor method, the sequence method, and the moving median method are reviewed. Then the GMM method is established for vehicle speed and length estimation. This is followed by the experimental tests to examine the effectiveness of the proposed approach. Finally, conclusions are provided at the end of this paper.

# 2. Methodology

## 2.1 Relationship among speed, vehicle length and on-time

Vehicles trigger loop inductance changes and are recorded during the time when the front bumper reaches the leading edge of the loop and the rear bumper leaves the following edge of the loop. Thus, as shown in Figure 1, the effective vehicle length ($L_E$) is defined as the sum of the actual vehicle length ($L_V$) and length of the loop detection zone, which is assumed to equal the loop length ($L_L$). The effective length is also a function of vehicle space-mean speed ($v$) and the on-time ($OT$).

$$\mathrm{L_E} = \mathrm{L_V} + \mathrm{L_L} = \mathrm{v} \cdot \mathrm{OT} \tag{1}$$

The on-time is the time slot between the vehicle front bumper reaches the loop and the vehicle rear bumper leaves the loop. Rearranging (1), the on-time and space-mean speed for a vehicle can be defined as

$$OT = \frac{L_E}{v} = \frac{L_V + L_L}{v} \tag{2}$$

$$v = \frac{L_E}{OT} = \frac{L_V + L_L}{OT} \tag{3}$$

Thus, with a known loop length ($L_L$), the space-mean speed within the distance of the effective vehicle length is directly related to the distribution of vehicle lengths and the distribution of on-time.

## 2.2 Previous speed estimation methods

As mentioned in Section 1, many previous studies concentrated on speed and classified vehicle volume estimation. In this section, three widely used methods, the conventional method, i.e. a constant g-factor method (Wang and Nihan, 2000), the sequence method (Coifman and Kim, 2009), and the moving median method (Coifman et al., 2003) are reviewed to provide the insights for the proposed GMM development.

### 2.2.1 Conventional method

In (2), loop length ($L_L$) is a constant, normally six feet for point detectors, and on-time ($OT$) can be measured from the single-loop detectors directly. However, both speed and vehicle length are unknown variables. Previous studies attempted to use the aggregate flow ($q$), and occupancy ($occ$) to reduce the speed estimation error (Mikhalkin et al., 1972; Pushkar et al., 1994; Dailey, 1999; Wang and Nihan, 2000; Coifman, 2001).

In conventional practice, a constant, $g$, is used to convert occupancy to density and the space-mean speed can be estimated by the fundamental speed, volume, and density formulation (Athol, 1965; Mikhalkin et al., 1972; Wang and Nihan, 2000).

$$v = \frac{q}{occ \cdot g} = \frac{N}{T \cdot occ \cdot g} \tag{4}$$

where $T$ is the time interval during which the volume measurements are aggregated by the loop detection system; $N$ is the number of vehicles during that time interval and $T \cdot occ$ equals to the average vehicle on-time during a time interval, $T$. Equation (4) is an extension of (3) (Coifman and Kim, 2009) and is used when the traffic count ($N$) and occupancy ($occ$) are available whereas (3) is used when the on-time is measurable.

Assuming that the vehicle length is constant, the space-mean speed (v) can be estimated if the loop length ($L_L$) and the measured on-time are available. The advantage of this conventional method is its easy implementation. However, it may introduce huge errors in some situations since this method assumes vehicle length is a constant even though the percentage of long vehicles may vary significantly during different time periods.

### 2.2.2 Sequence method

As an improved version of the conventional method, the sequence method does not use an assumed vehicle length for calculating the space-mean speed of the whole time period. Instead, it uses a sequence of vehicles to capture the space-mean speed for each small time period (Coifman, 2001; Neelisetty and Coifman, 2004). Then, the space-mean speed can be defined as

$$v = \frac{L_E}{OT} = \frac{L_V + L_L}{OT} \tag{5}$$

where $\overline{OT}$ is the moving average on-time. In the Coifman and Kim (2009) study, to estimate vehicle speed, the sequence method also separated the on-times between short vehicles and long vehicles.

### 2.2.3 Moving median method

To mitigate long vehicles' effect on the moving average, it would be desirable to filter out the on-times of very long vehicles. For example, at one detector station, Coifman (2001) found that 85% vehicles are between 15 and 22 feet, whereas some long vehicles are four-times as long as the median of these shorter vehicles. Therefore, Coifman et al. (2003) also recommended a moving median method for vehicle speed estimation. Instead of using a moving average on-time, they used a moving median in their estimation function. In this case, the speed can be estimated as

$$v = \frac{L_E}{median(OT)} = \frac{L_V + L_L}{median(OT)} \tag{6}$$

where $median(OT)$ is the moving median on-time for each short time period that the traffic condition does not change much. In Coifman and Kim (2009) study, a window of 33 vehicles is used in both the sequence and moving median methods.

## 2.3 Speed and vehicle length estimation using GMM

### 2.3.1 On-time characteristics and modeling

Previous studies have indicated that the vehicle on-time distribution is implicitly correlated with the combined distribution of different vehicle types (May et al., 2004; Coifman and Kim, 2009), such as a narrow short vehicle on-time distribution and wider long vehicle on-time distributions (Figure 2) or a short vehicle on-time distribution and multiple distributions for long vehicle (Figure 3). Note that in a dual loop, "M loop" refers to the upstream inductive loop detectors and "S loop" referred to the downstream inductive loop detectors.

Traffic flow is composed of different types of vehicles in terms of length. As illustrated in Figures 2 and 3, the on-time distributions are expected to be multi-modal for the mixed vehicle composition of traffic flow. The first peak, which is very high and narrow, represents the on-time distribution for short vehicles. The subsequent broader peaks include the on-times for different types of long vehicles. To approximate these unique features, a Gaussian Mixture Model (GMM) (McLachlan and Basford, 1988) can be applied to analyze the on-time distribution data and capture the on-time distribution patterns for different vehicle types. Assume that the on-time of $N_1$ adjacent vehicles joins a mixture of $K$ Gaussian distributions, which would make the probability of measuring the $i^{th}$ on-time as

$$P(OT_i) = \sum_{k=1}^{K} \omega_k f(OT_i, \mu_k, \sigma_k^2) \tag{7}$$

where $OT_i$ is the on-time of the $i^{th}$ vehicle observation; $K$ is the number of vehicle categories based on vehicle length, $K$=1 to 4 depending on the number of categories observable at the study site. (In this research, $K$=3 was used for the two study sites based on empirical observation and the efforts of trial and error;) $\omega_k$ is the weighting factor of the $k^{th}$ Gaussian distribution $f(OT_i, \mu_k, \sigma_k^2)$ with a mean of $\mu_k$and a variance of $\sigma_k^2$, $k$=1,2,..$K$ and $\sum_{k=1}^{K} \omega_j = 1$.The Gaussian distribution of on-time used here is only one-dimensional and is defined as:

$$f(OT_i, \mu_k, \sigma_k^2) = \frac{1}{\sigma_k\sqrt{2\pi}} e^{-(OT_i-\mu_k)^2/2\sigma_k^2}, \qquad OT_i > 0 \tag{8}$$

The weighting factor $\omega_k$ indicates the percentage of vehicles belonging to category $k$, whereas $\mu_k$ represents the mean on-time of vehicle category $k$. The value of the weighting factor is associated with local traffic composition and directly affected by the percentage of trucks in the traffic flow.

If we define the Gaussian distribution parameter set as $= \{\mu_1, \dots, \mu_K, \sigma_1^2, \dots, \sigma_K^2\}$, the posterior probability $P(k|OT_i, \psi)$ as the likelihood that the on-time of the $i^{th}$ vehicle follows the $k^{th}$ Gaussian distribution, then given Bayes' theorem, the posterior probability $P(k|OT_i, \psi)$ is:

$$P(k|OT_i, \psi) = \frac{P(k)f(OT_i,\mu_k,\sigma_k^2)}{\sum_{k=1}^{K} \omega_k f(OT_i,\mu_k,\sigma_k^2)} = \frac{\omega_k f(OT_i,\mu_k,\sigma_k^2)}{P(OT_i)} \tag{9}$$

Let $\hat{k}$ is the maximum-a-posteriori (MAP) estimation that maximizes $P(k|OT_i, \psi)$ (Stauffer and Grimson, 1999; Power and Schoonees, 2002), then

$$\hat{k} = \max_k P(k|OT_i, \psi) \Rightarrow \hat{k} = \max_k \omega_k f(OT_i, \mu_k, \sigma_k^2) \tag{10}$$

where the second equation holds true because $P(OT_i)$ in (9) is independent of $k$.

This GMM method uses multiple normal distributions to fit the vehicle on-time and weights each distribution based upon its influence in fitting the data set. The weight of different distributions can be estimated from the Expectation Maximization (EM) algorithm (Divo, 2008).

Figure 4 shows an example using the GMM to fix the on-time data collected on Lane 4, loop station 005es15652 (located at the milepost of 156.52) Southbound Interstate 5 from September 24th until September 29th, 2009. Figure 4.A shows the comparisons between the combination distribution of three distributions and the observed data, whereas the Figure 4.B is the comparison between three distributions and the observed data. Table 1 shows the Gaussian distribution parameters. Similarly, on-time distribution from other locations also can be described as the combination of several distributions (Wang et al., 2009).

### 2.3.2 Speed and vehicle length estimation based on an iterative procedure

Based on the dual-loop data, the vehicle speed can be calculated by the loop distance ($L$) between M loop and S loops divided by the arrival time difference between two loops. Typically, it is done at the controller or loop detector hardware level. The measured speed is defined as:

$$v_{di} = \frac{L}{\Delta t_{ai}} \tag{11}$$

where $v_{di}$ is the vehicle measured speed from dual-loop for the $i$ vehicle, $\Delta t_{ai}$ is the arrival time difference between M loop and S loop. For the single loop detectors, with an assumed average vehicle length for short vehicles, the average speed for a particular time period can be calculated based on the GMM. With a known average speed, the approximate vehicle lengths can also be calculated. Based on the equation (3), the average speed $v_a$ can be estimated as:

$$v_a = \frac{L_V + L_L}{mean(OT)} \tag{12}$$

where $mean(OT)$ is the mean of the on-time. Based on the Wang and Nihan's research (2000), the average length for short vehicles is 15.3ft. The mean on-time of short vehicle can be found using the GMM model. If the mean on time of short vehicles is defined as $\mu_1$, and the average short vehicle length as $L_{V1}$, then the average speed for short vehicles can be defined as

$$v_{a1} = \frac{L_{V1} + L_L}{\mu_1} \tag{13}$$

Assuming that the average s average speed in the traffic flow is similar to that of short vehicles, then the vehicle length for each vehicle can be calculated as

$$L_{Vi} = v_{a1} \cdot OT_i - L_L \tag{14}$$

where $L_{Vi}$ is the vehicle length for the $i^{th}$ vehicle.

After getting the approximate vehicle length, the short vehicles in the traffic flow can be roughly identified. Assume that $N_2$ adjacent vehicles have a consistent travel speed in the traffic flow. Calculate the mean on-time of the short vehicles in every $N_2$ adjacent vehicles. Then, equations (13-14) are computed again to get the average speed and vehicle length. The whole iteration process is shown in Figure 5. The iteration will continue until the difference between two iterations is small enough (0.01feet was used for the vehicle length difference in this study).

## 3. Study sites

Two locations on Mercer Island, Washington are chosen as the studies sites. The loop data of these two locations were collected from the loop stations 090es00720 (site A) and 090es00822 (site B), which are located on Interstate 90 (I-90) at mileposts (MP) 7.20 and 8.22, respectively. The particular data used for the on-time model was event data collected using the Advanced Loop Event Data Analyzer (ALEDA) system developed by the University of Washington STARLab (Cheevarunothai et al., 2005). At site A, the data were collected for 24 hours on Feb. 24th (Tuesday), 2009. Two lanes, including one outside lane (West lane 2) and one inside lane (West lane 3), were chosen for speed and length estimation. Similarly, at site B, the data were collected Mar. 3rd (Tuesday), 2009. Three lanes, including two outside lanes (West lane 1 and West lane 2) and one inside lane (West lane 3), were chosen for this study. The locations of these stations are shown in the red circles on Figure 6 (Ma et al., 2011). The data from M loop were used for single-loop speed and vehicle length estimation whereas the data from the corresponding dual-loop (M loop and S loop) were used for evaluation and validation.

## 4. Estimation results and performance evaluation

For comparison purposes, a conventional constant g-factor method, the sequence method, and the moving median, were also used for estimating vehicle speed and classification data. The estimation results were compared with the “ground-truth” data measured by dual-loop detectors. The speed and vehicle length measured by the dual-loop detectors may have some errors compared with the true values. However, the dual-loop measurement is the best estimation can be obtained in the test locations. Testing the GMM method with other ground-truth data is a further study direction. The average absolute error (AAE) is used for evaluating the accuracy of each method. The moving sample size for the sequence and moving median methods is 33 vehicles as recommended by Coifman and Kim (2009) while the moving sample size used for the GMM is $N_1$=100 vehicles. The moving sample size for estimating the mean on-time of short vehicles traveling at a consistent speed is $N_2$=10. Note that $N_2$ can be a little different from location to location. $N_2$ needs to be large enough to ensure there are at least 4 short vehicles within $N_2$ adjacent vehicles.

### 4.1 Speed and vehicle length estimation results

Figures 7-11 show the estimation results of four methods for the five data sets at the two locations on I-90. In each figure, the horizontal axis is the hour of day whereas vertical axis is the AAE of estimated speed in figure A or AAE of estimated length in figure B.

Figures 7 and 8 illustrate speed and vehicle length estimation for westbound traffic on Lane 2 and Lane 3 at Site A, respectively. Based on Figures 7 and 8, we found

- The estimation of speed and vehicle length using GMM method works well on both lanes during the whole day. The average AAE in estimated speed is about 4 mph and the average AAE in estimated vehicle length is less than 2 feet.
- The estimation results from GMM are stable during the whole day whereas the conventional method and sequence method have some fluctuations in some time periods. For example, the sequence method did not perform well during the midnight when traffic volume is low.
- The conventional method performed poorly during the peak hour 8am in this site. This is because the assumed mean on-time is much different from the actual on-time at 6 pm in the peak hours.

Figures 9-11 show the vehicle speed and length estimation for westbound at site B. Lane 1 is shown in figure 9; Lane 2 is shown in figure 10; lane 3 is shown in figure 11. From figures 9-11, it can be found

- The estimation of speed and vehicle length using GMM method works well on these three lanes during the whole day. The average AAE in estimated speed is about 6 mph and the average AAE in estimated vehicle length is about 2 feet.
- The estimation results from GMM are stable during the whole day whereas the other three methods have more fluctuations in some time periods. For example, the sequence method did not do well between 3am and 4 am in lane 1 and moving median did not do well after 8pm on lane 1. The AAE for the GMM method is also smaller than these for the other three methods.
- The conventional method did poorly during the peak hour, 6pm, at this study site. This is because the assumed mean on-time is much different from the actual on-time in the peak hour at 6pm.

## 4.2 Length based classification estimation

The vehicle length thresholds used for separating vehicles into different bins varied in different research and practice. For example, to separate the short vehicles from long vehicles, the thresholds range from 22 feet to 28 feet (Wang and Nihan, 2000; Coifman and Kim, 2009; Coifman, 2009). The latter two studies further separated the long vehicles into two classes: middle vehicles and long vehicles. Based on the characteristics of the 13 FHWA vehicle classes described in USDOT Final Report on Length based vehicle classification (Coifman, 2009), thresholds 22 ft and 40 ft are adopted here for separating three classes of vehicles: class 1 (short vehicles), class 2 (middle vehicles), and class 3 (long vehicles).

Figure 12 shows an example for classifying vehicles into three classes for westbound Lane 2 at site A. A total of 21591 vehicles were categorized. The reported vehicle length from dual-loop data versus the estimated length from four estimated methods are also shown in Figure 12. Assuming that the reported length is the ground truth vehicle length, then one can find that

- The other three methods do a much better job than the conventional method. For example, some vehicles with a length of about 70 feet reported by the dual loop are estimated as over 150 feet in length by the conventional method.
- The GMM method performs best. For example, some vehicles have a length about 80 feet in the reported length are estimated as less than 60 feet in the sequence method. There is also a vehicle which is about 70 feet in the reported length is estimated as over 100 feet in the moving median method. The GMM method has less underestimate and overestimate issues than any of the other three methods.

Table 2 shows the statistics results of vehicle classification for the four methods versus ground-truth data from dual-loop detectors. Five cases, two lanes at site A and three lanes at site B are analyzed in table 2. The sample size for these five lanes is all about 20,000. The vehicle number in each cell represents the number of vehicles classified into the corresponding class, for example 671 in the second cell of first row represents there are 671 vehicles classified as class 1 in the reported length from the dual-loop and also classified as class 2 in the estimated length from the conventional method. The percentage (P) of vehicles correctly classified in each class and the overall correctly classified rate (Rate) for all vehicles are also shown in Table 2. From Table 2, one can find that

- The GMM method performs best in all five cases based on the overall correctly classified rate, except in the last case the GMM method and sequence method have the same correctly classified rate 98.0%. This rate for the GMM method ranges from 97.6% to 99.6%.
- The overall correctly classified rate for the conventional method is not too low if the assumed speed used to calculate the g factor is similar to the travel speed during the data collection period. However, when congestion occurs and the travel speed drops, some short vehicles may be classified as long vehicles. For example, 331 short vehicles are identifies as long vehicles for lane 1 at site B.
- The correctly classified percentage for class 2 vehicles is a little sensitive since there are a relatively large number of vehicles with a length near the 22 feet threshold used to separate class 1 and class 2. For example, for the lane 2 at site B, there are 462 short vehicles are identified as middle vehicles in the sequence method, 1212 short vehicles are identified as middle vehicles in the moving median method, and there 210 middle vehicles are identified as short vehicles in the GMM method.
- Consider the percentage of vehicles correctly classified in each class, the GMM method also performed well in all three classes. Compared with the GMM method, the moving median method underperformed for the short vehicle classification, but it did better in the medium and long vehicle classification in some cases.

The proposed GMM method aims to implicitly extract and classify vehicles into appropriate categories based on their overall data patterns. The results shown in Table 2 reflect its robust and superior capability of classifying the majority of vehicles into an appropriate group. The unfavorable performance may be achieved by the proposed method for a particular group with a relatively small amount of vehicles to ensure its optimal overall estimation results. For example, in Case 1, the proposed method misclassifies 62 Class II-type vehicles into the other categories (52 in Class I and 10 in Class III) which are larger than 48 vehicles misclassified by the moving median method. However, its model specifications only result in 79 Class I-type vehicles incorrectly categorized into Class II, which is much smaller than 234 vehicles misclassified by the moving median method. The proposed method results in a 99.3% overall correctly classified rate, which is higher than the estimation rates from the other methods.

One should note that the proposed method initializes the estimation process by identifying short vehicle within "$N_1$" assuming short vehicle is dominant vehicle type on roadway. It is possible to have more long vehicles under certain circumstances. The methodology developed in this study is transferable to address these traffic situations although extra efforts are needed to calibrate the model by changing the parameters, $N_1$, $N_2$, and the weighting factor, $\omega_k$, the mean on-time of vehicle category $k$, $\mu_k$, *and* its variance, $\sigma_k^2$. When long vehicles become dominant, the corresponding values of these parameters, including the weight factor, need to be adjusted based on the data. The well-calibrated model can yield fairly good performance under such situations. In general, the proposed mode specification provides a robust framework to handle various conditions.

## 4.3 ANOVA test

To further statistically quantify the performance among different methods in this study, Analysis of Variance (ANOVA) test is conducted to verify whether the mean and variance of estimation errors from the various methods are significantly different. The null hypothesis $H_0$ here is that there is no difference among the mean of estimation errors from different methods. After the one-way ANOVA test, a post-hoc test is also needed to test which of these methods are significantly different from others (Gardener, 2012).

Table 3 shows post-hoc test results for the vehicle length estimation on Site B. On Lane 1, all four methods are significantly different from each other at the significant level of $p=0.05$. The GMM method performs the best, following by the sequence method, the moving median method, and the conventional method. On Lane 2, there is no significant difference between the sequence method and the moving median method. The GMM method significantly outperforms the other three methods. On Lane 3, there is no significant difference between the sequence method and the moving median method, and between the GMM method and the sequence method. The GMM method is significantly better than moving median method and the conventional method. Similarly, the test can be conducted for traffic data collected in Site A. The test results for Site A are similar with the results for Lane 1 in Site B. The ANOVA test results show that the estimation error from the GMM method is significantly smaller than the other three methods.

## 5. Conclusions

This research proposes a GMM-based approach to model vehicle on-times and iteratively estimate traffic speed and vehicle classification data. After the GMM is established to empirically model vehicle on-times measured by single-loop detectors, the optimal solution can be initially sought to separate length-based vehicle volume data. Traffic speed is estimated based on classified short vehicle volumes, which are then used to further correct vehicle length classification data for the next iteration of estimation. Based on updated classified vehicle volume estimations, traffic speed can be refined. The entire estimation procedure will iterate until the speed and classified vehicle volume estimation becomes statistically stable and convergent. The effectiveness of the proposed approach was examined by the data collected from several loop stations on Interstate-5 in the greater Seattle areas. The estimation results are compared with ground-truth data from dual-loop detector. The AAE and the overall correct classification rate are used to evaluate the performance of the four methods: the conventional constant g-factor method, the sequence method, the moving median method and the GMM mehod. The ANOVA test indicates that the AAE of the GMM method is significant lower than that of the other three methods in most cases. The total correct classification rate from the GMM method is all above 97.6% in all five lanes, which is also better than the other three methods. The test results show the proposed GMM approach outperforms the previous models, including conventional constant g-factor methods, sequence method, and moving median method, on traffic speed estimation and vehicle classification under various traffic conditions. Testing the GMM method performance with other ground-truth data, including the data manually collected or by using the in-car equipment, Bluetooth (Bachmann et al., 2012) and video sensors, is recommended for further studies. The values of parameters $N_1$ and $N_2$ in the proposal method may be different under different conditions. The sensitivity of the parameters $N_1$ and $N_2$ will be further studied in the future research.

## Acknowledgements

This research is partly funded by National Natural Science Foundation of China (NSFC) (Project No. 51028802). Authors would like to acknowledge NSFC for their kind support. Appreciation also goes to the Washington State Department of Transportation (WSDOT) for providing data for this study.

## Reference

(1) Athol, P. (1965). Interdependence of certain operational characteristics within a moving traffic stream. In Highway Research Record 72, HRB, National Research Council, Washington, D.C., 58–87.

(2) Bachmann, C., Roorda, M. J., Abdulhai, B. & Moshiri B. (2012). Fusing a Bluetooth traffic monitoring system with loop detector data for improved freeway traffic speed estimation. Journal of Intelligent Transportation Systems. Available online: 31 May 2012. DOI:10.1080/15472450.2012.696449.

(3) Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B, 57, 289–300.

(4) Cheevarunothai, P., Wang, Y., and Nihan, N. L. (2005). “Development of Advanced Loop Event Data Analyzer (ALEDA) for Investigations of Dual-Loop Detector Malfunctions.” Proceedings of the 12th World Congress of Intelligent Transportation Systems, paper 000662, November 2005.

(5) Coifman, B. (2001). Improved velocity estimation using single loop detectors. Transportation Research Part A: Policy and Practice 35(10), 863-880.

(6) Coifman, B., Dhoorjaty, S., & Lee, Z. (2003). Estimating median velocity instead of mean velocity at single loop detectors. Transportation Research Part C: Emerging Technologies 11(3-4), 211-222.

(7) Coifman, B., & Kim, S. (2009). Speed estimation and length based vehicle classification from freeway single-loop detectors. Transportation Research Part C: Emerging Technologies 17(4), 349-364.

(8) Coifman, B. (2009). Length based vehicle classification on freeways from single loop detectors. USDOT Region V Regional University Transportation Center Final Report.

(9) Corey, J., Lao, Y., Wu, Y., & Wang. Y. (2011). Detection and Correction of Inductive Loop Detector Sensitivity Errors Using Gaussian Mixture Models. Transportation Research Record: Journal of the Transportation Board. Volume 2256, 120-129.

(10) Dailey, D.J. (1999). A statistical algorithm for estimating speed from single loop volume and occupancy measurements. Transportation Research Part B: Methodological 33(5), 313-322. DOI: DOI: 10.1016/S0191-2615(98)00037-X.

(11) Divo, I. D. (2008). Expectation maximization and Mixture Modeling Tutorial. UC Los Angeles: Statistics Online Computational Resource. Retrieved from: http://escholarship.org/uc/item/1rb70972

(12) Gardener M. (2012). Using R for statistical analyses – ANOVA. Available at http://www.gardenersown.co.uk/Education/Lectures/R/anova.htm#anova_1

(13) Hall, F. L., & Persaud. B. N. (1989). Evaluation of Speed Estimates Made with Single-Detector Data from Freeway Traffic Management Systems. In Transportation Research Record 1232, TRB, National Research Council, Washington, D.C., 9–16.

(14) Hasan, B. A. S. and Gan. J. Q. (2009). Sequential EM for Unsupervised Adaptive Gaussian Mixture Model Based Classifier. Machine Learning and Data Mining in Pattern Recognition: 6th International Conference. Leipzig, Germany. 96-106

(15) Hellinga, B.R. (2002). Improving Freeway Speed Estimates from Single-Loop Detectors. Journal of Transportation Engineering 128(1), 58-67.

(16) JIN, S., WANG, D., & QI, H., (2010). Bayesian Network Method of Speed Estimation from Single-Loop Outputs. Journal of Transportation Systems Engineering and Information Technology 10(1), 54-58.

(17) Ju, Z., Liu, H., Zhu, X., & Xiong. Y., (2008). Dynamic Grasp Recognition Using Real Time Clustering, Gaussian Mixture Models and Hidden Markov Models. Advanced Robotics 23(10), 1359-1371.

(18) Li, B. (2009). On the recursive estimation of vehicular speed using data from a single inductance loop detector: A Bayesian approach. Transportation Research Part B: Methodological 43(4), 391-402.

(19) Ma, X., Wu Y., & Wang. Y., (2010). DRIVE Net: E-Science Transportation Platform for Data Sharing, Visualization, Modeling, and Analysis. Transportation Research Record: Journal of the Transportation Board. Volume 2215, 37-49

(20) May, A. D., Coifman, B., Cayford R., & Merritt. G. (2004). Automatic Diagnostics of Loop Detectors and the Data Collection System in the Berkeley Highway Lab, California PATH Research Report, UCB-ITS-PRR-2004-13.

(21) McLachlan, G.J. and Basford. K.E. (1988). Mixture Models: Inference and Applications to Clustering, Marcel Dekker, New York.

(22) McLachlan, G.J. and Peel, D. (2000). Finite Mixture Models, New York: Wiley. ISBN 0471006262

(23) Mikhalkin, B., Payne, H. J., & Isaksen. L. (1972). Estimation of Speed from Presence Detectors. In Highway Research Record 388, HRB, National Research Council, Washington, D.C., 73–83.

(24) NCSA (National Center for Statistics and Analysis). (2005). Traffic Safety Facts 2003. National Highway Traffic Safety Administration, U.S. Department of Transportation. Washington, D.C.

(25) Power, P.W., and Schoonees, J.A. (2002). Understanding background mixture models for foreground segmentation, in: Proceedings of the Image and Vision Computing New Zealand, 267–271.

(26) Pushkar, A., Hall, F. L., &Acha-Daza, J. A. (1994). Estimation of Speeds from Single-Loop Freeway Flow and Occupancy Data Using Cusp Catastrophe Theory Model. In Transportation Research Record 1457, TRB, National Research Council, Washington, D.C., 149–157.

(27) Reynolds, D. A., and Rose, R. C. (1995). Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Transactions On Speech and Audio Processing 3(1), 72-83

(28) Stauffer,C., and Grimson, W. (1999). Adaptive background mixture models for real-time tracking, in: Proceedings of IEEE conference on CVPR, 246–252.

(29) Sun, C., & Ritchie, S. G., (1999). Individual vehicle speed estimation using single loop inductive waveforms. Journal of Transportation Engineering, Vol. 125, No. 6, 531 - 538.

(30) Wang, Y., & Nihan, N. (2000). Freeway Traffic Speed Estimation with Single Loop Outputs. Transportation Research Record: Journal of the Transportation Board. Volume 1727, 120–126.

(31) Wang, Y., & Nihan, N. L. (2003). Can Single-Loop Detectors Do the Work of Dual-Loop Detectors? ASCE Journal of Transportation Engineering, 129(2), 169 - 176.

(32) Wang, Y., and Nihan, N. L. (2004). Dynamic estimation of freeway large-truck volumes based on single loop measurements, Journal of Intelligent Transportation Systems 8 (3), 133–141.

(33) Wang, Y., Corey, J., Lao, Y., & Wu. Y., (2009). Development of a Statewide Online System for Traffic Data Quality Control and Sharing. WSDOT report, TNW2009-12.

(34) Ye, R., Zhang, Y.L., & Middleton, D.R., (2006). Unscented Kalman filter method for speed estimation using single loop detector data, Transportation Research Record: Journal of the Transportation Board. Volume 1968, 117–125.

(35) Zhang, G., Wang, Y., & Wei, H., (2006). An Artificial Neural Network Algorithm for Length-Based Vehicle Classification Using Single-Loop Data. Transportation Research Record: Journal of the Transportation Research Board. Volume 1945, 100-108.

# Tables

**Table 1 M Loop Detector Gaussian Distributions from 005es15652 Lane 4 South**

| $k$=1,2,3 | Category 1 | Category 2 | Category 3 |
|---|---|---|---|
| $\omega_k$ | 0.925 | 0.052 | 0.023 |
| $\mu_k$ | 200.62 | 301.72 | 465.57 |
| $\sigma_k^2$ | 328.68 | 1054.94 | 88656.44 |

**Table 2 statistics results of vehicle classification for the four estimated methods versus reported results from dual-loop**

| | | Estimated Methods | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Conventional Method | | | | Sequence Method | | | | Moving Median Method | | | | GMM Method | | | |
| | | $C_1$ | $C_2$ | $C_3$ | P(%) | $C_1$ | $C_2$ | $C_3$ | P(%) | $C_1$ | $C_2$ | $C_3$ | P(%) | $C_1$ | $C_2$ | $C_3$ | P(%) |
| Case 1: West bound lane 2 at cabinet 090es00720, total sample size: 21591 | | | | | | | | | | | | | | | | | |
| Reported Length | $C_1$ | 19677 | 671 | 11 | 96.7 | 20164 | 195 | 0 | 99.0 | 20125 | 234 | 0 | 98.9 | 20280 | 79 | 0 | 99.6 |
| | $C_2$ | 97 | 375 | 15 | 77.0 | 90 | 386 | 11 | 79.3 | 27 | 439 | 21 | 90.1 | 52 | 425 | 10 | 87.3 |
| | $C_3$ | 0 | 11 | 734 | 98.5 | 0 | 16 | 729 | 97.9 | 0 | 2 | 743 | 99.7 | 0 | 2 | 743 | 99.7 |
| Rate (%) | | 96.3 | | | | 98.6 | | | | 98.7 | | | | 99.3 | | | |
| Case 2: West bound lane 3 at cabinet 090es00720, total sample size: 21702 | | | | | | | | | | | | | | | | | |
| Reported Length | $C_1$ | 20563 | 629 | 197 | 96.1 | 21333 | 56 | 0 | 99.7 | 21302 | 87 | 0 | 99.6 | 21348 | 41 | 0 | 99.8 |
| | $C_2$ | 72 | 91 | 5 | 54.2 | 37 | 129 | 2 | 76.8 | 18 | 146 | 4 | 86.9 | 36 | 127 | 5 | 75.6 |
| | $C_3$ | 0 | 17 | 128 | 88.3 | 0 | 9 | 136 | 93.8 | 0 | 5 | 140 | 96.6 | 0 | 6 | 139 | 95.9 |
| Rate (%) | | 95.8 | | | | 99.5 | | | | 99.5 | | | | 99.6 | | | |
| Case 3: West bound lane 1 at cabinet 090es00822, total sample size: 19978 | | | | | | | | | | | | | | | | | |
| Reported Length | $C_1$ | 18457 | 308 | 331 | 96.7 | 18802 | 294 | 0 | 98.5 | 18020 | 1075 | 1 | 94.4 | 18823 | 272 | 1 | 98.6 |
| | $C_2$ | 197 | 413 | 50 | 62.6 | 185 | 459 | 16 | 69.5 | 132 | 452 | 76 | 68.5 | 163 | 456 | 41 | 69.1 |
| | $C_3$ | 0 | 40 | 182 | 82.0 | 0 | 29 | 193 | 86.9 | 0 | 4 | 218 | 98.2 | 0 | 12 | 210 | 94.6 |
| Rate (%) | | 95.4 | | | | 97.4 | | | | 93.6 | | | | 97.6 | | | |
| Case 4: West bound lane 2 at cabinet 090es00822, total sample size: 20950 | | | | | | | | | | | | | | | | | |
| Reported Length | $C_1$ | 19257 | 216 | 301 | 97.4 | 19312 | 462 | 0 | 97.7 | 18562 | 1212 | 0 | 93.9 | 19527 | 247 | 0 | 98.8 |
| | $C_2$ | 161 | 281 | 71 | 54.8 | 152 | 335 | 26 | 65.3 | 92 | 371 | 50 | 72.3 | 138 | 338 | 37 | 65.9 |
| | $C_3$ | 0 | 18 | 645 | 97.3 | 0 | 22 | 641 | 96.7 | 0 | 5 | 658 | 99.2 | 0 | 10 | 653 | 98.5 |
| Rate (%) | | 96.3 | | | | 96.8 | | | | 93.5 | | | | 97.9 | | | |
| Case 5: West bound lane 3 at cabinet 090es00822, total sample size: 19932 | | | | | | | | | | | | | | | | | |
| Reported Length | $C_1$ | 18534 | 383 | 537 | 95.3 | 19320 | 134 | 0 | 99.3 | 19263 | 191 | 0 | 99.0 | 19323 | 131 | 0 | 99.3 |
| | $C_2$ | 194 | 66 | 122 | 17.3 | 251 | 125 | 6 | 32.7 | 233 | 135 | 14 | 35.3 | 259 | 113 | 10 | 29.6 |
| | $C_3$ | 0 | 6 | 90 | 93.8 | 0 | 6 | 90 | 93.8 | 0 | 3 | 93 | 96.9 | 0 | 5 | 91 | 94.7 |
| Rate (%) | | 93.8 | | | | 98.0 | | | | 97.8 | | | | 98.0 | | | |

$C_1$ Represents class 1 vehicles with length: $L_V$ <22feet; $C_2$ Represents class 2 vehicles with length: 22feet $\leq L_V$ <40feet; $C_3$ Represents class 3 vehicles with length: $L_V \geq$40feet; P represents the percentage of vehicles correctly classified in each class; Rate is the overall correctly classified rate in the three classes. The vehicle number in each cell represents the number of vehicle been classified in corresponding class, for example 19677 in the first cell represents there 19677 vehicles are classified as class 1 in the reported length from the dual-loop and also classified as class 1 in the estimated length from the conventional method.

**Table 3 Post-hoc test for the estimation error of vehicle length on Westbound Site B**

| | West bound lane 1 site B | | | | West bound lane 2 site B | | | | West bound lane 3 site B | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | diff | lower | upper | p.adj | diff | lower | upper | p.adj | diff | lower | upper | p.adj |
| G-C | -0.80 | -0.87 | -0.72 | 0.00 | -0.90 | -0.99 | -0.82 | 0.00 | -1.28 | -1.46 | -1.09 | 0.00 |
| M-C | -0.20 | -0.28 | -0.13 | 0.00 | -0.39 | -0.47 | -0.30 | 0.00 | -1.02 | -1.21 | -0.84 | 0.00 |
| S-C | -0.68 | -0.76 | -0.60 | 0.00 | -0.37 | -0.45 | -0.28 | 0.00 | -1.17 | -1.38 | -0.97 | 0.00 |
| M-G | 0.59 | 0.52 | 0.67 | 0.00 | 0.52 | 0.44 | 0.60 | 0.00 | 0.25 | 0.07 | 0.44 | 0.00 |
| S-G | 0.12 | 0.04 | 0.20 | 0.00 | 0.54 | 0.45 | 0.62 | 0.00 | 0.10 | -0.10 | 0.31 | **0.56** |
| S-M | -0.47 | -0.55 | -0.39 | 0.00 | 0.02 | -0.07 | 0.10 | **0.95** | -0.15 | -0.35 | 0.05 | **0.24** |

Note: G: GMM Method, C: Conventional method, M: Moving Median Method, S: Sequence Method, diff: difference lower: lower range, upper: upper range, and p.adj: adjust p-value (Benjamini and Hochberg, 1995)

# Figures

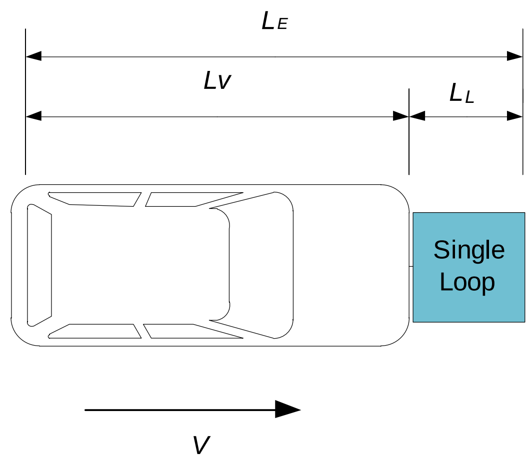
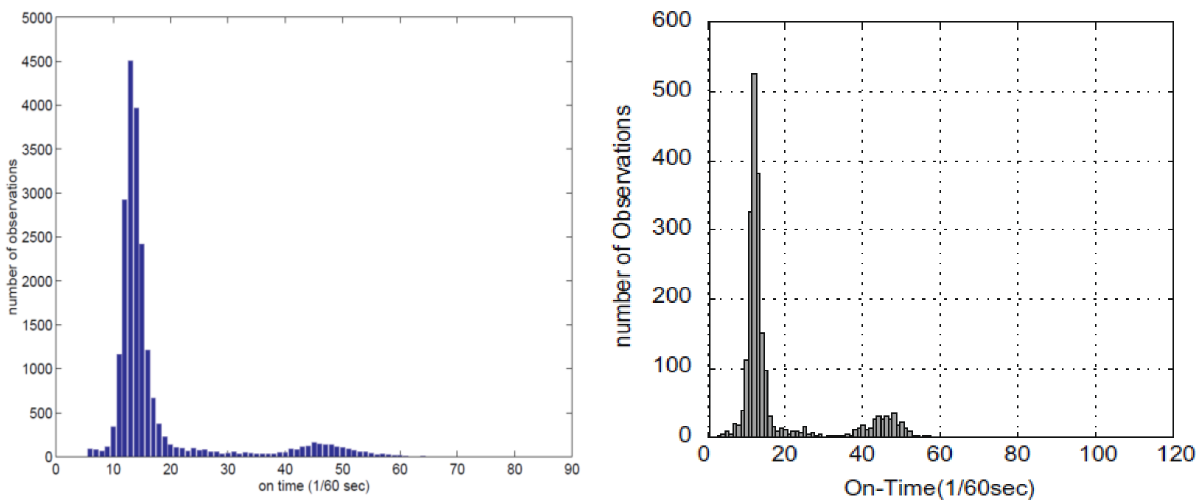



**Figure 1 Effective vehicle Length Diagram**




A (May et al., 2004) B (Coifman and Kim, 2009)

**Figure 2 Free Flow Dual-loop Detector On-Time (OT) Distribution**

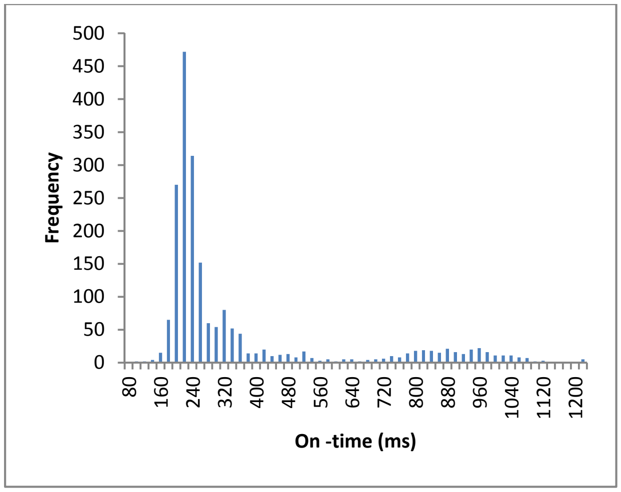
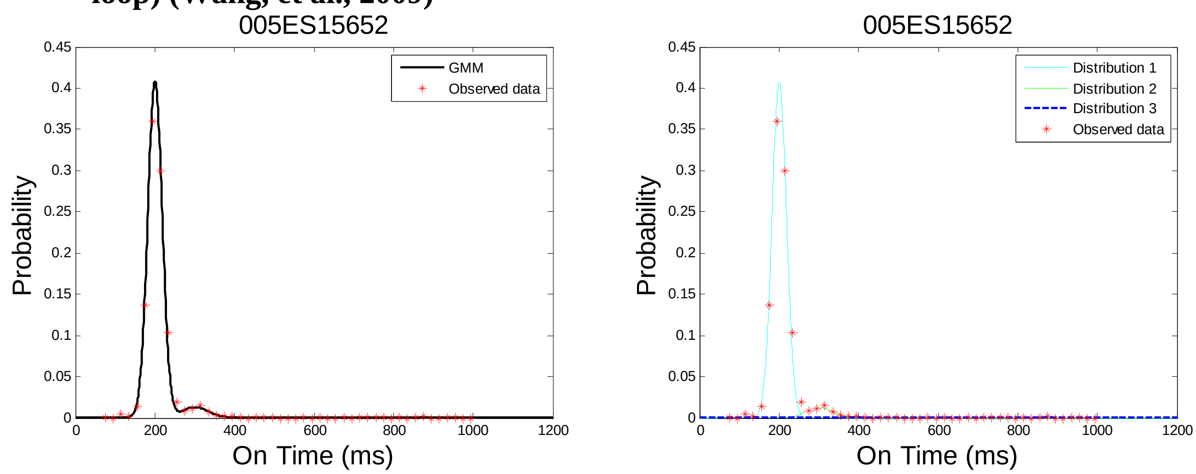



**Figure 3 On-time Distributions for the Loop Detectors 005ES16302 (North lane 1, M loop) (Wang, et al., 2009)**




A. GMM combination VS observed data B. Three distributions VS observed data

**Figure 4 GMM model fits with the On-time observed data on 005es15652 South Lane 4 M Loop**

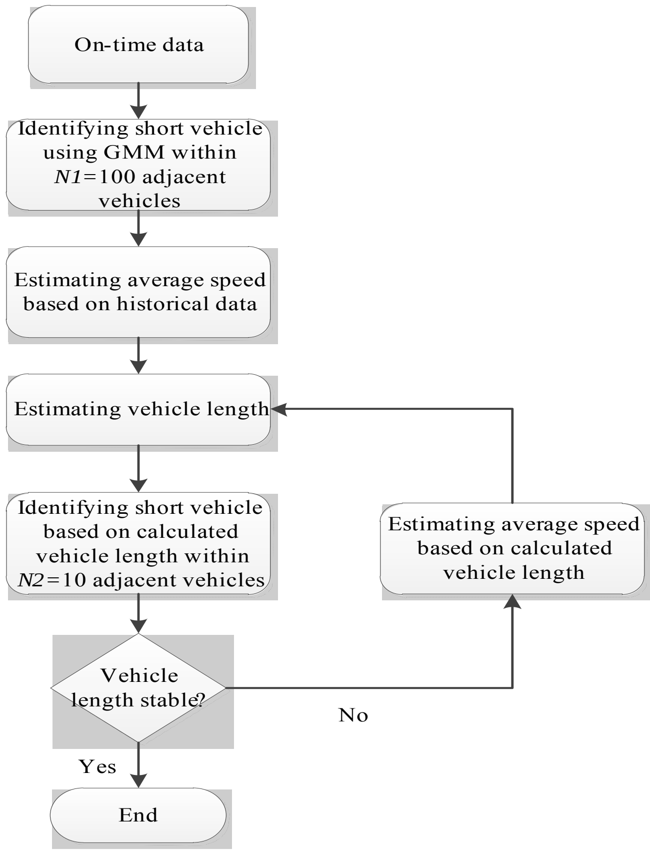



**Figure 5. Speed and vehicle length estimation process based on GMM**

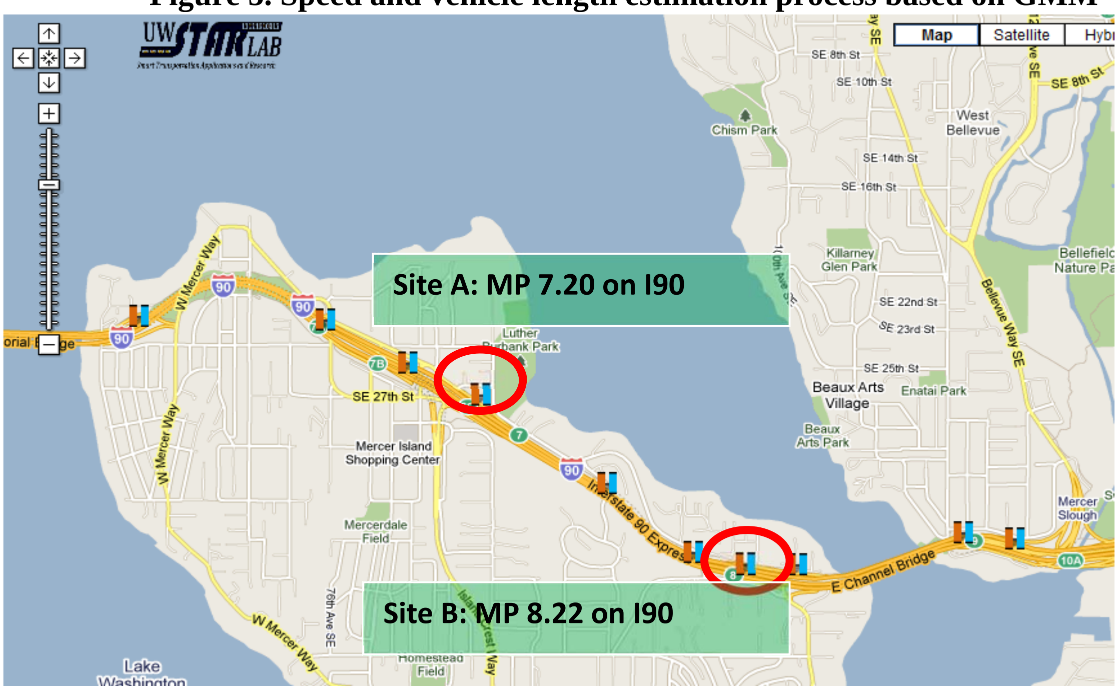
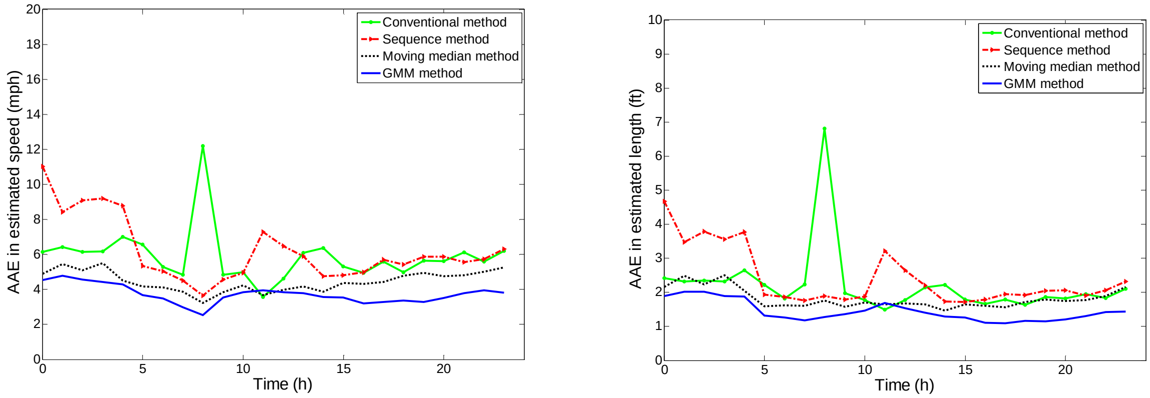
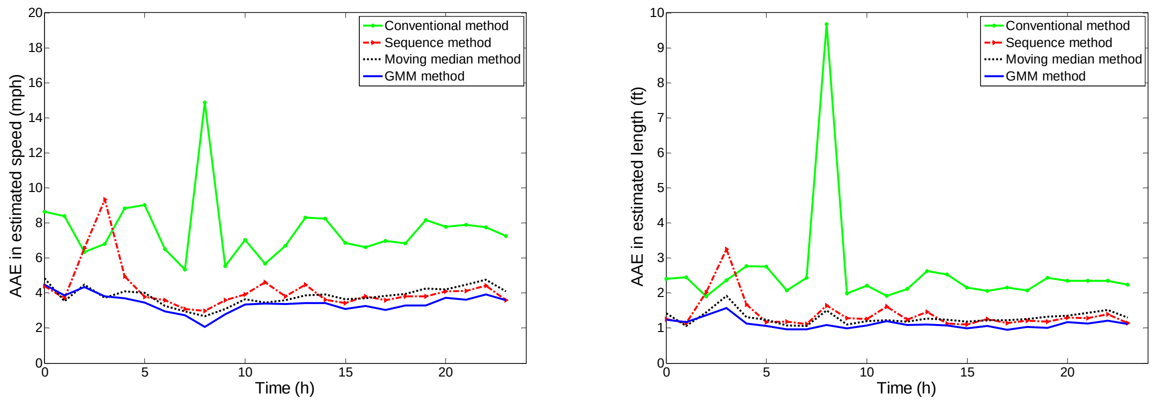



**Figure 6. Locations of Study sites**



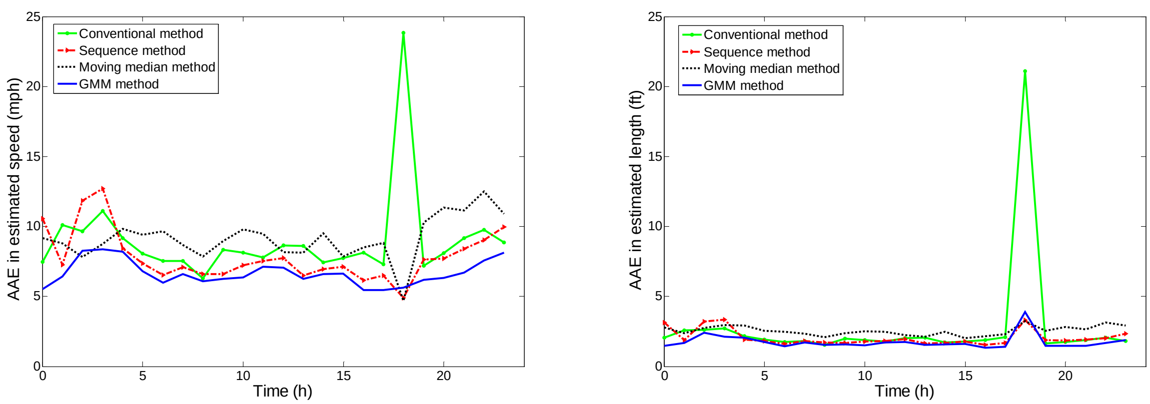

A. Speed estimation B. Vehicle length estimation

**Figure 7. Speed and vehicle length estimation for West bound lane 2 at site A**



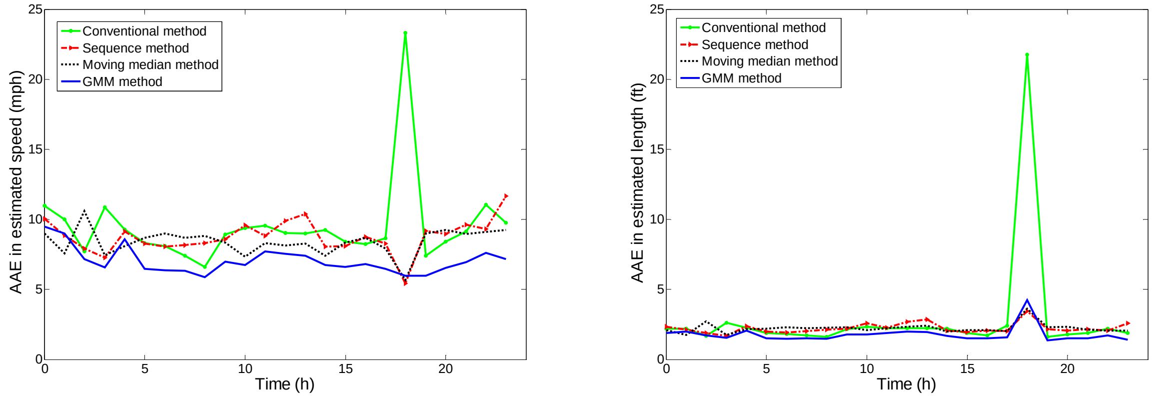

A. Speed estimation B. Vehicle length estimation

**Figure 8. Speed and vehicle length estimation for West bound lane 3 at site A**



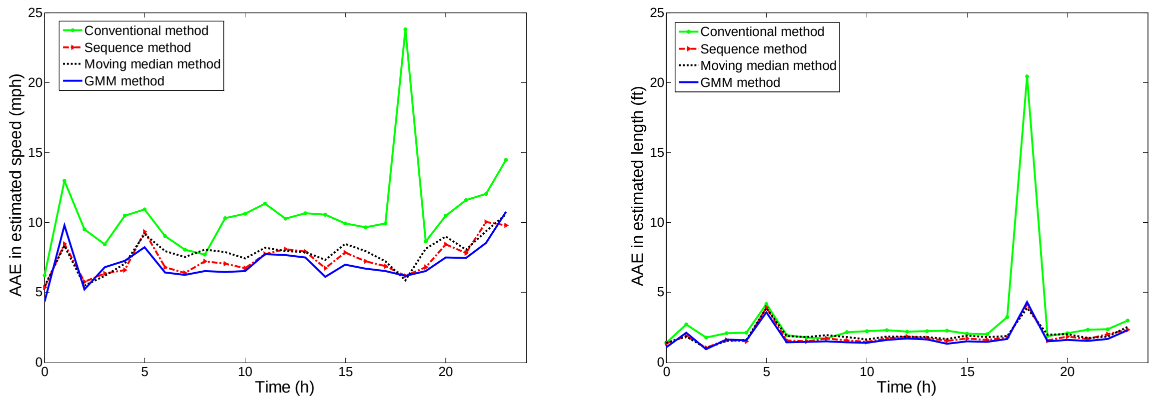

A. Speed estimation B. Vehicle length estimation

**Figure 9. Speed and vehicle length estimation for West bound lane 1 at site B**



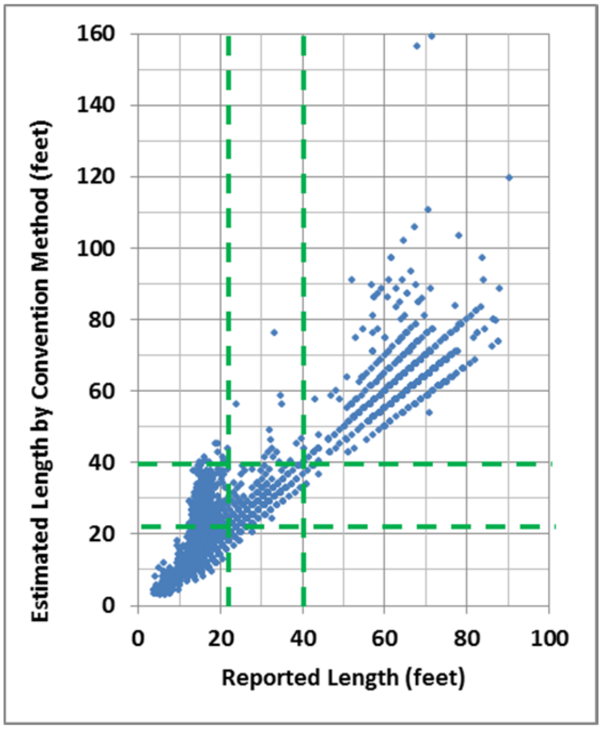

A. Speed estimation B. Vehicle length estimation

**Figure 10. Speed and vehicle length estimation for West bound lane 2 at site B**

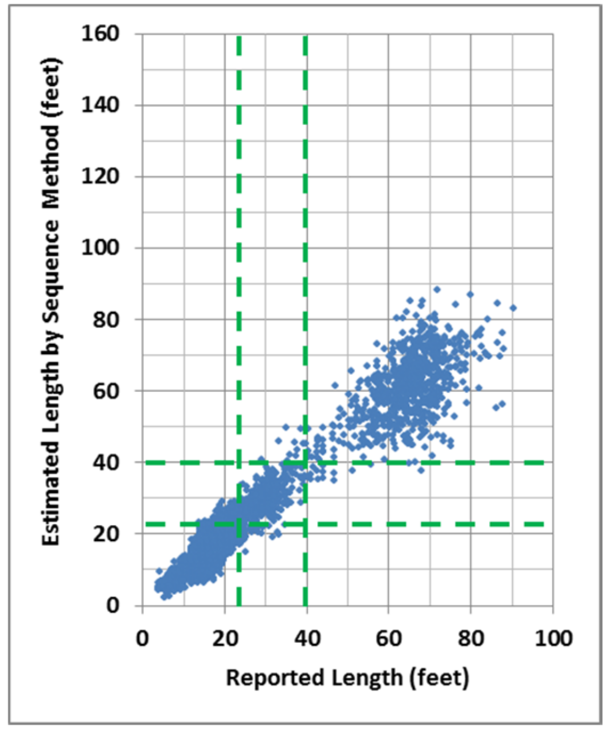



A. Speed estimation B. Vehicle length estimation

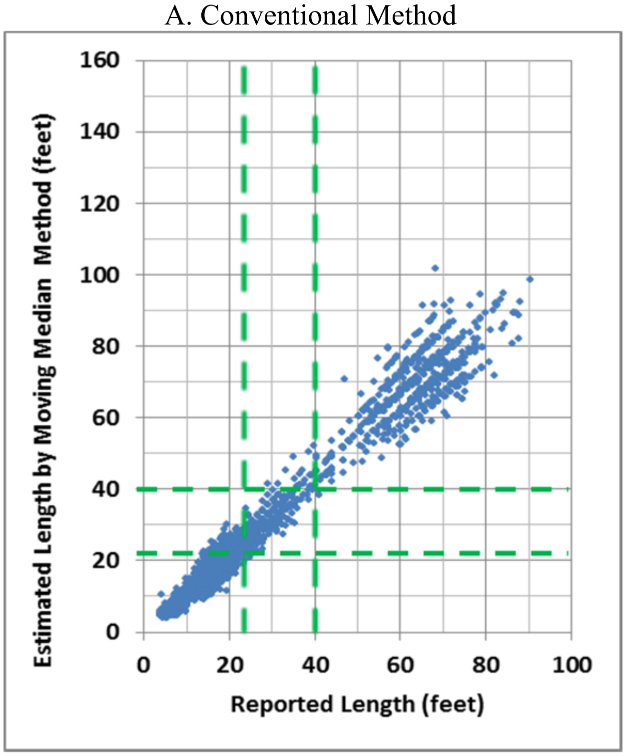
**Figure 11. Speed and vehicle length estimation for West bound lane 3 site B**



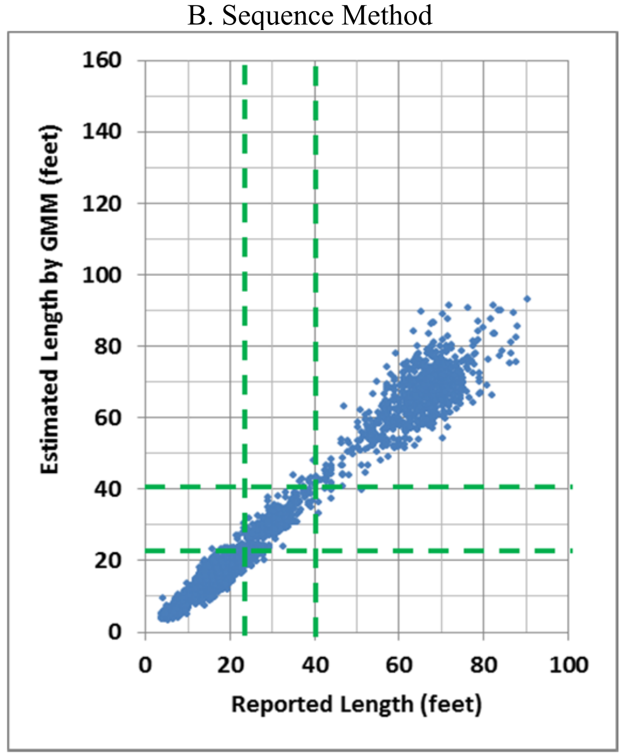

A. Conventional Method




B. Sequence Method




C. Moving Median Method




D. GMM Method

**Figure 12. Estimated length versus reported length from dual-loop within three classes for west bound lane 2 at site A**