Vehicle index estimation for signalized intersections using sample travel times

Peng Hao, Zhanbo Sun, Xuegang (Jeff) Ban, Dong Guo, Qiang Ji

Department of Civil and Environmental Engineering, Rensselaer Polytechnic Institute, 110 Eighth Street, Troy, NY 12180-3590, USA
Department of Computer Science, University of Southern California, 900 W 34th Street, Los Angeles, CA 90089, USA
Department of Electrical and Systems Engineering, Rensselaer Polytechnic Institute, 110 Eighth Street, Troy, NY 12180-3590, USA

Article history:
Received 21 May 2013
Accepted 30 June 2013

Abstract
We introduce in this paper the concept of vehicle indices in a cycle at a signalized intersection which are the positions of vehicles in the departure process of the cycle. We show that vehicle indices are closely related to the vehicle arrival and the departure processes at the intersection. Based on vehicle indices and sample travel times collected from mobile sensors, a three-layer Bayesian Network model is constructed to describe the stochastic intersection traffic flow by capturing the relationship of vehicle indices, and the arrival and departure processes. The non-homogeneous Poisson process and log-normal distributions are used respectively to model the stochastic arrival and departure processes. The methods of parameter learning and vehicle index inference are presented based on the observed intersection travel times. Simplification to the methods is discussed to reduce the computational effort of parameter learning and vehicle index estimation. The model is tested using data from NGSIM, a field test, and simulation with reasonable results.

1. Introduction and motivation
Advances in wireless communications and the wide deployment of mobile traffic sensors (those that move with the flow they are monitoring such as Global Positioning System (GPS) or smart phones) and other types of tracking devices have induced a heightened interest in the transportation field. Mobile sensors provide an important alternative way for traffic data collection. Compared with fixed-location sensor data (such as loop detector data), mobile data have some unique characteristics in that they usually provide a sample of the entire traffic flow. Such difference has inspired the recent developments of new modeling methods to effectively use mobile data for freeway traffic modeling (Leclercq et al., 2007; Herrera et al., 2010; Lu and Skabardonis, 2007) and arterial traffic modeling (Comert and Cetin, 2008; Ban et al., 2009, 2011; Cheng et al., 2010; Hao et al., 2012).

The sampling nature of mobile data, i.e., mobile data usually represent only a portion of the traffic flow, brings new challenges in especially arterial modeling such as arterial performance measurements: most existing methods based on fixed-location sensors assume volume, occupancy, or gap as the input which cannot be obtained directly from mobile data. In the past, certain assumptions have been used, based on which characteristics of the entire traffic flow can be reconstructed using mobile data. One example of this is the uniform arrival assumption that was used in Ban et al. (2009, 2011) and Cheng et al. (2010) to estimate real time signalized intersection performance measures. However, because such assumption cannot...
describe real world arterial traffic well, relatively large estimation errors could be possible as reported in Ban et al. (2011) for real time signalized intersection queue length estimation. The uniform assumption is also deterministic and ignores the stochasticity of arterial traffic flow, which could be very significant. Sun and Ban (2013) proposed a variational formulation (Daganzo, 2005) based method to estimate the traces of unsampled vehicles from traces of the sample vehicles (those equipped with mobile sensors) around a signalized intersection. The method is also purely deterministic. Another way to deal with the sampling issue is to model the penetration rate of mobile data explicitly (Argote et al., 2011). By assuming a given penetration rate, Comert and Cetin (2008) proposed a sampling based statistical model to estimate the queue length distribution based on probe vehicle data. The method attempted to model average signal performance instead of real time performance as studied in this paper.

The difficulty of inferring information of the entire traffic flow from mobile data is further exacerbated by other associated challenges. For example, the collection and use of mobile data have been associated from the very beginning with privacy concerns. To protect privacy, we cannot simply assume that we can collect and use detailed traces of sample vehicles for long period of time or long distance. This restricts the mobile data elements that can be used for extracting useful information. Furthermore, arterial traffic flow could be highly stochastic and dynamic. Most existing methods based on deterministic assumptions may not work well for such situations. In this paper, we aim at addressing those issues by (i) using intersection travel times from sample vehicles only which has shown to be able to protect privacy (Hoh et al., 2008) while at the same to be very useful for estimating signal performances (Ban et al., 2009, 2011; Sun et al., 2013) and (ii) developing a stochastic learning framework to capture the stochasticity and dependence of the difference processes of arterial traffic flow.

To better link the sample mobile data and the entire traffic flow, we propose a statistical learning approach to estimate the vehicle indices of sample vehicles obtained from mobile sensors. Vehicle index is defined as the position of a (sample) vehicle in the departure flow of a cycle. It indicates the sequence of the vehicle, in the entire traffic flow, when counted from the first discharged vehicle in a cycle. Here the beginning of a cycle is defined as the start of the red time. For mobile data, two consecutively sampled vehicles may not have consecutive vehicle indices if there are unsampled vehicles in between. However, knowing the indices of two sample vehicles in the same cycle, one can acquire the number of unsampled vehicles and the average arrival rate between them. Therefore, if the vehicle indices of all sample vehicles in a cycle can be accurately estimated, the linkage between mobile data and the entire traffic flow can be better established. As a result, this may make it possible to apply existing arterial traffic modeling methods to mobile sensor data. There are also other important considerations for why vehicle index is useful. For example, vehicle index is an important performance measure for an intersection: The index of a queued vehicle is also the position of the vehicle in the queue; if a vehicle is the last queued vehicle in a cycle, its index will be exactly the maximum queue size in the cycle. Therefore, vehicle indices of sample vehicles in a cycle can be considered as the basic information one can obtain/estimate from mobile data directly. Based on vehicle index information, other performance measures such as delays or queue lengths can also be estimated.

To estimate vehicle indices, we focus on the vehicle arrival and departure processes since traffic states or performance of a signalized intersection are largely determined by the arrival and departure processes, as well as the signal timing. Vehicle indices play a central role in bridging the arrival and departure processes at a signalized intersection. The key for vehicle index estimation is therefore how to properly integrate the two processes in an integrated framework by considering their stochastic nature.

For the arrival process, (homogeneous) Poisson distribution has been widely used before. However, for urban closely spaced intersections, arrivals do not follow Poisson well due to the frequent occurrence of platoons. In this paper, we adopt the non-homogeneous Poisson process (NHPP) recently developed in Kuhl et al. (2010). NHPP is a time dependent Poisson process, i.e., its arrival rate is a time varying function $\lambda(t)$, instead of a constant value $\lambda$ as for the homogeneous Poisson. More importantly, NHPP can consider both short-term variations and long-term periodic trends, which makes it suitable to deal with the periodic vehicle arrivals due to signals at nearby upstream intersections and/or traffic demand variations across the time of a day. The departure process at a signalized intersection is shown to be more stable (Roess et al., 2010, Chapter 17; Liu et al., 2009). Jin et al. (2009) presented a set of log–normal distributions to model the departure headways at each vehicle index (except the first vehicle). After validating the log–normal distribution using real data, we adopt it in this paper to model the departure process; the headway of the first vehicle is modeled as a normal distribution.

To systematically link the processes of vehicle arrivals and departures at traffic signals, especially to better represent the stochastic dependencies of these two processes and their relationship to vehicle indices, we develop a Bayesian Network (BN) model to estimate vehicle indices in this paper. BN is a stochastic learning approach (Jordan and Weiss, 2002; Neapolitan, 2003), which has already been used before to characterize traffic flow patterns and to model the temporal dynamics and spatial relationships in transportation networks. For example, Zhang and Taylor (2006) modeled the intuitive causal relationships between traffic pattern variables such as incidents and volumes, and introduced the dynamic BN to model the evolution of traffic patterns at different times. Sun et al. (2006) proposed a Gaussian BN framework to relate the traffic flow among adjacent road links. While the above references applied fixed–location sensor to BN, Herring et al. (2012) proposed a Hidden Markov Model, similar to the BN model, of the spatial congestion correlation between neighboring roads from a sparse GPS probe dataset of vehicle travel traces. Hofleitner et al. (2012) developed a BN-based arterial travel time prediction model which assumed that the intersection delay satisfied the uniform distribution. No attempt, to the best of the authors’ knowledge, has been done to link the stochastic arrival and departure processes of a signalized intersection using the BN method, especially in the real-time manner.
Here we notice that there have been recent research efforts to develop stochastic traffic flow models, noticeably the stochastic cell transmission models (CTM) to capture the stochasticity of traffic flow; see Sumalee et al. (2010), Jabari and Liu (2012). Such models largely rely on fixed-location sensor data and how to make them applicable to mobile data is still an open question. The BN model proposed in this paper is tailored directly to use mobile sensor data, particularly, the intersection sample travel times.

The proposed BN is a three layer model that integrates the vehicle arrival and departure. An index dependent sampling process is also integrated into the model to capture the penetration rate of the mobile data. All the conditional dependencies between different processes in the network are quantified with defined stochastic models. Solving the BN model involves two major steps: learning and inference. In parameter learning, the maximum likelihood method is used to estimate the parameters of the stochastic distributions of the arrivals (i.e., the NHPP distribution) and departures (i.e., log–normal and normal distributions). In particular, for the arrival process, we propose a constant global index difference method to approximate the NHPP. In the inference step, the method of calculating the conditional probability of the vehicle index combination is derived. Based on this method, the distribution of the vehicle index of a sample vehicle can be calculated. A most likely, deterministic vehicle index estimate can be derived using the Most Probable Explanation (MPE) and marginal distribution. A simplified model is also presented as an alternative to reduce the computational effort. The simplified model decomposes the original BN into sub-networks based on the unique characteristics of traffic flow during the departure process from a traffic signal. The BN-based index estimation model and its solution algorithm are tested using data from NGSIM, a field experiment, and microscopic traffic simulation, with reasonable results.

2. Preliminaries

In this section, terminology and basic assumptions of the model are first given.

2.1. Terminology

- **Sample vehicle**: a vehicle equipped with mobile sensors.
- **Virtual trip lines (VTL)**: a virtual location, upstream or downstream of a signalized intersection, to collect mobile data. In this paper, the only mobile data that are collected are the times that sample vehicles pass the VTLs, based on which sample intersection travel times can be derived (Hoh et al., 2008). The upstream and downstream VTLs are called VTL1 and VTL2 respectively.
- **Arrival time**: the time when a vehicle passes the upstream VTL.
- **Departure time**: the time when a vehicle passes the downstream VTL.
- **Local departure time**: time elapsed between the beginning of the green phase and a vehicle’s departure time.
- **Vehicle index**: the position of a vehicle in the departure sequence during the green time of a cycle. If a vehicle arrives and departs in different cycles under over-saturation conditions, the index of the vehicle is defined as its position in the departure process. Note that we group vehicles to each cycle based on their departure times instead of the arrival times. That is mainly because the departure process is much more stable.
- **Queued vehicle/free flow vehicle**: defined based on the vehicle’s delay experienced at the intersection. A vehicle is considered “queued” if its intersection travel time between VTLs exceeds a pre-defined threshold, called the minimum traverse time (MTT); see Ban et al. (2011). Otherwise, the vehicle is a free flow vehicle.

2.2. Assumptions

1. Each vehicle is sampled independently with a known penetration rate. If the penetration rate is unknown, we can estimate it using sample travel times. Details are provided in Appendix A.
2. VTL1 is far enough from the stop line so that queue never extends beyond VTL1. In this case the arrival time is not affected by the signals and queues in the intersection. This is the so-called “short-queue” scenario (Liu et al., 2009). Notice that since VTL is virtual and can be moved easily if needed, this assumption is not too unrealistic unless the approach is very congested so that the upstream intersection is blocked (i.e., when spillover happens).
3. There is no vehicle overtaking between VTL1 and VTL2, i.e., vehicles follow first-in-first-out (FIFO) between the two VTLs. Also “no-turn-on-red” is strictly enforced. This is generally true for single-lane streets. For multi-lanes streets, this assumption is not too unrealistic since the distance between VTL1 and VTL2 is usually very short (a few hundred feet). If overtaking does occur, e.g., a sample vehicle arrives after another sample vehicle, but departs before the second vehicle, we swap the arrival times of the two vehicles to ensure the FIFO rule. As a result, vehicle index on arrivals is the same as the index upon departure. This scheme is validated using the simulation data, in which the cycles with overtaking have nearly the same performance in index estimation as the other cycles if time-swapping is applied. Details are omitted here due to space limitation.
4. Arterial traffic flow satisfies certain Markov condition (Neapolitan, 2003, Chapter 1), which ensures that the BN approach can be used. The exact meaning of this assumption to arterial traffic flow will be discussed in more detail in Section 4.
3. Vehicle-index-dependent arrival and departure distributions

In this section, we show that the arrival and departure processes of a signalized intersection can be represented using vehicle-index-dependent distributions. First, Fig. 1 depicts the vehicle traces of a typical cycle in NGSIM data (Cambridge Systematics, 2007). The intersection under study is 500 ft apart from the upstream intersection. The traces of sampled and unsampled vehicles (at a 50% penetration rate) are shown using solid lines and dotted lines respectively. The two horizontal solid lines indicate the location of the upstream and downstream VTLs. The red time (denoted as a thick solid line) and the green time (denoted as a double line) are also illustrated at the location of the stop line. The indices of sample vehicles are labeled in the figure. This particular cycle has 13 vehicles with 8 of them sampled, i.e., their arrival and departure times are collected at the VTLs.

The figure shows that the arrival headways can have a very large variance. For example, the arrival headway between the 1st and 2nd vehicles in the cycle is about 6 s, which is three times as long as the headway between the 2nd and 3rd vehicle (about 2 s). The significant non-uniform characteristic of arrival flow can be explained by platoon dispersion models (Geroliminis and Skabardonis, 2005; Bonneson et al., 2008). Here we denote the upstream link as the link that connects the upstream intersection and the subject intersection. The traffic flow that enters the upstream link is significantly platoonized when queues are discharged with tight headways from the upstream intersection. The platoon spreads out by time due to lane changing behavior and the fluctuation of vehicle speeds. If two intersections are spaced far enough apart (i.e. more than a mile, see Bonneson et al., 2008), traffic arriving at the study intersection can be approximately considered as stationary or uniform. However, for close spaced intersections, e.g., the example in Fig. 1, the platoon is not completely dispersed. The average arrival headway of vehicles in the same platoon is much smaller than that of the vehicles between two different platoons. Thus, the homogeneous Poisson process cannot be directly applied to this case. Furthermore, mobile data cannot identify the exact arrival pattern between two sample vehicles, which could be free flow, in a platoon, or a mixture of the two. As a result, neither the homogeneous Poisson nor the log–normal distributions (which was used to describe the headway distribution within a platoon) can properly capture such uncertain arrival patterns between two sample vehicles. The non-homogeneous Poisson process (NHPP) is thus selected because it has a time-dependent arrival rate, and has higher variance that can accommodate the uncertainty in traffic arrivals. Details are given in Section 4.4 later.

According to the definition of queued vehicles (based on delays) in Section 2.1, there are 9 queued vehicles and 4 free flow vehicles in the cycle in Fig. 1. The departure time of a free flow vehicle is completely determined by its arrival time by adding a constant increment (i.e., the MTT) to its arrival time. By contrast, the departure time of a queued vehicle is affected by not only its arrival time but also the signal timing and queue. As can be observed, the departure headways of queued vehicles are usually small and stable. The index-dependent log–normal distributed (Tolle, 1976; Jin et al., 2009) is thus adopted here to model the headway distribution between two consecutive queued vehicles. This means that departure headways in different cycles have the same distribution for the same vehicle index. The departure headway between two non-consecutive queued vehicles consists of a few independent log–normal variables. It does not follow a log–normal distribution exactly, but can still be approximated as a log–normal distribution (Beaulieu et al., 1995). Although the headway distribution between queued vehicles can be well defined, there is no existing statistical model to describe the local departure time of the first queued vehicle. Same problem occurs when we estimate the local departure time of the first sample vehicle using mobile data. In this paper, based on the simulation and field test, we approximate the local departure time of the first sample queued vehicle using an index dependent Gaussian distribution.

Fig. 1. Vehicle trajectories in NGSIM data.
Since both arrival and departure times contain vehicle index information, the index is a bridge to connect the arrival and departure processes. Or reversely, the index can be estimated if both arrival and departure data can be properly utilized. A BN is developed for this purpose, as formally discussed in the next section.

4. Bayesian network model construction

4.1. Basics of a Bayesian network

A Bayesian network is a probabilistic graphical model that represents stochastic relationships among variables (Neapolitan, 2003, Chapter 1). It is a directed acyclic graph formed by nodes which represent random variables, and arcs which represent conditional dependencies. Fig. 2 illustrates a simple BN. In this network, the node A is defined as the parent of B and C; B and C are children of A.

Markov condition is a basic assumption in modeling BN, which requires that each variable in the network is conditionally independent of its non-descendents given its parents (Neapolitan, 2003, Chapter 1). In the case shown in Fig. 2, Node B and C are independent given their parent A,

\[ P(BC|A) = P(B|A) \cdot P(C|A) \]

Therefore the joint distribution in this example can be formulated by the chain rule.

\[ P(ABC) = P(A) \cdot P(BC|A) = P(A) \cdot P(B|A) \cdot P(C|A) \]

Although the BN model in this paper is far more complex than the above example, the Markov condition and chain rule are still applicable.

4.2. Bayesian network model description

Based on the example of Fig. 1, Fig. 3 illustrates its corresponding BN model graphically. The total number of sample vehicles in the cycle (denoted as \( M \)) is 8 and the total number of sample queued vehicles in the cycle (denoted as \( M_q \)) is 6. As assumed, the network follows the Markov condition so that a node is only conditional dependent to its immediate parents. The graph is composed of three horizontal layers and multiple vertical slices. The layers represent vehicle arrival times, indices, and departure times respectively from the top to the bottom. Each vertical slice represents a sample vehicle. The nodes in the \( i \)th slice represent the arrival time \( X_i \), the vehicle index, and the departure time from the top to the bottom, for the \( i \)th sample vehicle. As the departure time of a free flow vehicle can be determined by its arrival time (plus a traverse time), only the departure times (\( Y_i \)) of queued vehicles are shown in Fig. 3. All the nodes \( X_i, K_i, \) and \( Y_i \) are random variables. In this paper, \( k_i, x_i \) and \( y_i \) represent the values taken by \( K_i, X_i \) and \( Y_i \), respectively. \( \Delta k_i = K_i - K_{i-1} \) represents the index difference of two

![Fig. 2. A simple Bayesian network.](image)

![Fig. 3. Structure of Bayesian network.](image)
consecutively sampled vehicles. In the BN model, we assume that (i) the arrival time of a vehicle depends on its index, and the previous vehicle’s index and arrival time; (ii) the departure time of a queued vehicle depends on its index, and the previous vehicle’s index and departure time; and (iii) the index of a vehicle only depends on the previous vehicle’s index. They are the exact meaning of the Markov assumption for arterial traffic flow as indicated in Section 2. These assumptions mainly postulate that a vehicle is only dependent on the previous vehicles. This is consistent with the fact that vehicles usually respond to stimuli caused by vehicles in front of them, not those behind them. We postulate that they hold in general for arterial traffic. The directed arcs in the graph indicate the conditional dependency of the variables. The conditional distribution of each node, provided its parents are given, can be formulated. We will show next models for vehicle index (via sampling), and the arrival and departure processes.

4.3. Sampling process

The middle layer of the graph in Fig. 3 shows the indices of sample vehicles in a cycle. As assumed before, each vehicle is sampled independently with a given probability p, which represents the penetration rate of the mobile data. Then the index difference of two consecutively sample vehicles \( K_i - K_{i-1} \) follows a geometric distribution (Bain and Engelhardt, 1991, Chapter 3). The conditional probability that vehicle index \( K_i \) equals \( k_i \) given previous vehicle index \( K_{i-1} = k_{i-1} \) (\( k_i \) and \( k_{i-1} \) are integer and \( k_i > k_{i-1} \)) satisfies

\[
P(K_i = k_i|K_{i-1} = k_{i-1}) = p (1 - p)^{k_i - 1}, \quad i = 2, 3 \ldots M
\]

(1)

Here \( \Delta k_i = k_i - k_{i-1} \). Note that the first sample vehicle in the cycle does not have a parent, but a prior probability of its index can also be computed via geometric distribution.

\[
P(K_1 = k_1) = p (1 - p)^{k_1 - 1}
\]

(2)

4.4. Arrival process

As stated, a possible tool to better capture the arrival process is the NHPP. It is a Poisson process with a time dependent arrival rate \( \lambda(t) \). In this paper, the arrival rate is assumed to be a constant between two consecutive sample vehicles, but may vary between different sample vehicle pairs. Therefore, arrival rate is actually a step function in terms of time. This is obviously not the optimal way of using NHPP to model the arrival process, but is probably the best we can do to utilize the information from mobile sensor data, especially compared with the commonly used homogeneous Poisson process. Under NHPP, the arrival rate between the \((i - 1)\)th and the \(i\)th sample vehicles is

\[
\lambda(t) = \lambda_i \text{ if } x_{i-1} < t \leq x_i
\]

(3)

where \( x_{i-1} \) and \( x_i \) are true arrival times of the \((i - 1)\)th and the \(i\)th sample vehicle. Named as the local arrival rate \( \lambda_i \), it is the estimated average arrival rate during time interval \([x_{i-1}, x_i]\).

For an NHPP, the arrival headway of two consecutive vehicles in the time window defined by \([x_{i-1}, x_i]\) follows an exponential distribution. Such exponential distribution may vary for different time windows. Between two sample vehicles, there can be multiple headways (i.e., vehicles) with the same arrival rate. As gamma distribution can represent the sum of multiple independent exponentially distributed random variables with the same rate (Bain and Engelhardt, 1991, Chapter 3), it is reasonable to describe the arrival headway between the \((i - 1)\)th and \(i\)th sample vehicles using a gamma distribution \( \Gamma(\Delta n, 1/\lambda_i) \), where the shape parameter \( \Delta n \) is the number of headways between the two samples and \( \Delta n \) is the local arrival rate in. Since we only consider the case where the two sample vehicles are in the same cycle, \( \Delta n \) will be exactly the same as \( \Delta K \) in the gamma distribution. One can conversely estimate the index difference between the sample vehicles given the arrival time difference and the arrival rate.

In the NHPP process, the arrival time of the \(i\)th \((i \geq 2)\) sample vehicle depends on three factors: the arrival time of the previous sample vehicle \( X_{i-1} \), the amount of headway between two vehicles \( K_{i-1} - K_{i-1} \), and the local arrival rate \( \lambda_i \). The time difference between \( X_i \) and \( X_{i-1} \) follows a gamma distribution with shape parameter \( K_i - K_{i-1} \) and scale parameter \( 1/\lambda_i \):

\[
X_i - X_{i-1} \sim \Gamma(K_i - K_{i-1}, 1/\lambda_i), \quad i = 2, 3 \ldots M
\]

(4)

Thus \( X_i \) is considered as the child of \( X_{i-1}, K_{i-1}, \) and \( K_i \) as shown in the BN network in Fig. 3. The conditional probability density of \( X_i \) given \( X_{i-1}, K_{i-1}, \) and \( K_i \) has the following density distribution (Bain and Engelhardt, 1991):

\[
f(X_i = x_i|K_{i-1} = k_{i-1}, K_i = k_i, X_{i-1} = x_{i-1}) = \lambda_i^{k_i} (x_i - x_{i-1})^{k_i - 1} \frac{\exp(-\lambda_i (x_i - x_{i-1}))}{\Gamma(\Delta k_i)}, \quad i = 2, 3 \ldots M
\]

(5)

where \( \Gamma(\Delta k_i) = (\Delta k_i - 1)! \) is the gamma function.
4.5. Departure process

The bottom layer of the network indicates the departure process of traffic in the intersection. As shown in the graph, the departure time of a sample queued vehicle follows certain distribution, which depends conditionally on $Y_{i-1}$, $K_{i-1}$, and $K_i$ in the network. This is because the departure time difference, $Y_i - Y_{i-1}$, of the $(i-1)$th and $i$th ($i \geq 2$) sample queued vehicles follows an index dependent log–normal distribution (Jin et al. (2009)):

$$Y_i - Y_{i-1} \sim \ln N(\mu(K_{i-1}, K_i), \sigma^2(K_{i-1}, K_i)), \quad i = 2, 3 \ldots M_q$$ (6)

Here $\mu(K_{i-1}, K_i)$ is defined as the function of the location parameter given index pair $K_{i-1}$ and $K_i$; $\sigma(K_{i-1}, K_i)$ is the function of a scale parameter in a log–normal distribution. Hence the conditional probability density function of $Y_i$ given $Y_{i-1}$, $K_{i-1}$, and $K_i$ is (Ginos, 2009):

$$f(Y_i = y_i|K_{i-1} = k_{i-1}, K_i = k_i, Y_{i-1} = y_{i-1}) = \frac{1}{(y_i - y_{i-1})\sigma(k_{i-1}, k_i)\sqrt{2\pi}} \exp \left( -\frac{(\ln(y_i - y_{i-1}) - \mu(k_{i-1}, k_i))^2}{2\sigma^2(k_{i-1}, k_i)} \right), \quad i = 2, 3 \ldots M_q$$ (7)

As aforementioned, the time elapsed from the start of green (denoted as $T_g$) to the departure time of the first sample in the cycle ($Y_1$) is assumed to be normally distributed given its index:

$$Y_1 - T_g \sim N(\mu(K_1), \sigma^2(K_1))$$ (8)

where $\mu(K_1)$, $\sigma^2(K_1)$ are the mean and variance functions for the normal distribution. The conditional probability density function of $Y_1$ given index $K_1$ is

$$f(Y_1 = y_1|K_1 = k_1) = \frac{1}{(y_1 - T_g)\sigma(k_1)\sqrt{2\pi}} \exp \left( -\frac{(y_1 - T_g - \mu(k_1))^2}{2\sigma^2(k_1)} \right)$$ (9)

Note that the structure in Fig. 3 only illustrates the case when a cycle has both queued and free flow vehicles. If no free flow vehicle is sampled in a cycle (i.e. $M_q = M$), all slices in the network should have the departure time nodes. On the contrary, if we have no queued samples (i.e. $M_q = 0$), the departure layer of the Bayesian network will not exist.

The BN model in Fig. 3 is constructed for each cycle of a signal. The parameters of the models include the time varying arrival rate $\lambda_t$ in Eq. (4), the location and scale parameters for the log–normal distribution in Eq. (6), and the means and standard deviations of the normal distribution in Eq. (8). The BN model needs to be solved in two major steps: learning and inference. The learning step is to estimate the parameters of the model using historical data (e.g., data collected from multiple previous cycles). The inference step however is done based on the model and the collected mobile data in real time for each cycle. The input (or data) to the model includes: the penetration rate $p$, and the observed arrival and departure times of all sample vehicles $(x_i, y_i)$.

5. Parameter learning

5.1. Parameter learning for the departure process

As the departure headway at an intersection is stable for different cycles, the parameters can be learned from a 100% penetration training set. The maximum likelihood estimation method is used to estimate the parameters in the departure process. The location parameter $\mu$ and scale parameter $\sigma$ of a log–normal distribution in Eqs. (6) and (7) between the $h$th and $j$th queued vehicles ($1 \leq h < j$) can be estimated by the following formulas (Ginos, 2009).

$$\mu(h,j) = \frac{\sum_{n=1}^{N_j} \ln(Y^n_j - Y^n_h)}{N_j}$$ (10.1)

$$\sigma^2(h,j) = \frac{\sum_{n=1}^{N_j} [\ln(Y^n_j - Y^n_h) - \mu(h,j)]^2}{N_j}$$ (10.2)

Here $N_j$ is the number of “$j$-qualified” cycles defined as cycles that have at least $j$ queued vehicles. $Y^n_h$ and $Y^n_j$ are the departure times of the $h$th and $j$th vehicle respectively in the $n$th “$j$-qualified” cycles ($1 \leq h < j \leq N_j$).

The same dataset is used to estimate the mean $\mu'$ and variance $\sigma'^2$ of the normal distribution in Eq. (9) for the first sample vehicle. To estimate the parameters of the $j$th queued vehicle in a cycle, it is necessary to focus on the “$j$-qualified” cycles with at least $j$ queued vehicles. The sample mean of the time difference $Y^n_j - T^n_g$ is the estimator of the mean, $\mu'(j)$ (Bain and Engelhardt, 1991, Chapter 9).

$$\mu'(j) = \frac{\sum_{n=1}^{N_j} (Y^n_j - T^n_g)}{N_j}$$ (11.1)
where $T^n_i$ is the start of green, and $y^n_i$ is the departure time of the $j$th vehicle in the $n$th “$j$-qualified” cycle. Similarly, the sample variance is used to estimate the variance $\sigma^2(j)$ in the normal distribution.

$$
\sigma^2(j) = \frac{\sum_{n=1}^{N_j} (y^n_i - T^n_i - \mu(j))^2}{N_j - 1} \tag{11.2}
$$

### 5.2. Parameter learning for the arrival process

Although the BN model is capable of dealing with non-homogeneous arrival rates, it is not easy to learn the time-varying arrival rate in an urban arterial. In general, there are two types of data source for arrival rates learning. One is the training dataset which is collected vehicle by vehicle at the same intersection but at different time compared with the testing data. As mentioned, the traffic arrivals change rapidly and irregularly because of vehicle discharging from the upstream intersections, the effect of platoon dispersion, and the varying traffic demands in a day (e.g., peak hour vs. off peak). The parameters thus cannot be learned properly from the training dataset if the circumstances differ for the testing data.

The other data source is the mobile data collected in real time for the test. Such data reflect the real time arrival information, which however only provide a sample of the entire traffic flow. This incomplete information will bring confusion when estimating the arrival flow rate. For example, under a low penetration rate, a large vacancy between two sample vehicles may be caused by either the low arrival rate or several consecutively undetected vehicles. Two methods can be proposed to roughly estimate the arrival rate. A straightforward approach is to reduce the NHPP assumption to a regular (homogeneous) Poisson process with a constant arrival rate. Using this method, we only need to compute the average arrival rate from the sample dataset. Another way is named as the Constant Global Index Difference (CGID) method. Here global index is defined as the vehicle sequence in the complete dataset (including all the cycles). If the penetration rate is $p$, we assume the index of the $i$th sample vehicle is $i/p$. Then the index difference (i.e., number of vehicles) between the $(i-1)$th and $i$th sample vehicles is $1/p$, and the arrival time difference is $x_i - x_{i-1}$, where $x_i$ is the true arrival time of the $i$th sample vehicle. Therefore the arrival rate between the $(i-1)$th and $i$th sample vehicle is simply:

$$
\lambda_i = \frac{1}{p(x_i - x_{i-1})} \tag{12}
$$

Empirical results show that the CGID method generally outperforms the constant arrival rate method, especially when the penetration rate is relatively high. Details are omitted here due to space limitation.

In summary, we estimate the parameters of the departure distributions using historical data since the departure process is stable; same distributions are thus expected to hold for different cycles at different days. On the other hand, the parameters of the arrival distributions are estimated using mobile data collected in real time, NOT from historical data. This is because the arrival processes may vary significantly from cycle to cycle and across different days. Parameters estimated from historical data therefore may not describe well the traffic arrivals in real time (i.e., during the testing phase). This is one of the features of the proposed learning framework in this paper for which we aim to integrating proper traffic knowledge into the learning models.

### 6. Vehicle index inference

#### 6.1. Basic model

In this section, the inference method of the vehicle index is discussed. The purpose is to assign (probabilistic) indices to observed vehicle samples based on their arrival and departure times. The conditional probability of a vehicle index, given the observed arrival and departure times, is derived first. It is defined as $P(K = k | X = \bar{x}, Y = \bar{y})$. Here $K$, $X$, and $Y$ represent the sets of random variables $K_i$, $X_i$, and $Y_i$, respectively; $k = \{k_1, k_2, \ldots, k_M\}$ is the set of specific values taken by $K_i$; $\bar{x} = \{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_M\}$ and $\bar{y} = \{\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_M\}$ are the sets of observed values of $X$ and $Y$. In other words, the conditional probability of the vehicle index is the probability of assigning any specific combination of vehicle indices (i.e., $k$) to vehicle samples, based on their observed arrival and departure times (i.e., $\bar{x}$ and $\bar{y}$). Using the chain rule (Neapolitan, 2003, Chapter 1) and based on the graphical representation of the BN model in Fig. 3, Eq. (13) may be constructed:

$$
P(K = k | X = \bar{x}, Y = \bar{y}) = \frac{P(K = k | X = \bar{x}, Y = \bar{y})}{P(X = \bar{x}, Y = \bar{y})} = \alpha \cdot P(K_1 = k_1) \cdot f(Y_1 = \bar{y}_1 | K_1 = k_1) \prod_{i=2}^{M} P(K_i = k_i | K_{i-1} = k_{i-1})$$

$$
\prod_{i=2}^{M} f(X_i = \bar{x}_i | K_{i-1} = k_{i-1}, K_i = k_i, X_{i-1} = \bar{x}_{i-1}) \prod_{i=2}^{M} f(Y_i = \bar{y}_i | K_{i-1} = k_{i-1}, K_i = k_i, Y_{i-1} = \bar{y}_{i-1}) \tag{13}
$$

In Eq. (13), each term represents a type of conditional dependence in the proposed Bayesian Network; see Fig. 3. The index of the first sample vehicle $K_1$ does not have any parent, so we use its prior probability $P(K_1 = k_1)$ in the equation. The next term
\( f(Y_1 = y_1 | K_1 = k_1) \) corresponds to the conditional dependence between the first sample’s departure time and index, given in Eq. (9). For other vehicles (i.e., \( i \geq 2 \)), the last three terms formulated in product forms represent the conditional probability of the index, arrival time and departure time given their respective parents in sequence; they are given in Eqs. (1), (5), and (7) respectively. Also introduced is the normalizing constant in the equation. As the right hand side of the equation is the product of discrete probability and probability density, has to be used to make the sum of \( P(k|X,Y) \) equal to one.

By substituting Eqs. (1), (2), (5), (7), (9) to the conditional probability expression (13), one can calculate the conditional probability for any vehicle index combination.

The index estimation results from the BN model will thus be a probability distribution. For engineering applications, one is often interested in knowing the “best” possible solution in index estimation. For this purpose, the Most Probable Explanation (MPE) can be considered (Chan and Darwiche, 2006). It is a complete (and deterministic) index assignment with the highest probability given the current observations.

\[
    k^* = \{k'_1, k'_2, \ldots k'_M\} = \arg \max P(K = k|X = \bar{x}, Y = \bar{y})
\]

Another useful result is the marginal posterior distribution. For a certain vehicle in the cycle, say \( K_1 \), one can compute the marginal distribution of its index by summing up all of the conditional probability over \( \{K_2, K_3, \ldots, K_M\} \).

\[
P(K_1 = k_1 | X = \bar{x}, Y = \bar{y}) = \sum_{\{k_2, k_3, \ldots, k_M\}} P(K_1 = k_1, \{K_2, K_3, \ldots, K_M\} = \{k_2, k_3, \ldots, k_M\} | X = \bar{x}, Y = \bar{y})
\]

It is possible to enumerate all possible combination of indices and find the MPE or marginal distribution, but the complexity will be exponential. Although the network for one cycle does not have a very large size, it is still time consuming. This is especially so when a cycle has many sample vehicles. We next discuss how to reduce the computational effort based on the characteristics of intersection traffic.

### 6.2. Simplified model

To simplify the computation of the BN model, we develop a method to infer the conditional probability of vehicle index using a simplified network structure and some newly defined variables, e.g., the vehicle index (and arrival time or departure time) differences. As shown in Eqs. (1) and (5), the probability functions of the sampling and the arrival processes depend on the index difference and the arrival time difference instead of the absolute vehicle indices or arrival times. Eqs. (2) and (9) are particularly for the first sample vehicle and are related to \( K_1 \), the index of the first sample vehicle. The departure process in Eq. (7) is thus the only part of the network that depends on the absolute index (as well as the departure time differences).

Roess et al. (2010) showed that the vehicle departure headway, after the signal turns green, stabilizes at and fluctuates around the saturation flow rate after the fourth or fifth headway position. Therefore we can reasonably assume that after the fifth vehicle, the headway distribution should follow the same log–normal distribution, independent of the actual vehicle index. Since the index of the fifth sample queued vehicle is always greater than or equal to 5, it can be reasonably assumed that the \( \mu \) and \( \sigma \) functions used in Eq. (7) satisfy:

\[
    \begin{cases}
      \mu(k_{i-1}, k_i) = \mu(\Delta k_i) = \mu_k, & i = 5, 6, \ldots, M_Q \\
      \sigma(k_{i-1}, k_i) = \sigma(\Delta k_i), & i = 5, 6, \ldots, M_Q
    \end{cases}
\]

where \( \mu_k \) and \( \sigma(\Delta k_i) \) are the location parameter and scale parameter of the log–normal distribution after the headway stabilizes at the saturation headway, which only depend on vehicle index differences of the \( (i-1) \)th and the \( i \)th sample vehicles instead of their actual indices.

The basic BN can then be decomposed into types of independent sub-networks if the number of sample vehicles, \( M \geq 5 \), as shown in Fig. 4 (this decomposition will not be conducted if \( M < 5 \); in this case, the original procedure in Section 6.1 will be applied because the number of sample vehicles is small and the original method based on pure enumeration can work well). The first contains only one sub-network for the first four slices exactly the same as in the basic structure, which is for the first four sample vehicles. If the number of sample queued vehicles, \( M_Q < 4 \), the first sub-network will contain at least one free flow vehicle slice in which the departure time node and the arcs to the departure time node will be removed. The second type of sub-networks is for the other sample queued vehicles, if any. In particular, if \( M_Q > 4 \), it will include \( M_Q - 4 \) unconnected slices, one for each sample queued vehicle; if \( M_Q < 4 \), the second type of sub-networks will be omitted. The last type of sub-networks is for the other free-flow vehicles, if any. It includes at most \((M-M_Q)\) unconnected slices, one for each sample free flow vehicle. Furthermore, for the second type of sub-networks, the index difference \( \Delta K_i = k_i - k_{i-1} \) is taken as the variable instead of the absolute index \( K_i \) due to the above discussion. Similarly, the difference of arrival time \( \Delta x_i = x_i - x_{i-1} \) and the difference of departure time \( \Delta y_i = y_i - y_{i-1} \) are used as nodes in the upper and lower layers. \( \Delta x_i = x_i \) and \( \Delta y_i = y_i \) represent the true value of \( \Delta X_i \) and \( \Delta Y_i \). For the free flow sub-networks, the index difference and arrival time difference are used instead of their absolute values. The use of variable differences instead of the actual variables actually removes the dependence arcs between the sub-networks, making them independent. The reason is that in the original BN, dependences only exist between parents and their immediate children, which are already included when the variable differences are used. Therefore the new definitions of variable differences simplify the model by eliminating all the arcs which connect the slices between neighboring sub-networks. The conditional probability functions of the sampling, arrival
and departure processes are formulated in Eqs. (17)–(19) by substituting the index difference variable to Eqs. (1), (5) and (7) respectively.

\[
P(\Delta K_i = \Delta k_i) = p(1 - p)^{\Delta k_i - 1}, \quad i = 5, 6 \ldots M
\]

\[
f(\Delta X_i = \Delta x_i|\Delta K_i = \Delta k_i) = \frac{\exp(-\frac{\Delta x_i}{\Delta k_i})}{1!(\Delta k_i)}, \quad i = 5, 6 \ldots M
\]

\[
f(\Delta Y_i = \Delta y_i|\Delta K_i = \Delta k_i) = \frac{1}{\Delta y_i \sigma(\Delta k_i)\sqrt{2\pi}} \exp\left(-\frac{\left[\ln(\Delta y_i - \mu(\Delta k_i))^2\right]}{2\sigma^2(\Delta k_i)}\right), \quad i = 5, 6 \ldots M_Q
\]

Then the conditional probability of the vehicle index based on arrival and departure time observations, i.e. Eq. (13), becomes the following formulation when the number of sample vehicles, \(M \geq 5\):

\[
P(K = k|X = \bar{x}, Y = \bar{y}) = P_0 \cdot \prod_{i=5}^{M} P_i
\]

where \(P_0\) is the conditional probability derived from the first sub-network which consists of the first four vehicles (i.e., the first sub-network in Fig. 4).

\[
P_i = \begin{cases} P(\Delta K_i = \Delta k_i) \cdot f(\Delta X_i = \Delta x_i|\Delta K_i = \Delta k_i) \cdot f(\Delta Y_i = \Delta y_i|\Delta K_i = \Delta k_i) & \text{if } 5 \leq i \leq M_Q \quad \text{(i.e., queued vehicles)} \\ P(\Delta K_i = \Delta k_i) \cdot f(\Delta X_i = \Delta x_i|\Delta K_i = \Delta k_i) & \text{if } M_Q < i \leq M \quad \text{(i.e., free flow vehicles)} \end{cases}
\]

\[
ax_i P(\Delta K_i = \Delta k_i) \cdot f(\Delta X_i = \Delta x_i|\Delta K_i = \Delta k_i) \cdot f(\Delta Y_i = \Delta y_i|\Delta K_i = \Delta k_i) & \text{if } 5 \leq i \leq M_Q \\ ax_i P(\Delta K_i = \Delta k_i) \cdot f(\Delta X_i = \Delta x_i|\Delta K_i = \Delta k_i) & \text{if } M_Q < i \leq M
\]

For the samples arriving after the fourth observed vehicle, we use \(P(i = 5, 6 \ldots , M)\) to represent the conditional probability of the index difference of the \((i - 1)\)th and the \(i\)th sampled vehicles from the corresponding sub-network. If the \(i\)th vehicle is a queued vehicle (i.e. \(5 \leq i \leq M_Q\)), at least one of the second type of sub-networks in Fig. 4 will be generated. We need to calculate \(P_i\) given both arrival and departure time as shown in equation (20.2). Otherwise, we employ the third type of sub-networks to solve the conditional probability for free flow vehicles. In Eqs. (20.2) and (20.3), is the normalizing constant for each sub-network, which makes the sum of \(P(\Delta K_i|\Delta X_i, Y_i)\) or \(P(\Delta K_i|\Delta X_i)\) equal to 1 when \(i = 5, 6 \ldots , M\).

As the index difference model splits the basic BN into smaller pieces, the computation can be reduced, especially when the sample size is large. As the proposed model involves Poisson, geometric and log–normal distributions in a single BN, there is no existing BN package to directly solve this problem. We enumerate all possible combination of \(K_i (i = 1, 2, 3, 4)\) and \(\Delta K_i (i = 5, 6 \ldots , M)\) to compute the conditional probability of vehicle indices. Then Eqs. (14) and (15) are applied to calculate the MPE and posterior distribution.

7. Numerical experiments

The BN-based vehicle index estimation model and solution method are tested in this paper using multiple datasets, including data from NGSIM, a field experiment, and microscopic traffic simulation. We found that the results from the basic inference and simplified inference models are very similar. Therefore, we only implement and show the results of the simplified model in this section. We first show the stochastic features of the model through calculating the marginal probability of sample vehicles in the NGSIM data. Next, the performances of the index estimation are shown for all the three datasets. The most important performance measure in this analysis is the Mean Absolute Error (MAE), defined as follows:
where $k'_i$ is the estimation of the $i$th vehicle index in its cycle, and $k_i$ is the true value. $N$ is the total number of samples in the dataset. Note that the performance of absolute errors, i.e. MAE, is used here instead of the performance for relative errors (such as the mean absolute percentage error). This is because the same index error should mean the same thing no matter what the actual vehicle index is. If relative errors are considered, the same index error (such as 0.5) will mean very differently for different indices: it will mean 50% error for the 1st vehicle and 5% error for the 10th vehicle, which does not make sense at all.

7.1. NGSIM data

NGSIM data were collected on the Peachtree Street in Atlanta, Georgia which has 4 signalized intersections (Cambridge Systematics 2007). The two datasets used here are the ones collected for the intersection of the 14th Street NE and Peachtree Street. The dataset from 12:45 pm to 1:00 pm is considered as the training set; the test dataset was collected from 4:00 pm to 4:15 pm. The signal timing parameters of both datasets are available for the index inference. In the test dataset, there are in total 9 cycles with 51 queued vehicles and 25 free flow vehicles.

First, the stochastic feature of index estimation is shown. Focus is on the traffic of the 7th cycle using a randomly generated 30% penetration rate. In this cycle, there are 6 samples; the last two vehicles are traveling at free flow speed. The true (observed) indices in the traffic stream in the cycle are: 1, 2, 3, 8, 10, 13. Fig. 5(a) shows the marginal probability of sample vehicles in the cycle. Fig. 5(b) shows the log–likelihood which is the natural logarithm of the probability in Fig. 5(a). The log–likelihood figure can help better illustrate the estimation when there is some probability with very low values. In the two figures, asterisks represent the estimated indices; data points for a particular vehicle are connected with a dotted line to clearly show the distribution of the estimated index. Circles represent the true indices of the sample vehicles. As shown in the two figures, the model correctly estimates the indices of the first four queued vehicles because the true indices of them are exactly at the peak of the estimated distributions. However, the indices of the fifth and sixth vehicles are underestimated. The index distribution of a free flow vehicle has a “fatter” shape than a queued vehicle. That is mainly because the departure process for a free flow vehicle is not affected by the signal and queue, but determined by its arrival process. The departure headway model does not apply to the index inference of free flow vehicles. As a result, the variance of the free flow vehicle indices is higher than the queued vehicle indices.

Fig. 6(a) depicts the MPE for the test data using the same 30% penetration rate dataset. We show the true index of each sample vehicle with circles and the estimated MPE with crosses. A cross overlaps with the corresponding circle in the figure when the estimation perfectly matches the true index. Because the vehicle index is defined as the position in the departure sequence, the x-coordinate of the data point refers to the departure time of the vehicle. The start of red time is depicted by a solid line and the start of green time is shown with a dotted line. Note that vehicle departure time is the time at VTL2, so that a vehicle may pass VTL2 after the green time ends. Fig. 6(a) shows that the index estimation performs well on queued vehicles. About half of the estimations of queued vehicle indices (depart right after the start of the green times) correctly match the true index. To measure the accuracy on index inference, the MAE of the estimated vehicle indices is calculated by comparing with the true values. In this dataset, the MAE of all vehicles is 1.19, and the MAE of queued vehicles is 0.82. The MAE of free flow vehicles is 1.8, larger than queued vehicles by a factor of roughly two.

![Fig. 5. (a) Marginal probability of vehicle index and (b) log–likelihood of vehicle index.](image-url)
For each sample vehicle, we also show the uncertainty of the estimated marginal index distribution. In Fig. 6(a), the two solid piece-wise curves depict the range within one standard deviation from the mean value, and dashed curves depict the range within two standard deviations from the mean. Note that for each sample, the mean value of marginal index distribution...
bution may not equal to the MPE estimate. In this dataset, 69.6% sample indices fall within the range of one standard deviation from the mean value, and 87.0% are within two standard deviations. The rate is approximately the same for either queued vehicles or free flow vehicles. However, as shown in the figure, the estimation of free flow vehicles (often depart right before the start of the red times) usually have much wider ranges, and deviates more significantly from the true indices. This also explains the higher MAE of free flow vehicles in the MPE estimation.

To further test how the model performs under different penetration rates, shown in Fig. 6(b) are the box plots of MAE with penetration rates from 5% to 100% for the NGSIM data. The bottom and top of each box indicate the 25th and 75th percentile. The (red) line segment near the middle of the box represents the median. We also use plus signs to show the outliers, and whiskers to show the minimum and maximum of all the data except those outliers. As shown in the figure, the median MAE of all vehicles decreases with the increase of the penetration rate. At 10% penetration rate, the MAE is about 2.0. At 100% penetration rate, the estimated index is exactly the same as the true index (i.e., MAE is 0). The variability of performance (i.e., MAE) also decreases almost monotonically when the penetration increases.

7.2. Field test data

The one-hour field experiment was conducted in Albany, NY (Ban et al., 2011). The focus was on the left turn traffic of an actuated intersection. Nine test drivers repeatedly made a left turn at the intersection. Each car was equipped with a GPS logger so that the arrival and departure times were collected when the vehicle passed VTL1 and VTL2 of the intersection. Meanwhile, all vehicles were monitored by two video cameras deployed at VTL1 and VTL2 respectively. After the test, the travel times of all unsampled vehicles were manually extracted. The penetration rate of the equipped vehicles is about 30% (126 out of 441). In the full dataset observed by the cameras, about 20% of vehicles are free flowing. We use the first 15-min data for training and the rest for testing.

Fig. 7(a) depicts the estimated and true indices along with one/two standard deviation curves under 30% penetration rate.

Fig. 7(b) shows a “zoom-in” version of Fig. 7(a) from 16:33 to 16:41. Similar to the NGSIM data, the queued vehicles outperform the free flow vehicles on index inference. The MAE of queued vehicles is 0.81, while the MAE of free flow vehicles is

Fig. 8. Index estimation performance in micro-simulation: (a) estimated index, true index, and standard deviation curves; (b) index inference for 16:44–16:53; and (c) MAE under different penetration rates.
1.42. On average the MAE of all vehicles is 0.94. One standard deviation accounts for 57.8% samples and two standard deviations account for 86.3% samples. Fig. 7(c) shows the MAE box plot in terms of the penetration rates. The median MAE drops from 1.2 to 0.142. On average the MAE of all vehicles is 0.94. One standard deviation accounts for 57.8% samples and two standard deviations account for 89.1% samples. For each sample, the standard deviation of marginal index estimation in the simulation is smaller than the NGSIM and field test. Fig. 8(c) depicts the box plot of MAE with varying penetration rates. As shown in the figure, the median MAE does not change too much when the penetration increases, but the variability decreases significantly. The index inference method has a very small error (less than 0.3 in average), even when the penetration rate is 5 or 10%. A possible reason for the high accuracy using simulation data is that, besides the highly congested situation, in simulation the departure process is usually implemented to follow certain distribution, which makes it easier to conduct the estimation using stochastic learning approaches. In real life, driver behaviors are more random and even unpredictable, which may cause more variations in the arrival/departure processes and thus larger errors when conducting the estimation.

To address the significance of the proposed BN-based method, we compare the new method with a simple method which assumes constant departure headway for queued vehicles and constant arrival headway for free flow vehicles. For the same 30% penetration rate scenario, the BN-based method generally outperforms the simple method on index estimation. The improvement is at least 15–20% in term of MAE for most cases. Another advantage of the BN method is that it can consider traffic stochasticity so that the distribution of index estimation can be produced. The simple method however does not have the capability to integrate such stochasticity and can only produce deterministic estimates.

8. Conclusions

A Bayesian network (BN) model was presented in this paper to estimate the index of a sample vehicle at a signalized intersection, i.e., the position of the vehicle in the departure process. The non-homogeneous Poisson Process (NHPP) was applied to model vehicle arrivals and log–normal distributions were used to model the vehicle departure process. As both of them are dependent on vehicle indices, a three-layer BN model was constructed to capture the relationships of the vehicle indices and the stochastic arrival and departure processes of the traffic flow. The BN model thus provides a graphical representation of such relationships that enables a systematic investigation of their conditional dependences. The parameters of the stochastic distributions used in the BN model were learned from real world mobile data: parameters of the departure distributions are estimated from historical data and the parameters of the arrival distributions are estimated from real time mobile data (not historical data). This is because the departure process is more stable and the estimated parameters can be applied to different cycles across different days. The arrival process however may not be stable across different cycles or days, and thus cannot be estimated from historical data. The vehicle indices were then estimated using stochastic inference approaches. To improve computational efficiency, the BN model was decomposed into three types of sub-networks for the first 4 sample vehicles, the remaining queued vehicles (if any), and the remaining free flow vehicles (if any), respectively. The decomposition is based on the special structure of the BN model, more importantly, the unique feature of the departure process from a traffic signal which tends to stabilize after the first 5 or 6 vehicles. The proposed BN model thus integrates important traffic knowledge into the stochastic learning framework to make it more suitable to solve traffic problems. It is our understanding that such consideration, i.e., to integrate proper transportation knowledge with learning methods, is critical when applying advanced mathematical models to mobile data in dealing with real world traffic/transportation applications.

The model was tested using data from NGSIM, a field test, and microscopic traffic simulation. The results of the index estimation method were presented in both deterministic (i.e., MPE) and stochastic (e.g., marginal distribution) forms. The performance for queued vehicles was found to be usually satisfactory, but that for the free flow vehicles had relatively larger errors. The results also indicated that the estimation performance for congested intersections usually outperforms that of less congested intersections.

This paper mainly provides some initial testing and validation results for the proposed BN model for arterial traffic flow. There are several issues that need to be resolved as well as some related future research directions.
• The penetration rate of mobile sensor data is assumed and integrated into the BN model; see Eqs. (1) and (2) in Section 4.3. In practice, such a rate might not be available and could vary from location to location or from time to time even at the same location. In the appendix, we propose a method to estimate the penetration rate from mobile data. However, how to accurately estimate the penetration rate and study how sensitive the BN model is to varying penetration rates will be an importation future research topic.
• Overtaking issues are discussed in Section 2.2. A simple time-swapping scheme is applied to deal with overtaking on multiline streets. We will work to develop more sophisticated algorithms in detecting and resolving the overtaking conditions in the future.
• The time-dependent arrival rate \( \lambda(t) \) in the NHPP for modeling the arrival process is estimated using the global index difference method and is assumed as constant between any two consecutively sampled vehicles. This works fine if the two sampled vehicles are in the same platoon but may result in large errors if they are at two different platoons. To address this issue, studying the platoon effects of the upstream intersection is critical especially for closely spaced intersections. The challenge here is how to properly model platoon forming and dispersion using mobile data only. We will study this issue in future research.
• The performance on free flow vehicle index estimation needs to be improved. The major issue is that the arrival process is difficult to learn because of the mixed effect of low penetration and incomplete information from mobile data. As shown in Section 5.2, the prior arrival rate is roughly estimated based on a constant global index difference assumption. More information on the upstream intersection (such as platoons) may be helpful when learning parameters for free flow arrivals.
• To exploit the advantages of the proposed stochastic learning models, more datasets will be needed especially those collected from real world congested intersections. This will be pursued in future research.
• The proposed BN model and index inference method focus on individual signalized intersections in this paper, which may provide a useful framework to estimate the performance measures of a signalized intersection. As mentioned in Section 1, the estimated index reconstructs the intersection traffic flow. So it is possible to apply the intersection performance estimation methods and algorithms previously developed for fixed-location sensor data to mobile sensor data. More sophisticated methods could be possible for this purpose by investigating how to connect the estimated vehicle index with specific performance measures in a more robust and systematic manner. For example, to estimate the queue length, we can localize the problem around the end of the queue by recognizing that the queue length is the index of the last queued vehicle of the cycle. Using the Bayes’ rule, we can then relate the queue length of a cycle to the hidden variables of the (possibly) undetected last queued vehicle. This will enable us to develop similar BN-based models for queue length estimation. Furthermore, the model and methods may even be used for performance evaluation and optimal control of arterial corridors and/or arterial networks. For example, the BN models of different individual intersections may be integrated together by properly developed stochastic platoon forming and dispersion models to form a larger stochastic learning framework for arterial corridors or networks. This will enable the performance estimation and optimal control of arterial corridors/networks. The authors are investigating these issues and results may be reported in subsequent papers.

Acknowledgements

The authors thank the three anonymous referees for their helpful suggestions on an earlier version of the paper. The work of Peng Hao, Xuegang (Jeff) Ban, and Zhanbo Sun is based on research supported by the National Science Foundation under Grants CMMI-1031452 and CMMI-1055555. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix A. Penetration rate estimation algorithm

If the penetration rate is unknown, we can estimate it by computing the percentage of the sample queued vehicles in the total queued vehicles. The number of sample queued vehicles is collected directly, and the total number of queued vehicles can be derived via estimation and summation of the queue lengths of all cycles.

A simple queue length estimation algorithm is proposed here, assuming that the queue discharges with the saturation flow rate. If the \( i \)th cycle has at least one sample queued vehicle, the index of the last sample queued vehicle in the cycle (denoted as \( k_i \)) is determined by its departure time at VTL2 (denoted as \( y_i \)).

\[
\begin{align*}
k_i &= \frac{y_i - T_g^{i} - \tau_2}{h_d} + 1
\end{align*}
\]

(22)

where \( T_g^{i} \) is the start of the effective green time, \( \tau_2 \) is the minimum traverse time from the stop line to VTL2, and \( h_d \) is the constant departure headway at VTL2 which is the reciprocal of the saturation flow rate. In this paper, \( h_d \) and \( \tau_2 \) are calibrated using the training data. Note that \( T_g^{i} + \tau_2 \) is the departure time of the first sample queued vehicle in the \( i \)th cycle. \( \frac{y_i - T_g^{i} - \tau_2}{h_d} \) is the...
index difference between the first sample queued vehicle and the last sample queued vehicle. Thus Eq. (22) is applied to estimate the index of the last sample queued vehicles.

Let $h_a$ be the average arrival headway, the intersection delay will reduce at a constant reduction rate $h_a - h_d$ for two consecutively queued vehicles. The index difference between the last sample queued vehicle and the exact queue rear (which has no delay) is estimated by $\frac{y_i - x_i - \tau}{h_a - h_d}$, where $x_i$ is the arrival time of the last sample queued vehicle and $\tau$ is the minimum traverse time from VTL1 to VTL2. The queue length of the $i$th cycle is

$$q_i = k_i + \frac{y_i - x_i - \tau}{h_a - h_d}$$

Then the estimated penetration rate is the number of sample queued vehicle (denoted as $N_s$) divided by the summation of queue lengths of all cycles

$$\bar{p} = \frac{N_s}{\sum q_i}$$

Fig. 9 shows the performance of the penetration rate estimation algorithm using box plots. For the field dataset, the median of the estimated penetration rate is close to the exact value, and most of the estimation falls within ±5% interval. The NGSIM data have relatively large errors on penetration estimation, specifically at low penetration rates, mainly because the duration of the test is only 15 min and the sample size is too small for the estimation.

References


