CALIBRATION OF A MACROSCOPIC TRAFFIC FLOW MODEL WITH STOCHASTIC SATURATION RATES

Li Jin
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
jnl@mit.edu

Saurabh Amin
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
amins@mit.edu

Word Count: 1053 words + 9 figure(s) + 2 table(s) = 3803 words
Submission Date: November 15, 2016
ABSTRACT

It is well known that freeway capacity (saturation rate) is highly uncertain and that a nominal value alone may not suffice for traffic management purposes. To analyze the impact of capacity disruptions, we consider a macroscopic traffic flow model with stochastic saturation rates, called the stochastic switching cell transmission model (SS-CTM). The SS-CTM enables analysis of traffic delay and throughput drop due to uncertainty of capacity. In this article, we present a method for calibration of parameters of the SS-CTM using real traffic data from the Caltrans Performance Measurement System (PeMS). Nominal parameters except saturation rates are estimated using a linear regression-based method. The stochastic saturation rate model is developed by clustering of observed saturation rates (mode identification) and constructing a Markov chain governing the switches between the identified modes. We also apply the proposed method to a segment of the US Route 101. The results imply that the saturation rate may deviate from the nominal value for a remarkable fraction of time. In addition, correlation between saturation rates at adjacent locations is significant and is thus relevant for delay and throughput analysis.

Keywords: Freeway capacity, Saturation rate, Disruption, Cell transmission model.
1 INTRODUCTION

Capacity is a critical factor for performance (typically throughput and delay) evaluation and traffic assignment. The 2000 Highway Capacity Manual (1) defines freeway capacity as "the maximum hourly rate at which persons or vehicles reasonably can be expected to traverse a point or a uniform section of a lane or roadway during a given time period under prevailing roadway, traffic, and control conditions," which suggests the use of a design or nominal value. Traditional dynamic traffic flow models (2, 3) largely adopt this definition and assume a nominal capacity and a deterministic flow-density relation.

However, considering only a nominal capacity does not allow estimation of expected traffic delay and throughput loss due to capacity disruptions. It is well known that capacity, or saturation rate,1 can be time-varying and highly uncertain (4–10). Possible reasons for fluctuation of saturation rate include traffic incidents, weather, and uncertainty in driver behavior (11). Hall and Agyemang-Duah (4) showed that saturation rate approximately follows a normal distribution. They also discussed how the random nature of freeway capacity affects throughput and delay. Brilon et al. (8) examined freeway capacity from the perspective of breakdown (i.e. speed drop due to congestion) probability. They considered capacity as a random threshold for incoming flow to cause breakdown.

A major consequence of the fluctuation of saturation rate is build-up of traffic queues. For example, if the saturation rate is less than the incoming flow over a period of time, then a traffic queue will be produced. Another consequence is reduction in a freeway’s capability of serving incoming traffic. To estimate the induced delay and throughput loss, we need to know how the saturation rate evolves with respect to time. Specifically, we are interested in how often the actual saturation rate deviates from the nominal (maximum) value, and how long the periods of reduced saturation rate last. Several papers have studied the traffic delay due to capacity disruptions with specified, deterministic durations (12–14). In contrast, we model saturation rate as a stochastic process that captures the random fluctuation with time.

A straightforward way of constructing such a stochastic process is to assume that the saturation rate at various time steps are independently and identically distributed (IID) random variables (rv s). Based on this assumption, Sumalee et al. (10) devised a macroscopic traffic flow model that admits stochastic traffic dynamics and is capable of estimating the delay due to unreliable saturation rates. However, since the objectives of (10) are different from ours, the authors did not consider the spatial and temporal correlations of saturation rates. The IID assumption is not likely to be valid in the situation of disruptions due to capacity-reducing events. In addition, the IID assumption allows unrealistic temporal patterns of saturation rate, such as highly frequent oscillations. Thus, the IID assumption may lead to inaccurate conclusions.

In contrast to (10), we assume that saturation rate is governed by an underlying stochastic process rather than IID over time steps. In our previous work (15, 16), we developed the stochastic switching cell transmission model (SS-CTM), which is based on the classical cell transmission model (CTM, see (3, 17)). Daganzo (3) allowed the saturation rate of the CTM to be time-varying to capture capacity disruptions. The SS-CTM captures this feature by incorporating a Markov chain-based model for saturation rates. In the SS-CTM, the saturation rate of a freeway section

---

1 In this article, we use these two terms interchangeably.
(cell) stochastically switches between a finite set of values (or modes) according to a Markov chain. The stochastic switches capture capacity disruptions (due to arbitrary reasons) in reality. Since the SS-CTM focuses on saturation rate disruptions, only saturation rates are stochastic; the other traffic characteristics, including free flow speed, congestion wave speed, critical density, and jam density, are constant.

It is worth mentioning that the stochasticity of the SS-CTM does not result from heterogeneity of individual vehicles’ behavior, but stochastic switching of the traffic dynamics (due to e.g. incidents). The former aspect has been extensively discussed in the literature (see e.g. (10, 18–21)). Newell (22) argued that both the former and the latter contribute to traffic delay; however, the former is typically relevant for analysis on the scale of seconds, while the latter on the scale of minutes/hours. In this sense, the SS-CTM captures the aggregate behavior of traffic, which may also be subject to uncertainty.

By adopting the Markov chain model, the SS-CTM considers saturation rate as a finite-state stochastic process. As a comparison, Jia et al. (9) argued for a continuous, autoregressive model for saturation rates. However, although that model enables simulation-based analysis, it relies on real-time estimate of traffic conditions, and does not allow tractable analysis of the induced delay. In contrast, our Markov chain model is easy to implement, and facilitates theoretical analysis (16, 23). Another advantage of the discrete Markov chain model is its clear relation to actual disruptions: if a cell’s saturation rate switches from a large value to a small one, then an accident may have occurred in that cell. If a cell’s saturation rate switches from a small value to a large one, then an accident may have just got cleared.

The main focus of this article is a method for calibration of parameters of the SS-CTM, especially the Markov chain model for saturation rates. The SS-CTM calibrated with real traffic data allows analytical and/or simulation-based analysis of delay due to capacity disruptions and design of disruption-aware freeway traffic management (15, 16). It also allows analysis of interaction between capacity-reducing events at various locations. In addition, the SS-CTM is relevant for analysis of incident response or capacity recovery operations (24, 25), which are related to the frequency and duration of capacity disruptions.

The calibration method proposed in this article only requires traffic flow data, which is increasingly available because of the growing deployment of traffic sensors. Our approach focuses on capacity disruptions observed from traffic data. However, we can easily incorporate the information on cause of disruptions into our framework. In this sense, our saturation rate model is data-driven, which captures the inherent stochastic nature of freeway capacity. A future work would be to relate the observed temporal pattern to additional information, e.g. incident records by highway patrol or weather reports, which is beyond the scope of this article.

The calibration of the SS-CTM consists of two parts: (i) calibration of traffic characteristics, and (ii) estimation of the underlying Markov chain. The former has been systematically addressed by Dervisoglu et al. (26) and we will build on their approach. Calibration of traffic characteristics is based on the triangular/trapezoidal flow-density relation of the SS-CTM. Therefore, this part essentially determines the boundaries, slopes, and intercepts of a piecewise-affine (PWA) model. Ferrari et al. (27) proposed a more general method for calibration of PWA systems. The approach in (26) is largely consistent with that in (27), but the approach in (26) is simpler since it determines the boundaries of the PWA model based on insights from traffic flow theory.
To estimate the Markov chain, we first need to estimate the saturation rate at every discrete time. There is a consensus in the literature that saturation rate can be estimated only at bottlenecks (4–8). A bottleneck is essentially a cross-section of the freeway such that the section upstream thereof is congested (due to insufficient saturation rate at this cross-section) and the section downstream is not. Based on the above definition, several methods for saturation rate/capacity calibration have been proposed in the literature (6, 8, 28). However, those methods do not focus on the temporal pattern and thus do not lead to an underlying stochastic process governing the fluctuation of saturation rates. In contrast, we propose a method that is simple to implement and facilitates the construction of the stochastic process (Markov chain). Next, we identify the modes by approximating the empirical, continuous probability density function (pdf) of saturation rates with discrete probability mass functions (pmf). We conduct this discretization by clustering the observed saturation rates to the nearest peak (local maximum) of the empirical pdf. With the modes identified, the stochastic matrix of the Markov chain can be estimated using a standard maximum-likelihood method as described in (29).

This article is organized as follows. As a background, Section 2 introduces the SS-CTM, including the saturation rate model and the traffic flow model. Section 3 presents the main contribution of this paper, i.e. the details of the calibration of the SS-CTM. In Section 4, we present results from implementation of our calibration method on a freeway segment of the US Route 101. Section 5 gives further analysis and potential future directions.
2 TRAFFIC FLOW MODEL WITH STOCHASTIC SATURATION RATES

As a background, we introduce the SS-CTM in this section. This section is intended to familiarize the readers with the model and the notations. A detailed presentation is available in (16).

Consider a freeway consisting of \( n \) cells, as shown in Figure 1. Let \( \mathbb{Z}_+ = \{0, 1, 2, \ldots\} \) be the set of discrete times. The time step \( \delta \) should be no greater than the time that is needed for a vehicle to travel through a cell “in light traffic” (3). The saturation rate (or capacity, in vehicles per hour, \( vph \)) of the \( k \)-th cell at discrete time \( t \in \mathbb{Z}_+ \) is denoted by \( F_k[t] \), and let \( F[t] = [F_1[t], \ldots, F_n[t]]^T \).

One can interpret \( F_k[t] \) as the maximum rate at which cell \( k \) can discharge traffic to downstream at time \( t \). In the rest of this section, we will first introduce a Markov-chain model that specifies \( F[t] \) and then a traffic flow model that links \( F[t] \) to traffic evolution.

FIGURE 1 The SS-CTM with \( n \) cells. The warning signs indicate that the saturation rates at cell boundaries may stochastically vary with time.

2.1 Stochastic saturation rate model

The SS-CTM assumes that the vector of saturation rates \( F[t] \) is a quantity that switches between a finite set of values. Specifically, let \( H \) be a finite set of modes of the freeway and \( m = |H| \).

Each mode \( i \in H \) is associated with a vector of saturation rates \( F^i = [F^i_1, \ldots, F^i_n]^T \). We define \( F^\text{max}_k := \max_{i \in H} F^i_k \) as the nominal saturation rate of cell \( k \).

A mode as defined above corresponds to a particular “profile” of saturation rates at various locations (cells). Unless otherwise stated, we do not restrict the set of possible transitions in our analysis. Particularly, saturation rates at various locations may or may not be statistically independent. Typical causes of transitions include:

1. Occurrence of primary incidents: Previous literature suggests that the occurrence of incidents can be modeled as “Poisson arrivals” (30–34).

2. Occurrence of secondary incidents: Secondary incidents are traffic incidents induced by primary incidents, typically due to the congestion upstream of primary incidents. Secondary incidents also happen in a Poisson-like manner, with the arrival rate depending on the occurrence of primary incidents (12, 35–37).

3. Clearance of incidents: Empirical evidence suggests that incident duration is approximately exponentially (in continuous time, or geometrically in discrete time) distributed (15, 32, 38).

Some authors also report that incident rate may depend on additional information such as traffic density and traffic flow speed (33, 39). We do not consider this complication in this article.

However, our model can be extended to account for more sophisticated models of incident rates.

With a slight abuse of notation, we use \( i[t] \) to denote the mode of the freeway at time \( t \). We call \( i[t] \) the discrete state of the SS-CTM. We assume that the process \( \{i[t]; t \in \mathbb{Z}_+\} \) is specified by a
Markov chain defined over $H$ and with the *stochastic matrix*
\[
P = \begin{bmatrix}
p_{11} & p_{12} & \cdots & p_{1m} 
p_{21} & p_{22} & \cdots & p_{2m} 
\vdots & \vdots & \ddots & \vdots 
p_{m1} & p_{m2} & \cdots & p_{mm}
\end{bmatrix}
\]

such that
\[
\text{Pr}\left\{ i[t+1] = j \mid i[t] = i \right\} = p_{ij}, \quad i, j \in H, \ t \in \mathbb{Z}_+.
\]

The stochastic matrix satisfies
\[
0 \leq p_{ij} \leq 1, \ \forall i, j \in H, \ \sum_{j=1}^{m} p_{ij} = 1, \ \forall i \in H.
\]

Furthermore, the *holding time* $U_i$ in a mode $i$ is geometrically distributed with the probability mass function (pmf) (40)
\[
p_{U_i}(u) = p_{ii}^{u-1}(1 - p_{ii}), \quad u = 1, 2, \ldots
\]

A key assumption that we make for our model is that the saturation rate does not depend on density; i.e. we assume that the saturation rate $\{F[t] : t \in \mathbb{Z}_+\}$ is a Markov process that does not depend on traffic density $\rho[t]$. This is an assumption inherited from Daganzo’s CTM (3). Admittedly, this assumption has been challenged by some authors. For example, Hall and Agyemang-Duah (4) showed that the saturation rate may suddenly drop after traffic density enters the congestion regime. Furthermore, Brilon (8) reported that saturation rate $F[t]$ may depend on not only $\rho[t]$, but the time series $\{\rho[\tau] : t \leq \tau\}$, since $F[t]$ may evolve hysteretically as congestion begins and ends. Though, the triangular/trapezoidal fundamental diagram has been widely used for practical traffic analysis and provides adequate results for traffic assignment (41). In addition, models more sophisticated than the triangular/trapezoidal one requires a larger amount of data for calibration, which is not always available. Therefore, in this paper, we assume that $F[t]$ is independent of $\rho[t]$ and leave the above complications for future work.

Another point worth mentioning here is that we do not impose restrictions on the correlation of saturation rates at various locations. In our modeling setup, a cell’s saturation rate may be statistically independent of the adjacent cells’ saturation rates, or may be strongly correlated to them. Our goal is to capture the extent of correlation from the traffic data, since, according our previous study (16), the extent of correlation may significantly impact the induced traffic delay and throughput.

### 2.2 Stochastic switching cell transmission model

We now describe the SS-CTM. Let $\rho_k[t]$ denote the *traffic density* (in vehicles per mile, or vpm) in the $k$-th cell at time $t$. Traffic density $\rho_k[t]$ is non-negative and upper bounded by $\overline{\rho}_k$, the $k$-th cell’s *jam traffic density*. The SS-CTM’s *continuous state* is the $n$-dimensional vector $\rho[t] = [\rho_1[t], \rho_2[t], \ldots, \rho_n[t]]^T$. The *continuous state space* of the SS-CTM is the set $P = \prod_{k=1}^{n} [0, \overline{\rho}_k]$. The *hybrid state* of the SS-CTM is $(i[t], \rho[t])$ at time $t$, and the *hybrid state space* is $H \times P$. 
Assuming homogeneous cell lengths, the traffic density evolves as follows (17):

$$\rho_k[t+1] = \rho_k[t] + \frac{\delta}{l} \left( f_{k-1}[t] + r_k[t] - f_k[t] - s_k[t] \right), \quad k = 1, 2, \ldots, n,$$

(3)

where $\delta$ is the time step size, $l$ is the cell length, $f_k[t]$ is the flow from the $k$-th to the $(k+1)$-th cell, $r_k[t]$ is the on-ramp flow into the $k$-th cell, $s_k[t]$ is the off-ramp flow out of the $k$-th cell. Flows are in vph. For each cell $k$, we define the sending flow $S_k$ and the receiving flow $R_k$ as follows:

$$S_k[t] = \min\{v_k\rho_k[t], F_k[t]\}, \quad k = 1, \ldots, n$$

(4)

$$R_k[t] = w_k(\rho_k - \rho_k[t]), \quad k = 1, \ldots, n,$$

(5)

where $v_k$ denotes the free-flow speed and $w_k$ the congestion wave speed, both in miles per hour (mph). Following (23), we assume that $s_k[t]$ is proportional to $f_k[t]$ and

$$\frac{f_k[t]}{f_k[t] + s_k[t]} = \beta_k, \quad t \in \mathbb{Z}_+.$$

(6)

The flow $f_k[t]$ is given by the flow function (i.e., the so-called fundamental diagram):

$$f_k[t] = \begin{cases} \min\{r_0, R_1[t]\}, & k = 0. \\ \min\{\beta_k S_k[t], R_{k+1}[t]\}, & 1 \leq k \leq n - 1 \\ \beta_n S_n[t], & k = n. \end{cases}$$

(6)

The SS-CTM is specified by (2)–(6). Note that only the saturation rate $F_k[t]$ is mode-dependent and stochastic; the other parameters are constant and independent of the mode. For more information about the model and its properties, see (15, 16).
3 CALIBRATION OF SS-CTM

In this section, we introduce a method of calibrating the parameters of the SS-CTM defined in Section 2. We assume availability of the following data:

1. Flow rate $q_k[t]$ over the $k$-th cell for all times $t$ in a set $T \subset \mathbb{Z}_+$, for $k = 1, \ldots, n$, in vehicle per hour (vph);
2. Traffic density $\rho_k[t]$ of the $k$-th cell for all times $t \in T$, for $k = 1, \ldots, n$, in vehicles per mile (vpm).

Typically these data can be retrieved or derived from field sensors’ measurement when the freeway is appropriately partitioned; the cell boundaries are selected such that every cell contains at least one sensor. Alternatively, aggregate flow speed $u_k$ can replace either $q_k$ or $\rho_k$, since these three quantities are related by

$$q_k = u_k \rho_k.$$  \hspace{1cm} (7)

Note that $q_k$ here is not exactly the same as $f_k$ in (6). One can view $q_k$ as the space-average flow within the $k$-th cell, while $f_k$ is defined as the flow crossing the boundary between the $k$-th and the $(k+1)$-th cells. As a Godunov discretization of partial differential equation-based traffic flow models (3, 42, 43), the CTM (and thus the SS-CTM) to a large extent approximates $q_k$ with $f_k$. In addition, the aggregate flow speed $u_k[t]$ is different from the free flow speed $v_k[t]$.

As Figure 2 shows, calibration of the SS-CTM consists of two parts, viz. calibration of the traffic characteristics (nominal fundamental diagram) and estimation of the Markov chain. The former largely follows from (26). However, (26) only considers nominal capacity and does not involve saturation rates at every time step. We will develop a different method for estimation of saturation rates based on insights from previous work on freeway capacity (4, 8) and the CTM (3, 17). To construct the Markov chain model, we first identify the modes with a simple clustering applied to the saturation rates. Then, the stochastic matrix is estimated based on the method discussed in (29).

**FIGURE 2 Calibration procedure.**
3.1 Calibration of traffic characteristics

We follow the method proposed by Dervisoglu et al. (26) to calibrate the traffic flow model. Each cell is calibrated independently. The parameters obtained in this subsection, including free flow speed \( v_k \), congestion wave speed \( w_k \), critical density \( \rho_k^c \), and jam density \( \overline{\rho}_k \), are independent of mode.

Consider the \( k \)-th cell. Note that the SS-CTM is based on the triangular flow-density relation

\[
q_k = \begin{cases} 
vp_k, & 0 \leq \rho_k \leq \rho_k^c, \\
wp_k(\overline{\rho}_k - \rho_k), & \rho_k^c < \rho_k \leq \overline{\rho}_k,
\end{cases}
\]

where

\[
\rho_k^c = \frac{w_k}{v_k + w_k} \overline{\rho}_k
\]

is the critical density of cell \( k \). By (7), we see that

\[
u_k = \begin{cases} 
v, & 0 \leq \rho_k \leq \rho_k^c, \\
w(\overline{\rho}_k / \rho_k - 1), & \rho_k^c < \rho_k \leq \overline{\rho}_k.
\end{cases}
\]

In words, when \( \rho_k \leq \rho_k^c \), the aggregate flow speed \( u_k \) is equal to the free flow speed \( v_k \); when \( \rho_k > \rho_k^c \), \( u_k < v_k \).

As illustrated in Figure 3, we partition the data points \( \{(\rho_k[t], q_k[t]); t = 1, \ldots, T\} \) into three categories. The first category (“Free flow” in Fig 3) includes those points such that \( u_k[t] \geq \tilde{v} \), where \( \tilde{v} \) is the minimum speed for free flow. We choose \( \tilde{v} \) to be 55 mph for all cells, which is similar to the thresholds used in (6, 26). We denote the set of indices of this category of data points as \( T_1 \). The second category (“Congested flow” in Fig 3) includes points such that \( \rho_k[t] > \rho_k^c \), which means congestion. We denote the set of indices of this category of data points as \( T_2 \). The remaining data points are placed in the third category (“Other” in Fig 3), denoted by \( T_3 \).
The set $T_1$ is given by

$$T_1 = \{t \in T : q_k[t]/\rho_k[t] \geq \hat{v}\}.$$

The free flow speed $\hat{v}_k$ is estimated by linear regression over the data set \{(\rho_k[t], q_k[t]) : t \in T_1\}.\footnote{In this article, we use \(\hat{x}\) to denote the estimate of \(x\).}

Throughout this article, we use the letter with hat to denote the estimated value of a variable or parameter. The nominal capacity $\hat{Q}_k$ is estimated by

$$\hat{Q}_k = \max_{t \in T_1} q_k[t].$$

Thus, the estimated critical density is

$$\hat{\rho}^c_k = \frac{\hat{Q}_k}{\hat{v}_k}.$$ 

The set $T_2$ is given by

$$T_2 = \{t \in T \setminus T_1 : \rho_k[t] \geq \hat{\rho}^c_k\}.$$

Note that the congestion wave speed $w$ satisfies

$$q_k - Q_k = -w(\rho_k - \hat{\rho}^c_k), \quad \rho^c_k \leq \rho_k \leq \rho_k.$$ 

Therefore, $\hat{w}$ can be estimated via linear regression over the data set \{(\rho_k[t] - \hat{\rho}^c_k, q_k[t] - \hat{Q}_k) : t \in T_2\}. The jam density is estimated by

$$\hat{\rho}^j_k = \frac{\hat{Q}_k}{\hat{w}_k} + \hat{\rho}^c_k.$$ 

This completes the calibration of the nominal fundamental diagram.

Note that $\hat{Q}_k$ is the “nominal” capacity in the sense of (26), but is not the saturation rate $F_k$ in the SS-CTM. $\hat{Q}_k$ is essentially the maximum flow rate observed in the time interval $T$, but is not the saturation rate that is available at every time $t \in T$.

### 3.2 Estimation of Markov chain model

To estimate the Markov chain, we first estimate the saturation rate at every time step, then cluster the saturation rates at various times to a finite set of discretized values, which gives the set of modes and the time series of realized modes, and finally estimate the stochastic matrix.

**Saturation rate**

Estimation of saturation rate is not straightforward, since it can only be done at bottlenecks (4). Specifically, the flow is not saturated in either of the following cases: (i) The demand is low and the traffic stays in free flow; (ii) The downstream section is congested and prevents the section in question to discharge traffic.
Our method is intuitive and easy to implement. It is based on the flow-density relation of the CTM and practical interpretation of saturation rate. In addition, it naturally results in an underlying stochastic process, i.e. the Markov chain model of the SS-CTM (see the next subsection for details). As a benchmark, Polus and Pollatschek (6) proposed a more sophisticated method to estimate saturation rate. Their method builds on linear regression and random sampling of the data. However, because of focuses different relative to ours, their method does not directly apply to the SS-CTM due to two reasons. First, they focus on those data points close to the nominal fundamental diagram, while in our setting those points far from the nominal fundamental diagram are of significant interest. Second, their method does not consider the saturation rate as a signal (i.e. a time series) and thus does not enable the construction of a stochastic process. Several other methods for saturation rate estimation have similar drawbacks (8, 9, 28), since they were developed for objectives different from that in this paper.

Recall that the flow $f_k[t]$ is given by (6), which states that the saturation rate $F_k[t]$ cannot be explicitly observed unless (i) there is enough upstream incoming traffic, i.e. $\rho_k[t] \geq F_k[t]/v_k$, and (ii) there is no downstream congestion, i.e. $\rho_{k+1}[t] \leq \rho_{k+1} - F_k[t]/w_{k+1}$. In other words, one can only estimate saturation rates at bottlenecks.

Based on the above arguments, we assume the following:

1. The saturation rate $F_k[t]$ is lower-bounded by the realized flow $f_k[t]$.
2. If this cell is a bottleneck, then $F_k[t] = f_k[t]$.
3. If this cell is not a bottleneck, then $F_k[t] = F_k[t-1]$, unless the $F_k[t]$ estimated in this way is less than $f_k[t]$.
4. The average flow $q_k[t]$ approximates $f_k[t]$.

Algorithm 1 summarizes the specific procedure derived from the above assumptions.

**Algorithm 1** Estimation of saturation rate

$$\hat{F}_k[1] = q_k[1]$$

for $t = 2 : T$

if $\rho_k[t] \geq \rho^c_k$ and $\hat{w}_{i+1}(\hat{\rho}_{i+1}[t]) > q_k[t]$ then

$$\hat{F}_k[t] = q_k[t]$$

else

$$\hat{F}_k[t] = \max\{q_k[t], \hat{F}_k[t-1]\}$$

end if

end for

Note that the above method does not apply to the case where a cell is always in free-flow. The reason is that, in this case, one cannot determine whether a cell is saturated or not. Nevertheless, the CTM and the SS-CTM are largely intended for analysis of traffic congestion, which is of limited relevance to a roadway always staying in the free-flow regime.

Mode identification

Recall that the modes of the SS-CTM are associated with vectors of saturation rates rather than saturation rates of individual cells. Therefore, an ideal method for mode identification would be based on the joint empirical distribution of $F_1, F_2, \ldots, F_n$. However, estimation of the joint
distribution may be challenging, especially when the number of cells is large. Estimation of such a complex distribution requires a very large data set, which is not always available. In addition, an excessively complex model is not always practically useful. Therefore, we develop a two-stage method for mode identification. In the first stage, we identify the “cell modes”, i.e. the saturation rate values that each cell can take. In the second stage, we combine the “cell modes” of individual cells to obtain the set of modes of the freeway segment.

With the estimated saturation rates \( \{ \hat{F}_k[t]; t \in T \} \), we can compute the empirical (marginal) pdf \( f_{\hat{F}_k} \) of \( F_k \). Identification of modes is essentially approximation of \( f_{\hat{F}_k} \) with a discrete pmf \( p_{\hat{F}_k} \). To determine the number of modes, we assume that every peak (local maximum) of \( f_{\hat{F}_k} \) will correspond to a mode (more precisely, a value that \( F_k \) can take). This assumption is reasonable in that, if the empirical pdf turns out to have for example two peaks, then it is quite natural to consider it as the superposition of two normal-like distributions, which center at the two peaks, respectively.

One way to avoid excessive peaks is to use the histogram of \( F_k[t] \) with appropriately selected bin size instead of the continuous pdf \( f_{\hat{F}_k} \) for mode identification (see Figure 5). Generally speaking, the larger the bin size \( \Delta F \) is, the fewer the peaks. Empirically, the bin size \( \Delta F \) should be selected such that the histogram has no more than four or five peaks; otherwise the estimation of the stochastic matrix may be challenging, especially when the number of cells is large. Estimation of such a complex distribution requires a very large data set, which is not always available. In addition, an excessively complex model is not always practically useful. Therefore, we develop a two-stage method for mode identification. In the first stage, we identify the “cell modes”, i.e. the saturation rate values that each cell can take. In the second stage, we combine the “cell modes” of individual cells to obtain the set of modes of the freeway segment.

To identify the cell modes, we use a method similar to the one described in Section 3. Let \( F_k[t] \) be the discretized saturation rate at time \( t \), where \( F_k[t] \) is given by

\[
\hat{F}_k[t] = \min_{\hat{F}_k} \left| F_k[t] - \hat{F}_k \right|
\]

Suppose that the discretized saturation rate \( \hat{F}_k \) of the \( k \)-th cell can take values in the finite set \( \mathcal{F}_k \). Then the possible set of vectors of saturation rates \( \hat{F} = [\hat{F}_1, \ldots, \hat{F}_n]^{\top} \) of the freeway segment is

\[
\mathcal{F}_k = \prod_{k=1}^{n} \mathcal{F}_k.
\]

However, due to the correlation between the cells, the actual set \( \mathcal{F} \) of \( F \) may be a proper (i.e. strict) subset of \( \mathcal{F} \). Specifically, \( \mathcal{F} \) is given by

\[
\mathcal{F} = \left\{ \hat{F} \in \mathcal{F} : (\exists t \in T) \hat{F}[t] = F \right\}.
\]

Then \( \mathcal{F} \) immediately leads to the set \( H \) of modes of the freeway segment.

Recall that we use \( F^i \) to refer to the vector of saturation rates associated with \( i \in \mathcal{H} \). The mode \( i[t] \) at time \( t \) is thus given by \( i \in H \) such that \( \hat{F}[t] = F^i \).

**Stochastic matrix**

Let \( m = |H| \) and \( P \in \mathbb{R}^{m \times m} \) be the transition matrix of the discrete-time Markov chain. According to (29), the transition probabilities are estimated using the maximum likelihood method as follows.
Let $N_{ij}$ be the observed number of transitions from mode $i$ to mode $j$. Let $N_i = \sum_{j \in H} N_{ij}$, i.e. the number of transitions from $i$. Given a transition matrix $\hat{P}$, the likelihood of the observation is

$$L = \log \text{Pr}\{i[0]\} + \sum_{i \in H} \sum_{j \in H} N_{ij} \log p_{ij},$$

subject to $\sum_{j \in H} p_{ij} = 1$ for any $i \in H$. The transition probabilities that maximize the likelihood are given by

$$\hat{p}_{ij} = \frac{N_{ij}}{N_i}, \quad i, j \in H.$$ 

Note that the denominator includes self-transitions, i.e. from $i$ to $i$ itself. This completes the estimation of the parameters of the SS-CTM.

To sum up, the calibration of the SS-CTM consists of two parts. The first is the calibration of traffic characteristics using a linear-regression-based method. The second is the construction of the underlying Markov chain, where the saturation rates are estimated, the modes are identified, and the transition rates are estimated.
4 CASE STUDY: A SEGMENT OF U.S. ROUTE 101

In this section, we apply the estimation method described in Section 3 to a segment of the U.S. Route 101. We will first introduce the data set to use, then present the results, and finally discuss about the results.

4.1 Setting and data

The particular freeway segment to be studied is the southbound portion of US101 from post-miles 433 to 420, i.e. from downtown San Francisco to the San Francisco International Airport. This is a busy freeway segment in the U.S. 35 sensors are installed along this stretch, including inductive loops, radar, and magnetometers (45).

The CTM for this stretch of US101-S is illustrated in Figure 4. The cell boundaries are selected such that on-ramps are located at the upstream end and off-ramps at the downstream end of the cells. In addition, every cell includes at least one mainline sensor.

FIGURE 4 The US 101 section from downtown San Francisco (A) to the San Francisco International Airport (B).

PeMS provides measurements of traffic flow $q_k$ and aggregate flow speed $u_k$ every five minutes; each sensor reports 288 records per day. This article uses a data set spanning the full year of 2012. PeMS does not explicitly report density $\rho_k$, which can be computed by (7).

4.2 Results

First, Table 1 lists the traffic flow parameters estimated using the method in Section 3.1. Note that the jam densities and maximum flows are average over all lanes. Figure 3 shows the calibration of an example cell. We refer to (26) for a benchmarking of the estimated values.

Second, the modes of individual cells are identified, following Section 3.2.2. Figure 5 shows three examples. Figure 5(a) shows an empirical distribution of saturation rate with a single peak, which implies a single mode. Figure 5(b) shows a two-peak distribution and implies two modes; one mode may correspond to the nominal condition and the other may correspond to disruptions. Figure 5(c) shows a three-peak distribution and implies three modes, which may correspond to
TABLE 1 Estimated traffic characteristics.

<table>
<thead>
<tr>
<th>Cell</th>
<th>No. of Lanes</th>
<th>Free-flow speed $v_k$ [mph]</th>
<th>Congestion wave speed $w_k$ [mph]</th>
<th>Jam density $\rho_k$ [vpm per lane]</th>
<th>Maximum flow $Q_{\text{max}}^k$ [vph per lane]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>62.93</td>
<td>12.58</td>
<td>205.2</td>
<td>2151</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>66.10</td>
<td>9.76</td>
<td>225.7</td>
<td>1920</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>61.80</td>
<td>16.63</td>
<td>149.7</td>
<td>1961</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>63.83</td>
<td>9.11</td>
<td>241.8</td>
<td>1929</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>66.03</td>
<td>9.14</td>
<td>225.1</td>
<td>1807</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>65.17</td>
<td>8.55</td>
<td>297.6</td>
<td>2250</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>63.83</td>
<td>10.19</td>
<td>252.8</td>
<td>2220</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>63.69</td>
<td>15.61</td>
<td>175.1</td>
<td>2196</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>65.64</td>
<td>8.33</td>
<td>258.2</td>
<td>1908</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>64.30</td>
<td>9.89</td>
<td>216.3</td>
<td>1854</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>64.56</td>
<td>9.09</td>
<td>265.5</td>
<td>2115</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>66.30</td>
<td>14.14</td>
<td>100.6</td>
<td>1173</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>63.72</td>
<td>9.87</td>
<td>208.9</td>
<td>1785</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>61.13</td>
<td>6.79</td>
<td>267.5</td>
<td>1635</td>
</tr>
<tr>
<td>15</td>
<td>6</td>
<td>62.87</td>
<td>15.25</td>
<td>120.9</td>
<td>1484</td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>62.18</td>
<td>11.93</td>
<td>183.6</td>
<td>1838</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>63.70</td>
<td>10.75</td>
<td>275.4</td>
<td>2532</td>
</tr>
</tbody>
</table>

nominal condition, minor disruptions, and major disruptions.

FIGURE 5 Cells with various numbers of peaks, which implies various numbers of modes.

Third, the time series of modes is generated; 1211 distinct modes are observed in the time series. Based on the set of “modes” of individual cells, the number of modes could be as large as 13824; however, only $1211/13824 = 8.8\%$ thereof are observed in the data. Following the procedure in Section 3.2.2, we estimated the transition matrix $P$ of the freeway segment. The $1211 \times 1211$ matrix has 5289 non-zero elements, which means 4078 non-self transitions; only $0.36\%$ of the elements in $P$ are non-zero. The small number of possible modes and the sparsity of the transition matrix imply relatively strong correlation between the modes or saturation rates of various cells.

Further scrutiny of the time series implies further reduction of the set of modes. In fact, the
majority of the 1211 observed modes happen rarely. Figure 6 shows the proportion of time that
the freeway segment spends in an increasing set of modes; the mode associated with the maximum
proportion of time is first included, and then the mode with the second, third, and so forth maximum
proportion of time. It is observed that the 50 most frequent modes account for approximately 80%
of the time, and only the top 78 modes happen with an average frequency higher than 2 minutes
per day. If we only consider these 78 modes and estimate again the stochastic matrix, then 2479
out of 78² elements in the matrix are non-zero.

4.3 Discussions
Several findings from the results worth further elaboration.

First, the calibration of the fundamental diagram (see Figure 3) strongly suggests that nominal
values for traffic flow parameters, especially nominal capacity, is not sufficient for effective traffic
management. Although the nominal flow-density curve provides a good envelop (upper bound) of
the data, only part of the data points lie near this curve. For a large fraction of time (see Table 2,
column “Fraction of time with maximum saturation rate”), the nominal saturation rate is available
for approximately 80% of the time on average, and can be as low as 38% for a particular cell.
Consequently, if a traffic manager solely considers the nominal model, he/she will not be aware of
the potentially significant congestion due to fluctuation of saturation rate.

Second, the Markov chain model is able to adequately capture the fluctuation in freeway satu-
ration rates. Figure 7(a) illustrates how the discrete model approximates the actual saturation rate.
To quantitatively evaluate the discretization, we compare the mean and the standard deviation of
the empirical distribution and those of the discretized distribution. Table 2 lists the results. Gener-
ally speaking, the model’s statistics is fairly consistent with those of the data. Some deviations are
observed for the cells with only one mode. For such cells, more sophisticated method for mode
identification is needed, which is an interesting topic for future work.

Figure 7(b) shows the fitted and empirical CDFs of the holding times of the three most fre-
quently observed modes. According to the Markov chain model, the holding times should be ge-
TABLE 2 Statistics of saturation rate model

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>5740</td>
<td>6039</td>
<td>1792</td>
<td>1970</td>
<td>0.81</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4871</td>
<td>4962</td>
<td>1926</td>
<td>2028</td>
<td>0.79</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>5168</td>
<td>5336</td>
<td>2434</td>
<td>2686</td>
<td>0.72</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>4853</td>
<td>4943</td>
<td>1578</td>
<td>1530</td>
<td>0.61</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>4975</td>
<td>5335</td>
<td>1473</td>
<td>1698</td>
<td>0.87</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>4629</td>
<td>4704</td>
<td>1677</td>
<td>1794</td>
<td>0.64</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>4682</td>
<td>6000</td>
<td>1920</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>5026</td>
<td>4649</td>
<td>1350</td>
<td>1132</td>
<td>0.91</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>5735</td>
<td>5471</td>
<td>1709</td>
<td>1538</td>
<td>0.89</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>4002</td>
<td>4042</td>
<td>1562</td>
<td>1707</td>
<td>0.76</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>5323</td>
<td>5462</td>
<td>2296</td>
<td>2436</td>
<td>0.38</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>3559</td>
<td>3594</td>
<td>1638</td>
<td>1910</td>
<td>0.65</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>3934</td>
<td>4000</td>
<td>1464</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>3328</td>
<td>4000</td>
<td>1555</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>6218</td>
<td>7000</td>
<td>1746</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>5758</td>
<td>5073</td>
<td>2227</td>
<td>1943</td>
<td>0.81</td>
</tr>
</tbody>
</table>

1 ometerically/exponentially distributed, which corresponds to the solid curves. The empirical CDFs roughly stays near the fitted CDFs.

FIGURE 7 Comparison between the Markov chain model and the estimated saturation rates.

Third, saturation rates of successive cells can be strongly correlated. To see this, note that the structure of H and P depends not only on the behavior of individual cells, but also on the
correlation between the cells. If the saturation rates of various cells are statistically independent,
then \( \mathcal{F} = \hat{\mathcal{F}} \), and every element in \( P \) is strictly positive. However, the independence assumption
is usually questionable. From an intuitive perspective, if a cell is experiencing reduced saturation
rate (due to e.g. bad weather), the adjacent cells is also likely to have reduced saturation rate. Our
data analysis suggests that \( |\mathcal{F}| \) is usually significantly less than \( |\hat{\mathcal{F}}| \), and that \( P \) is usually sparse.

\[ r_i = \frac{\text{Cov}(F_i, F_{i+1})}{\text{Var}(F_i)\text{Var}(F_{i+1})}. \]

The results are illustrated in Figure 8. Intuitively, all correlations are positive. The strongest
correlation (0.85) happens between cells 5 and 6, and the lowest correlation (0.28) between cells 3
and 4. 11 out of 15 correlation coefficients are above 0.5. This result is consistent with the intuition
that saturation rates of consecutive sections of freeway can be strongly correlated, the implication
being that the assumption of independence (such as that in (10)) has to be carefully justified.

Fourth, the results further justify the several assumptions adopted by the SS-CTM. The traffic
data strongly suggests that saturation rate is a rv spanning a fairly large range rather than a constant.
A finite set of saturation rate values is able to adequately approximate the fluctuation. The Markov
chain model gives reasonable approximation in terms of mean value and standard deviation.

Finally, as a benchmark, we plot the spatial distribution of incidents reported by the California
Highway Patrol (CHP) in Figure 9. The report was retrieved from PeMS (45). The incidents
include collisions, hit-and-runs, traffic hazards, floods, etc. As the figure indicates, the average
number of incidents that a cell experiences per day is less than three. Recall that the nominal
saturation rate is not available for 20%–60% of the time. Therefore, incidents seem to account
for only a subset of the saturation rate disruptions. Due to lack of information, it is challenging
to identify the cause of every disruption. Though, the calibrated stochastic switching model itself
sometimes suffices for planning and proactive operations purposes, and a traffic manager does not
necessarily need to figure out the cause of every disruption. However, the causes of disruptions
FIGURE 9  Spatial distribution of traffic incidents reported by highway patrol.

are relevant for reactive operations against particular types of disruptions, such as diversion during incidents and response team dispatch (24).
5 CONCLUSIONS

This article proposes a method for estimation of the SS-CTM, a traffic flow model with stochastic saturation rates developed in our previous work. The estimation method consists of two steps. In the first step, traffic characteristics are calibrated using an automatic algorithm. In the second step, the saturation rate at every time step is estimated based on the notion of bottlenecks, and the Markov chain model for the saturation rates is then estimated, which involves identification of modes and estimation of transition probabilities.

This paper can be extended in several directions. First, if data associated with the on-ramp and off-ramp flows are available, their impact can be incorporated in the calibration of the SS-CTM. Second, more sophisticated flow-density relation can be considered in the presence of sufficient data. Third, the calibrated SS-CTM can be used to evaluate and design traffic management strategies in the wake of fluctuating saturation rate.
ACKNOWLEDGMENT

This work was supported by NSF Grant No. 1239054 CPS Frontiers: Foundations Of Resilient Cyber-Physical Systems (FORCES) and MIT CEE Graduate Fellowship. We sincerely appreciate the data provided by the California Department of Transportation (Caltrans) and Mr. J. Liu’s help with data collection. We also thank the anonymous reviewers for their valuable suggestions.
REFERENCES


