Intersection SPaT Estimation by means of Single-Source Connected Vehicle Data

Majid Rostami-Shahrbabaki*
Institute for Intelligent Transportation Systems
Bundeswehr University Munich, Neubiberg, Germany
E-mail: majid.rostami@unibw.de
ORCiD: 0000-0002-8129-4519

Klaus Bogenberger
Chair of Transportation Engineering
Institute for Intelligent Transportation Systems
Bundeswehr University Munich, Neubiberg, Germany
E-mail: klaus.bogenberger@unibw.de
ORCiD: 0000-0003-3868-9571

Ali Akbar Safavi
Department of Power and Control Engineering
School of Electrical and Computer Engineering
Shiraz University, Shiraz, Iran
E-mail: safavi@shirazu.ac.ir
ORCiD: 0000-0002-2265-8300

Armaghan Moemeni
Institute of Artificial Intelligence (IAI)
Faculty of Computing, Engineering and Media
De Montfort University, Leicester, United Kingdom
E-mail: armaghan@dmu.ac.uk

*corresponding author

Word Count: 6,500 words + 4 tables = 7,500 words

Submitted [31.07.2019]

ABSTRACT

Current traffic management systems in urban networks require real-time estimation of the traffic states. With the development of in-vehicle and communication technologies, connected vehicle data has emerged as a new data source for traffic measurement and estimation. In this work, a machine learning-based methodology for signal phase and timing information (SPaT) which is highly valuable for many applications such as green light optimal advisory systems and real-time vehicle navigation is proposed. The proposed methodology utilizes data from connected vehicles travelling within urban signalized links to estimate the queue tail location, vehicle accumulation, and subsequently, link outflow. Based on the produced high-resolution outflow estimates and data from crossing connected vehicles, SPaT information is estimated via correlation analysis and a machine learning approach. The main contribution is that the single-source proposed approach relies merely on connected vehicle data and requires neither prior information such as intersection cycle time nor data from other sources such as conventional traffic measuring tools. A sample four-leg intersection where each link comprises different number of lanes and experiences different traffic condition is considered as a testbed. The validation of the developed approach has been undertaken by comparing the produced estimates with realistic micro-simulation results as ground truth, and the achieved simulation results are promising even at low penetration rates of connected vehicles.

Keywords: Clustering, Connected vehicle, Machine learning, Traffic light, Signal phase and timing estimation

INTRODUCTION

Many traffic management systems such as dynamic traffic management (DTM) and Advanced Traveler Information Systems (ATIS) require reliable, high-accuracy, and real-time estimates of the current and future traffic states. Traffic signals have been an indispensable part of the current transportation networks and there is no evident proof for a probable change in their functions in the near foreseeable future (*1*). Traffic lights, by continual interruption of traffic flow, strongly impact vehicle movement and consequently fuel consumption in cities. In arterial driving, the complex and unknown switching pattern of traffic signals makes accurate travel time estimation or optimal routing often impossible even with modern traffic-aware in-vehicle navigation systems. The knowledge of the status of traffic lights at urban intersections such as cycle time, signal phase and timing information (SPaT) is valuable for many applications such as traffic management and real-time vehicle navigation.

In an ideal situation in which the SPaT information is available, the drivers can adjust their speed for timely arrival at the green phase. One can expect considerable fuel savings in city driving with such predictive cruise control algorithms via so-called Green Light Optimal Speed Advisory (GLOSA) systems (*2*). Alternatively, the onboard navigation system may suggest an efficient detour that will save the driver from multiple stops and long waits at the upcoming red lights. In practice, however, direct access to the signal timing information and real-time state of the traffic lights may be probably accessible only for signals on a corridor or a small-scale network. Whereas collecting such information for large areas (e.g. region or nationwide) directly from controllers can be very challenging (*3*).

Today, many companies such as Google, TomTom, and BMW collect and use data from vehicles and cellular phone probes. It seems that the most attractive way to estimate signal timing information is via direct usage of the data that they have already collected. Recently, the usage of connected vehicle data (i.e. vehicles equipped with GPS sensors and V2I communication technology) has been received many attentions for the traffic state estimation problems (*4*, *5*).

The focus of this present work is the real-time estimation of SPaT information of a signalized intersection, including cycle time, phase order, and green time of each phase. The main contribution of this paper is that the developed methodology only assumes that connected vehicle data at short sampling time is available. Estimating such information without the need for any other data source provides a straightforward approach for attaining such valuable information which can be used by many traffic applications, as discussed above. The single-source proposed methodology requires no prior knowledge of intersection cycle time or signal phases. In addition, no additional information from other conventional traffic measuring tools such as loop detectors is needed. The methodology is based on, first, estimating the queue tail location at each signalized link approaching the intersection. Then, the high-resolution (second-by-second) estimation of outflows is carried out. Link outflows are indeed a good indicator of the green and red phases of an approach. Since traffic flow and subsequently outflows at each link demonstrate periodic behavior, periodicity analysis of the obtained patterns is an important data mining task which provides useful insights into the physical events at intersections. It is shown that auto- and cross-correlation of outflows reveal important information about traffic light behavior. Learning signal timing information through outflows is carried out via a proper machine learning algorithm. In this paper, spectral clustering, as it is explained in the methodology section, is used.

In summary, the autocorrelation of outflows is used for cycle time and green time estimation. Dominant peaks of the autocorrelation of each continuous outflow signal reveal the main period of the outflow signal, and consequently, the cycle time of the intersection. The dominant peaks are clustered form all peaks via spectral clustering. Spectral clustering is also used for clustering the pulse widths of quantized outflows which are used for green time estimation. Phase order estimation is carried out via cross-correlation analysis of outflows.

The rest of the paper is organized as follows. In the next section, a review of related works is provided. The section following that presents necessary methodologies for estimation of queue tail location, vehicle accumulation within the queue, link outflow, and SPaT information. Afterward, the results of the conducted simulation are presented to demonstrate the efficiency and accuracy of the developed estimation

algorithms. Several key issues concerning practical considerations are carefully investigated. Finally, the main conclusions are summarized, and future works are outlined.

RELATED WORKS

Early research on signal timing estimation implemented some supplementary equipment in order to learn traffic or drivers' behavior. In (*6*) signal phases are estimated from the delay patterns using sampled travel times. Data for estimation is provided by two consecutive sets of wireless traffic sensors installed upstream and downstream of the subject intersection. In another work, a software service, namely SignalGuru, that relies on the collection of images captured by mobile phones to detect and predict the traffic signal schedule, is introduced (*7*). Wang and Jiang (2012) focused on the estimation of traffic signal phases by floating car data. Given a signalized intersection, by approximating and analyzing the velocity-time curves of all vehicles passing by this signal, the cycle length and the green time of each movement were calculated.

The possibility of using intersection travel times, i.e., those collected between upstream and downstream locations of an intersection, to estimate signal timing parameters is also studied in (*3*). The proposed method is a combination of traffic flow theory and estimation algorithms and it is shown that produces promising results for a relatively high penetration rate of travel time data. By letting drivers share their velocity profiles in a digital cloud, and in return benefiting from smart algorithms evaluating the collected data, an algorithm in (*9*) is presented which provides information like traffic light phase schedules for the drivers. If the exact cycle time of a traffic signal is known, then there is a minimum necessary historical data of vehicles trajectories which makes the learning of the cyclic plans of the pre-timed traffic lights possible (*10*, *11*). The procedure is done by projecting trajectories into the range of the cycle length which actually requires a huge data history and the result would be completely inaccurate if the cycle length is unknown or has a small error (*12*, *13*).

In (*14*), GPS data of smartphones on vehicles are used to detect stop and go events of vehicles, and accordingly movement features of vehicles. Then, by adopting the shockwave techniques, signal timings are inferred. The feasibility of estimating traffic signal phase and timing from statistical patterns is also explored in (*1*) in which a few days of low-frequency bus data traversing fixed-time traffic lights are utilized. The traffic light information is estimated in (*15*) by equipping an intersection with magnetic vehicle detectors at the stop bar, at advance locations, and in the departure lanes, which provide an accurate count of turns. The analysis uses two months of data including 36,000 cycles. A cycle length analysis approach based on GPS trajectory data is proposed in (*16*), which was validated by a set of 7,000 taxi GPS trajectories travelling for two days.

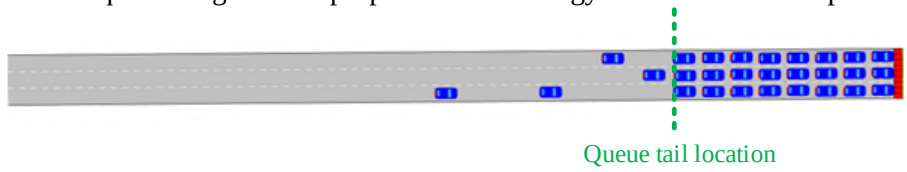
The scope of this paper is developing a methodology for SPaT information estimation, more precisely, estimation of cycle time and phase order of a traffic light as well as green time for each driving direction. Since traffic flow at signalized intersections has some periodic features, the proposed approach learns the periodic pattern of each link outflows by the combination of correlation analysis and spectral clustering algorithm. The outflows are estimated only based on the availability of connected vehicle data. This means that, unlike some similar works, it is a single-source data approach which does not need additional information form loop detectors or other traffic sensing tools such as magnetic detectors. In addition, prior knowledge of the intersection traffic light such a cycle time or phase order is not required. It is shown that even for a one-lane link, the results are still trustworthy. This is promising since it suggests the possibility of green time estimation of the dedicated lanes of the main streams. Another significant contribution of this paper over similar works is that it requires a relatively short time horizon for data collection. The conducted simulation lasts for just one hour and produced estimates are in fact promising. This fact indicates that this approach can be applied for adaptive control strategies with a small moving window even though the considered signal control in this paper is fixed-time. The proposed scheme is thoroughly tested and demonstrated in a realistic microscopic simulation environment with different penetration rates of connected vehicles. The performance of the developed estimation scheme is tested and validated under different conditions through realistic simulations using the AIMSUN (*17*) micro-simulator as ground truth.

METHODOLOGY

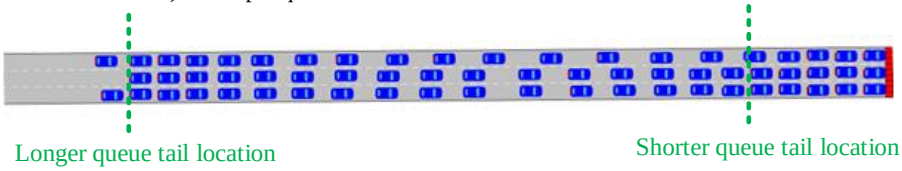
A procedure based on the following sequences is proposed for the SPaT information estimation of a signalized intersection. First, sufficient data (location and speed of connected vehicles) at each sampling time $k = 1, 2, ...,$ where we have for the time $t = kT$, with T the sampling period (e.g. $T = 1\ s$) is collected. Then, based on such time-stamped information, queue tail location, vehicle accumulation and outflow of each signalized link at corresponding intersection are estimated. High-resolution (second-by-second) outflow estimates are further enhanced using some heuristics from crossing information of connected vehicles and are then quantized to discrete values to comply with the discrete green/red traffic signals. Since outflows feature periodic behavior, their auto- and cross-correlation are used along with a machine learning approach for the estimation of the traffic light cycle time and the green time of each approach. The methodology for each step is described below in more details. The algorithms described in the following subsections use a relatively short time horizon which makes it applicable and producing valid estimates not only for the fixed-time but also for the actuated and adaptive signal controls.

Queue Tail Estimation

The first step of the necessary procedure is the estimation of the queue tail location. For this preliminary stage, the algorithm developed by authors is utilized (*18*). Unlike the previous study, where the main goal was the estimation of the number of vehicles within the queue, the longest queue tail as the target estimate was considered, here we consider the tail of the shortest sub-queue. For the simple queue formation, as occurs in undersaturated traffic conditions, both approaches deliver the same result. While the situations occurring in oversaturated traffic conditions, due to the possible existence of multiple sub-queues downstream of the principal queue tail, especially when the link length is relatively high, is different. Illustrative comparison of both cases in a sample signalized link is depicted in Figure 1. Since the average speed of connected vehicles is used for the outflow estimation, some moving vehicles between a former queue and a new queue may produce some wrong nonzero outflows during the red phase of the traffic light. Thus, the shortest sub-queue length in the proposed methodology is considered as depicted in Figure 1.



a) Simple queue formation in an undersaturated traffic condition



b) Mutiple queue formation in an oversaturated traffic condition

Figure 1. Single and multiple sub-queue formations. In a), due to undersaturated traffic condition, one queue exists and thus, one queue tail can be estimated, while in b), two sub-queues exist and the shorter queue tail location is considered in this paper while in (18), the longer one was used.

To this end, at every time step, the measurements from connected vehicles are appropriately treated. Specifically, based on a defined velocity threshold, any connected vehicle is assigned to the group of (virtually) stopped or the group of moving vehicles. The distance of stopped connected vehicles from the stop bar is, in fact, the criterion for queue estimation. The farthest stopped connected vehicle in the shortest sub-queue is obviously closest to the corresponding sub-queue tail and may be considered to provide a first rough estimate of the queue tail length. This first rough queue tail estimate, L_q, is calculated as follows:

$$L_q = \max_i(d_i)\ , \quad \text{where } i \in I = \{n | v_n \leq v_{min}\} \quad \text{for} \quad n = 1, \dots, N_c \tag{1}$$

where v_{min} is the speed threshold that designates vehicles to either stopped or moving groups, N_c is the number of connected vehicles in the sub-queue, and d_i and v_i are the distance of the i^{th}-connected vehicle

measured from the downstream end of the link and its corresponding speed, respectively. Obviously, in low penetration of connected vehicles, this may be an underestimate, since there may be farther, non-connected, vehicles queuing behind the last connected vehicle.

This rough estimate of the queue tail location may be elaborated further, using probability considerations, to improve estimation accuracy and robustness, particularly in situations with a low penetration rate of connected vehicles as explained in (*18*). The probabilistic approach developed by the authors partially compensates for such misidentification of the queue tail and render the algorithm more robust to low penetration rates and hence more practical. It was shown that adding the mean value of the derived probability distribution function of the queue tail dislocation error to equation (1) produces a bias-free estimation with minimum variance for the final resulting estimation error.

Outflow Estimation

Having estimated the queue (or sub-queue) tail, the number of vehicles in the queue (or sub-queue), or equivalently vehicle accumulation, can be estimated subsequently. Fortunately, in the presence of connected vehicles, the average speed downstream of the queue tail is available from corresponding vehicle reports and may be used for proper estimation. A physical model based on traffic flow theory proposed in (*18*) is used here. Equation (2) for vehicle accumulation estimation $\widehat{N}(k)$, only requires two simple tuning parameters with physical interpretation, i.e., average headway of queuing vehicles L_v, and the queue wave speed A. Such values can be reasonably adjusted even without elaborative tuning (*19*).

$$\widehat{N}(k) = \frac{\lambda A \hat{L}_q(k)}{L_v(v(k)+A)} \quad (2)$$

where $\hat{L}_{qi}(k)$ is the estimate of the queue tail after compensation of the queue tail dislocation error, λ is the number of lanes, and $v(k)$ is the average speed of connected vehicles inside the queue. The next step is the estimation of the link outflow. To this end, we assume that traffic conditions in the downstream of the queue tail are homogenous, and thus, the outflow of each section, $\hat{q}(k)$, is calculated as:

$$\hat{q}(k) = \widehat{N}(k)v(k). \quad (3)$$

The high-resolution estimates produced by (3) may not be of enough accuracy compared to ground truth outflow values but definitely demonstrate the periodic characteristic inherited from signal timing patterns. Thus, a learning-based approach may capture these periodic features and estimates the required SPaT information. Before utilizing these outflow estimates, it is also possible to enhance them by additional information from crossing connected vehicles. This complementary step is explained in the next section.

Outflow Estimation Enhancement and Quantization

Apart from location and speed of all connected vehicles within the link which are already used for estimations in the previous sections, one may use the information of each specific connected vehicle passing the stop bar as a relevant indicator of green time. Since signal timing in the proposed methodology is estimated based on the outflow estimates, zero outflows may deliver a non-real red phase. Specifically, it may happen during a long green phase. If a queue is completely dissipated during a green phase, and hence, no outflow is estimated via equation (3), this may end up to a shorter green time estimate. This underestimation of green time can be mitigated via direct use of speed and acceleration information of connected vehicles passing the intersection at such corresponding times. To this end, a Virtual Trip Line (VTL) (*20*) is placed at the downstream end of the link which records the speed of traversing connected vehicles. The acceleration, a_i, can be retrieved then, based on the two consecutive values of the speed of the corresponding vehicle. Some heuristics, as explained below, are used along with such data to improve the outflow estimates. Three thresholds for speed as high S_h, medium S_m, and low S_l and two thresholds for acceleration as high A_h, and low A_l, are considered. GT_{min} is assumed as the minimum green time of each approach which is common among traffic control engineering (*21*).

- If a connected vehicle i passes the VTL at time t with a speed value close to the free-flow speed, i.e., $v_i(t) > S_h$ and with acceleration $a_i(t) < A_l$, thus, it indicates a free flow stream which is a sign of a long green time. Therefore, one may say that the signal time before the passing time t was green for a period of time ΔT (e.g., $\Delta T = GT_{min}$).

- If the speed of a connected vehicle traversing the VTL at time t is between the high and medium value, i.e., $S_h < v_i(t) < S_i$, and corresponding acceleration is low, i.e., $a_i(t) < A_l$, it may probably be at the middle of the green time, and hence, there is a period of green time before and after the passing time t (e.g., $\Delta T = \pm GT_{min}/2$).
- If a vehicle passes the VTL with a speed value of $S_l < v_i < S_m$ and at the same time, the corresponding acceleration is high, i.e., $a_i > A_h$, then it indicates an early acceleration of a stopped vehicle at the begging of the green time. Hence, one may say that a certain period of time ΔT after t is also green (e.g., $\Delta T = GT_{min}$).

In all of the above cases, the outflow during the period ΔT is set to half of the saturation flow rate. Simulation results indicate that this outflow enhancement certainly improves and facilitates the learning approach for determining the cycle time and greet time of the traffic signal. For green time estimation in this work, the continuous values of outflow may not be in our interest. Therefore, instead of the continues values of outflow, a threshold-based quantization of outflow is used where the continuous values of outflows are projected to either zero or one as the new discrete values.

Spectral clustering

As stated earlier, a machine learning approach is utilized for learning the periodic characteristics of outflow estimates. Machine learning is the scientific study of algorithms and models based on sample data which provides systems the ability to automatically learn and improve from experience without being explicitly programmed (*22*). Since the provided database for machine learning, i.e., outflow estimates, contains only inputs and no desired output labels, an unsupervised clustering machine learning algorithm is utilized.

Different clustering techniques make different assumptions on the structure of the data, often defined by some similarity metrics. Spectral clustering, which is found to be more suitable for our data, is a technique with roots in graph theory. The approach is used to identify groups of nodes in a graph based on the edges connecting them which is, in fact, identifying groups of "similar behavior" in the dataset. Spectral clustering uses information from the eigenvalues (spectrum) of special matrices built from the graph or the data set. Considering c as the number of clusters, the algorithm for spectral clustering is explained as follows (*23*):

- Given a set of data points $x_1, \dots, x_n$, a similarity graph $G = (V, E)$ whose each vertex v_i represents a data point x_i is developed. Two vertices are then connected with the edge e_{ij}, if the similarity $s_{ij} = s(x_i, x_j)$ between the corresponding data points x_i and x_j is positive or larger than a certain threshold. The corresponding edge is consequently weighted by s_{ij}.
- Having the adjacency matrix with entries representing the edge weights of graph G, the normalized graph Laplacian $L \in \mathbb{R}^{n \times n}$ is then calculated.
- The next step is constructing the matrix $U \in \mathbb{R}^{n \times k}$ with columns corresponding to the first c normalized eigenvectors of L.
- The final step is performing the k-means clustering algorithm on row data of U and assigning them to clusters $C_1, \dots, C_c$.

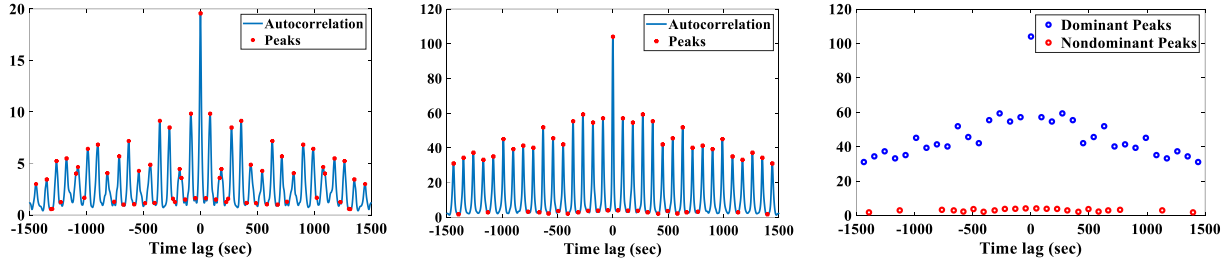
The usage of the spectral clustering for learning periodic features of traffic signals is explained in the next section.

Spat Information Estimation

SPaT information that is estimated in this paper includes intersection cycle time, phase sequences, and green time of each signalized link. Such information provides a full and realistic image of signal timings based on a relatively short history of connected vehicle data. As already explained, link outflows are estimated using connected vehicle data. The outflows are periodic time series which inherently contain the signal timing information. Retrieving such information form a periodic signal can be done via time-domain or frequency-domain analyses. To proceed, correlation analysis of outflow signals along with a clustering approach is utilized for the SPaT estimation.

Cycle Time Estimation

In mathematics, autocorrelation is a representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals. The autocorrelation sequence of a periodic signal has the same cyclic characteristics as the signal itself. Thus, autocorrelation can help to verify the presence of cycles and determining their durations. Hence, the autocorrelation of outflows with sufficient delay time is computed. At such time lags that the link outflow demonstrates high similarity with its delayed version, the autocorrelation has a high value. The highest peak would be consequently at zero lag.



a) Autocorrelation of outflow estimates

b) Autocorrelation of enhanced outflow estimates

c) Clustering peaks of the autocorrelation signal

Figure 2. Autocorrelation of a sample outflow signal where peaks show its high similarity with the delayed version

Autocorrelation of a sample outflow is shown in Figure 2.a. Any red peak $p_i(x_i, t_i)$ represents a time lag when there is a high similarity between the signal and its delayed version. Since the outflow estimates are not very exact, there are some small peaks that first needed to be filtered out. The peaks should be then clustered into dominant and nondominant peaks. Dominant peaks represent the main period of the signal. As discussed earlier, the outflow estimation can be enhanced via information of connected vehicles crossing the intersection. Autocorrelation of the enhanced estimate of the sample outflow is illustrated in Figure 2.b. As it can be easily seen, a lot of small nondominant peaks are now removed and, in addition, there is a bigger vertical distance between the dominant and nondominant peaks which provides better clustering accuracy. The utilized spectral clustering algorithm clusters the peaks into dominant and nondominant peaks and then filters out the nondominant one as depicted in Figure 2.c. Finally, the cycle time of outflow autocorrelation signal T which is the cycle time of the intersection would be estimated based on the average time differences between two consecutive dominant peaks as:

$$T = \frac{\sum_{i=1}^{P-1} \Delta t_i}{P} \qquad with \quad \Delta t_i = t_{i+1} - t_i \tag{4}$$

where P is the number of dominant peaks and t_i is the corresponding time of p_i.

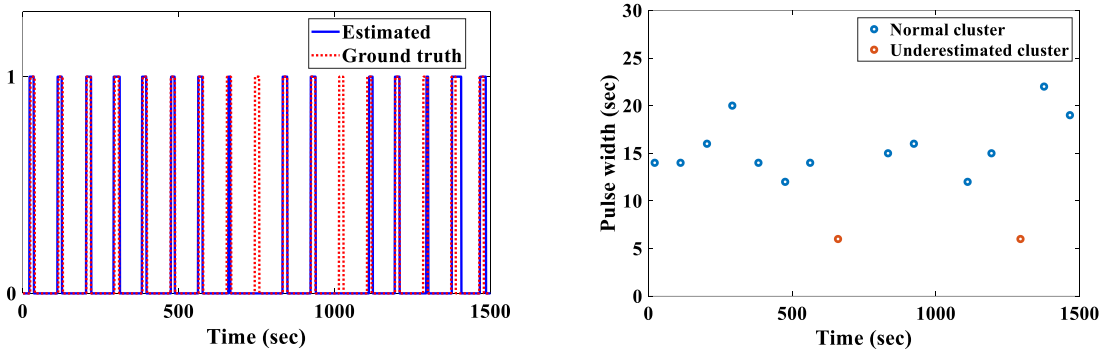
Phase Order Estimation

In the next step, cross-correlation analysis of links outflows is used for phase sequence estimation. In signal processing, cross-correlation is a measure of similarity of two time series as a function of the displacement of one relative to the other. Having estimated the cycle time and restricting the lags equal to $\pm T/2$, the cross-correlation of every two outflows reveals their corresponding delay time. The maximum cross-correlation is achieved at the relative delay time. Since the duration of green times for different links are different, the achieved delay time is not the delay of the starting moment of the green phase (i.e., offset between two green signals). It is actually the difference between the median of two green phases.

Green Time Estimation

The final step in SPaT information estimation is the green time estimation of different approaches. To this end, the quantized time series of outflows based on a threshold, which designates outflow values either to one or zero, is used. Measured pulse widths of such quantized signals is a representation of the corresponding green time. Simulation results show that the rising edge time is well estimated and corresponds to the green phase start time while the falling edge might be under- or over-estimated. In undersaturated traffic conditions, the falling edge of the pulse width might be estimated lower than the actual end of green time, specifically due to the fast dissipation of the queue. On the contrary, in very

oversaturated traffic conditions, one may receive longer green time estimation resulted from the positive average speed of queueing connected vehicles (see Equation (4)). For the considered time horizon, the algorithm estimates the pulse widths for different cycles and the learning algorithm clusters them into two or three clusters, depending on the traffic condition. It should be noted that the traffic conditions can be understood based on the tracing the movement of the queue tail location. Finally, the clusters associated with under- and/or oversaturated traffic conditions are eliminated and the average pulse width of the remaining clusters is calculated which is the estimation of the corresponding approach green time. This procedure for a sample quantized outflow signal is illustrated in Figure 3. As it is evident, the green phases of two cycles are not estimated, due to the lack of any connected vehicles data for such a period of time. However, these zero-length pulse widths are not considered for the estimation and thus have no effect on the final result. Other pulse widths are clustered based on their value and very underestimated values are eliminated from the final estimation as depicted in Figure 3.b.



a) Estimated green time pulses versus the ground truth b) Clustering the pulse widths

Figure 3. Green time estimation results for a sample link

SIMULATION RESULTS AND DISCUSSION

The performance of the proposed methodology is tested in this section using Aimsun (*17*) micro-simulator data as ground truth. For a comprehensive and complete analysis, a four-leg intersection is considered. Time-varying demands are considered to produce different queue lengths and dynamics at each link. In addition, in order to empirically analyze the effect of the number of lanes on the produced estimates, each link has a different number of lanes which encounter different number of connected vehicles at the same penetration rate, comparing to the others. A fixed-time control strategy with four phases and 90 s of cycle time is assumed. Detailed information about the simulation model is given in Table 1. To evaluate the effect of the market penetration rate on the produced estimate, different penetration rates of connected vehicles ranging from 5% to 50% are considered. The intersection model is simulated for one hour with five minutes of warm-up time. At each sampling time, location and speed of all connected vehicles are collected and used for the estimation algorithm. The simulation is run for 10 different replications. The presented results are the average of produced estimates in all replications.

Table 1. Detailed information of the considered intersection

Link number	Number of lanes	Green time (sec)	Offset (sec)	Traffic condition
1	1	10	38	Oversaturated
2	2	14	24	Undersaturated
3	3	24	0	Oversaturated
4	4	42	48	Undersaturated

It is assumed that each link has its own right of way, and thus, there is a distinct phase for each approach. Note that this assumption makes the estimation procedure more complex, specifically for the low penetration rate of connected vehicles. In case that two approaches have the same right of way, and consequently, the same green time, which is also very common at real intersections, the redundancy of information is twice compared to a single-approach case. The concurrent outflow estimates of two links derived from simultaneous connected vehicle data would provide more accurate greet time estimation.

To proceed, the results of SPaT estimation for the considered intersection are presented separately in the following subsections.

Cycle Time Estimation

As explained in the methodology section, the cycle time can be estimated through the autocorrelation of each outflow signal. Autocorrelation of each estimated link outflow signal at each penetration rate is computed and the corresponding dominant peaks are then selected via spectral clustering as depicted in Figure 2. The cycle time is then estimated based on equation (4). Hence, at each penetration rate, four values of cycle time are estimated based on each link outflow. The produced estimates for each link at different penetration rate are given in Table 2.

The results of Table 2 reveal that the cycle time estimation has very high accuracy even at very low penetration rate (i.e., 5%), if the number of lanes is three or more. At links with a lower number of lanes, the produced estimates have very good accuracy for 10% of penetration rate and more. As stated in the Introduction section, if the cycle time is known, then the connected vehicle trajectories can be projected into one cycle time, and thus, related green time can be estimated accordingly. This approach has two main drawbacks: first, it needs a very huge amount of data history; and second, if the cycle time is inaccurate then the produced outcomes are not reliable. Our investigation also showed that even a small error (e.g. 1 s) in the cycle time, makes the trajectory projection completely inaccurate, and thus, produces untrustworthy results. Therefore, even with such relatively precise cycle time estimates as given in Table 2, we did not employ the trajectory projection.

Table 2. Cycle time estimation based on enhanced outflow estimates (sec)

Link Number	Penetration rate%				
	5	10	15	25	50
1	96.24	90.03	90.01	90.00	90.00
2	93.26	89.93	90.00	90.01	90.00
3	89.87	89.87	89.87	89.99	90.00
4	90.15	89.85	90.00	90.00	90.00

Phase Order Estimation

Phase order or phase sequence estimation is carried out based on the cross-correlation of the enhanced outflows. Since the phase order, and not the exact delay time, is required here, the result of the delay for just one penetration rate (i.e., 10%) is given in Table 3. Although the estimated delays are not completely exact, the phase order can be truly retrieved based on the relative delays. It should be noted that if the phase sequence is known, then the offset of each signal can be derived from green times estimates which are produced in the next step.

Table 3. Relative delays of intersection signals based on 10% penetration rate (sec)

Link Number	1	2	3	4
1	0	-14	-34	25
2	14	0	-20	40
3	34	20	0	58
4	-25	-40	-58	0

Green Time Estimation

The final step for the proposed methodology is green time estimation. The measured pulse widths of quantized outflows are used for green time estimation. Pulse widths related to under/over-saturated traffic conditions are filtered out via spectral clustering as shown in Figure 3. The green time for each approach is then the average of pulse width in the selected cluster. Produces estimates for each link at different percentages of penetration rates of connected vehicles are given in Table 4. Comparing the results presented in Table 4 with ground truth indicates that the accuracy improves constantly with the penetration rate increment. For link 1, with just one lane, green time cannot be estimated at low penetration rate since the number of connected vehicles within the link are quite few, and thus, the input data are not sufficient. For link 3, that encounters a highly oversaturated traffic condition, there is a significant error at low penetration

rate as well. The acceptable results are produced with 15% of pentation rate and more. For link 2 and 4 at 10% penetration rate, the acquired estimates feature high accuracy

Table 4. Green time estimates based on enhanced outflow estimates (sec)

Link Number	Penetration rate					Ground truth
	5	10	15	25	50	
1	15.85	15.53	13.33	12.56	11.87	10
2	14.14	13.27	13.86	13.04	14.16	14
3	37.16	35.33	26.32	25.28	23.05	24
4	38.17	40.04	42.50	43.04	41.13	42

It should be noted again that for the real intersections in which usually two approaches have the same right of way, the two concurrent outflows are used for one green time estimation and thus the data redundancy is twice. Thus, there is a potential for more precision at the low pentation rate or for one-lane links such as mainstream dedicated lanes. Considering a real intersection with real-world data is, in fact, the subject of authors' ongoing work.

DISCUSSION AND CONCLUSIONS

With recent technology developments, connected vehicle data has become a new data source for traffic state estimation. In this paper, a single-source learning approach for SPaT information estimation at the urban signalized intersection has been proposed. Based on the location and speed of some randomly distributed connected vehicles, queue length and the number of vehicles within the queue were first estimated. The approach at the second step estimated the link outflows and further enhanced the produced estimates with data from connected vehicles crossing the intersection. Correlation analysis of the enhanced outflow estimates was then used for extracting the periodic behavior of the outflows and a machine learning approach was used afterward for signal timing estimation. It has been shown that the cycle time estimation has very high precision even at a very low penetration rate of connected vehicles. Whereas green time estimation requires higher penetration rates or a greater number of lanes to have sufficient accuracy. For real intersections that two approaches might have the same green time, the redundancy of information is twice, and thus, we expect higher accuracy even for one-lane links at low pentation rates. This is indeed the subject of the authors' future work.

SPaT information, in practice, may not be available for wide areas which makes the proposed single-source estimation approach beneficial for many standalone companies. There is a big potential for such companies to use their own data for signal timing estimation since direct access to the signal timing information and real-time state of the traffic lights may be probably inaccessible in practice

REFERENCES

1. Fayazi, S. A., A. Vahidi, G. Mahler, and A. Winckler. Traffic Signal Phase and Timing Estimation from Low-Frequency Transit Bus Data. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 16, No. 1, 2015, pp. 19–28.
2. van Leersum, J. Implementation of an Advisory Speed Algorithm in TRANSYT. *Transportation Research Part A: General*, Vol. 19, No. 3, 1985, pp. 207–217.
3. Hao, P., X. Ban, K. P. Bennett, Q. Ji, and Z. Sun. Signal Timing Estimation Using Sample Intersection Travel Times. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 13, No. 2, 2012, pp. 792–804.
4. Ramezani, M., and N. Geroliminis. Queue Profile Estimation in Congested Urban Networks with Probe Data. *Computer-Aided Civil and Infrastructure Engineering*, Vol. 30, No. 6, 2015, pp. 414–432.
5. Bekiaris-liberis, N., C. Roncoli, and M. Papageorgiou. Highway Traffic State EstimationWith Mixed Connected and Conventional Vehicles. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 17, No. 12, 2016, pp. 3484–3497.
6. Ban, X. (Jeff), R. Herring, P. Hao, and A. M. Bayen. Delay Pattern Estimation for Signalized Intersections Using Sampled Travel Times. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2130, No. 1, 2010, pp. 109–119.
7. Koukoumidis, E., and M. Martonosi. SignalGuru: Leveraging Mobile Phones for Collaborative Traffic Signal Schedule Advisory. *ACM MobiSys*, 2011, pp. 127–140.
8. Wang, C., and S. Jiang. Traffic Signal Phases' Estimation by Floating Car Data. *2012 12th International*

Conference on ITS Telecommunications, ITST 2012, 2012, pp. 568–573.

9. Kerper, M., C. Wewetzer, A. Sasse, and M. Mauve. Learning Traffic Light Phase Schedules from Velocity Profiles in the Cloud. *2012 5th International Conference on New Technologies, Mobility and Security - Proceedings of NTMS 2012 Conference and Workshops*, 2012, pp. 1–5.
10. Protschky, V., C. Ruhhammer, and S. Feit. Learning Traffic Light Parameters with Floating Car Data. *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, Vol. 2015-Octob, 2015, pp. 2438–2443.
11. Protschky, V., S. Feit, and C. Linnhoff-Popien. On the Potential of Floating Car Data for Traffic Light Signal Reconstruction. *IEEE Vehicular Technology Conference*, Vol. 2015, 2015, pp. 1–5.
12. Axer, S., and B. Friedrich. A Methodology for Signal Timing Estimation Based on Low Frequency Floating Car Data: Analysis of Needed Sample Sizes and Influencing Factors. *Transportation Research Procedia*, Vol. 15, 2016, pp. 220–232.
13. Axer, S., and F. Pascucci. Estimation of Traffic Signal Timing Data and Total Delay for Urban Intersections Based on Low Frequency Floating Car Data. *Proceedings of the 6th mobility TUM (2015)*, No. July, 2015.
14. Chuang, Y. T., C. W. Yi, Y. C. Tseng, C. S. Nian, and C. H. Ching. Discovering Phase Timing Information of Traffic Light Systems by Stop-Go Shockwaves. *IEEE Transactions on Mobile Computing*, Vol. 14, No. 1, 2015, pp. 58–71.
15. Ibrahim, S., D. Kalathil, R. O. Sanchez, and P. Varaiya. Estimating Phase Duration for SPaT Messages. *IEEE Transactions on Intelligent Transportation Systems*, Vol. PP, 2018, pp. 1–9.
16. Yang, Q., J. Yu, and J. M. Han. Traffic Signals Timing Cycle Length Learning: Using Taxi GPS Trajectories. *Proceedings - International Conference on Machine Learning and Cybernetics*, Vol. 1, No. 1, 2018, pp. 13–18.
17. Barceló, J., and J. Casas. Dynamic Network Simulation with AIMSUN. *Simulation approaches in transportation analysis*, 2005, pp. 57–98.
18. Rostami Shahrbabaki, M., A. A. Safavi, M. Papageorgiou, and I. Papamichail. A Data Fusion Approach for Real-Time Traffic State Estimation in Urban Signalized Links. *Transportation Research Part C: Emerging Technologies*, Vol. 92, No. November 2017, 2018, pp. 525–548.
19. Diakaki, C., V. Dinopoulou, K. Aboudolas, M. Papageorgiou, E. Ben-Shabat, E. Seider, and A. Leibov. Extensions and New Applications of the Traffic-Responsive Urban Control Strategy: Coordinated Signal Control for Urban Networks. *Transportation Research Record*, Vol. 1856, 2003, pp. 202–211.
20. Hoh, B., M. Gruteser, R. Herring, J. Ban, D. Work, J.-C. Herrera, A. M. Bayen, M. Annavaram, and Q. Jacobson. Virtual Trip Lines for Distributed Privacy-Preserving Traffic Monitoring. 2008.
21. Koonce, Peter and Rodegerdts, L. *Traffic Signal Timing Manual (No. FHWA-HOP-08-024)*. United States. Federal Highway Administration, 2008.
22. Shalev-Shwartz, S., and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press, 2014.
23. Von Luxburg, U. A Tutorial on Spectral Clustering. *Statistics and Computing*, Vol. 17, No. 4, 2007, pp. 395–416.