Traffic State Estimation for Urban Road Networks Using a Link Queue Model

Yiming Gu, Zhen (Sean) Qian, and Guohui Zhang

Traffic state estimation (TSE) is used for real-time estimation of the traffic characteristics (such as flow rate, flow speed, and flow density) of each link in a transportation network, provided with sparse observations. The complex urban road dynamics and flow entry and exit on urban roads challenge the application of TSE on large-scale urban road networks. Because of increasingly available data from various sources, such as cell phones, GPS, probe vehicles, and inductive loops, a theoretical framework is needed to fuse all data to best estimate traffic states in large-scale urban networks. In this context, a Bayesian probabilistic model to estimate traffic states is proposed, along with an expectation–maximization extended Kalman filter (EM-EKF) algorithm. The model incorporates a mesoscopic traffic flow propagation model (the link queue model) that can be computationally efficient for large-scale networks. The Bayesian framework can seamlessly integrate multiple data sources for best inferring flow propagation and flow entry and exit along urban roads. A mesoscopic link queue model with reasonable resolution and that is computationally efficient is used to model flow propagations for large-scale urban networks.

Two categories of traffic states are defined here: the traffic infrastructure state (TIS) and the traffic flow state (TFS). TFS is a highly complex, stochastic, and heterogeneously distributed random variable. From a mesoscopic point of view, TFS is represented by the flow rate, speed, and density of vehicles for each link, which are influenced by the current condition of the roads, the traffic signal system, the TFS of its neighboring links, and other factors. TFS also implies the behavior of each vehicle, which is influenced by the characteristics of the drivers, the behavior of its neighboring vehicles, and the current TIS. The proposed model infers both TIS and TFS for urban networks.

Since the introduction of the TSE problem (4), decades of research have sought models and inference algorithms that accurately estimate traffic states. Following is a review of the TSE literature along the three dimensions of research subjects, modeling methods, and inference algorithms.

There are three general research subjects: (a) estimating TFS, (b) estimating TIS, and (c) estimating TFS and TIS together. With respect to estimating only the TFS, early applications were reported in traffic monitoring systems for short interdetector distances (4, 5). Research focused on applying Kalman filter–based methods (6–8). For estimating TIS, most research has focused on the real-time detection of traffic incidents, dating back to the California algorithms (9). Later, various incident detection methods were developed, such as probe-based methods (10, 11), machine learning–based methods (12, 13), and signal processing–based methods (14). Some studies estimated TIS and TFI simultaneously (15, 16).

Most research does not explicitly separate modeling and inference. Instead, the modeling process is usually embedded inside an inference framework. This paper separates modeling and inference by their respective natures: flow and infrastructure modeling approximates the flow propagation constrained by the infrastructure capacity, whereas inference reliably estimates the parameters in the model by using data.

For the modeling methods, TSE can be categorized in three ways: (a) direct observation, (b) hidden state space, and (c) hierarchical Bayesian. For the direct observation model, the data, as well as the traffic states to be estimated, are modeled by directly applying a data-driven regression or classification model (17–20). Although use of these data-driven models leverages the established framework of distributed computing, such as the Streaming Spark framework used by Hunter et al. (20), the drawback is obvious: without the underlying mechanism of a flow model, it is difficult to estimate most traffic states that can be hardly covered by sparse or limited historical observations. The most used model in traffic state estimation...
is the hidden state space model \((8, 16)\). Wang and Papageorgiou used a hidden state model to successfully estimate vehicle densities on a segment of highway \((8)\). Hidden state space models have two layers of random variables: a layer of hidden states and a layer of observations. Traffic states to be estimated are assumed to be hidden states. The transition of hidden states is modeled by various traffic flow models, usually classified as microscopic model and mesoscopic models. Microscopic models have been proposed to describe movements of individual vehicles \((21–23)\). In contrast to microscopic models, mesoscopic models describe the traffic states at an aggregated level, such as a segment of road (typically a few hundred feet). The most used mesoscopic models are the Lighthill–Whitham–Richards (LWR) model \((24, 25)\) and gas kinetic models \((26)\), where the basic computational unit is a small segment of a road, referred as a cell. LWR can be numerically solved with the cell transmission model (CTM) \((27)\). The CTM has been calibrated and applied to estimate traffic states in a subsection of highway \((8)\). There are mesoscopic flow models that assume a larger segment of road spanning from intersections to intersections \((28, 29)\). The third category, the hierarchical Bayesian model, is a more general case of a hidden state space model. It comes with more than two layers of random variables to be learned from data \((30)\). Also, Bayesian inference has been applied on the particle filtering–based LWR model to illustrate that the Bayesian method can quickly adapt to an accident on I-880 and can accurately estimate the change in the traffic flow regime \((31)\).

There are three main types of inference algorithms: \((a)\) Gaussian approximate methods, \((b)\) Monte Carlo methods, and \((c)\) variational inference methods. Most research uses Gaussian approximate methods, exemplified by the Kalman filter and the extended Kalman filter (EKF) \((8)\). The major drawback of Gaussian mixture models is their inability to estimate a posterior distribution being close to non-Gaussian. Recently, Monte Carlo methods have gained popularity because they are unbiased estimators, such as the particle filter \((32)\). However, Monte Carlo methods tend to use a lot of computational resources (in both time and space), and certain Monte Carlo methods such as Markov chain Monte Carlo may not necessarily converge. It prevents Monte Carlo methods from scaling for a large-scale network. The third category, variational inference, solves the target posterior distribution by assuming a simpler variational distribution and iteratively optimizing it. Hofleitner et al. used an expectation–maximization algorithm to estimate traffic states and parameters \((30)\). The benefit in the use of variational inference is that it is more computationally efficient. However, variational inference is biased, and there is a lack of theory to support the asymptotic properties of variational inference.

The goal for this research was to develop a traffic state estimator that

1. Can estimate flow characteristics in a large-scale urban road network considering the flow entry and exit on the streets and complex flow propagation,
2. Uses a general model to integrate heterogeneous data sources effectively for the best estimation of traffic flow characteristics, and
3. Is computationally efficient for large-scale networks while being reasonably accurate.

**PROBLEM DEFINITION**

Given the observations of TFS (such as vehicle speed, travel time, or density) as well as observations of TIS (such as weather and incident), the key issues are as follows:

1. How to estimate the traffic states on those road segments that are not covered by sensors, for current and previous time periods;
2. How to predict the traffic states on those road segments covered by sensors, for future time periods;
3. How to predict the traffic states on those road segments that are not covered by sensors, for future time periods; and
4. How to calibrate the model to adapt to the unknown traffic incidents and consequently detect these unknown traffic incidents to perform a better estimation.

To answer these questions in a well-defined mathematical form, first, the traffic network is modeled as random variables in a Bayesian network, as shown in Figure 1. In the figure, the blue circles indicate

![FIGURE 1 Graphical representation of problem of traffic state estimation (q = flow; d = density; v = space–mean speed; i = time; j = location; w = weather; c = incident).](attachment:image.png)
the random variables, and the lines indicate the correlation between two random variables. In a typical traffic network, there are random variables such as traffic flow, traffic density, and traffic speed. All these random variables have two dimensions: time and space. Time and space are discretized into integers. For the set of a time sequence,

\[ T = \{1, 2, 3, \ldots , j, \ldots , M \} \]  

and for the set of a space sequence,

\[ D = \{1, 2, 3, \ldots , i, \ldots , N \} \]  

Each random variable of traffic state has two indexes. For example, the traffic flow at time \( i \) and location \( j \) is noted as \( q^{i,j} \). Similarly, two other TFS are written as follows: traffic density, \( d^{i,j} \), and space–mean speed, \( v^{i,j} \). Also, for TIS, there are weather, \( w^{i,j} \), and incident, \( c^{i,j} \). The discretized time and space are shown in Figure 1 as the horizontal and vertical axes, respectively.

\[ \{q^{i,j}, d^{i,j}, v^{i,j}, w^{i,j}, c^{i,j}\} \quad \forall i \in T; \forall j \in D \]  

Some of the random variables have been observed in this traffic network. The set of the observed variables is denoted \( X \), with a realized value \( X = x \). The set of unobserved variables is denoted as \( Y \). The problem in this research is simple: given the observed traffic states \( X = x \), estimate \( a \) the probability distribution of unobserved traffic states \( Y \) and \( b \) the probability distribution of the unknown model parameters, \( \alpha \). The general definition of unobserved variables and observed variables indicates a broad application: the unobserved states could be the TFS or traffic incidents at any locations and time periods, and the observed states could be the other TFS or incident at any other locations and time periods. Formally, the objective for this research was to find the posterior conditional distribution of

\[ Y, \alpha | X = x \]  

Definition 1. Temporal estimation. When the observed TFS \( (X) \) are at all locations from time \( 0 \) to \( j \), and the unknown TFS \( (Y) \) are at all locations from time \( j + 1 \) to \( j + k \) \((k \) is a positive integer), and when all TIS are known and stable over time, the estimation is called “temporal estimation.”

Definition 2. Spatial–temporal estimation. When the observed TFS \( (X) \) are at locations \( 0 \) to \( i \) from time \( 0 \) to \( j \) and \( k \), and at locations \( i \) to \( i + z \) from time \( 0 \) to \( j \), and the unknown TFS \( (Y) \) are at locations \( i \) to \( i + z \) from time \( j + 1 \) to \( j + k \) \((k \) and \( z \) are positive integers), and when all TIS are known and stable over time, this estimation is called “spatial–temporal estimation.” Temporal estimation is a special case of spatial–temporal estimation.

Definition 3. Incident adaptation. During spatial–temporal estimation, one TIS, traffic incidents, is unknown and changes over time, causing the closure and reopening of lanes in the traffic network. Under this condition, the problem of incident adaptation requires not only spatial–temporal estimation but also inferring the occurrence of the traffic incident (equivalently the change in flow capacity of each link) in the traffic network.

The task of temporal estimation answers Question 2, the task of spatial–temporal estimation answers Questions 1 and 3, and the task of incident adaptation answers Question 4. To solve the three estimation tasks in an urban network with flow entry and exit, a series of methodologies and algorithms is developed in the next section. Then, a small-scale experiment validates the accuracy and the performance of the methodology. Finally, some conclusions are provided.

METHODOLOGY

This section consists of two parts. The first part, traffic flow dynamics, models the random variables of flow rate, flow density, average speed, weather, and incident in detail. The second part, inference, infers the unknown random variables given observation data and solves the three questions discussed above.

Traffic Flow Propagation: Fundamental Diagrams and a Link Queue Model

This section recaps a link queue model proposed by Jeong et al. with some minor modifications (14). It specifies the spatiotemporal relationships among flow density \( d^{i,j} \), flow rate \( q^{i,j} \), and average speed \( v^{i,j} \), as well as incident \( c^{i,j} \), at any location \( j \) and time \( i \). Specifically, fundamental diagrams are used to model the relationship among flow density \( d^{i,j} \), flow rate \( q^{i,j} \), and average speed \( v^{i,j} \), as well as incident \( c^{i,j} \), at different locations and different times.

Because fundamental diagrams model the correlations of the variables at the same time and location, for simplicity they are written as \( d, q, v, \) and \( c \). This paper uses triangular fundamental diagrams defined as follows:

Definition 4. Fundamental diagrams:

\[ q_{d} = \min (F_{f}, \min (w(J - d)) \quad d \in [0, J] \]  

where

\[ J = \text{jam density}, \quad F_{f} = \text{free-flow speed, and} \]  

\[ w = \text{backward wave speed}. \]

The flow rate attains its maximum, flow capacity \( C \), at the critical density \( d_{c} \), that is, \( C = q_{a}(d), \forall d \).

The preceding fundamental diagrams are centered on the density, meaning knowing only the density is enough to calculate the other two variables of TFS. As for the impact of traffic incident to the fundamental diagrams, traffic incidents such as accidents, road work, or police activity can change the number of lanes available and thus pose a bottleneck. Therefore, a simple relationship is proposed between fundamental diagrams and the traffic incidents as follows.

Definition 5. Incident impact on capacity reduction. Traffic incidents change the number of lanes (or flow capacity in general) on a link,

\[ q_{a} = \min (F_{f}, \min (w(J - d)) \quad d \in [0, J] \]  

where \( \gamma \) is the decreasing factor related to incidents.

The relationships between the TFS \( (F^{i,j}) \) across all space \( j \) and time \( i \) are modeled with the link queue model (14). The elementary
unit is a link. The reasons the link queue model was chosen as the TFS transition model are as follows:

1. Compared with the cell transmission model, the link queue model is computationally inexpensive, meaning the link queue model may scale better to large-scale networks.
2. The link queue model rigorously describes interactions among links by using link demands, supplies, and junction models, which is consistent with mesoscopic merging and diverging behavior.

Now the specific form of the TFS transition function is written with the link queue model. In the link queue model, traffic on each link is considered a queue, and the state of a queue is its density (the number of vehicles per unit length). Based on the fundamental diagrams of the link, the link queue model defines the demand (maximum sending flow) and supply (maximum receiving flow) of a queue. Then the outfluxes of upstream queues and influxes of downstream queues at a junction are determined by mesoscopic merging and diverging rules.

The link queue model introduces two intermediate variables in the TFS in time \(i\) and space \(j\): demand \(D^{(i)}_j\) and supply \(S^{(i)}_j\).

**Definition 6. Demand:**

\[
D^{(i)}_j = q_{d,j} \left( \min \left\{ d^{(i)}_j, d_{c,j} \right\} \right) \tag{7}
\]

\[
= \begin{cases} 
q_{d,j}(d^{(i)}_j) & \text{if } d^{(i)}_j \in [0, d_{c,j}] \\
C_j & \text{otherwise } d^{(i)}_j \in \left[ d_{c,j}, d_{a,j} \right] 
\end{cases} \tag{8}
\]

where

- \(q_{d,j}\) = jam density at link \(j\),
- \(d_{c,j}\) = critical density at link \(j\),
- \(C_j\) = maximum flow rate at link \(j\), and
- \(d_{a,j}\) = maximum density at link \(j\).

**Definition 7. Supply:**

\[
S^{(i)}_j = q_{d,j} \left( \max \left\{ d^{(i)}_j, d_{a,j} \right\} \right) \tag{9}
\]

\[
= \begin{cases} 
C_j & \text{if } d^{(i)}_j \in [0, d_{c,j}] \\
q_{d,j}(d^{(i)}_j) & \text{otherwise } d^{(i)}_j \in \left[ d_{c,j}, d_{a,j} \right] 
\end{cases} \tag{10}
\]

where \(d_{a,j}\) is the maximum density at link \(j\).

Also for each link are defined the influx (the traffic flow that enters the link \(j\)), \(f_r\), and outflux (the traffic flow that leaves the link \(j\)), \(g_j\). Further defined are traffic junctions \(J\) as the intersection among traffic links, such as highway merging and diverging sections. At junction \(j\), outfluxes, \(g_j\), and influxes, \(f_r\), can be computed from upstream demands, \(D_j\), and downstream supplies, \(S_j\). For a junction with multiple inflow links and outflow links, \(g_j\) and \(f_r\) can be vectors of multiple random variables. To determine the transmission within traffic junctions, the flux function is defined.

**Definition 8. Flux function \(f_j()\) is a function that governs**

\[
(g_j, f_r) = f_j(D_j, S_j) \tag{11}
\]

A general flux function (Figure 2) with \(m\) inflows and \(n\) outflows at junction \(J\) was given by Jeong et al. (14).

1. From the definition of the demand and supply, all the upstream demand and downstream supply can be calculated by the flow density of upstream links \((d_i : d_m)\) and downstream links \((d_m : d_n)\).
2. The turning proportion \(K_{a,ab}\) from an upstream link \(a \in \{1 : m\}\) to a downstream link \(b \in \{m + 1 : m + n\}\) is an independent variable of route choice.
3. The outflux of upstream link \(a\) is

\[
g_a = \min \{D_a, 0, C_a\} \tag{12}
\]

where \(C_a\) is the maximum flow rate and \(0\) is the critical demand level, uniquely solves the following minimum–maximum problem:

\[
\theta_j = \min \max D_a, C_a, \min \max C_a, S_a, \sum_{a \in \{1 : m\}} K_{a,ab} A_j
\]

where \(A_j\) is a nonempty subset of \(\{1 : m\}\)

4. The influx of the downstream link \(b\) is

\[
f_b = \sum_{a \in \{1 : m\}} g_a K_{a,ab} \tag{14}
\]

To get the relationship between the flux and the traffic density, the conservation law of traffic flow is used.

**Definition 9. The conservation law of traffic flow at time \(i\) and space \(j\) is**

\[
\frac{dd^{(i)}_j}{dt} = \frac{1}{L_j} (f^{(i)}_j - g^{(i)}_j) \tag{15}
\]

The preceding definitions describe the link queue model: Item 1 calculates the supply and demand at each link from the density at each link, Item 2 calculates the influx and outflux among links from the supply and demand calculated before, and Item 3 calculates the new density at each link after knowing the influx and outflux. In general, the definition, along with the fundamental diagrams, specifies a nonlinear system from all traffic densities at time \(i\), \(d^{(i)}\) to all traffic at time \(i + 1\), \(d^{(i+1)}\).
Definition 10. The link queue model specifies a nonlinear function such that

\[ d_{k+1}^{u_0} = F(d_{k}^{u_0}) \]  

(16)

**Inference: EKF and Expectation–Maximization**

The preceding section showed how random variables in the traffic network are related. In other words, given the boundary conditions such as the flow rate at all the links in the boundary regions of the network, and the initial conditions such as the flow density at all the links in the traffic network at time step 0, the traffic flow in the entire network could be simulated.

Definition 11. The task of inference in the traffic state estimation is to estimate the unknown traffic states \( Y \) and unknown model parameters \( \alpha \) given observations \( X \) and to estimate the following distribution:

\[ Y, \alpha | X \]  

(17)

This estimation is pursued in the framework of expectation–maximization. Without mathematical proof given, the algorithm of expectation–maximization is written as the repetition of the expectation step (E-step) and the maximization step (M-step) until convergence. Specifically, expectation–maximization is performed as follows:

1. E-step: estimate \( Y|X, \alpha \).
2. M-step: estimate \( \alpha|Y, X \).
3. Repeat 1 and 2 until convergence.

The E-step is performed as the EKF. In the context of the link queue model, the EKF is described as follows. First, it is assumed that there is only one state variable, traffic density \( d \), at time \( i \) for all links.

Also, there are observations about the travel speed on every link \( v \) at time \( i \). All other model parameters are assumed to be known. Therefore, the inference task is simplified to find the posterior of traffic densities given the initial conditions, boundary conditions, and the observations, up to the \( k \)th step:

\[ d_{i,k} | v_{i,k} \]  

(18)

From the link queue model and fundamental diagrams, the state transition function \( F() \) and the observation function \( h() \), where \( v = h(d) \), are known. Specifically,

\[ d_i = F(d_{i-1}) + w_{i-1} \]  

(19)

\[ v_i = h(d_i) + v_i \]  

(20)

where \( v_i \) and \( w_i \) are zero-mean independent and identically distributed Gaussian noise with known variance. The initial state \( d_0 \) with mean \( \mu_0 \) and variance \( P_0 \) and the boundary conditions are also known.

\[ E[d_i|v_{i,k}] = d_i^* \]  

\[ = d_i + K_i(y_i + \hat{\gamma}_i) \]  

(21)

(22)

where

- \( d_i^* \) = posterior mean of the state variable \( d_i \),
- \( d_i = \) state estimation from the previous state,
- \( K_i = \) Kalman gain to be calculated,
- \( y_i = \) observation, and
- \( \hat{\gamma}_i = \) projected estimation from \( d_i^* \).

The steps of the EKF are as follows:

1. **Initialization:**

\[ d_0 \sim N(\mu_0, P_0) \]  

(23)

2. **Model forecast.** Run the link queue model on the initial condition and boundary conditions:

\[ d_i = F(d_{i-1}) \]  

(24)

\[ P_i = j_i(d_{i-1})P_{i-1}j_i^T(d_{i-1}) + Q_{i-1} \]  

(25)

where \( j_i \) is the Jacobian matrix of the function \( F \) and \( Q_{i-1} \) is the known covariance matrix of the state transition errors.

3. **Data correction:**

\[ d_i = d_i^* + K_i(y_i - h(d_i)) \]  

(26)

\[ K_i = P_i j_i^T(d_i^*) (j_i(x_i)P_i j_i^T(x_i) + R_i)^{-1} \]  

(27)

\[ P_i = (I - K_i j_i(d_i^*)) P_i^* \]  

(28)

where \( j_i \) is the Jacobian matrix of function \( h() \) and \( R_i \) is the observation error covariance matrix.

Following the steps above, the state estimation can be solved sequentially, such that

\[ d_{i,k} | v_{i,k} \sim N(d_{i,k}^*, P_{i,k}) \]  

(29)

Equation 29 is an example of the EKF solution of \( Y|X, \alpha \) in the E-step. After the E-step, observation \( X \), model parameter \( \alpha \), and estimated unknown traffic states \( Y \) are available. The M-step is used to minimize the conflict between the estimated unknown traffic states \( Y \) and observations \( X \) by modifying the model parameter \( \alpha \).

The fundamental diagrams and link queue model are used to obtain the observation function \( h() \), giving the pseudo-observation

\[ X^- = h(Y) \]  

(30)

Therefore, minimizing the difference between \( X^- \) and \( X \) can lead to a new iteration of model parameter \( \alpha \) as defined in the expectation–maximization algorithm. The M-step is

\[ \alpha^* = \arg\min_{\alpha} \|X - X^-\|^2 \]  

(31)

where \( \alpha^* \) is the new iteration of \( \alpha \) and \( \| \ldots \|^2 \) means the Euclidean norm. Here, \( \alpha \) incorporates the incident adaptation. Any incident detection, quantified as the capacity reduction, can be added to the equation as the additional constraints of \( \alpha \). In this paper, this
for each time step, the entire network has four TFS variables, \( i \), \( d \), \( T \), duration is \( T \) to generate the ground truth for the network flow. The simulation in the beginning, all the links are empty. Vissim simulation is used to test the accuracy of EKF without expectation maximization, a vanilla temporal estimation is defined. With all the model parameters, initial conditions, and boundary conditions given, the details of these four estimation tasks in this experiment are defined as follows:

1. Vanilla temporal estimation. Given the average speed observation of Links 1 through 4 at time 0 to 1.05 h, the task is to estimate the flow density and flow rate of Links 1 through 4 at time 0 to 1.05 h.
2. Temporal estimation. Given the average speed observation of Links 1 through 4 at time 0 to 0.5 h, the task is to estimate the flow density and flow rate of Links 1 through 4 at time 0 to 1.05 h.
3. Spatial–temporal estimation. Given the speed observation of Links 1, 2, and 4 at time 0 to 1.05 h and speed observation of Link 3 at time 0 to 0.5 h, the task is to estimate the flow density of Link 3 at time 0 to 1.05 h.
4. Incident adaptation. Under spatial–temporal estimation, the number of lanes at Link 3 drops from two to one at time point 0.5 h. However, this is unknown to the model. The task is to estimate the flow density of Link 3 at time 0.5 to 1.05 h as well as the number of lanes at Link 3.

Three categories of estimation are investigated: temporal estimation, spatial–temporal estimation, and incident adaptation. Also, to test the accuracy of EKF without expectation maximization, a vanilla temporal estimation is defined. With all the model parameters, initial conditions, and boundary conditions given, the details of these four estimation tasks in this experiment are defined as follows:

1. Vanilla temporal estimation. Given the average speed observation of Links 1 through 4 at time 0 to 1.05 h, the task is to estimate the flow density and flow rate of Links 1 through 4 at time 0 to 1.05 h.
2. Temporal estimation. Given the average speed observation of Links 1 through 4 at time 0 to 0.5 h, the task is to estimate the flow density and flow rate of Links 1 through 4 at time 0 to 1.05 h.
3. Spatial–temporal estimation. Given the speed observation of Links 1, 2, and 4 at time 0 to 1.05 h and speed observation of Link 3 at time 0 to 0.5 h, the task is to estimate the flow density of Link 3 at time 0.5 to 1.05 h.
4. Incident adaptation. Under spatial–temporal estimation, the number of lanes at Link 3 drops from two to one at time point 0.5 h. However, this is unknown to the model. The task is to estimate the flow density of Link 3 at time 0.5 to 1.05 h as well as the number of lanes at Link 3.

Illustrating the result of vanilla temporal estimation, the ground truth and the estimation of the traffic state variables are shown in Figure 4, a and b. Figure 4a illustrates the result of density estimation for Links 1 through 4 in 0 to 1.05 h given the average speed for Links 1 through 4 in 0 to 1.05 h. Figure 4a shows that the estimated traffic density approaches the true density as time passes because EKF uses the input of the average speed (the observations of true traffic density) to correct its estimation. The mean absolute error of the traffic states estimation of all four links is 24.8 vehicles per mile. With respect to flow rate (Figure 4b), the finding is similar, and the estimated flow rate from EKF approaches the true flow rate as it goes.

A comparison of vanilla temporal estimation and temporal estimation is shown in Figure 4, c and d. In vanilla temporal estimation, the average speed for all links at 0.5 to 1.05 h is available, but in temporal estimation, these data are not available. Figure 4, c and d, shows that these extra data in vanilla temporal estimation have a positive influence on the estimation results in both density and flow rate. The mean absolute error for vanilla temporal estimation is 34.9 vehicles per mile and that of temporal estimation is 37.5 vehicles per mile. Specifically, the estimation results of temporal estimation and vanilla temporal estimation overlap at 0 to 0.5 h because the data given in these two estimations are the same. After 0.5 h, since the observation data is on the average speed in Links 1 through 4 available in vanilla temporal estimation but not in temporal estimation, the result of vanilla temporal estimation is better than the result of temporal estimation. The reason for that is clear: the input of extra observation data helps vanilla temporal estimation to continuously

NUMERICAL AND REAL-WORLD EXPERIMENTS

Synthetic Small Network

The synthetic network of the experiment is shown in Figure 3. In the beginning, all the links are empty. Vissim simulation is used to test the accuracy of EKF without expectation maximization, a vanilla temporal estimation is defined. With all the model parameters, initial conditions, and boundary conditions given, the details of these four estimation tasks in this experiment are defined as follows:

1. Build a microscopic simulation model (Vissim) as shown in Figure 3.
2. Load the microscopic simulation model with constant vehicle input of \( d_i(t) = 7,020 \) vph.
3. Collect the flow rate, average speed, and flow density from the Vissim model by using the COM interface of Vissim.
4. Use the link queue model to approximate the flow propagation function. This link queue model approximates the state transition function \( F \) in EKF. Also, the fundamental diagrams specified earlier serve as an observation function \( h(t) \) in EKF.
5. Use the observations from the simulated model, the state transition function \( F(t) \) and observation function \( h(t) \), and the specified initial condition and boundary conditions above to run the EM-EKF algorithm.
6. Compare the state variables from the simulation to those derived from EM-EKF.

![FIGURE 3 Experiment setup.](image)
correct its estimates. But for temporal estimation, only the base model with given parameters can be used to make the estimation.

The results of the spatial–temporal estimation for Link 3 are shown in Figure 5, a and b. In both figures, the first half shows the traffic state estimation with observation and the second half shows the traffic state estimation without observation, which satisfies the definition of temporal–spatial estimation. The mean absolute error for spatial–temporal estimation of density is 18.5 vehicles per mile. It achieves a better result than both vanilla temporal estimation and temporal estimation. Comparing the spatial–temporal estimation in Figure 5a with vanilla temporal estimation in Figure 4a, according to the definitions of vanilla temporal estimation and spatial–temporal estimation, shows that the spatial–temporal estimation with EM-EKF achieves a better result with even fewer data. The reason is that with the expectation–maximization algorithm, one not only can make an optimal estimation of the unknown traffic states but can estimate optimal parameters for flow dynamics, such as the number of lanes (or capacity reduction).

Figure 5, c and d, illustrates the results of the last task: incident adaptation. As defined above, at the time point of 0.5 h, the number of lanes in Link 3 drops from two to one, which results in a drop in density in Link 3 and more congestion in the network. Figure 5c shows that the adaptability of the expectation–maximization algorithm can be exploited to make an accurate estimation of flow density even without knowledge that the number of lanes actually drops from two to one. As compared with the spatial–temporal estimation in Figure 5a, the mean absolute error of density estimation increases from 18.5 vehicles per mile to 21.4, because without knowing about the change in the number of lanes, the EM-EKF needs a transition period to react to this lane closure, and this transition period could introduce more error. This conclusion is verified in Figure 5d, which explicitly shows the change of the model parameter, number of lanes, in the EM-EKF. Figure 5d shows that the number of lanes drops from two to nearly one gradually in about 15 min.

The analysis in Figure 5, e and f, illustrates the sensitivity of the EM-EKF algorithm to the location of the data source. In Figure 5, c and d, the average speed data from Links 1, 2, and 4 is given. In Figure 5, e and f, however, average speed data from Links 1 and 2 and from Links 1 and 4 are given, respectively (average speed data for Link 3 at 0 to 0.5 h are always given by the definition of incident adaptation). A comparison of Figure 5, c, e, and f, shows that: (a) the result in Figure 5c is better than the results in both Figures 5e and 5f,
FIGURE 5  Link 3: (a) spatial–temporal estimation (density in vpm), (b) spatial–temporal estimation (flow rate in vph), (c) incident adaptation (density in vpm), (d) incident adaptation (number of lanes), (e) incident adaptation (given data from Links 1 and 2 only), and (f) incident adaptation (given data from Links 1 and 4 only) (spa.–temp. = spatial–temporal).
meaning that in this case having more data is also preferred; (b) the
result in Figure 5e is better than the result in Figure 5f, meaning that
with respect to estimating the flow density in Link 3, obtaining the
data from Link 4 is more useful than the data from Link 2. It is also
shown that the model can perform robustly with missing and partial
data, and the amount of data available has a direct influence on the
performance of this data-driven model—the estimation error of the
model tends to increase when the coverage of data is decreasing.

Figure 6 illustrates the performance of EM-EKF with and with-
out data. As in Figure 5, the dark blue line shows the ground truth
and the green line shows the EM-EKF estimation of flow density
given the speed data at Links 1, 2, and 4 at 0 to 1.05 h and the
speed data at Link 3 at 0 to 0.5 h. However, if the speed data at
Links 1, 2, and 4 are given only for 0 to 30 min, 0 to 36 min, and
0 to 42 min, the result of EM-EKF is shown in red, light blue, and
purple, respectively. The first observation is that the ability of inci-
dent adaptation in the EM-EKF algorithm is enabled by constantly
measured data: if there are measured data, EM-EKF can adapt to the
incidents, and if there are no measured data, EM-EKF can act as a
tuned simulator, to predict the future traffic states given the current
model parameters and traffic states. For instance, in the scenario
labeled “given data 0–30 min” (red line), because there is no incident
at 0 to 30 min, and there are no data after 30 min, the traffic states
estimation after 30 min can be predicted only with the current model
parameters, which assume there is no incident. Therefore, although
there is a lane reduction at 30 min, EM-EKF cannot adapt to this
incident without data and so makes a less reasonable traffic state
estimation compared with the ground truth. During the period of
30 to 42 min, EM-EKF gradually learns from the measured data and
adapts to the incident. As shown by the purple line in Figure 6,
if the measurement is discontinued from 42 min, EM-EKF can still
make a comparatively fair estimation since the model has been fully
adapted to the incident. If one compares this with the light blue line
in Figure 6, where only 0 to 36 min of data are given, in the scenario
labeled “given data 0–36 min,” the model is only halfway toward
the full incident adaptation, and consequently the estimation result
lies between the red line (incident not adapted) and the purple line
(incident fully adapted).

Real-World Application

To test the proposed methodology in the real world, the EM-EKF
algorithm was applied for a small arterial network in the Washington,
D.C., area. Figure 7 shows the configuration of the small arterial net-
work, which consists of 11 arterial links, in red. In this network, there
are two microwave-based speed detectors, D1 and D2, in blue, which
are located on Link L1 and Link L11, respectively. The speed data
were collected every 1 min from 9:00 to 18:00 every day. Additionally,
GPS-based probe vehicle speed data from L2 to L11 were provided by
INRIX. Compared with microwave-based speed data, INRIX probe
data are sparse during the day. Also, data are available for speed limits
and lane numbers for Links L1 to L11. These physical parameters are
used to construct the initial fundamental diagrams. Finally, the initial
turning proportions are set to be proportional to the number of lanes
and are assumed to be weekly periodical.

This experiment used speed data from L1 to L10 for February 1,
2015, to March 28, 2015, to train the model parameters such as
the parameters in fundamental diagrams and turning proportions.
Also, to mitigate the influence of the traffic lights, all the speed
data were smoothed by a moving average function over a 15-min
window before use. The goal on March 29, 2015, was to estimate
the travel speed at Link L11 given the travel speed at L1 to L10
in real time, without using any historical data for Link L11. The
estimated travel speed at L11 and the observed travel speed at L11
from D2 are compared in Figure 8. The figure shows that, in general,
the estimation is reasonably accurate, with an average error rate of 8.5%. Although some peaks are not fully captured by the estimator, the general decreasing or increasing trend of travel speed is captured by the EM-EKF algorithm. This small experiment showed that EM-EKF could estimate traffic states in an urban arterial network.

CONCLUSIONS

In this research, the problem of traffic state estimation was formulated in a Bayesian probabilistic framework to find the posterior distribution of unknown traffic states and unknown model parameters given sampled observation data. With use of the general framework of a Bayesian probabilistic model, various data sources can be integrated in the model to make the estimation on all traffic flow and infrastructure characteristics. Extending this definition of traffic state estimation in a Bayesian probabilistic framework, three estimation tasks were proposed: (a) temporal estimation, (b) spatial–temporal estimation, and (c) incident adaptation. To solve these three tasks, a mesoscopic traffic flow propagation model, the link queue model proposed by Jeong et al. was used (14). The link queue model is computationally more efficient than the classical cell transmission model. It can model a more complex large-scale urban network with flow entry and exit at any location of streets. Based on the link queue model, the developed EM-EKF algorithm estimates traffic states and infers parameters given limited observations.

To validate the effectiveness of the EM-EKF algorithm, an experiment based on the synthetic network and simulation data from microscopic simulation software (Vissim) was performed. The three estimation tasks—temporal estimation, spatial–temporal estimation, and incident adaptation—were fully performed and tested. First, with the EM-EKF, all three estimation tasks can be completed with less than 10% error. For the temporal estimation, the results show that although having more observations over the period of unknown traffic states can have a positive influence on the overall performance, EKF can perform the temporal estimation for future traffic states even without knowing any observations in the future. By comparing spatial–temporal estimation from EM-EKF and temporal estimation from EKF, because of the optimization properties of expectation–maximization, the EM-EKF algorithm can achieve better results even with fewer data than EKF alone. For incident adaptation, it was shown that EM-EKF can adapt to unknown traffic disruptions, such as lane closures caused by traffic incidents, by continuously correcting and optimizing the model parameters. Moreover, in reaction to a sudden disruption such as a lane closure, EM-EKF gradually changes its model parameters. In this experiment, EM-EKF takes about 15 min to fully adapt to a sudden lane closure, learned by itself. Also, different data provided to the EM-EKF algorithm can lead to different results: although having more data always has a positive influence on estimation accuracy, observation data from some locations are more critical than those from other locations in terms of estimation accuracy. Finally, EM-EKF was applied to a real-world urban arterial network in the Washington, D.C., area. The results show that the EM-EKF algorithm can reliably estimate traffic states with an error rate of 8.5%.

ACKNOWLEDGMENTS

This research was funded in part by Carnegie Mellon University’s Traffic 21 Institute and Technologies for Safe and Efficient Transportation, a U.S. Department of Transportation University Transportation Center for Safety.

REFERENCES


The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented. The U.S. government and state governments assume no liability for the contents or their use.

The Standing Committee on Traffic Flow Theory and Characteristics peer-reviewed this paper.