Connected vehicle penetration rate for estimation of arterial measures of effectiveness

Juan Argote-Cabañero a,⁎, Eleni Christofa b, Alexander Skabardonis a

a Institute of Transportation Studies, University of California, Berkeley, McLaughlin Hall 416E, Berkeley, CA 94720, United States
b Department of Civil and Environmental Engineering, University of Massachusetts, Amherst, 130 Natural Resources Road, Amherst, MA 01003, United States

ARTICLE INFO

Article history:
Received 11 March 2015
Received in revised form 21 August 2015
Accepted 24 August 2015

Keywords:
Connected vehicles
Measures of effectiveness
Minimum penetration rate
Signalized arterials
Traffic estimation

ABSTRACT

The Connected Vehicle (CV) technology is a mobile platform that enables a new dimension of data exchange among vehicles and between vehicles and infrastructure. This data source could improve the estimation of Measures of Effectiveness (MOEs) for traffic operations in real-time, allowing to perfectly monitor traffic states after being fully adopted. However, as with any novel technology, the CV adoption will be a gradual process. This research focuses on determining minimum CV technology penetration rates that would guarantee accurate MOE estimates on signalized arterials. First, we present estimation methods for various MOEs such as average speed, number of stops, acceleration noise, and delay, followed by an initial assessment of the penetration rates required to accurately estimate them in undersaturated and oversaturated conditions. Next, we propose a methodology to determine the minimum CV market penetration rates to guarantee accurate MOE estimates as a function of traffic conditions, signal settings, sampling duration, and the MOE variability. A correction factor is also provided to account for small vehicle populations where sampling is done without replacement. The methodology is tested in a simulated segment of the San Pablo Avenue arterial in Berkeley, CA. The outcomes show that the minimum penetration rate required can be estimated within 1% for most MOEs under a wide range of traffic conditions. The proposed methodology can be used to determine if MOE estimates obtained with a portion of CV equipped vehicles can yield accurate enough results. The methodology could also be used to develop and assess control strategies towards improved arterial traffic operations.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

The Connected Vehicle (CV) technology is a promising mobile data source that can provide real-time information necessary for evaluating traffic conditions on a network. Such real-time information, which includes position, speed, acceleration and direction of motion, can be used to design and evaluate signal control and other arterial management strategies to improve the efficiency of urban traffic networks. However, the ability to use such data depends on the penetration rate of equipped vehicles and the underlying traffic conditions. Those two factors determine the number of available data points to estimate a certain metric during a given period of time. It is therefore critical to develop a comprehensive methodology that considers the effects of the CV technology's penetration rate as well as traffic conditions on the accuracy of the estimation of various Measures of Effectiveness (MOEs).

⁎ Corresponding author.
E-mail addresses: juan.argote@berkeley.edu (J. Argote-Cabañero), christofa@ecs.umass.edu (E. Christofa), dromeas@berkeley.edu (A. Skabardonis).

http://dx.doi.org/10.1016/j.trc.2015.08.013
0968-090X/© 2015 Elsevier Ltd. All rights reserved.
Multiple studies have been conducted in estimating arterial MOEs using CV data (alone or in combination with other data sources). The majority of these studies have focused on travel time estimation (Srinivasan and Jovanis, 1996; Chen and Chien, 2000; Cetin et al., 2005; Vasudevan, 2008; Li et al., 2008). Other MOEs that have been estimated with the exclusive use of CV data are average speeds (Cheu et al., 2002; Vasudevan, 2008), midblock flows, and turning percentages (Vasudevan, 2008). A big part of the literature has focused on queue length estimation with the exclusive use of CV data (Comert and Cetin, 2009; Argote et al., 2011; Venkatanarayana et al., 2011; Hao and Ban, 2011; Cheng et al., 2011; Comert, 2013) or by fusing CV and detector data (Li et al., 2013). Queue length has been a MOE of interest because it is one that cannot be easily estimated with the use of stationary sensors.

Macrosopic characteristics of traffic such as average flow, average density, and space mean speed have also been estimated with information from a combination of probe and stationary detector data for freeways (Nanthawichit et al., 2003). More recently, CV data alone have been utilized to estimate network densities and the macroscopic fundamental diagram (MFD) of signalized networks (Gayah and Dixit, 2013; Nagle and Gayah, 2014).

The upcoming CV technology has also motivated studies on trajectory reconstruction (Sun and Ban, 2013) and traffic signal control systems. The signal control strategies have focused on queue spillback control (Christofa et al., 2013), minimization of vehicle delay and stops (Güler et al., 2014) or vehicle delay and queue length (Feng et al., 2015), and real-time signal control strategies for multimodal transportation systems (He et al., 2012, 2014; Hu et al., 2015).

A big part of the literature has centered their efforts on identifying penetration rates necessary for obtaining reliable MOE estimates. However, most of these efforts have focused on a single MOE at a time. Several examples that have focused on travel time estimation on urban arterials using probe or CV data have concluded that even low penetration rates can provide satisfactory travel time estimates for freeways and congested arterials (Srinivasan and Jovanis, 1996; Vasudevan, 2008). Cheu et al. (2002) concluded that a minimum of 4–5% penetration rate is required for estimating average link speeds within 5 km/h at least 95% of the time while Vasudevan (2008) found that the accuracy of the estimation varied for different MOEs and it improved with higher penetration rates of the CV technology. Cetin et al. (2005) studied the impact of probe penetration on the accuracy of travel time estimates but did not provide an analytical equation for direct estimation of minimum penetration rates while Comert (2013) investigated the impact of probe penetration and volume to capacity ratio on the accuracy of queue length estimates using numerical examples.

A thorough review of the literature revealed that only a few studies have attempted to present methodologies for directly determining the minimum number of vehicles necessary to accurately estimate a wide variety of MOEs. Chen and Chien (2000) proposed a heuristic for calculating the minimum number of probes needed to obtain accurate travel time estimates for freeways within a certain relative error, but no analytical method was developed for estimating travel time from probe data. In addition, the relationship between traffic flow and the minimum number of vehicle requirements was captured via simulation and it was not parametrically modeled. Cheu et al. (2002) utilized a modified version of the standard deviation formulation by Quiroga and Bullock (1998) for determining minimum probe vehicle sample requirements for arterial speed estimation. However, their method was iterative in nature and no closed form solution was presented. Moreover, these studies did not considered the fact that when vehicle sampling is performed without replacement1 and the population sizes are small, the standard error correction factor needs to be added to the equation used for estimating the minimum penetration rate to ensure accurate MOE estimation (Steel and Torrie, 1960).

Despite the plethora of recent research efforts in the area of MOE estimation that use CV or probe vehicle data, none of the studies have provided a comprehensive description of the estimation process for a wide variety of MOEs and traffic conditions. A recent effort was performed by the authors (Argote et al., 2011) but was restricted to only undersaturated traffic conditions. In addition, the penetration rate requirements that ensure accurate MOE estimates for real-time traffic control applications have not been identified in the literature.

In this paper we explicitly address these issues. We develop and test algorithms for estimating a variety of MOEs that can be obtained as the first moment of a certain parameter using CV trajectory data, and develop equations for estimating minimum penetration rate requirements for certain levels of MOE estimation accuracy as a function of traffic conditions and signal settings as well as the sampling interval duration and the variability of the MOE estimate. The proposed methodology also reveals the importance of including a correction factor that accounts for small population sizes when sampling is done without replacement.

In particular, four MOEs are investigated in detail: average speed, average delay per unit distance, average number of stops, and acceleration noise. The importance of such mesoscopic and macroscopic MOEs for developing management strategies for signalized arterial networks and assessing its performance has been made clear through recent research efforts. An important advancement in transportation has been the introduction of the MFD that allows us to relate the average network flow with the average network density (Daganzo and Geroliminis, 2008; Geroliminis and Daganzo, 2008). Applications of such macroscopic models include perimeter control (Aboudolas and Geroliminis, 2013; Geroliminis et al., 2013), gating strategies (Keyvan-Ekbatani et al., 2012, 2013), congestion pricing (Gonzalez and Daganzo, 2012; Gonzales, 2015), and parking management systems (Geroliminis, 2015). In addition, MOEs such as average speed and average delay are useful for traveler information systems, prediction of travel vehicle arrival times for cars and buses, and design of real-time signal

---

1 In general, algorithms that use CV or probe data to compute MOEs include a vehicle identifier, which precludes using data from a unique vehicle repeatedly in a specific calculation.
control length was not included in the MOEs investigated in this study because its estimate is not obtained as an average of individual vehicle characteristics. Instead the most efficient queue length method depends on the position of the single last detected vehicle in queue as shown in Argote et al. (2011).

Even though the focus of this paper is on CV data, the methodology can be used for any type of data source that can provide data on individual vehicle characteristics (e.g., probe vehicles or mobile phone data), which can be used for obtaining MOEs of interest. The choice of working with CV data is due to the fact that this methodology is expected to be most relevant for this type of data. Probe vehicles are generally regarded as a tool used for sporadic data collection. On the other hand, CVs will offer the possibility of continuous access to very rich streams of data, once they are fully adopted. However, during the early stages of the CV technology, it will be essential to know minimum penetration rate levels required to guarantee accurate MOE estimates so that CV-generated data can safely and reliably facilitate real-time applications in urban networks. The proposed methodology can be used to estimate both the minimum number of CVs and the minimum penetration rate required for accurate estimation of various MOEs. However, the focus of this paper is on the minimum penetration rate since it is easier to communicate to decision-makers, which could translate these requirements into effective policies.

This paper is organized as follows. First, the three datasets used in this study are briefly described. Next, the methods for estimating arterial MOEs are presented. Then, the minimum penetration rate requirements for accurate estimates are identified in two scenarios: undersaturated and oversaturated traffic conditions. The next section presents the methodology that generalizes the determination of minimum penetration rates for a variety of flow and green ratios of the critical intersection and arterial direction under consideration. The results from implementing the proposed algorithm on a three-intersection arterial segment are presented next. The last section summarizes the study findings and outlines future research.

2. Test sites

The impact of the CV technology penetration rate on the accuracy of the MOEs has been estimated with the use of actual vehicle trajectories collected at Peachtree Street, Atlanta, GA and obtained from the Next Generation Simulation (NGSIM) program (Alexiadis et al., 2004), simulated trajectories from El Camino Real, a major arterial in the San Francisco Bay Area, and simulated trajectories from San Pablo Avenue in Berkeley, CA. The NGSIM dataset was used to test the proposed estimation procedures in undersaturated traffic conditions: conditions where queues are completely dissipated by the end of each signal cycle. No real data were available for oversaturated traffic conditions: conditions where demand cannot be served by the intersection in a cycle and residual queues form. In view of this, we resorted to a simulation of a stretch of the El Camino Real arterial, south from San Francisco, to initially test the estimation of arterial MOEs in oversaturated traffic conditions. Finally, a three-intersection segment of San Pablo Avenue was simulated in a multitude of traffic states. These simulations produced trajectories that were used to test the performance of the MOE estimation algorithms and determine how different traffic volumes affect the minimum penetration rate level necessary to obtain accurate MOE estimates. The simulation outputs were also used to validate the analytical model developed to determine minimum CV penetration rates necessary for various levels of accuracy in the MOE estimation. A more detailed description of the three datasets follows.

2.1. NGSIM field data: Peachtree Street, Atlanta, GA

The NGSIM arterial dataset consists of vehicle trajectories collected at Peachtree Street, Atlanta, GA, for four signalized intersections, starting at 10th Street NE and ending at 14th Street NE in the northbound direction; see Fig. 1(a), source: Cambridge Systematics Inc. (2007). The dataset includes vehicle ID, time, position, lane, speed, and acceleration. Trajectory data are available at a resolution of 0.1 seconds. In addition, vehicle type, vehicle length, and their origins and destinations are available.

All signals operate as actuated and the signal cycle length is not constant. The data used in this study were collected from 12:45 to 13:00 and present only undersaturated conditions. Since the MOEs estimated in this paper describe the traffic conditions for a given direction, only the trajectories of vehicles traveling northbound on the main arterial were considered. This resulted in a total of 228 vehicle trajectories being processed.

2.2. Simulated data: El Camino Real, San Francisco Bay Area, CA

A microsimulation model of El Camino Real in VISSIM (Fellendorf, 1994) was used to obtain data for oversaturated traffic conditions, as depicted in Fig. 1(b). The selected arterial section includes three intersections starting at Cambridge Avenue and ending at Page Mill Road in the southbound direction.

All signals operate as actuated coordinated with a cycle length of 130 seconds. The simulation included a 30-min warm-up period where the arterial was gradually loaded until reaching oversaturated conditions. The next 15 min were used in the analysis. The trajectories were exported with the same fields and resolution as in the NGSIM dataset. A total of 1332 vehicle trajectories were processed.
2.3. Simulated data: San Pablo Avenue, Berkeley, CA

A microsimulation model of San Pablo Avenue in Berkeley, CA, was built in AIMSUN (Barceló and Casas, 2005). This model was used to test the analytical methodology to determine minimum CV penetration rates for accurate MOE estimation in a variety of traffic conditions; see Fig. 1(c). The arterial section under consideration includes four intersections from Dwight Way to University Avenue. Trajectories were extracted for the northbound direction, which is the heaviest direction for the evening peak period.

All signals operate as fixed-time coordinated with a common cycle length of 80 seconds. The critical intersection for the northbound direction is the intersection of San Pablo and University Avenues. This arterial segment has a minimum green ratio of 0.5. The green ratio is defined as the effective green time allocated to the approach, \( G \), over the cycle length, \( C \), and is denoted throughout this paper as \( G = C \).

In the base scenario, the critical intersection is characterized by a flow ratio of 0.35, which is defined as the arrival rate for that direction, \( q \), over the saturation flow, \( s \), denoted as \( q/s \). The three intersections upstream of the critical one all have green and flow ratios equal to 0.65 and 0.4 respectively. These \( q/s \) values were modified to compile four different simulation scenarios to test the effect of traffic demand on the minimum penetration rate levels for a given green ratio, where \( q/s \in \{0.35, 0.4, 0.5, 0.6\} \). These scenarios resulted in a total of 313, 365, 447, and 456 vehicle trajectories being processed respectively.

All simulations also included a 30-min warm-up period to gradually load the arterial. The 15 min following the warm-up period were used in the analysis. The trajectories were exported using the same resolution as in the other two scenarios for all four links in the arterial section considered.

3. MOE estimation algorithms and preliminary analysis

This section presents the algorithms used to calculate the arterial-based MOEs using only CV data for two distinct scenarios: under- and oversaturated traffic conditions. For each one of these scenarios, various market penetration rate values \( p \in \{0.01, 0.025, 0.05, 0.10, 0.125, 0.15, 0.20, 0.25, 0.35, 0.50, 0.75, 0.90, 1\} \) are tested. Then, a simple visual method to determine the minimum penetration rate that would yield accurate estimates is introduced. The results illustrate the disparity
that exists in the minimum \( p \) values necessary to obtain accurate estimates based on the underlying traffic conditions. In this preliminary analysis we consider an estimate to be accurate when it is within 10% of the underlying ground truth value.

3.1. Sampling method and minimum penetration rate

For each of the \( p \) values considered, 10,000 different sampling runs are generated. In each run, to determine if a vehicle is CV-equipped, a Bernoulli trial with a probability of success \( p \) is performed. This process reduces the number of original trajectories to a random sample of CV-equipped trajectories as if the CV penetration rate was effectively \( p \). These sampled trajectories are then used to determine the different MOE estimates.

Box plots are used to reveal the arterial MOE estimate variation for the different \( p \) values and identify the minimum \( p \) required to accurately estimate a given MOE for both undersaturated and oversaturated conditions. On each individual box, the central red mark represents the median value of the 10,000 estimates, while the edges of the box capture the 25th and 75th percentiles; see for example Fig. 2(a). These plots also include a whisker range that extends 1.5 times the distance between the 25th and 75th percentiles at both extremes of the box. A penetration rate was considered acceptable if the whisker range, which corresponds to approximately \( \pm 2.7 \) the standard deviation of the estimate, was within 10% of the average value for the ground truth. If the estimation errors are assumed normally distributed, this criterion ensures that the MOE estimate would lay within 10% of the ground truth with a probability of 0.9965. This is a simple methodology to determine if a given \( p \) would yield accurate estimates of a certain MOE but it reveals a clear difference between the results obtained in under- and oversaturated traffic conditions.

3.2. MOE estimation, preliminary results

This subsection focuses on a handful of different arterial MOEs. The algorithms applied to obtain the MOE estimates are introduced first and the results obtained following the aforementioned sampling procedure are presented next. The algorithms are applied to both under- and oversaturated datasets, revealing large disparities in the accuracy of the results for a given penetration rate.

3.2.1. Average speed

The average speed for all lanes in the observed direction in a particular sample is obtained using Edie’s generalized average speed definition (Edie, 1963). Edie postulates that the average speed for any region \( A \) in the time–space plane, \( \langle v(A) \rangle \), is equal to the ratio of the total distance traveled by the vehicles in that region, \( d(A) \), to the total time spent by those vehicles in that region, \( t(A) \): \( \langle v(A) \rangle = d(A)/t(A) \). With CV data, this ratio can be obtained as:

\[
\langle v \rangle = \frac{\sum_{i=1}^{S} l_i}{\sum_{i=1}^{S} t_i}
\]

where \( S \) is the total number of CV trajectories sampled, \( l_i \) is the total distance traveled by vehicle \( i \) in the sample, and \( t_i \) is the total time spent by that vehicle in the arterial portion of interest for which the average speed is being calculated.

Fig. 2(a) and (b) show box plots of the average speeds estimated under different \( p \) values for both the NGSIM and El Camino Real data. The solid horizontal black lines in both figures provide the \( \pm 10\% \) range from the underlying ground truth value (i.e., the result obtained with full market penetration rate, \( p = 1 \)). That range can provide a guideline to determine if the measured average speeds would be accurate enough at a certain CV market penetration rate level. Fig. 2(a) and (b) reveal that, from the array of possible penetration rates considered, a value of at least 35% would be necessary to provide speed estimates within 10% of the ground truth in over 99% of the cases for the arterial data in undersaturated traffic conditions. On the other hand, the dataset with oversaturated traffic conditions requires only a 5% penetration rate level to achieve the same accuracy level.

This disparity shows that contrary to other more traditional methods that can be utilized with data from more conventional sensing technologies, e.g., detector data, CV-based MOE estimation is not only influenced by the technology’s penetration rate but also by the underlying traffic conditions. The reason behind this is clear. If an observer takes a sample of constant duration with a given penetration rate level \( p \), the observer will be more likely to see more CV-equipped vehicles in oversaturated than in undersaturated traffic conditions. This should in turn reduce the error in the estimation results for oversaturated conditions. The following subsections confirm this observation for various other MOEs.

3.2.2. Average delay per unit distance

Delay is defined as the difference between the actual travel time to traverse an arterial section and the corresponding travel time under free-flow conditions. The average delay per unit distance for a given sample can be obtained using CV data per the following expression:

\[
D = \frac{1}{S} \sum_{i=1}^{S} \frac{1}{t_i} \left( t_i - \frac{l_i}{v_i} \right)
\]
where $\bar{D}$ is the average delay per unit distance for a single sample of trajectories and $v_f$ is the free-flow speed on the arterial corridor. The free-flow speeds used for Peachtree Street and El Camino Real are 40 km/h (25 mph) and 64 km/h (40 mph) respectively. This MOE can also be interpreted as the difference between the sampled average pace and the free-flow pace.

As before, the box plot in Fig. 3(a) shows that samples obtained with low penetration rates present a significant number of outliers, which could lead to a misinterpretation of the traffic conditions. Moreover, the two figures reveal once again the disparity in the variability of the measurements at a given $p$ value for the two datasets. For this MOE, if we follow the accuracy approach defined earlier, penetration rates higher than 75% would be necessary for accurate estimates in undersaturated conditions. On the other hand, Fig. 2 (b) shows that only a 15% penetration rate would be required in oversaturated conditions.

### 3.2.3. Average number of stops

The average number of stops, $\bar{n}_s$, is simply the average number of times that CV-equipped vehicles in the sample traveling in the direction of interest have to halt their forward motion. To characterize vehicle stops it is necessary to observe the vehicles’ trajectory data. Imagine that for every time step $j$ the vehicle speed can be obtained, as it is the case with CV data. Then a vehicle stop can be characterized as the moment in which a vehicle speed transitions from a value greater than a certain speed threshold to a value below that threshold. In this particular implementation, the speed threshold was considered to be 5 km/h (3.1 mph).\(^2\)

As with the previous two MOEs, the minimum acceptable $p$ values differ greatly for the two sets of traffic conditions considered. Fig. 4 (a) and (b) show that $p$ values higher than 50% and 20% would be necessary to obtain results that would lie close to the average number of stops ground truth value for the undersaturated and oversaturated scenarios respectively.

### 3.2.4. Average acceleration noise

Acceleration noise, $\sigma_a$, is defined as the standard deviation of a vehicle’s acceleration measurements along its trajectory (Jones and Potts, 1962). This MOE is widely used as a proxy for the smoothness of traffic flow along signalized arterials; see Fig. 5.

The results show that, for this particular MOE, penetration rates on the order of 10% and above should be sufficient to obtain reliable standard deviations for the acceleration for undersaturated conditions, but only 1% is necessary for oversaturated conditions. Again the figures reveal the same tendency observed earlier: oversaturated traffic conditions would require lower minimum $p$ values to yield accurate estimates.

### 3.3. Preliminary analysis discussion

Note that the results presented for under- and over saturated conditions are not directly comparable due to the use of test sites that differ in their geometric and signal setting characteristics in addition to the level of traffic. However, the preliminary results presented in this section are useful for pointing out the need to account for traffic conditions when estimating the minimum penetration rate required for accurate MOE estimates.

---

\(^2\) One could argue that 5 km/h may be too high, however the real NGSIM data revealed that using lower speed thresholds could lead to an overestimation of the number of stops. The reason is that when waiting at traffic lights, some drivers have a tendency to incrementally get closer to their preceding car or the intersection entry. This behavior could trigger new stops if the speed threshold is too low.
Fig. 3. Box plots of the average unitary delay estimates for various CV market penetration rates, considering two 15-min scenarios using CV data.

Fig. 4. Box plots of the average number of stops estimates for various CV market penetration rates, considering two 15-min scenarios using CV data.

Fig. 5. Box plots of the average acceleration noise estimates for various CV market penetration rates, considering two 15-min scenarios using CV data.
There are many more metrics that one could calculate using CV data, such as the average turning ratios for a given intersection approach, the average proportion of stopped time for all vehicles in an arterial segment, and the average velocity gradient in an arterial segment. However, all these examples as well as the ones presented in this paper share a common characteristic: their estimation accuracy will depend on the total number of CV-equipped vehicles observed. That number in turn will not only depend on the CV technology market penetration rate \( p \), but also on the underlying traffic conditions. Therefore, one should expect minimum penetration rate requirements for accurate estimates to be less restrictive in oversaturated conditions than in undersaturated conditions, ceteris paribus, since equal duration samples will contain more observable CV-equipped vehicles when the arterial carries more cars.

The following section presents an analytical method to determine minimum penetration rate values. This method captures the direct relationship between minimum \( p \) values for accurate estimates and the underlying traffic conditions on the arterial of interest expressed through the flow and green ratios for the intersection approach of interest.

4. Analytical model for estimating minimum penetration rates

The previous analysis provides basic insights on the minimum penetration rate requirements for undersaturated and oversaturated traffic conditions. To generalize this, the proposed methodology relates the minimum penetration rate with the number of CV-equipped vehicles required to estimate a MOE with certain accuracy for a certain level of traffic demand. The methodology is based on the notion of a 1 - \( \alpha \) confidence interval; i.e., a region such that the probability an estimate falls within is 1 - \( \alpha \). The methodology also considers the sampling duration, \( \Delta t \), and the traffic conditions as expressed by the vehicle flow observed at an intersection during a cycle, because both parameters determine the maximum number of vehicles that can be sampled.

4.1. Methodology

The objective of this methodology is to estimate the mean value of a MOE \( \bar{\theta} \) (\( \theta \) could refer to any MOE such as speed or delay), in a vehicle population of size \( N \). Let us initially assume that the population size is known. This assumption will be relaxed later, by taking into account the underlying traffic conditions in the arterial. Unfortunately, with \( p < 1 \) it is not possible to observe the true underlying mean value \( \bar{\theta} \) of the entire population. Instead, with the CV technology one could observe \( \bar{\theta}_n \), which is the mean estimate obtained based on the information reported by \( n \) CV-equipped vehicles in the population of interest.

\[
\bar{\theta}_n = \frac{1}{n} \sum_{i=1}^{n} \theta_i,
\]

Another value that is also observable is \( S_n \), the CV sample standard deviation, which can be calculated as:

\[
S_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (\theta_i - \bar{\theta}_n)^2}.
\]

Note that \( S_n \) is not necessarily equal to the population standard deviation but it is a necessary input in the formulation to predict minimum penetration rate values.

Since the true underlying parameter of interest is unobservable with a partial vehicle sample, we resort to confidence intervals to provide an estimate range, which is likely to contain the sought parameter \( \theta \). As we mentioned earlier, the standard deviation for the population of interest is not known. In this case, the distribution of the sample mean, \( \bar{\theta}_n \), can no longer be considered normal but t-Student with \( n - 1 \) degrees of freedom, \( \bar{\theta} \) as mean, and a standard deviation that is equal to the sample standard deviation \( S_n \) divided by \( \sqrt{n} \) and multiplied by a correction factor. Thus, the 1 - \( \alpha \) confidence interval for \( \theta \), CI\(_{1-\alpha} \), can be expressed as:

\[
CI_{1-\alpha} = \bar{\theta}_n \pm T_{n-1}^{-1}(1-\alpha/2) \frac{S_n}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.
\]

where \( T_{n-1}^{-1} \) is the Student’s \( t \)-inverse cumulative distribution function with \( n - 1 \) degrees of freedom. The term \( \sqrt{(N-n)/(N-1)} \) is the aforementioned standard error correction factor to be considered when sampling from a finite population; see Steel and Torrie (1960). This correction factor accounts for the fact that the population size will be relatively small and sampling is performed without replacement. Based on this expression, the minimum number of vehicles, \( n_{\text{min}} \), that one would need to observe in order to have a 1 - \( \alpha \) confidence interval that extends less than \( m \) units from the CV sample average (\( CI_{1-\alpha} \leq \bar{\theta}_n \pm m \)) would be:

\[
n_{\text{min}} = N \frac{(T_{n_{\text{min}}-1}^{-1}(1-\alpha/2)S_n)^2}{m^2(N-1) + (T_{n_{\text{min}}-1}^{-1}(1-\alpha/2)S_n)^2},
\]

as given in the equation above.
\[ n_{\text{min}} \approx \left( T_{n_{\text{min}}-1}^{-1} \left( 1 - \frac{z}{2} \right) \frac{S_n}{m} \right)^2 \approx \left( \Phi^{-1} \left( 1 - \frac{z}{2} \right) \frac{S_n}{m} \right)^2. \]  

Eq. (6a) provides the minimum number of CV-equipped vehicles one would have to observe to have a confidence interval within \( m \) units of the sampled mean at a \( 1 - \alpha \) probability level when the population size \( N \) is small. If the population size is large (e.g., \( N > 200 \) vehicles) and the CV penetration rate \( p \) is small, Eq. (6a) can be further simplified into (6b). Note that for any desired degree of accuracy, \( m > 0 \), there will be at least one integer \( n_{\text{min}} \in (0, N) \) such that the absolute value of the difference between the left hand side and the right hand side of Eqs. (6) is less than 1.

However, Eq. (6a) has an interdependency between the sought result \( n_{\text{min}} \) and \( T_{n_{\text{min}}-1}^{-1} \), since \( n_{\text{min}} \) determines the degrees of freedom of the \( t \)-Student distribution. This interdependency can be easily solved using the secant method, a simple root-finding numerical method similar to the Newton–Raphson method in which the derivative of a function is numerically approximated; see Dennis and Schnabel (1996). In this case, we seek to find the root of a version of Eq. (6a):

\[ f(n) = n - N \left( T_{n_{\text{min}}-1}^{-1} \left( 1 - \frac{z}{2} \right) S_n \right)^2 \left( \frac{m^2(N - 1)}{2n_{\text{min}}(1 - \frac{z}{2})S_n} \right). \]

Note that his method will converge since there is only one root for this function for values of \( n \geq 0 \). The first part of the function increases linearly with \( n \) and the term \( T_{n_{\text{min}}-1}^{-1} \left( 1 - \frac{z}{2} \right) \) on the second part is a strictly decreasing function of \( n \). The pseudo-code for the secant method is shown below:

1. Select two initial \( n_{\text{min}} \) candidates \( n_0 \) and \( n_1 \) as \( n_0 = \left( \Phi^{-1} \left( 1 - \frac{z}{2} \right) \frac{S_n}{m} \right)^2 \) and \( n_1 = n_0 + \delta \) with \( \delta \) being a reasonable integer value (e.g., \( \delta = 5 \)).
2. If \( |f(n_i)| < \epsilon \approx 1 \) for \( i \in \{0, 1\} \) stop, and \( n_{\text{min}} = n_i \), otherwise set \( i = 1 \) and continue.
3. If \( |f(n_0)| < |f(n_1)| \) swap \( n_0 \) and \( n_1 \) using:
   - Set \( n_{\text{temp}} = n_1 \).
   - Set \( n_1 = n_0 \).
   - Set \( n_0 = n_{\text{temp}} \).
4. Find a new \( n_{\text{min}} \) candidate with the secant method’s iteration rule:
   \[ n_{i+1} = n_i - \frac{n_i - n_{i-1}}{f(n_i) - f(n_{i-1})}. \]
5. If \( |f(n_{i+1})| < \epsilon \approx 1 \) stop and \( n_{\text{min}} = n_{i+1} \) else set \( i = i + 1 \) and return to step 4.

Numerical tests have shown that the iteration converges in a small number of steps, generally on the single digits, if reasonable initial \( n_0 \) and \( n_1 \) candidates are used. In practice, for values of \( n_{\text{min}} \) sufficiently large, this iteration procedure can be ignored replacing \( T_{n_{\text{min}}-1}^{-1} \left( 1 - \frac{z}{2} \right) \) with \( \Phi^{-1} \left( 1 - \frac{z}{2} \right) \), the inverse standard normal cumulative density function.

Now, in order to relate Eqs. (6a) and (6b) with the CV market penetration rate, it is necessary to consider how the technology’s \( p \) value will affect the number of CV-equipped vehicles that one can expect to sample at a arterial during a certain time interval \( \Delta t \). If the traffic conditions are stationary and the arterial characteristics (i.e., flow ratio and green ratio) are known, the expected total number of vehicles \( N \) (population size) and those that will be CV-equipped \( n_{CV} \) (sample size) will be given by the following equations:

\[ N = \Delta t \min(q, sG/C), \]
\[ n_{CV} = p \Delta t \min(q, sG/C). \]

where \( q \) is the total upstream vehicle demand, \( s \) is the saturation flow for the arterial approach of interest, \( G \) is the effective green time for the approach (i.e., the most restrictive green time in the direction of interest), and \( C \) is the cycle length for all intersections in the arterial. After reaching stationary conditions, the vehicle flow observed downstream of the intersection entrance is equal to the demand flow, \( q \), if the signalized approach operates in undersaturated conditions, and equal to \( sG/C \) if the signalized approach operates in saturated or oversaturated conditions. So, the minimum penetration rate requirement, \( p_{\text{min}} \) can be obtained by combining Eqs. (6a), (6b), (7a), and (7b) as follows:

\[ p_{\text{min}} = \frac{\left( T_{n_{\text{min}}-1}^{-1} \left( 1 - \frac{z}{2} \right) S_n \right)^2}{m^2(\Delta t \min(q, sG/C)) + \left( T_{n_{\text{min}}-1}^{-1} \left( 1 - \frac{z}{2} \right) S_n \right)^2}, \]

\[ n_{\text{min}} = \Phi^{-1} \left( 1 - \frac{z}{2} \right) \frac{S_n}{m}. \]

\[ 3 \] For the sake of simplicity, it is assumed that the intersection approaches in the arterial stretch and direction of interest have common saturation flow, \( s \), and green ratio, \( G/C \). If that is not the case, Eqs. (7a) and (7b) should consider the most restrictive intersection approach, i.e., one that has the lowest \( sG/C \) value.
Eq. (8a) is the result of considering the correction factor for small populations and sampling without replacement, while Eq. (8b) approximates these results for large populations and low \( p \) values. The term \( \Phi \) in Eq. (8b) refers to the normal cumulative distribution function. If an observer is interested in obtaining average MOE estimates of stationary traffic conditions during a period of duration \( \Delta t \), the previous formulas reveal the minimum CV market penetration rate \( p \) that should be required to have a symmetric confidence interval of 2 \( m \) units from the measured mean \( \bar{v}_n \) at a confidence level \( 1 - \alpha \).

Fig. 6 presents the results of a sensitivity analysis to investigate the impact of traffic demand and MOE variability on the minimum penetration. In particular, it compares the \( p_{\text{min}} \) predictions that Eqs. (8a) and (8b) provide in an idealized scenario. For illustration purposes we consider an arterial approach with saturation flow of \( s = 1800 \) vehicles/h and an effective green ratio of \( G/C = 0.5 \). We also assume in this comparison that the duration of the sampling period is \( \Delta t = 15 \) min and that the sought \( m \)-value is 2 units (e.g., 2 km/h if the MOE of interest is average speed). Fig. 6 reveals that for high vehicle demand and low MOE sample standard deviation values the differences between the two expressions is almost insignificant, yielding \( p_{\text{min}} \) values within 0–5% of each other; see Fig. 6(c). On the other hand, the approximate expression should not be used when traffic demand is low and the MOE standard deviation value is high, as showcased by unfeasible \( p_{\text{min}} \) values over 1 in Fig. 6(b).

Finally, note that the assumptions used when deriving Eqs. 6(a) and 6(b) do not apply to the average speed MOE presented earlier. Instead of the arithmetic mean, the average speed MOE in Section 3.2.1 is the harmonic mean\(^4\) if all vehicles sampled travel the entire length of the arterial segment of interest, \( l \). In that case, the average speed in Eq. (1) can also be expressed as:

\[
\bar{v} = \frac{N l}{\sum_i^N t_i} = \bar{t} = \frac{1}{\bar{t}} \frac{1}{\sum_i^N \frac{1}{t_i}}
\]

As shown in Eq. (9), the generalized average speed can be also expressed as the ratio of the arterial segment length to the average travel time, \( \bar{t} \), for those vehicles that traverse the entire segment of interest, \( l \). Taking this into consideration, it can be shown that the minimum penetration rate to guarantee a \( 1 - \alpha \) confidence interval of width 2\( m \) around \( \bar{v} \) would be given by an expression that is different from Eq. (8a) as follows:

\[
P_{\text{min}} \approx \frac{(\bar{a}_n S_n)^2}{2m^2 \bar{t}_n^4 \Delta t \min (q, \frac{1}{\bar{v}})}
\]

For a more detailed derivation of this formula please refer to Appendix A. Note that in this case, the minimum penetration rate not only involves the sample travel time standard deviation, \( S_n \), but also the sample travel time mean, \( \bar{t}_n \). This expression differs from Eq. (8a) and leads to significantly different \( p_{\text{min}} \) predictions.

4.2. Validation via simulation

To validate the minimum penetration rate methodology presented in the previous section we resort to a Monte Carlo simulation approach. While the validation of the proposed methodology is performed with simulated data, there is nothing that precludes using real-world data. The only necessary inputs for this methodology are (1) the accuracy sought, (2) the underlying variability of the MOE of interest, (3) the sampling interval duration and (4) the arterial capacity and its traffic conditions. All of them can be either pre-determined or obtained from historical data through various sensing technologies. Simulation was used in this study in order to allow for altering the flow ratios and therefore, testing the performance of the proposed methodology for a variety of traffic conditions.

The simulated San Pablo Avenue test site described in Section 2.3 is used in this section. Four different demand scenarios are considered to generate the ground truth population samples at different traffic demand levels. A unique flow ratio level, \( q/s \in \{0.35, 0.4, 0.5, 0.6\} \), characterizes each scenario.

For each one of these four traffic scenarios, vehicle trajectories are sampled 1000 times at each penetration rate level, \( p \).

The sampling simulations start at 1% and they are progressively increased by 1% until the accuracy criterion established for a certain MOE is satisfied. That is the first \( p \) value for which the average MOE obtained in at least 950 samples falls within \( m \) units of the underlying ground truth, to reflect a 95% confidence interval. These values are then compared with the \( p_{\text{min}} \) predicted by the proposed methodology.

The different \( m \) values for each one of the four MOEs analyzed are summarized in Table 1. Note that the \( m \) values have been arbitrarily chosen. In reality, the choice of a particular acceptable range for a MOE should reflect some objective consideration, which could be of technical nature or linked to an underlying policy or regulation. For example, the \( m \) values could be lower for applications that do not require very high accuracy (e.g., trip travel time estimation) but higher values

\(^4\) The harmonic mean speed \( \bar{v} \) is used here because it is part of the fundamental traffic equation that relates flow, \( q \), and density, \( k \): \( \bar{v} = q/k \).
of \( m \) would be necessary for implementing real-time signal control and transit signal priority at specific signalized intersections.

Table 2 presents the MOE standard deviation values, \( S_n \), used for each flow ratio. Note that the average speed’s \( p_{\text{min}} \) value is calculated based on the underlying average travel time, i.e., harmonic mean speed. These \( S_n \) values were obtained as the observed population standard deviations for the four different traffic conditions, under perfect information (i.e., 100% penetration rate). In a real-world setting, these values would not be available but they could be replaced by values obtained during a preliminary data gathering effort or they could be bootstrapped using the actual sampled data.

Table 3 presents the observed and predicted minimum penetrations rates for various flow ratios and the four MOEs under consideration. These results were obtained with the use of the Monte Carlo simulation approach presented earlier and the proposed analytical method for estimating minimum penetration rate as a function of traffic conditions, green ratio, and sampling duration.

As one would expect from the preliminary data analysis, lower penetration rates are required for lower flow ratios when estimating average acceleration noise and average speed. However, when the average unitary delay is estimated, we observe that lower penetration rates are needed for medium flow ratios (\( \approx 5\% \)), medium penetration rate levels are needed for low flow ratios (\( \approx 10\% \)), and high penetration rates (15%) for oversaturated conditions. This is due to the fact that the sample

\( \text{Fig. 6. Minimum penetration rate values in an idealized single lane arterial approach with } sG/C = 900 \text{ veh/h, sampling period } \Delta t = 15 \text{ min, and a 95\% confidence interval with } m = 2 \text{ units for various demand and MOE sample standard deviations, } S_n, \text{ levels.} \)
standard deviation, $S_n$, remains fairly constant when conditions are undersaturated and it spikes in the simulation of highly oversaturated traffic flow.

The minimum penetration rates required to estimate the number of stops follow a different pattern. In this case, higher flow ratios also require higher minimum penetration rates and lower flow ratios lower ones. This is due to the sample standard deviation, $S_n$, for this particular metric, which increases significantly with the flow ratio increases; see Table 2. This noise increase is most likely due to the fact that more congestion leads to more variability in the number of stops that vehicles will have to make while traveling through the arterial. Finally, note that the average speed presented here is the space mean speed, which depends on the inverse of the vehicle’s travel time. In that case, the equation to obtain the minimum penetration rates is a different one and it is presented in Eq. (10). This is the reason why the trend seems to be counter-intuitive when one compares the average speed with the other MOEs.

A comparison of the two sets of results indicates that the proposed methodology, which estimates the minimum CV penetration rate necessary to accurately determine MOEs, closely matches the observed results obtained from the Monte Carlo sampling simulations for the MOEs that are estimated as arithmetic means (i.e., average delay per unit distance, average number of stops, and average acceleration noise). For these MOEs the predicted minimum penetration rate is within 1% of the observed Monte Carlo value for all traffic demand levels. Note that in the proposed Monte Carlo-based testing framework this corresponds to the best possible result, since the step in the Monte Carlo simulations considers an increase of 1% CV penetration rate and therefore, it is not possible to obtain a better precision.

On the other hand, the results obtained for the estimation of the harmonic mean speed reveal bigger discrepancies on the order of 6–7% for low traffic demand levels. This result is not surprising because the methodology used for this metric assumes that the population size is large enough to ignore the sampling without replacement correction factor; see Appendix A. The effect of this simplification for low traffic demand levels is that our predicted minimum penetration rate errs on the side of caution by predicting a higher $p_{\text{min}}$ level than is actually necessary. On the other hand, the results obtained for high traffic demand scenarios present a better alignment of the predicted and the observed minimum CV penetration rates. This is positive, since precise MOE estimation under heavy traffic is likely to be more important than when traffic demand is low.

5. Conclusions

This paper presented algorithms for MOE estimation using data from CV-equipped vehicles. Estimation algorithms were developed for a variety of MOEs such as average speed, average delay per unit distance, average number of stops, and average acceleration noise, and were tested with the use of two different datasets that varied in their traffic conditions and signal settings. The results of this preliminary analysis indicated that the level of estimation accuracy achieved for various MOEs at different penetration rates of the CV technology varies depending on the prevailing traffic conditions.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Acceptable $m$ values that define an accurate MOE estimation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average speed (km/h)</td>
<td>Average unitary delay (s/m)</td>
</tr>
<tr>
<td>2</td>
<td>0.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2</th>
<th>$S_n$ proxy values used in the $p_{\text{min}}$ prediction for the simulated traffic demand levels.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q/s = 0.35$</td>
<td>Travel time $S_n$ (s)</td>
</tr>
<tr>
<td>21.58</td>
<td>0.27</td>
</tr>
<tr>
<td>30.33</td>
<td>0.22</td>
</tr>
<tr>
<td>34.49</td>
<td>0.26</td>
</tr>
<tr>
<td>33.29</td>
<td>0.44</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Observed and predicted minimum penetration rate values for various MOE estimates using simulated data for four different traffic demand levels.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOE</td>
<td>$q/s = 0.35$</td>
</tr>
<tr>
<td></td>
<td>Observed (%)</td>
</tr>
<tr>
<td>Average speed (harmonic mean) (km/h)</td>
<td>23</td>
</tr>
<tr>
<td>Average unitary delay (s/m)</td>
<td>9</td>
</tr>
<tr>
<td>Average number of stops</td>
<td>4</td>
</tr>
<tr>
<td>Average acceleration noise (m/s$^2$)</td>
<td>13</td>
</tr>
</tbody>
</table>
Therefore, a methodology was developed to analytically and precisely determine the minimum penetration rate necessary for accurate MOE estimation. The methodology revealed that minimum CV penetration rate levels for accurate MOE estimation depend on four factors: (1) the accuracy sought, (2) the underlying variability of the MOE of interest, (3) the sampling interval duration and (4) the arterial capacity and its traffic conditions and signal settings. Specifically, the relationship between traffic flow and the MOE variability is essential. Some MOEs are less variable at higher traffic flows than lower ones while others behave in the opposite way. Understanding those nuances is essential. Transportation practitioners should also be aware that those factors might be correlated. For instance, higher traffic volumes could lead to a higher number of CV-equipped vehicles observed during a particular sampling duration, which should in turn decrease the necessary penetration rate for accurate estimation.

However, if higher traffic is also associated with higher MOE variability (as it was the case for the average number of stops in the simulation performed), there might be an increased need for higher CV penetration rates even for highly congested traffic states. In addition, our methodology does not exploit the possibility that in some cases, e.g., congested conditions, MOE values in consecutive samples could be correlated. Future work could be devoted to develop alternative estimation methods, for example drawing from Bayesian estimation principles, to try to take advantage of possible correlation among consecutive samples and lower the $p_{\text{min}}$ boundaries.

It is also necessary to highlight the importance of the desired accuracy levels. In this work accuracy levels were arbitrarily chosen. In real-world implementations, the accuracy thresholds should reflect technical considerations or should comply with established regulations for the specific applications the MOE estimates are intended to be used for.

In addition, the methodology presented includes a correction factor to account for errors that could occur when one deals with small population sizes and sampling is performed without replacement. This correction factor proved to be relevant in yielding accurate minimum penetration rate predictions in low traffic states. The proposed methodology was tested for a variety of traffic conditions and was validated with simulated data. The results showed that it is possible to predict the necessary penetration rate for most MOEs and traffic conditions within 1%. The results also revealed that in general, relatively low penetration rates (smaller than 15%) are required to accurately estimate multiple MOEs.

The methodology presented can be used in multiple ways. As an example, once a particular sample is collected, a transportation authority could easily determine the minimum penetration rate necessary to probabilistically guarantee that the MOE estimates derived from the sample meet a certain level of accuracy. Then, they should compare this value with the expected penetration rate in the field in order to decide whether the estimate is valid or not. Another possibility would be to perform a preliminary analysis of the relationship between traffic flow levels and MOE variability in an arterial and then use the proposed methodology to establish minimum CV penetration rate requirements that are necessary for accurate estimation of each MOE, well in advance of the technology’s implementation.

Further work on the use of CV data to estimate MOEs should also consider how to best implement this estimation methodology in real-world systems. Multiple questions on this domain could be addressed via further research. For example, it could be interesting to develop feasible architectural approaches to retrieve CV data from individual vehicles and relay them to central servers where the data could be aggregated into the desired MOE values. One possibility could be to cache the data in the vehicles and communicate them at particular roadside locations like a traffic signal while the vehicle is in proximity of a receiver. How to do so in a way that is efficient by minimizing the necessary communication broadband, CPU-usage, and memory requirements would be a problem worth studying. Other aspects of great interest could be how to identify and respond to potential cyber attacks that could yield undesired and inaccurate estimates.

Finally, even though this work did not focus on data specification, we want to convey the importance of establishing such data standards in the early stages of a technological development. Experience in other transportation areas such as public transportation has revealed the importance that unified data specifications have in fostering innovation and synergies across multiple agencies and service providers. In that particular area, the implementation of the Global Transit Feed Specification (GTFS) has facilitated the apparition of a great number of applications that benefit from a clear and unified data framework. Establishing a clear specification for CV data should be a priority. The agencies in charge of transportation management and regulation should recognize that potential and embark on a path to establish a standardized framework that could open up all the potential that the CV technology has to offer.

**Acknowledgments**

This work was partly sponsored by the Federal Highway Administration (FHWA) of the United States Department of Transportation (US DOT) as part of the Exploratory Research Program (EAR)-Advanced Signal Control Algorithms project. The authors would like to thank Dr. David Yang, FHWA contract monitor and Professor Carlos Daganzo of the University of California at Berkeley for their comments and suggestions. The comments provided by the external reviewers are also greatly appreciated.

**Appendix A. Minimum penetration rate estimation for the arterial speed harmonic mean**

Note that the speed harmonic mean can be defined as the ratio of the arterial segment length and the average travel time, $\bar{t}$; see Eq. (9). Let us also denote $\bar{t}_s$ as the sampled travel time mean. The $1 - \alpha$ confidence interval for the average travel time estimate will then verify the following equation:
Then, a confidence interval of $1 - \alpha$ level for $l/t$ will verify:

$$\Pr \left[ \frac{l}{T_{a_n-1} - (1 - \alpha/2) \frac{s_n}{\sqrt{n}} \sqrt{\frac{N - n - \alpha}{N}} \leq \frac{l}{T_{a_n} - (1 - \alpha/2) \frac{s_n}{\sqrt{n}} \sqrt{\frac{N - n}{N}}} \leq \frac{l}{T_{a_n} - (1 - \alpha/2) \frac{s_n}{\sqrt{n}} \sqrt{\frac{N - n}{N}}} \right] = 1 - \alpha. \tag{A.1}$$

Moreover, the width of the confidence interval should not exceed $2m$ units. Thus, the following inequality arises:

$$m \geq \frac{(a_n S_n)^2 \left( \bar{t}^2 + 2(m t_n)^2 + 4(m t_n)^2 \right)}{2m^2 t_n^2}. \tag{A.3}$$

As an approximation, let us also assume that $(N - n) / (N - 1) \rightarrow 1$ (a realistic assumption for heavy traffic conditions and low penetration rates). Then, after using some algebraic manipulations, the minimum sample size, $n_{\text{min}}$, can be obtained as the solution of a radical equation. For the sake of brevity, let us denote the product $T_{a_n^{-1}} (1 - \alpha/2) S_n$ as $a_n$, then:

$$n_{\text{min}} = \frac{(a_n S_n)^2 \left( \bar{t}^2 + 2(m t_n)^2 + 4(m t_n)^2 \right)}{2m^2 t_n^2}. \tag{A.4}$$

The approximate minimum penetration rate required for an accurate estimation of the harmonic mean speed could then be simply obtained by dividing Eq. (A.4) by $N = \Delta t \min \{q, sG/C\}$.

References


