Lane-Based Traffic Arrival Pattern Estimation Using License Plate Recognition Data

Chengchuan An, Xiaoyu Guo, Rrongrong Hong, Zhenbo Lu, and Jingxin Xia*

Abstract—Understanding the traffic arrival process and its patterns are of vital importance for the delay and queue analysis at intersections. Installation of advance loop detectors to detect vehicle arrivals could be costly and biased. Utilization of sampled vehicle trajectory data to reconstruct the traffic arrival flow might suffer from small sample sizes. The license plate recognition (LPR) data commonly available at intersections in cities of China are promising to overcome such limitations. This study aims to estimate the lane-based traffic arrival pattern using LPR data collected at both downstream and upstream intersections. The proposed method develops a probability model with an assumption of a two-stage piecewise arrival process for upstream merge movements. Given the observations of vehicle arrivals provided by matched vehicles in LPR data, the model is able to estimate second-based mean arrival rates to each lane at the downstream intersection. The proposed method is validated using actual LPR data collected at two adjacent intersections in Kunshan City, China. The results demonstrate that the proposed method can well describe the traffic arrival patterns of upstream merge movements with either two-stage or uniform arrival processes in different traffic scenarios. Besides, the proposed method is more robust and reliable than an average-based benchmark method, in terms of revealing actual traffic arrival patterns under different match rates.

Index Terms—License Plate Recognition Data, Probability Model, Traffic Arrival Pattern, Traffic Arrival Rate

I. INTRODUCTION

Traffic arrivals refer to vehicles that depart from the upstream intersection. As traffic signal controls the departure, vehicles are commonly observed to form moving platoons when approaching the downstream intersection. The size and arrival pattern of platoons may vary under different traffic conditions and signal schemes throughout a day. Understanding the traffic arrival process and its patterns are of vital importance for the procedures of delay and queue analysis [1], which is the foundation to develop effective traffic signal timing plans.

Detection of traffic arrivals is the first step to analyze traffic arrival patterns. Conventionally, it is achieved by installing advance loop detectors at fixed upstream locations of the stop bars. SCOOT [2] places advance loop detectors at the entries of each link to detect vehicle arrivals. Cyclic Flow Profile is predicted to evaluate vehicle delay and stops at the downstream intersection for signal timing optimization. Day et al. [3] utilized historical data collected by advance loop detectors to calculate the percent on green (POG) for the coordinated movements, and developed an arterial offsets optimization algorithm via maximizing the total number of arrivals in the green intervals. Zheng et al. [4] reconstructed arrival vehicle trajectories in a time-space diagram from shockwave profiles. The shockwave profiles were estimated using signal event-based data collected by advance loop detectors and signal controller at the downstream intersection. Uniform arrival process is assumed when advance detectors are occupied by standing queues. Indeed, the locations of advance loop detectors are critical for the accuracy of vehicle arrival detection. If the location is too far away from the downstream intersection, vehicle lane changing and road geometry (e.g., extension of lanes on an approach) could introduce bias in determining actual vehicle arrivals for different movements. If the location is too close to the downstream intersection, the observation of vehicle arrivals would be frequently interrupted by the extended queues.

With the emergence of connected vehicle technology and ride-hailing service, the mobile-based vehicle trajectory data has become a promising data source to provide traffic arrival information without the need of installing advance loop detectors. The trajectory data naturally contains the vehicle arrival-departure information as it is able to track vehicle locations in time and space. However, the penetration rate is still low in current situations, meaning that only a small fraction of traffic flow (e.g. no more than one connected vehicle arrived per signal cycle) could provide their trajectory data. Nevertheless, the sampled vehicle trajectories are still able to provide certain traffic arrival information. The stop location (i.e. distance to the stop bar) of a sampled vehicle can be converted to the number of vehicle arrivals before that vehicle,
given the average vehicle stop space headway. Although it is difficult to observe real-time vehicle arrivals with less sufficient vehicle trajectories, it is still possible to infer the traffic arrival patterns considering the vehicle arrival process is stationary for a particular period. Zheng & Liu [5] proposed a time-dependent Poisson Process model to describe the changes in arrival rates over the signal cycle. Information from both stopped and non-stop vehicles (sampled by connected vehicles and taxis) are utilized to estimate the model parameters (i.e. arrival rates) by an Expectation Maximum Algorithm. However, as the authors pointed out their model was limited to estimate arrival rates for under-saturation conditions. Besides, the arrival rates of different upstream merge movements were not distinguished. Other studies also utilized low-penetration-rate vehicle trajectory data to estimate traffic volumes [6][7], but the traffic arrival process within signal cycles was not explicitly modeled in their studies.

The License Plate Recognition (LPR) data is another emerging data source on urban roads. In China, the LPR sensors are commonly installed at intersections with the primary usage of law enforcement (i.e. red-light violation detection) to enhance traffic safety. The LPR sensors are able to record timestamps when vehicles passing through the stop bars on each lane of an approach, meanwhile recognize the vehicle license plates through video imaging processing. With the rich information provided by LPR data, various applications were proposed in literature including discharge headway analysis[8], degree of saturation estimation[9], queue length estimation [10][11][12], travel mode analysis [13][14], and OD pattern extraction [15][16]. The most appealing feature of the LPR data is that vehicles can be tracked from the stop bar to stop bar between two adjacent intersections through the matching of their unique license plates. However, there inevitably exist unmatched vehicles, whose license plates cannot be successfully matched between the upstream and downstream LPR data. This can be caused by the failures or errors in the license plate recognition process, mid-block merge vehicles from driveways, and the absence of sensors at the upstream intersection. From the perspective of cumulative departure and arrival curves, the LPR data is able to provide a complete departure curve, but only a partial arrival curve with unknown arrival timestamps of unmatched vehicles. Imputation of the missing information of unmatched vehicles is the key to fully describe the vehicle arrival process. Zhan, Li & Ukkusuri [10] modeled the arrival flow rate at the upstream intersection by a piecewise function and developed an interpolation method based on the Gaussian Process model to estimate the missing arrival timestamps of unmatched vehicles. Mo, Li & Zhan [17] further improved the interpolation method by mending the erroneously recognized license plates, and introduced unmatched vehicles as hidden variables in the model. Their studies provide a way to obtain lane-based vehicle arrival information using LPR data. However, the interpolation method is conducted in a cycle-by-cycle manner, which might be prone to the varying numbers of matched vehicles in each cycle, and may not reveal the traffic arrival pattern for a particular traffic scenario.

Inspired by [5] and [10], this study aims to estimate the lane-based traffic arrival pattern using LPR data collected at both downstream and upstream intersections. The proposed method develops a probability model based on the Non-homogeneous Poisson Process. Given the observations of vehicle arrivals provided by matched vehicles in LPR data, the model is able to estimate the second-based mean arrival rates to each lane at the downstream intersection. The proposed method has been validated by field-collected LPR data. The results show both the effectiveness and robustness of the proposed method.

The rest of the paper is organized as follows: Section 2 gives a detailed description of the problem by introducing the opportunities provided by LPR data to model the traffic arrival process. Section 3 explains the methodology of the proposed method, including the probability model and its assumptions, the model parameter estimation method, and the second-based mean arrival rate estimation. Section 4 presents results found in a field experiment followed by conclusions in Section 5.

II. PROBLEM STATEMENT

Fig. 1 illustrates a typical scenario of LPR data collection at two adjacent intersections. The arrival traffic to each lane of the downstream intersection comes from four upstream merge movements (see the arrows in Fig. 1): through, left, and right turning movements at the upstream intersection, and mid-block turning movement on the link. The right and mid-block turning movements are commonly not monitored by LPR sensors since they are not controlled by a signal phase (e.g. right-on-red is permitted). The through and left turning movements are monitored by LPR sensors, and would discharge vehicles following the indications of signal lights. Here, these two groups of movements are referred to as uncontrolled and control movements respectively. The actual signal timings are supposed to be given in this study.

![Fig. 1. A typical LPR data collection scenario.](image_url)

The right side of Fig. 1 gives a time-space diagram of vehicle trajectories. Each circle represents a single LPR record. The solid circles represent the matched vehicles, and the hollow ones are unmatched vehicles. Denote the upstream and downstream intersections as $i$ and $i+1$ respectively, and the indexes of the matched vehicles as $\{j, j+1, j+2, \ldots\}$. $t_{i,j}$ represents the discharge timestamp of matched vehicle $j$ at the intersection $i$. Suppose the First-in-First-Out (FIFO) rule is always satisfied, thus for matched vehicles $j$ and $j+1$, the
number of vehicles departed between \( t_{i+1,j} \) and \( t_{i+1,j+1} \) at the downstream intersection would equal to the number of vehicles arrived at the upstream intersection between \( t_i \) and \( t_{i,j+1} \). In other words, the number of arrivals can be observed between two paired matched vehicles. Note that the observed vehicle arrivals could be contributed by one or more upstream movements. For instance, in Fig. 1, the vehicle arrivals between the matched vehicle \( j \) and \( j+1 \) are only from the through movement, whereas the vehicle arrivals between the matched vehicle \( j+1 \) and \( j+2 \) are from the right, mid-block, and left turning movements.

Given the observations of vehicle arrivals on each lane of an approach, it is possible to estimate the lane-based arrival rates associated with different upstream movements. For a particular traffic scenario, the estimated arrival rates would reveal the underlying pattern of the traffic arrival process. It is worth mentioning that the arrival pattern is perceived in the sense of FIFO condition. In this study, all the FIFO violators are considered as unmatched vehicles, and they would be regarded as the vehicles arrived either later or earlier than their actual arrival times. The effects of FIFO violations on arrival rate estimation could be negligible as long as the majority of traffic follows the FIFO condition.

### III. ARRIVAL PATTERN ESTIMATION

#### A. Model Assumptions

The vehicle arrival processes of the controlled and uncontrolled movements are assumed differently. For a controlled movement, two-stage piecewise arrival rates are assumed, shown in Fig. 2. In general, the first stage has a higher arrival rate than the second one because more vehicles will be discharged at the beginning of the green interval. The uncontrolled movements are assumed to have a uniform arrival rate within the signal cycle. Note that the traffic arrival rates are defined over the signal cycle at the upstream intersection.

For notation convention, \( \lambda_d^1 \) and \( \lambda_d^2 \) represent the arrival rates of movement \( d \) at the first and second stages respectively, and \( \tau_d \) is the switch point within the green interval when the arrival rate changes from \( \lambda_d^1 \) to \( \lambda_d^2 \), \( d \in \{l,s,r,u\} \) represents the left, through, right, and uncontrolled turning movements respectively.

With the assumptions of the vehicle arrival processes, the observations of the number of vehicle arrivals are further assumed to follow the Non-homogeneous Poisson Process.

\[
n_{j,j+1} \sim \text{Poisson} \left( \int_{t_{i,j}}^{t_{i,j+1}} \lambda(t) \, dt \right) = \text{Poisson}(\lambda_{j,j+1}) \tag{1}\]

where \( n_{j,j+1} \) is the number of arrivals observed by matched vehicles \( j \) and \( j+1 \), \( \lambda_{j,j+1} \) is the integration of arrival rates (vehicle per second in unit) from \( t_{i,j} \) to \( t_{i,j+1} \).

#### B. Likelihood Function Derivation

With the probability model defined by (1), the log-likelihood function of independent observations can be expressed as:

\[
\ln p(n | \lambda) = \sum_{j=1}^{n} \ln \text{Poisson} \left( n_{j,j+1} | \lambda_{j,j+1} \right) \tag{2}\]

where \( n = (n_{1,2}, \ldots, n_{i,j+1}, \ldots, n_{M-1,M})^T \) and \( \lambda = (\lambda_{1,2}, \lambda_{j,j+1}, \ldots, \lambda_{M-1,M})^T \), \( M \) is the total number of matched vehicles. Note \( \lambda_{j,j+1} \) will take a complex form which depends on \( t_{i,j} \) and \( t_{i,j+1} \). There are two cases to calculate \( \lambda_{j,j+1} \), described as follow.

Case 1: \( t_{i,j} \) and \( t_{i,j+1} \) are both in the same signal phase. Without loss of generality, suppose \( t_{i,j} \) and \( t_{i,j+1} \) are in the same through phase. Case 1 can be further divided into three subcases according to the different locations of \( \tau_s \). For computational convenience, \( t_{i,j} \) and \( t_{i,j+1} \) are all defined as the relative time from the start of the phase.

Subcase 1-1: \( \tau_s \geq t_{i,j+1} \), and the arrival rate between \( t_{i,j} \) and \( t_{i,j+1} \) is constant as \( \lambda_s^1 + \lambda_u \), as shown in Fig. 3(a). Then,

\[
\lambda_{j,j+1} = (\lambda_s^1 + \lambda_u)(t_{i,j+1} - t_{i,j}) \tag{3}\]

Subcase 1-2: \( \tau_s \leq t_{i,j} \), and the arrival rate between \( t_{i,j} \) and \( t_{i,j+1} \) is constant as \( \lambda_s^2 + \lambda_u \), as shown in Fig. 3(b). Then,

\[
\lambda_{j,j+1} = (\lambda_s^2 + \lambda_u)(t_{i,j+1} - t_{i,j}) \tag{4}\]

Subcase 1-3: \( t_{i,j} < \tau_s < t_{i,j+1} \), and the arrival rate between \( t_{i,j} \) and \( t_{i,j+1} \) switches at \( \tau_s \), as shown in Fig. 3(c). Then,

\[
\lambda_{j,j+1} = \lambda_s^1(\tau_s - t_{i,j}) + \lambda_s^2(t_{i,j+1} - \tau_s) + \lambda_u(t_{i,j+1} - t_{i,j}) \tag{5}\]

Case 2: \( t_{i,j} \) and \( t_{i,j+1} \) are in different phases or signal cycles. Without loss of generality, suppose \( t_{i,j} \) and \( t_{i,j+1} \) are in the through and left-turn phases respectively. Depending on the locations of \( \tau_s \) and \( \tau_t \), Case 2 can be further divided into four subcases.

Subcase 2-1: \( \tau_s \leq t_{i,j} \) and \( \tau_t \geq t_{i,j+1} \), as shown in Fig. 3(d). Then,

\[
\lambda_{j,j+1} = \lambda_s^1 t_f^j + \lambda_s^2 t_{f+1}^j + \lambda^+ \tag{6}\]

\[
\lambda^+ = \sum_d n_d(\lambda_s^1 \tau_d + \lambda_s^2 (G_d - \tau_d)) + \lambda_u(t_{i,j+1} - t_{i,j}) \tag{7}\]

where \( t_f^j \) is the time from \( t_{i,j} \) to the end of the through phase, \( t_{f+1}^j \) is the time from the beginning of the left-turn phase to \( t_{i,j+1} \), \( n_d \) and \( G_d \) are the number of phases and their green durations of movement \( d \in \{l,s,r\} \) between the through and left-turn phases, and \( \lambda^+ \) is used to account for the arrivals from phases between the through and left-turn phases.

Subcase 2-2: \( \tau_s \leq t_{i,j} \) and \( \tau_t < t_{i,j+1} \), as shown in Fig. 3(e). Then,
\[
\lambda_{j,j+1} = \lambda^2 \tau_{j}^+ + \lambda_1 \tau_i + \lambda^2 (t_{i,j+1}^{b} - \tau_i) + \lambda^+ \tag{8}
\]

Subcase 2-3: \( \tau_s > t_{i,j} \) and \( \tau_l \geq t_{i,j+1} \), as shown in Fig. 3(f). Then,
\[
\lambda_{j,j+1} = \lambda^2 (\tau_s - t_{i,j}^{b}) + \lambda_2 (G_s - \tau_s) + \lambda_1 \tau_i + \lambda^2 (t_{i,j+1}^{b} - \tau_i) + \lambda^+ \tag{9}
\]

Subcase 2-4: \( \tau_s > t_{i,l} \) and \( \tau_l < t_{i,j+1} \), as shown in Fig. 3(g). Then,
\[
\lambda_{j,j+1} = \lambda^2 (\tau_s - t_{i,j}^{b}) + \lambda_2 (G_s - \tau_s) + \lambda_1 \tau_l + \lambda^2 (t_{i,j+1}^{b} - \tau_i) + \lambda^+ \tag{10}
\]

In Subcase 2-2, 2-3, and 2-4, \( \lambda^+ \) is calculated as (7).

---

**C. FIFO Violation Remove**

Recall that \( n_{j,j+1} \) is only valid under the FIFO condition. To guarantee the FIFO condition, a FIFO check process should be applied to the LPR data, including three steps. First, obtain the cumulative departure sequence \( \{I\} \) for all matched vehicles by their timestamps collected at the downstream intersection. Second, sort the matched vehicles by their timestamps collected at the upstream intersection, and get the updated sequence \( \{I'\} \). Third, check the sequence \( \{I'\} \) using the Longest Increasing Sequence Algorithm [18], and obtain the final sequence \( \{I''\} \). The matched vehicles corresponding to \( \{I''\} \) would satisfy the FIFO condition. Note that the procedures described above are based on a fact that the majority of matched vehicles follow the FIFO condition.

---

**D. Model Parameter Estimation**

Due to the complexity of the likelihood function \( p(n|\lambda) \), the model parameters cannot be easily determined by directly solving a maximum likelihood problem. In this study, the Metropolis-Hastings (M-H) algorithm of the Markov Chain Monte Carlo (MCMC) technique is used to approximate the model parameters. The M-H algorithm is a widely used sampling algorithm to solve complex parameter estimation problems based on Bayesian Theory. Given the initial values of model parameters \( \theta^{(0)} \) and their prior distributions, at each iteration step, the M-H algorithm will accept or reject a new sample generated by the previous sample according to their likelihoods. With a sufficient amount of running time, the sampling process will finally converge to the posterior distribution. Then, the posterior samples can be used to approximate the means or credibility intervals of the unknown parameters. Here, we would like to make the introduction of the M-H algorithm concise. More information about the M-H algorithm and MCMC technique is referred to [19].

Reasonable prior distributions are helpful to achieve efficient and effective sampling of the M-H algorithm. As the arrival rates (i.e., \( \lambda^1_d, \lambda^2_d \) and \( \lambda_u \)) are all positive continuous variables, the exponential distribution is qualified as their prior distributions. The switch points (i.e., \( \tau_d \)) can be viewed as positive discrete variables, so the discrete uniform distribution is adopted as their prior distributions. The prior distributions of the model parameters are defined as follows.

\[
\lambda^1_d \sim \text{Exponential}(a^1_d) \\
\lambda^2_d \sim \text{Exponential}(a^2_d) \\
\lambda_u \sim \text{Exponential}(a_u)
\]

\[
\tau_d \sim \text{DiscreteUniform}(a, b) \tag{11}
\]

where \( a^1_d, a^2_d, a_u, a \) and \( b \) are parameters of the prior distributions, also known as the hyper-parameters.

The settings of the hyper-parameters will affect the initial values of model parameters sampled by the M-H algorithm. Suitable choices of the hyper-parameters may speed up the convergence process of the algorithm. For arrival flow rates, what we can know from the LPR data is the departure flow rate at the downstream intersection. Thus, the mean departure flow rate is used to define \( a^1_d \) and \( a^2_d \) Note. The mean of the exponential distribution is the inverse of its parameter, then \( a^1_d \) and \( a^2_d \) would equal to the inverse of the mean departure flow rate. As \( \lambda_u \) is usually much smaller than \( \lambda^1_d \) and \( \lambda^2_d \), \( \lambda_u \) is chosen to be ten times larger than \( \lambda^1_d \) and \( \lambda^2_d \). As for the switch points, the only information we can know is that they are within the green interval of a signal phase, then the range \([a, b]\) is set to be \([0, G_d]\).

**E. Arrival Rate Estimation**

The posterior samples of model parameters produced by the M-H algorithm could be viewed as the “observations” from the joint distribution of model parameters. According to the law of large numbers, the mean, variance, and credibility interval can be calculated using the posterior samples by simple arithmetical
operations. For our problem, the goal is to obtain second-based mean arrival rates over the signal cycle length. Therefore, it is necessary to convert the posterior samples of model parameters to second-by-second arrival rates.

Without loss of generality, we take the estimation of arrival rates in the through phase as an example. According to the model assumption, the arrival rates in the through phase are contributed by the through and uncontrolled movements. Let \((\lambda_1^{(i)}, \lambda_2^{(i)}, \tau_u^{(i)}, \lambda_u^{(i)})\) be the \(i\)th posterior sample (other unrelated model parameters are not listed). The second-based mean arrival rates in the through phase can be estimated using (12).

\[
\bar{\lambda}(t) = \frac{\sum_{t_\tau \geq t} \lambda_1^{(i)} + \lambda_u^{(i)}}{N} + \frac{\sum_{t_\tau < t} \lambda_2^{(i)} + \lambda_u^{(i)}}{N}
\]

where \(\bar{\lambda}(t)\) is the mean arrival rate of the \(i\)th second in the through phase, \(t \in [1, G_t]\), \(N\) is the total number of posterior samples, \(i \in [1, N]\).

Give a time \(t\), the posterior samples are divided into two parts according to the values of \(t_\tau^{(i)}\), see the numerator of (12). The first part is the set of posterior samples where \(t_\tau^{(i)}\) is after time \(t\), then the arrival rate at time \(t\) equals to \(\lambda_1^{(i)} + \lambda_u^{(i)}\). The second part is the set of posterior samples where \(t_\tau^{(i)}\) is before time \(t\), then the arrival rate at time \(t\) equals to \(\lambda_2^{(i)} + \lambda_u^{(i)}\). In fact, the numerator of (12) is the summation of \(N\) possible arrival rates at time \(t\).

Similarly, it is also convenient to obtain the credibility interval of arrival rate at time \(t\) based on the posterior samples. For example, with a credibility level of 95\%, we just need to calculate the 2.5\% and 97.5\% percentiles from the \(N\) possible values of the arrival rate at time \(t\).

IV. FIELD EXPERIMENT

A. Test Site Description

Two intersections on an arterial named Qianjin Road at Kunshan City in China were selected to test the proposed method. The layout of lanes and LPR sensors is shown in Fig. 4(a). The upstream intersection is defined at the one of Qianjin Road & Zhongshan Road, and the downstream intersection is defined at the one of Qianjin Road & Tinglin Road. The lanes on the eastbound approach of the downstream intersection were investigated for their traffic arrival patterns. The exclusive right-turn lane was not considered as vehicles are allowed to turn on red. Denote the three test lanes as LT, TH1, and TH2 respectively.

Both intersections are operating under fixed-timing signal control with an identical signal timing scheme during the daytime. Fig. 4(b) illustrates the signal timing plan for the upstream intersection. All lanes at both intersections are monitored by LPR sensors. The field data were collected on five weekdays of a typical week from June 4\textsuperscript{th} to June 8\textsuperscript{th}, 2018. Three traffic scenarios were investigated to test the proposed method. Each scenario lasts for 15 minutes: the morning peak (MP) from 8:15 to 8:30 am, the off-peak (OP) from 2:15 to 2:30 pm, and the afternoon peak (AP) from 5:15 to 5:30 pm. The LPR sensors at the test site did not have a good performance, and a significant amount of license plates were unrecognized or erroneously recognized. The match rate of LPR data is 61.6\% on average for the downstream lanes. It is not uncommon that the performance of LPR sensors would degrade after years of usage. Video data were also collected in the corresponding periods and processed manually to fix the unrecognized and erroneously recognized vehicle license plates. The fixed dataset was of match rate at 97.8\% and used to calculate the ground truth for model evaluation.

B. Model Implementation

In our test sites, the right turning movement at the upstream intersection is considered as a controlled movement since the though traffic frequently restricts the free right turning of vehicles on the shared lane. Thus, ten independent parameters are defined to implement the proposed method at the test sites: three parameters (i.e. \(\lambda_1\), \(\lambda_2\), and \(\tau_d\)) for each of the three controlled movements (i.e. through, left and right turning movements), and one parameter \(\lambda_u\) for the uncontrolled movement (i.e. mid-block turning movement). The proposed method is coded in Python with the help of a stochastic programming framework PyMC [20]. The M-H algorithm is configured to collect 5000 posterior samples, where the first 2000 samples would be discarded as a burn-in stage. According to our test, 5000 times of sampling are sufficient to ensure the convergence of the algorithm.

C. Benchmark Method

To better illustrate the performance of the proposed method, a benchmark method is introduced by directly calculating the observed mean arrival rates. For each observation of vehicle arrivals, the arrival rates are assumed to be uniformly distributed between the arrival times of two matched vehicles at the upstream intersection. The mean arrival rates at each unit time of the signal cycle are then calculated by averaging the observed rates provided by all pairs of matched vehicles.
D. Estimation Results

Fig. 5 illustrates the estimation of the mean arrival rate at each second of the signal cycle for the three test lanes using actual LPR data. All three scenarios are investigated for each test lane. The ground truth of the mean arrival rate is calculated using the fixed LPR dataset by dividing the total number of arrivals to the number of signal cycles.

It is shown that the proposed method successfully captures the underlying arrival patterns in the three scenarios for all three test lanes. The two-stage pattern is clearly illustrated, see the through movements (within the range from 0 to 57 second in the cycle). The proposed method also well adapts to the uniform arrival pattern where \( \lambda_d^1 \) and \( \lambda_d^2 \) are mostly identical to each other, see the left turning movements (within the range from 135 to 175 second in the cycle) to the LT and TH1 lanes in the off-peak scenarios. A limitation of the proposed method is also spotted at the start of each signal phase, especially at the start of the through phase. When the signal phase turns to green, the actual arrival rates experienced an increasing trend before reaching a stable higher arrival rate. This could be related to the characteristic of vehicle start-up lost time. The assumption of the two-stage arrival processes restricts the model from capturing such a trend. However, the effects could be minor since the duration of such a stage is relatively short.

By observing the arrival patterns among different scenarios, the most significant change is found in the afternoon peak scenario of the LT lane. This may suggest developing a different signal timing scheme to better accommodate such a change. It is also interesting to find that the utilization of the TH1 lane by traffic is much higher than the TH2 lane. In fact, give the lane-based arrival rates, it is possible to derive a Movement-Lane based OD matrix. This would be helpful to obtain more information about actual traffic behaviors, such as the lane utilization ratio and non-homogenous turning ratio for different merge movements.

E. Impacts of Match Rates

In real applications, the match rates in LPR data could vary significantly. Therefore, it is important to investigate the impacts of different match rates on the performance of arrival pattern estimation. For our test sites, different matching rates were tested for both benchmark and proposed methods. The matching rates are set to be varying from 5% to 90% with 5% as an increment. The matched vehicles are randomly selected in the fixed LPR dataset with probability equals to the match rate. The Root Mean Square Error (RMSE) is used to measure the estimation errors of arrival rates, which is calculated by (13).

\[
RMSE = \sqrt{\frac{1}{C} \sum_{t=1}^{C} (\lambda(t) - \hat{\lambda}(t))^2}
\]  

(13)

where \( C \) is the cycle length, \( \lambda(t) \) and \( \hat{\lambda}(t) \) are the ground truth and estimated arrival rate at time \( t \) respectively. To account for the randomness in the data selection process, the estimation was conducted for 10 repetitions with different random seeds, and the distribution of the RMSE will be investigated.

Fig. 6 illustrates the RMSE of both benchmark and proposed methods for the three lanes and three scenarios. The error bars represent the estimated 95% confidence interval of RMSE generated by the 10 repetitions. The lines connecting the centers of the error bars represent the means of RMSE.
For both methods, it is clear to see that the mean of RMSE is decreasing and the confidence interval is gradually shrinking to a narrow range as the increase of match rate. For the TH1 and TH2 lanes, it is observed that the curves of the two methods intersect at a certain level of match rates (approximately between 40% and 50%). Before the point of intersection, the proposed method performs better with a lower RMSE than the benchmark method. After the intersection, the proposed method quickly converges to a stable performance, whereas the benchmark method is still able to improve slightly and ends up with a lower value of RMSE than the proposed method. For the LT lane, there is no intersection exists between the valuesable of RMSE from the two methods, but the values become closer as the match rate increases.

To get more insights into such results, two typical examples of TH2 lane in the afternoon-peak scenario are provided in Fig. 7. Fig. 7(a) and (b) are corresponding to the estimation results at 15% and 85% match rates respectively. In the 15% match rate example, the benchmark method gives a significantly smoothed arrival rate curve over the signal cycle. This is because the observed arrival rates may cover a relatively long interval at a low match rate. In contrast, the low match rate has less impact on the proposed method. The proposed method is still able to identify the major “peaks” in the actual arrival curves. This could benefit from the nature of Bayesian-based probability models which are able to obtain reliable results even when observations are not sufficient. In the 85% matching rate example, both methods are able to well describe the actual arrival rate curves. The proposed method eliminates the fluctuations in the actual arrival rates, which is helpful to perceive the real patterns of arrival traffic by avoiding to present unnecessarily overfitted results.

![Graph showing RMSE for TH2 lane](image)

**Fig. 7.** Examples of estimation results for TH2 lane in the afternoon-peak.

V. CONCLUSIONS

This study proposed a lane-based traffic arrival pattern estimation method using the LPR data, which provides an opportunity to interpret the patterns of second-based traffic arrival rates of upstream movements to each lane at the downstream intersection. The proposed method models the vehicle arrival process at the upstream intersection as the Non-homogeneous Poisson Process. The likelihood function is derived explicitly according to the observations of the number of vehicle arrivals provided by paired matched vehicles in LPR data. The rate parameters are sampled by the Metropolis-Hasting (M-H) algorithm and further used to estimate the mean arrival rates at each second of the signal cycle of the upstream intersection. The proposed method is validated using actual LPR data collected at two adjacent intersections in Kunshan City, China. The results show that the proposed method can well describe the traffic arrival patterns of upstream merging movements with either two-stage or uniform arrival processes in different traffic scenarios. The impacts of different match rates on arrival rate estimation are also investigated. A simple average-based benchmark method is introduced for comparison. Both benchmark and proposed methods degrade as the match rates decrease. However, the performance of the proposed method is found to be more robust and reliable at low match rates, which is a desirable feature for practical applications when varying quality of LPR data may present.

Several further works can be done to further improve the performance of the proposed method. First, in the paper the model parameters are pre-defined, and it is unnecessary to define the switch point variable for uniform arrival flow, thus there is an opportunity to incorporate the parameter selection process in the model to improve the robustness. Second, the vehicle trajectory data could be integrated into the model to provide additional vehicle arrival information, which may be helpful when some minor movements at the upstream intersection are absent of LPR sensors. Third, the applications of the proposed method can be studied, such as traffic lane choice behavior, vehicle trajectory reconstruction, and signal offset fine-tuning and evaluation.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 71871055, and the Key Research and Development Program of Jiangsu under Grant BE2017027.

ABOUT THE AUTHORS

**Chengchuan An** received his Ph.D. degree in Transportation Engineering from Southeast University, Nanjing, China, in 2019. From 2014 to 2016, he was a visiting scholar at the Department of Civil and Architectural Engineering and Mechanics, University of Arizona, USA. Since 2020, he has been a post-doctor in the Intelligent Transportation System Research Center at Southeast University. His current research interests include intelligent traffic signal control systems and traffic data mining.

**Xiaoyu Guo** received her B.S. degree in Mathematics with a minor in Statistics from SUNY-Buffalo in 2017, and her M.S. degree from the Zachry Department of Civil and Environmental Engineering in 2019, Texas A&M University, where she is currently pursuing the Ph.D. degree. She is currently working
as a Research Assistant for the Connected Infrastructure Group under the System Reliability Division at Texas A&M Transportation Institute. Her research interests include connected infrastructures and vehicles, traffic modeling and simulation, traffic operations, safety, and traffic flow theory.

**Rongrong Hong** received her B.E. degree in Industrial Design and M.S. degree in Mechanical Engineering from Xinjiang University, Urumqi, China, in 2012 and 2012, respectively. She is currently pursuing the Ph.D. degree in Transportation Engineering at Southeast University, Nanjing, China. Her research interests include network traffic state estimation and traffic behavior modeling.

**Zhenbo Lu** received the Ph.D. degree in Traffic Information Engineering and Control from Southeast University, Nanjing, China, in 2011. He is currently an Associate Professor with the Intelligent Transportation System Research Center, Southeast University. His main research interests include transportation planning, traffic simulation, and intelligent transportation systems.

**Jingxin Xia** is a professor at the Intelligent Transportation System Research Center, Southeast University, Nanjing, China. He received the Ph.D. degree in Transportation Engineering from the University of Kentucky, USA in 2006. He has published more than forty peer-reviewed papers so far, and his main research interests include traffic flow theory, transportation network modeling, traffic signal control, and intelligent transportation systems.

**REFERENCES**


