Mixture Kalman Filter Based Highway Congestion Mode and Vehicle Density Estimator and its Application

Xiaotian Sun, Laura Muñoz, and Roberto Horowitz

Abstract— In this paper, we present our latest results on developing and implementing a traffic congestion mode and vehicle density estimator for a segment of Interstate 210 in Southern California. Using a mixture Kalman filtering (MKF) algorithm on the switching-mode traffic model, the estimator is able to provide estimated vehicle densities at unmeasured locations, as well as the congestion statuses (free-flow or congested), which are not directly observed. The program runs efficiently, thus making it possible to carry out estimation in real time.

I. INTRODUCTION

In today's metropolitan areas, highway traffic congestion occurs regularly during rush hours. In addition, nonrecurrent congestion often takes place as a result of incidents or road work. The congestion causes inefficient operation of highways, waste of resources, increased air pollution, and intensified driver fatigue.

One of the methods used to prevent and/or relieve highway traffic congestion is on-ramp metering. Many strategies have been proposed and deployed to regulate the demands on the highway by limiting the rate at which vehicles enter the highway. These ramp-metering strategies can be classified into two categories: traffic responsive and nontraffic responsive. Traffic responsive ramp metering strategies utilize real-time measurements of the traffic state, such as vehicle densities and traffic congestion status (freeflow or congested), while non-traffic responsive ones often operate under a fixed time-of-the-day table, and ignore the current traffic conditions completely. It is believed that traffic responsive metering strategies are more effective and more robust than non-traffic responsive ones.

Currently, traffic data are usually obtained from inductive loop detectors embedded in the pavement. These loop detectors provide percent occupancy (percentage of the time that the detector is occupied by vehicles) and volume (number of vehicles that have passed over the detector), from which other macroscopic quantities can be derived when accurate vehicle lengths are available. These macroscopic quantities include vehicle density (number of vehicles per unit length of highway), flow (number of vehicles per unit time), and speed. One effort to provide such macroscopic traffic data through a centralized database is the Freeway Performance Measurement System (PeMS) [1].

Ideally, one would like to have access to these traffic data continuously across time and space. However, because of the high cost and difficulty of installing and maintaining loop detectors, oftentimes traffic data are not available at all desired locations at all times. These missing data thus have to be estimated using the available data. On the other hand, the traffic flow mode, *i.e.*, whether the traffic is in free-flow (where traffic moves freely at high speeds) or congestion (where traffic moves slowly and is restricted by downstream conditions), cannot be measured directly and has to be inferred from other quantities. In our previous work [2], we developed a traffic state estimator based on the switching mode traffic model [3] and the mixture Kalman filter (MKF) [4]. This estimator is able to estimate the vehicle densities at unmeasured locations, as well as to determine the traffic congestion mode, in a highway section. It was tested on a short section of highway and its performance was evaluated using the measured data. It was shown that on average, a mean percentage error of $\sim 10\%$ was achieved for vehicle density estimation at unmeasured locations.

In this paper, we describe our recent work, which involves implementing this congestion mode and vehicle density estimator on our entire selected test site, a 14-mile long segment of Interstate 210 Westbound in Pasadena, California, and interfacing the estimator with traffic simulation programs that have been calibrated to the test segment, including a macroscopic simulator—the modified cell transmission model [5], and a microscopic simulator—VISSIM [6]. We will first briefly review the switching-mode model and the estimation method in Section II. Then in Section III, we will describe the test site and discuss the constraints and other considerations for the cell configuration. The main results will be presented and discussed in Section IV. In the last section, we will give some concluding remarks.

II. METHOD

A. Cell Transmission-Based Switching Mode Model

Highway traffic models can be categorized into two groups: microscopic models and macroscopic models. Microscopic models describe the behavior of each individual vehicle on the highway, while macroscopic models describe the evolution of aggregated traffic quantities, such as vehicle densities and flows, based on the conservation principle. Due to their complexity, the microscopic models, which require an enormous number of equations to model each individual vehicle's behavior and the interaction between

This work has been supported by California Partners for Advanced Transit and Highways (PATH) under Task Order 4136.

The authors are with the Department of Mechanical Engineering, University of California at Berkeley, Berkeley, CA 94720-1740. E-mail addresses: {sunx,lmunoz,horowitz}@me.berkeley.edu.

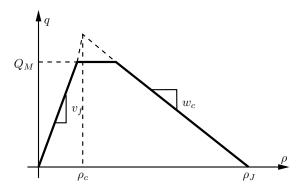


Fig. 1. A trapezoidal fundamental diagram (flow-density relation) used by the cell transmission model.



Fig. 2. A highway section divided into cells in the cell transmission model.

these vehicles, often can only be solved numerically to simulate the traffic evolution. On the other hand, macroscopic models often consist of a few dynamic equations, along with some constitutional relation between certain traffic quantities, such as the so-called fundamental diagram between flow and density. As an example, a widely used macroscopic model is the Lighthill-Whitham-Richards (LWR) model, which consists of a first-order partial differential equation based on conservation of vehicles, and a flowdensity relationship. Because of their simplicity compared to microscopic models, macroscopic models are often used to quantitatively analyze the effects of changing conditions on traffic behavior.

The cell transmission model (CTM) [7], [8] is a macroscopic model. It can be viewed as an approximation and simplification of the LWR model obtained by discretizing the PDE of the LWR model in both space and time, and using a simple trapezoidal fundamental diagram as shown in Fig. 1. The CTM divides a highway into small segments, which are called *cells*, as shown in Fig. 2. The traffic flow q_i going into a cell *i* is considered constant between two consecutive time points *t* and *t* + 1, and is determined by

$$q_i(t) = \min\{v_{f,i-1}\rho_{i-1}(t), w_{c,i}(\rho_{J,i} - \rho_i(t)), Q_{M,i-1}\}, \quad (1)$$

where for a cell *i*, $\rho_i(t)$ is the average vehicle density between times *t* and *t* + 1, $\rho_{J,i}$ is the jam density, *i.e.*, the maximum vehicle density allowed in cell *i*, $v_{f,i}$ is the free-flow speed, $w_{c,i}$ is the backward congestion wave propagation speed, and $Q_{M,i-1}$ is the flow capacity, *i.e.*, the maximum possible flow. These quantities are also illustrated in Fig. 1.

The three terms involved in the minimum in (1) can be interpreted as follows. The first term $v_{f,i-1}\rho_{i-1}(t)$ is the flow that can be supplied by cell *i*-1. The second term is the flow that can be absorbed by cell *i*. And the third term $Q_{M,i-1}$

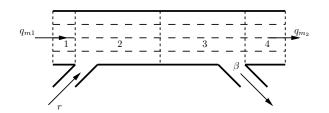


Fig. 3. A schematic plot of a 4-cell highway section with one on-ramp and one off-ramp.

is the maximum possible flow from cell i - 1 to i. Each of these three terms is shown as a thick line segment in Fig. 1 and the minimum is the outline of the trapezoid.

Despite the simplifications that are made in the cell transmission model, it still captures many important traffic phenomena, such as queue build-up and dissipation, backward propagation of congestion waves, etc. The simplicity and accuracy of the CTM makes it very desirable for traffic engineering. However, due to the nonlinear nature of the flow-density relationship in the CTM, it can be quite difficult to analyze the model and use it as a basis for designing traffic controllers. Therefore, we have piecewise linearized the CTM and derived a switching-mode model for each section of the highway, based on the traffic congestion status (congested or free-flow) in each section [3].

This switching-mode model includes several modes. In each mode, the vehicle densities in the cells evolve according to a different set of linear difference equations. Among these modes, two are of greatest importance: pure free-flow and full congestion. For the purpose of designing the traffic estimator, we further simplify this switching-mode model by considering only these two modes. For a typical section of highway that consists of 4 cells and contains an on-ramp and an off-ramp, as shown in Fig. 3, we write down this simplified switching-mode model as follows.

When the highway section is in free-flow mode, the first term in (1) dominates, and the difference equations are

$$\begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \rho_4 \end{bmatrix} (t+1) = \begin{bmatrix} 1 - \frac{v_{f1}T_s}{l_1} & 0 & 0 & 0 \\ \frac{v_{f1}T_s}{l_2} & 1 - \frac{v_{f2}T_s}{l_2} & 0 & 0 \\ 0 & \frac{v_{f2}T_s}{l_3} & 1 - \frac{v_{f3}T_s}{l_3} & 0 \\ 0 & 0 & (1-\beta)\frac{v_{f3}T_s}{l_4} & 1 - \frac{v_{f4}T_s}{l_4} \end{bmatrix} \begin{bmatrix} \rho_1 \\ \rho_2 \\ \rho_3 \\ \rho_4 \end{bmatrix} (t)$$

$$+ \begin{bmatrix} \frac{T_s}{l_1} & 0 & 0 \\ 0 & 0 & \frac{T_s}{l_2} \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} q_{m1} \\ q_{m2} \\ r \end{bmatrix} (t)$$

$$= A(1)o(t) + B(1)o(t)$$

$$(2)$$

$$= A(1)\rho(t) + B_q(1)q(t),$$
(3)

where q_{m1} and q_{m2} are the mainline entering and exiting flows, respectively, *r* is the on-ramp flow, β is the split ratio of the off-ramp flow, l_i is the length of cell *i*, and T_s is the sampling time.

When the section is in congested mode, the second term

in (1) dominates, and the difference equations are

where ρ_{Ji} is the jam density (maximum allowable density) in cell *i*.

It is important to point out that the dynamics of the cell vehicle densities in different modes are dramatically different, as evidenced by the observability and controllability in each mode. In free-flow mode, the densities in the section are observable through a measurement that is downstream of the section and controllable by an on-ramp upstream of the section; while in congestion mode, the densities are observable through a measurement upstream of the section and controllable by an on-ramp downstream of the section. These observations, especially those for congestion mode, are counter-intuitive at a first glance. However they agree with the understanding in highway traffic theory that when during congestion, a high density wave, or in other words, information, propagates backwards along the highway.

B. Improved Mixture Kalman Filter

In the switching-mode model, the mode (free-flow or congested) is determined by the flow condition in the section. However, there is no direct measurement or observation of the current traffic congestion mode in a highway section. The congestion mode can only be inferred from measured quantities, for example, the speed. The general practice in traffic engineering is to set an upper threshold and a lower threshold for the speed. When the speed in a section is above the upper threshold, the section is considered to be in free flow; when the speed is below the lower threshold, the section is in congestion; when the speed is between the two thresholds, the section is considered somewhat likely to be in congestion. The problem with this kind of method is two-fold: 1) The selection of the thresholds is based on experience and, to a certain degree, is arbitrary, and 2) When the speed falls between the two thresholds, the mode of the section cannot be determined.

Therefore, we assume that we do not have direct observation of the mode and that the mode jumps between possible values following a discrete-time Markov chain with a certain transition probability. Under these assumptions, the switching-mode model falls into a special class, called *Markov jump linear systems (MJLS)*. The previous four-cell example has a continuous state $x = [\rho_1 \quad \rho_2 \quad \rho_3 \quad \rho_4]^T$ and possible discrete modes 1 (free-flow) and 2 (congested).

It is known that it is difficult to estimate the states and the mode when the mode itself is not observed. The difficulty lies in the fact that the sample space S^t of the mode sequence grows exponentially as time *t* increases, where *S* is the set of possible discrete modes.

The *mixture Kalman filter (MKF)* [4] approximately solves this difficult probability inference problem by employing a Monte Carlo approach that approximates the exponentially growing sample space by a fixed finite number, M, of mode sample sequences $s_t^{(m)}$, where m = 1, ..., M. A weight is associated with each of the sample sequences to represent the *a posteriori* probability of that sample sequence,

$$\xi_t^{(m)} := \frac{p\left(\mathbf{s}_t^{(m)} \mid \mathbf{y}_t, \mathbf{u}_t\right)}{\sum_{m=1}^M p\left(\mathbf{s}_t^{(m)} \mid \mathbf{y}_t, \mathbf{u}_t\right)},\tag{6}$$

where a symbol in boldface, for example, s_t , represents a sequence from time 0 to time *t*. After the new measurement y_{t+1} is available, these weights are updated by

$$\xi_{t+1}^{(m)} = \frac{\xi_t^{(m)} \zeta_{t+1}^{(m)}}{\sum_{m=1}^M \xi_t^{(m)} \zeta_{t+1}^{(m)}},\tag{7}$$

where the incremental weight

$$\boldsymbol{\zeta}_{t+1}^{(m)} \coloneqq p\left(\boldsymbol{y}_{t+1} \mid \boldsymbol{y}_t, \boldsymbol{u}_t, \boldsymbol{s}_t^{(m)}\right), \tag{8}$$

represents the likelihood of the new measurement for a given mode sample sequence.

On each of these mode sample sequences, a (timevarying) Kalman filter is implemented to estimate the continuous states. The state estimates on all mode sample sequences are then "mixed" (averaged) by the weights, and this weighted average approximates the *a posteriori* estimate of the continuous states, *i.e.*,

$$\hat{x}_{t|t} = \sum_{m=1}^{M} \xi_t^{(m)} \hat{x}_{t|t}^{(m)}, \tag{9}$$

where $\hat{x}_{t|t}^{(m)}$ is the *a posteriori* state estimate from each of the Kalman filters.

The accuracy of this Monte Carlo method is improved by a predictive sampling technique, in which the current mode is sampled according a predictive probability

$$\mu_{t+1}^{(m)}(s) \propto p\left(s_{t+1}^{(m)} = s, y_{t+1} \mid s_t^{(m)}, y_t, u_{t+1}\right),$$
(10)

which favors the mode with higher likelihood given the current measurements and the previously sampled modes.

The weight update procedure is recursive. The entire history from time 0 influences the current weight. It is often found in implementation that most of these weights approach 0 while only a few remain of modest magnitudes. This phenomenon reduces the effective number of sample sequences and introduces an underflow risk for



Fig. 4. An aerial photo of the test segment of Interstate 210 Westbound, from Vernon to Fair Oaks, in Pasadena, California (Image source: U.S. Geological Survey digital orthophoto quarter quadrangle, 1996-12-03).

the weights when implemented on a machine with finite floating point precision. Therefore, a forgetting and weight underflow prevention scheme [2] has been introduced in our implementation. In this scheme, the weights of the sample sequences are bounded from below,

$$\xi_{t+1}^{(m)} = \max\left\{\xi_{t+1}^{(m)}, \xi\right\},\tag{11}$$

and are re-normalized after the bounding step. This simple procedure not only prevents the underflow, but in effect limits the influence of the early history on the weights and makes the weights recover more quickly when their corresponding sample sequences are favored by the current measurements.

In addition to the mixture estimate of the continuous states, the mixture Kalman filter also provides an approximate maximum *a posteriori* (MAP) estimation of the congestion mode:

$$\hat{s}_{t,\text{MAP}} = \arg \max p\left(s_t = s \mid \boldsymbol{y}_t, \boldsymbol{u}_t\right), \quad (12)$$

where

$$p\left(s_{t}=s\mid\boldsymbol{y}_{t},\boldsymbol{u}_{t}\right)\approx\sum_{m=1}^{M}\xi_{t}^{\left(m\right)}\boldsymbol{1}_{s}\left(s_{t}^{\left(m\right)}\right),\tag{13}$$

and $1_s(s_t)$ is the indicator function. This is particularly important to our application because different control schemes will be used in different congestion modes (free-flow or congested), based on the highway density controllability properties [3].

III. TEST SITE

A segment of Interstate 210 Westbound in Pasadena, California, as shown in an aerial photo in Fig. 4, has been selected as the testbed for new ramp-metering algorithms. It is approximately 14 miles long, from Vernon (Mile Post 39.159) to Fair Oaks (Mile Post 25.4). This site was selected as the preferred test location for the following reasons:

- 1) It is a heavily used freeway segment that experiences severe recurring congestion and can benefit greatly from ramp-metering strategies that reduce congestion and improve total travel time.
- I-210 has most of the necessary infrastructure, such as loop detectors, on-ramp metering signal controllers, and a centralized advanced traffic management system (ATMS), that is needed to test new ramp-metering designs.

- 3) New ramp metering schemes, such as System Wide Adaptive Ramp Metering (SWARM), have undergone testing and evaluation on I-210.
- 4) District 7 of the California Department of Transportation (Caltrans D7), which manages this highway segment, continues to support researchers in exploring innovative traffic management techniques.

Therefore, our efforts in traffic modeling and model calibration have been concentrated on this highway segment.

A. Cell Configuration

To apply the switching-mode model to the test highway segment, we first need to divide it into cells. There are several constraints and considerations that we must take into account during this process.

- The cell transmission model requires that the cell length be larger than the product of the free-flow speed and the time step. In other words, vehicles cannot travel across more than one cell within one time step.
- 2) With the measurements (density and flow) from mainline loop detectors, the section of highway between two mainline detectors can be decoupled from the rest of the highway. Therefore, we generally divide the highway into sections at the mainline loop detectors. We skip detectors that are believed to be malfunctioning and combine the neighboring sections into one.
- Within one section, cells are divided at the mergepoints of on-ramps and the diverge-points of offramps.

Based on the above considerations, as well as the nominal value of the free-flow speed, 63 miles per hour, the geometry of the highway and the locations of the mainline detectors, we have chosen the sampling time to be 2 seconds.

Fig. 5 on the next page shows the cell configuration for the chosen test segment of I-210. In the schematic, the cells are numbered at the corner of each cell, while their lengths, in feet, are marked in the center, in *italic* typeface. The arrows indicate on- or off-ramps. The entire test segment is divided into 73 cells, and the cells are grouped into 18 sections divided by mainline loop detectors, as shown by thick solid lines in the figure. Some of the mainline loop detectors, shown by thick dashed lines, are not used because they are believed to be malfunctioning.

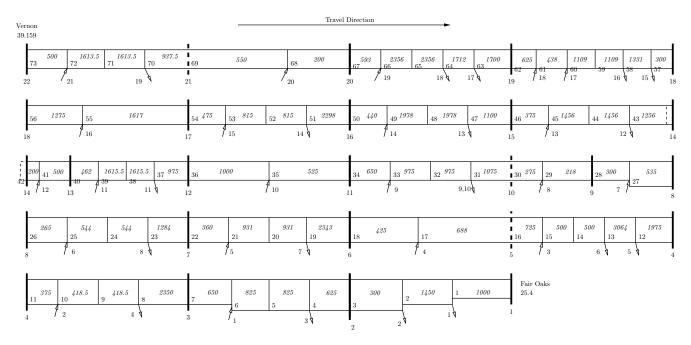


Fig. 5. The cell configuration for the test segment of I-210 West. The cell numbers are marked at the corner of each cell, while the cell length (in feet) are indicated in italic type in the center of each cell. The schematic is not to scale.

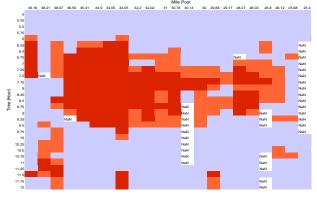
IV. RESULTS AND DISCUSSION

This mixture Kalman filter-based congestion mode and vehicle density estimator has been implemented for the entire 14-mile long test segment. The program was written in the C language for reasons of efficiency and portability. Not only can the estimator run on collected traffic data sets, it has been successfully interfaced with a calibrated VISSIM microscopic traffic simulator [6], through the VISSIM DDE (Direct Data Exchange) interface, and with a calibrated macroscopic cell transmission model [5]. The estimator is running synchronously with these traffic simulators. The running time of the estimator with 10 mode sample sequences for a 7-hour (from 5AM to 12 noon) time period and the full 14-mile segment is less than one minute on a 1.4 GHz Pentium M computer.

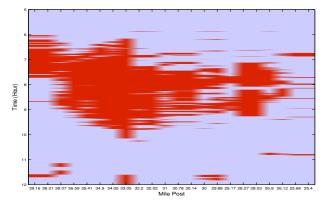
The traffic data were extracted from the PeMS [1] database. The traffic flows and vehicle densities are available every 30 seconds, while the speeds are available every 5 minutes. We interpolated the 30-second flow and density data into 2-second intervals and passed the interpolated data through a low-pass filter to reduce the amount of noise in the original 30-second traffic data. The estimator produced the vehicle density for each cell, as well as the congestion mode for each section, using these 2-second interpolated and filtered data.

Fig. 6(b) shows an example of the MAP congestion mode estimation from the estimator. In this example, data from January 10, 2002 were used. In the plot, blue indicates free-flow mode, while red indicates congestion.

For comparison, a contour plot of the PeMS-derived 15minute average speeds for that day is given in Fig. 6(a). In this plot, blue indicates an average speed of 55 miles per

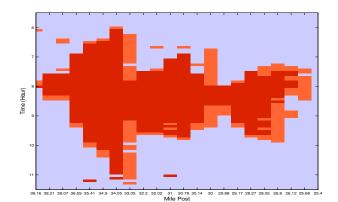


(a) PeMS 15-minute average speed contour plot (Red: <40 mph; orange: 40–55 mph; blue: >55 mph).

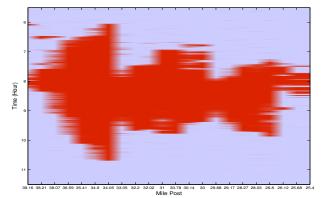


(b) MKF maximum *a posteriori* estimation (Red: congestion; blue: free-flow).

Fig. 6. Congestion mode estimation for the test segment of Interstate 210 Westbound in Pasadena, California (January 10, 2002).



(a) VISSIM 5-minute average speed contour plot (Red: <40 mph; orange: 40–55 mph; blue: >55 mph).



(b) MKF maximum *a posteriori* estimation (Red: congestion; blue: free-flow).

Fig. 7. Congestion mode estimation with a VISSIM microsimulation model for the test segment of Interstate 210 Westbound in Pasadena, California.

hour and above, which is generally considered to indicate free-flow conditions in traffic engineering. Red indicates an average speed of 40 miles per hour and below, in which the traffic is considered to be in congestion. Orange indicates an average speed between 40 and 55 miles per hour, in which the traffic is somewhat likely to be in congestion. In the plot, white indicates unavailable data.

As another example, similar plots for the VISSIM simulation model of the same I-210 segment are shown in Fig. 7. In this example, the measurements for the MKF-based estimator were from the traffic flows and vehicle densities simulated by VISSIM. The estimated congestion modes, as shown in Fig. 7(b), are compared with a contour plot (Fig. 7(a)) of the speeds that were simulated by VISSIM.

It can be seen from the plots that in general, the congestion mode estimation by the MKF-based estimator agrees with the speed contour plot. However, the MKF-based congestion mode estimation is preferable for the following reasons.

1) As mentioned earlier, the thresholds for the speed are determined empirically and can vary from location to

location.

- It is not clear whether a section is in congestion or not when the speed is between the upper and lower thresholds.
- 3) The speed data usually are not available as frequently as the density and flow data when the data are collected using single loop detectors, which is usually the case.
- 4) More importantly, the MKF-based estimation provides the statistically most probable mode that directly corresponds to one of the possible dynamic models, while the speed-based estimate itself does not have this direct correspondence.
- 5) The MKF-based estimator also provides vehicle density estimation for all the cells where no measurements are available.

V. SUMMARY

In the paper, we presented our latest results on developing and implementing a traffic congestion mode and vehicle density estimator for a segment of Interstate 210. Using the mixture Kalman filtering algorithm on the switching-mode traffic model, the estimator is able to provide the estimated vehicle densities at unmeasured locations, as well as the traffic congestion modes (free-flow or congested), which are not directly observed. The program runs efficiently, thus making it possible to carry out estimation in real time. The availability of the congestion modes enables us to design more effective ramp metering algorithms, utilizing the appropriate switching-mode model dynamics under different flow conditions.

REFERENCES

- [1] "The freeway performance measurement system." [Online]. Available: http://pems.eecs.berkeley.edu/
- [2] X. Sun, L. Muñoz, and R. Horowitz, "Highway traffic state estimation using improved mixture Kalman filters for effective ramp metering control," in *Proceedings of the 42nd IEEE Conference on Decision and Control*, Maui, Hawaii, USA, Dec. 9–12 2003, pp. 6333–6338.
- [3] L. Muñoz, X. Sun, R. Horowitz, and L. Alvarez, "Traffic density estimation with the cell transmission model," in *Proceedings of the* 2003 American Control Conference, Denver, CO, USA, June 2003, pp. 3750–3755.
- [4] R. Chen and J. S. Liu, "Mixture Kalman filters," *Journal of the Royal Statistical Society, Series B—Statistical Methodology*, vol. 62, pp. 493–508, 2000.
- [5] L. Muñoz, X. Sun, D. Sun, G. Gomes, and R. Horowitz, "Methodological calibration of the cell transmission model," in *Proceedings of the 2004 American Control Conference*, Boston, MA, USA, June 30– July 2 2004, to appear.
- [6] G. Gomes, A. D. May, and R. Horowitz, "Calibration of VISSIM for a congested highway," in *The 83rd Annual Meeting of the Transportation Research Board*, Washington, D.C., Jan. 2004.
- [7] C. F. Daganzo, "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory," *Transportation Research Part B: Methodological*, vol. 28, no. 4, pp. 269–287, Aug. 1994.
- [8] —, "The cell transmission model, part II: Network traffic," *Transportation Research Part B: Methodological*, vol. 29, no. 2, pp. 79–93, Apr. 1995.