Identification and Analysis of Queue Spillovers in City Street Networks

Nikolas Geroliminis and Alexander Skabardonis

***Abstract*—We propose a methodology for identifying queue spillovers in city street networks with signalized intersections using data from conventional surveillance systems, such as counts and occupancy from loop detectors. The key idea of the proposed methodology is that when spillovers from a downstream link block vehicle departures from the upstream signal line, queues discharge at rates smaller than the saturation flow. The application of the methodology on an arterial site and the comparison with field data show that it consistently identifies spillovers in urban networks with signal-controlled intersections. The method is extended to account for the variations in vehicle lengths. We also investigate the significant effect of spillovers in congestion and show that a macroscopic diagram that connects spillovers with vehicle density exists in large-scale congested urban networks.**

***Index Terms*—Arterial networks, performance measures, queues, spillbacks, traffic congestion.**

I. Introduction

SUBSTANTIAL research has been devoted to developing accurate and reliable techniques for estimating performance measures such as queue lengths and travel times on arterials and networks controlled by traffic signals. In addition, several studies have been undertaken for the traffic control and queue management of oversaturated arterials (e.g., [1]–[4]). However, the literature is limited in methods for identifying queue spillovers in city street networks with signal-controlled intersections. Spillovers occur when growing queues at the downstream signal block the arrivals from the directly upstream signal and vehicles cannot depart, although the signal phase is green. Spillovers may also occur when turning vehicles fill up the available storage length of turn bays and block the through movements. When these physical queues exceed the link length, departures from the upstream link are blocked and lead the entire system to restricted mobility and service inefficiency.

Spillovers are the result of oversaturated conditions but, simultaneously, are the cause of intensely increasing queued traffic and the creation of additional congestion. It is already known, for example, that spillovers past merges can lead to gridlocks on ring roads and other networks with closed loops [5]. Modeling of congestion dynamics and spillovers for pedestrians has attracted interest from a microscopic trajectory prospective [6] or using queuing theory tools [7]. Osorio and Bierlaire [8] developed an analytic finite-capacity queuing network model to capture blocking traffic for roads with signalized intersections. Recently, a model for travel times estimation in signalized arterials [9], which integrates the effect of spillovers in link capacity, has observed that when spillovers occur, the travel delay can increase by 50%–100% for short distance links between successive intersections. The practical implications of identifying spillovers in a consistent and fast way are important; for example, it can be a smart approach to preempt congestion by predicting congestion locations/times and have remedial strategies in place to restrict access to highly congested areas (as opposed to the current state of the art, where traffic management strategies react to congestion). It can also be integrated in real-time traffic management schemes, either at intersection scale (actuated control, e.g., [10]) or at larger complex urban systems (e.g., [11]).

This paper presents the development and application of a robust method for identifying spillovers on arterials based on surveillance data. The proposed methodology can readily be applied in large urban areas and provides important information with regard to oversaturated links and active bottlenecks in city street networks. Another main novelty of this paper is the investigation of the macroscopic effect of spillovers in urban mobility at the network level. We observe that spillovers are not only local system disturbances, but they might spread as well. We also observe that spillovers influence the spatial distribution and magnitude of congestion pockets in a network and significantly decrease the overall system performance, as this expressed by network flows and speeds. We address these questions based on a network-based approach, motivated by recent findings in the macroscopic modeling of traffic in cities [12], [13]. We show that the total number of vehicles at spillovers is a key variable of urban congestion in a city network, which reveals clear functional relationships rather than having qualitative descriptions of congested traffic.

In this paper, Section II describes the methodology for predicting spillovers in city street networks with signalized intersections and the extension of the methodology for populations of vehicles with different vehicle lengths. Next, we present the results from the application of the method on a real-world arterial in Section III. Section IV analyzes the effect of spillovers in congestion by providing results from a simulation

Manuscript received February 8, 2011; revised March 10, 2011 and April 4, 2011; accepted April 5, 2011. The Associate Editor for this paper was S. Tang.

N. Geroliminis is with the Urban Transport Systems Laboratory and the School of Architecture and Civil Engineering, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland (e-mail: nikolas.geroliminis@epfl.ch).

A. Skabardonis is with California Partners for Advanced Transportation Technology, Institute of Transportation Studies, University of California, Berkeley, Berkeley, CA 94720 USA (e-mail: skabardonis@ce.berkeley.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2011.2141991

of the San Francisco downtown network. Finally, we discuss the study findings and propose ideas for future work in Section V.

II. Methodology

Loop detectors are the predominant sensor technology for surveillance and control along signalized arterials and networks. Data from single inductive loop detectors include vehicle count and occupancy (the proportion of time that the loop is occupied). The raw data are typically reported back to a transportation management center (TMC) at 20- or 30-s intervals. The average speed of vehicles can be obtained from the count and occupancy data using (1) [14]

$$\bar{u}_{s_i} = \frac{N_i \cdot L_{eff}}{T \cdot o_i} \tag{1}$$

where $\bar{u}_{s_i}$ is the space mean speed for a time interval i, N_i and o_i are the measured volume and occupancy in i, respectively, T is the length of time interval (20 or 30 s), and L_{eff} is the average effective vehicle lengths (EVLs) of the traffic stream (the average vehicle length plus the detector length).

In practice, L_{eff} has been assumed to have a constant value; for example, the Washington State Department of Transportation uses $L_{eff} = 20 - 25$ ft [15]. In reality, L_{eff} varies as the average EVL changes with vehicle composition. Hellinga [16] proposed an algorithm that uses dual-loop measured vehicle lengths to calculate L_{eff} and applied that value to estimate speed at adjacent single-loop stations. Wang and Nihan [17] studied the relationship between lane occupancy and speed and concluded that L_{eff} can be considered constant only when all vehicle lengths were approximately equal.

Loop detectors for traffic surveillance along arterials are usually sufficiently placed upstream from the intersection stop line (system detectors) so that measured flows and occupancies are not affected by the presence of queues at the traffic signal. This assumption is violated in cases of heavy traffic or oversaturated conditions, because growing queues extend past the detector location. We present a method of predicting spillovers in city street networks with signal-controlled intersections by using the detectors' count and occupancy data. The analysis is based on the selection of data in time intervals of one cycle but can be applied for more disaggregated data.

We estimate the queue dynamics according to the kinematic wave [Lighthill–Whitham–Richards (LWR)] theory, which was originally proposed by Lighthill and Whitham [18], [19] and Richards [20], to explicitly consider the temporal and spatial formation of queues. We assume a piecewise linear flow–density relationship (fundamental diagram) with parameters u_f (free-flow speed), k_j (jam density), and u (congested wave speed). According to the LWR theory, shock waves are generated by the traffic signal, which cause congested conditions to develop near the stop line during the red interval and capacity conditions to occur in the period during which the queue discharges at the saturation flow rate. When the queue has dissipated, the rest of the platoon that arrives during the green time crosses the intersection stop line with no interference from the traffic signal. Fig. 1 shows the shockwaves at an isolated signal under the assumed form of the flow–density relationship. Because the system detector is placed at distance L_d upstream the stop line, the effect of the red phase and queues is identified with a time lag.

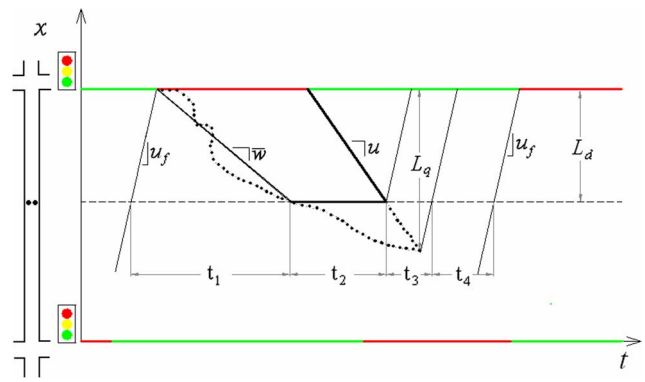

Fig. 1. Shockwaves in an isolated signal.

The key idea of the developed methodology is that, when spillovers from the downstream signal block vehicle arrivals from the upstream signal line, the queue discharges at rates smaller than the saturation flow during the green phase. The first part of this section sets out the relation between growing queues and high values of occupancy, whereas the second part explicates the threshold values of occupancy for which spillovers occur.

A. Identification of Growing Queues

When queues extend past the detector location, the platoon arrivals are not known, and for an amount of time (interval t_2 in Fig. 1), the measured vehicle count is zero, and the detector occupancy is close to 1. Under the assumed triangular flow–density relationship, drivers arrive from the upstream signal and reach the detector at free-flow speed u_f if they are not constrained by queues. Based on (1), the critical value of occupancy $\bar{o}_{cr}$ for which the queue length extends past the detector is given by (2), where $\bar{q}$ is the average flow measured by the detector

$$\bar{o}_{cr} = \frac{L_{eff} \cdot \bar{q}}{u_f}. \tag{2}$$

The importance of this simple formula is that if the variations of the free-flow speed between drivers are small, this value is stable, independent of the time interval at which the data are collected.

The cycle time c is divided into four time intervals by using the detector position as a reference space point. Interval t_1 is the time that queues do not reach the detector, whereas vehicles arrive during the red phase, t_2 is the time that the detector is occupied because of the congested conditions developed near the stop line during the red phase, t_3 is the time that the queue discharges at the saturation flow, and t_4 is the remaining green time that platoons cross the stop line with no interference from the traffic signal. L_d is the distance between the detector and the stop line, and L_q is the maximum queue length.

When the average occupancy $\bar{o}$ is greater than $\bar{o}_{cr}$, growing queues extend past the detector $(t_2 > 0)$, and the average

occupancy $\bar{o}$ is given by (3). The first term is the free-flow occupancy, whereas the second term expresses the congested conditions at the detector. The term o_{stop} is the jam occupancy with a value close to 1, and N is the vehicle count during that cycle. The free-flow speed u_f can be obtained from data under undersaturated conditions, and the EVL is approximately equal to the average vehicle length plus the detector length (about 20–25 ft). We have

$$\bar{o} = \frac{\sum_{i=1,3,4} t_i \frac{L_{eff} q_i}{u_f} + t_2 \cdot o_{stop}}{c} = \frac{\frac{L_{eff}}{u_f} \cdot N + t_2 \cdot o_{stop}}{c} = \frac{L_{eff} \cdot \bar{q}}{u_f} + \frac{t_2 \cdot o_{stop}}{c}. \quad (3)$$

The time interval t_2 that the detector is occupied, assuming $o_{stop} \approx 1$, is

$$t_2 = \frac{c}{o_{stop}} \cdot \left(\bar{o} - \frac{L_{eff} \cdot \bar{q}}{u_f} \right) \cong c \cdot \left(\bar{o} - \frac{L_{eff} \cdot \bar{q}}{u_f} \right). \quad (4)$$

The estimation of t_2, which is based only on values of observable quantities, is an important tool for identifying traffic states. First, when t_2 takes negative values, queues do not reach the detector at any point in the cycle, i.e., the link is uncongested, and the vehicle accumulation in that link never exceeds the value $L_d \cdot k_j$, where L_d is the distance between the detector and the stop line, and k_j is the jam density. Furthermore, as $t_2 > 0$ increases, queue reaches the detector, and the arrival rate during that cycle is more intense. Thus, the rate that queues grow in that link is faster, as expressed by the average shockwave speed $\bar{w}$ in Fig. 1. Note that (2)–(4) hold, although the arrivals at the traffic signal are not homogeneous in time. In the following section, we present the methodology for spillover identification from the estimated value of t_2 using the detector count and occupancy data.

B. Identification of Spillovers

Spillovers are a result of oversaturated conditions but, simultaneously, are the cause of increasing intensely queued traffic and the creation of more congestion as a "chain reaction." This case is illustrated in Fig. 2, which shows a time–space diagram for three consequent signals. Vehicles arrive at the first upstream signal at rate A, and during the red phase for the through movements of all signals, turning traffic enters the through direction with rate $B < A$. It is easily recognizable that the upstream signal operates at undersaturated conditions, because residuals queues do not occur. However, growing queues at the downstream signal block the arrivals from the middle signal, and after two cycles, the spillover reaches the upstream undersaturated signal. Spillovers may also occur when turning vehicles fill up the available storage length of turn bays and block the through movements.

The key concept in spillover identification is the observation that the queue discharges at a lower rate than the saturation flow s. Thus, the methodology focuses on simultaneously recognizing the presence of queues and discharging rates smaller than saturation flow s. To achieve this goal, the time interval t_2 is considered, as described in (4). If there is no blocking traffic,

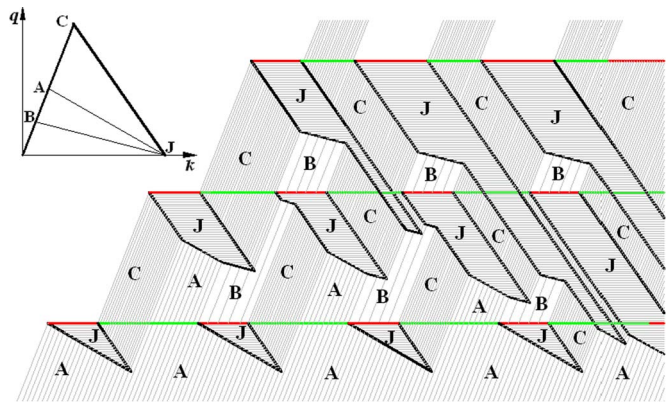

Fig. 2. Time–Space diagram for three successive traffic signals.

then the maximum value of t_2 is equal to the duration of the red phase r, because the maximum speed of the shockwave is the congested wave speed u, which occurs when the arrival rate is equal to the saturation flow s. By applying this observation ($t_2 = r$) to (3), we estimate the critical value for the "blocking occupancy" o_{sp} as

$$o_{sp} = \frac{L_{eff} \cdot \bar{q}}{u_f} + \frac{r}{c}. \quad (5)$$

Therefore, if the measured occupancy is greater than the critical value o_{sp}, we conclude that spillovers from the downstream link block the departures from the first link upstream. This condition is sufficient for the occurrence of spillovers but is not necessary. It is a way of identifying the spillovers when oversaturated conditions take place and the queues extend past the detector location. Consequently, the aforementioned methodology captures the existence of "active" spillovers, i.e., spillovers that result in congestion problems and the underutilization of the green phase. Such spillovers cause a drastic decrease in vehicle speeds and a possible extension of spillovers to more links upstream if the demand continues to be high. For example, if the aforementioned methodology is applied in the traffic conditions presented in Fig. 2, blocking traffic for the middle signal is identified with a delay of one signal cycle. The use of more disaggregated data can identify spillovers without the delay described in the previous example.

C. Extension of the Method for Different Vehicle Lengths

The aforementioned methodology predicts the existence of spillovers, assuming a constant EVL L_{eff}. In this section, we extend the methodology to account for the variation in vehicle lengths. Higher values of detector occupancies are obtained not only because of long queues that pass the detector location but because of LVs that travel in free-flow speed and occupy the detector more time because of their length as well. Our analysis shows that the effect of vehicle length variation in the estimation of o_{sp} is not significant.

The distribution of vehicle lengths generally follows a bimodal distribution with two separated peaks: one higher peak for passenger cars and a smaller peak for long vehicles (LVs), e.g., trucks and buses, as shown in Fig. 3.

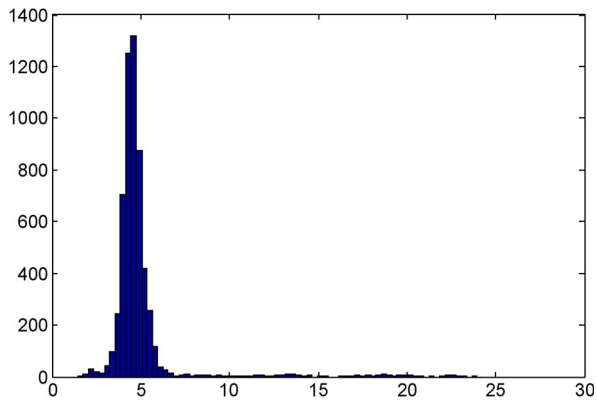

Fig. 3. Histogram of vehicle length distribution for U.S. 101 (the horizontal axis refers to the length (in meters), and the vertical axis refers to the frequency) [1].

Yeo *et al.* [21] found that both small-vehicle (SV) and long-vehicle (LV) lengths are normally distributed. Therefore, SV lengths are assumed to follow the $N(\mu_1, \sigma_1^2)$ distribution (called x_1), and LV lengths are assumed to follow the $N(\mu_2, \sigma_2^2)$ distribution (called x_2), where μ_1 and σ_1 are the mean and standard deviation of SV lengths, and μ_2 and σ_2 are the mean and standard deviation of LV lengths, respectively. We can approximate the vehicle length density function π as

$$\pi = \xi x_1 + (1-\xi)x_2 \tag{6}$$

where ξ is the Bernoulli distribution with probability of success p, where p is the proportion of SVs. Assuming that ξ, x_1, and x_2 are independent, the average vehicle length is simply $p\mu_1 + (1-p)\mu_2$. With respect to the variance of π, we have

$$\begin{aligned}
&\mathrm{var}\,[\xi x_1 + (1-\xi)x_2] \\
&\quad = \mathrm{var}\,[\xi x_1] + \mathrm{var}\,[(1-\xi)x_2] + 2\mathrm{cov}\,(\xi x_1, (1-\xi)x_2) \\
&\quad = \mathrm{var}[\xi x_1] + \mathrm{var}\,[(1-\xi)x_2] \\
&\qquad + 2\mathrm{cov}(\xi x_1, x_2) - 2\mathrm{cov}(\xi x_1, \xi x_2) \\
&\quad = \mathrm{var}[\xi x_1] + \mathrm{var}\,[(1-\xi)x_2] - 2\mathrm{cov}(\xi x_1, \xi x_2).
\end{aligned} \tag{7}$$

Using statistical analysis and after some manipulations, for $i = 1, 2$, we have

$$\begin{aligned}
\mathrm{cov}(\xi x_1, \xi x_2) &= \mathrm{E}[\xi^2 x_1 x_2] - \mathrm{E}[\xi x_1]\cdot\mathrm{E}[\xi x_2] \\
&= \mathrm{E}[\xi^2]\mathrm{E}[x_1 x_2] - \mathrm{E}^2[\xi]\cdot\mathrm{E}[x_1 x_2] \\
&= (p - p^2)\mu_1\mu_2 \qquad (8)\\
\mathrm{var}[\xi x_i] &= \mathrm{E}\left[\xi^2 x_i^2\right] - \mathrm{E}^2[\xi x_i] \\
&= \mathrm{E}[\xi^2]\cdot\mathrm{E}[x_i^2] - \mathrm{E}^2[\xi]\cdot\mathrm{E}^2[x_i] \\
&= \mathrm{E}[\xi]\cdot\left(\mathrm{E}^2[x_i] + \mathrm{var}[x_i]\right) - \mathrm{E}^2[\xi]\cdot\mathrm{E}^2[x_i] \\
&= p\cdot\left(\mu_i^2 + \sigma_i^2\right) - p^2\cdot\mu_i^2. \qquad (9)
\end{aligned}$$

Using (7)–(9), the variance of the vehicle length distribution π is

$$\begin{aligned}
&\mathrm{var}\,[\xi x_1 + (1-\xi)x_2] \\
&\quad = p\cdot\left(\mu_1^2 + \sigma_1^2\right) - p^2\cdot\mu_1^2 + (1-p)\cdot\left(\mu_2^2 + \sigma_2^2\right) \\
&\qquad - (1-p)^2\cdot\mu_2^2 - 2(p - p^2)\mu_1\mu_2 \\
&\quad = p\cdot\left(\sigma_1^2 - \sigma_2^2\right) + \sigma_2^2 + (p - p^2)\cdot(\mu_1 - \mu_2)^2.
\end{aligned} \tag{10}$$

TABLE I
CRITICAL BLOCKING OCCUPANCY WITH OR WITHOUT VARIATIONS IN VEHICLE LENGTHS FOR DIFFERENT VALUES OF p AND g/c ($\mu_1 = 6$ m, $\sigma_1 = 0.7$ m, $\mu_2 = 13$ m, $\sigma_2 = 2$ m, $u_f = 15.65$ m/s $= 35$ mi/h, $\bar{q} = (0.5$ veh/s$)\cdot$(g/c))

p \ g/c	No variation in lengths				With variation in lengths			
	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
0.99	.839	.678	.516	.355	.842	.683	.523	.362
0.95	.841	.681	.522	.362	.847	.690	.533	.375
0.90	.843	.686	.528	.371	.851	.697	.543	.388
0.85	.845	.690	.535	.380	.855	.704	.552	.400

Fig. 4. Schematic of the study site (Lincoln Avenue, Los Angeles).

Table I shows the estimated value of "blocking occupancy" with or without variation in L_{eff}. In the first case, (5) was applied using average values for L_{eff}. In the second case, the 95th percentile value of vehicle length was used for L_{eff}. The results show that the variation in vehicle lengths is not critical for estimating the blocking occupancy (the absolute error is less than 2% in most of the cases); therefore, the proposed methodology can be applied in cases where vehicle lengths vary. Note that, in arterials with higher speed limits or grades, the free-flow speed of LVs may be smaller than the free-flow speed of SVs. In this case, we suggest applying the aforementioned equations per lane using the free-flow speed of LVs for the rightmost lane and the speed of SVs for the remaining lanes.

III. METHOD VERIFICATION USING REAL-WORLD DATA

The selected test site is a 1.42-mi-long stretch of Lincoln Avenue, which is a major urban arterial near the Los Angeles International Airport (see Fig. 4). The study section includes seven signalized intersections, with the link lengths varying from 500 ft to 1600 ft. The number of lanes for the through traffic per link is three lanes per direction. Additional lanes for turning movements are provided at intersection approaches. The free-flow speed is 40 mi/h. Traffic signals are all multiphase operating as coordinated under the traffic-responsive control. The system cycle length ranges from 100 s early in the analysis period (6:00–6:30 A.M.) to a maximum of 150 s during the periods of the highest traffic volumes (7:30–8:30 A.M.). System loop detectors are located on each lane approximately 250 ft upstream the intersection stop line. Detector data every 30 s (vehicle count and occupancy) and signal timing data for the study period were obtained from the Los Angeles Automated Traffic Surveillance and Control central traffic control system database. Floating-car runs were performed at 7-min headways. Vehicle location and speed were recorded on each second using Global Position System (GPS) units. The study period enabled us to obtain data for a wide range of traffic conditions (from low-volume off-peak conditions to peak-period conditions). For

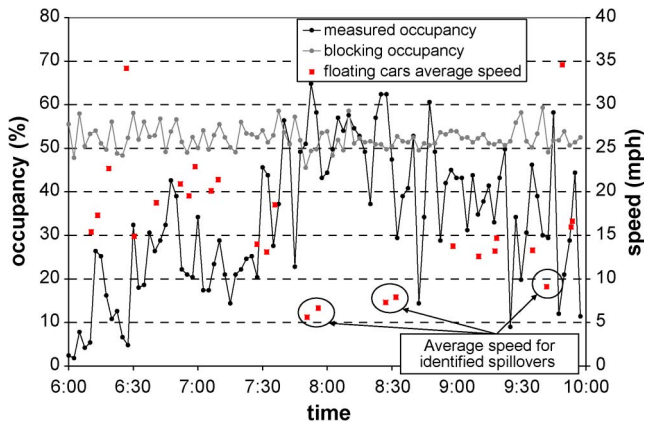

Fig. 5. Time series of measured and blocking occupancy at the Washington intersection (left axes) and the average speed of floating cars (right axes).

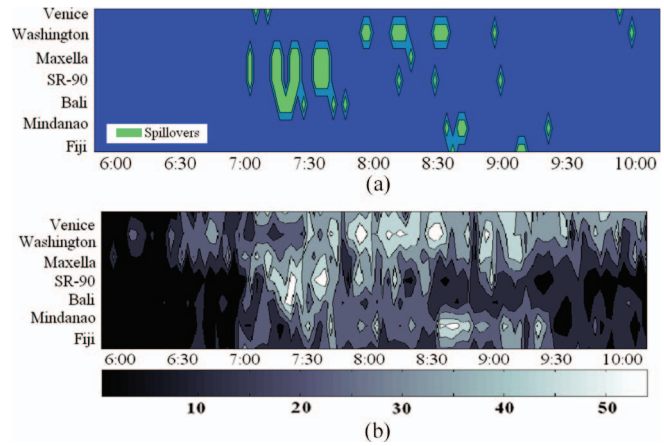

Fig. 6. (a) Spillover plot in space and time. (b) Occupancy contour plot in the study site.

more information about the field study, the reader should refer to [22].

Traffic demand is high, particularly during the peak hour. Traffic volumes are heavily directional, with the higher volumes in the northbound direction. The average travel speeds on the test section are 25 mi/h during the off-peak times and drop to about 10 mi/h during the peak hour in the heavily traveled northbound direction. The proposed methodology is applied to the aforementioned site by using the occupancy and count data to spatially and temporally identify the spillovers. In addition, average speeds per cycle for each link are calculated using the GPS data from floating cars for comparison. The results show that spillovers cause very low observed speeds during the peak hour (7:00–9:00). Because cameras were not available to spatiotemporally identify spillover occurrences, the floating car data were utilized.

Fig. 5 illustrates the following three factors: 1) values of the measured occupancy from the detectors; 2) critical occupancy that was estimated from (5) at the Washington intersection; and 3) average speeds of probe vehicles. We observed that floating-car average speeds at this intersection were measured less than 9 mi/h when spillovers identified. In addition, we observed that floating cars repetitively stopped during the green phase of the signal for cycles with spillovers.

Fig. 6(a) plots different traffic states in the Lincoln site for the 4-h study period. The green color indicates the existence of spillovers, whereas blue indicates the opposite. The interesting result is that locations of spillovers are very clearly presented. By considering the effect of spillovers in delays and the decrease of mobility, the aforementioned method detects the critical congested intersections in a network. An intervention in these critical points could result in the avoidance of long queues and the efficient operation during the peak hour. As shown in Fig. 6(b), a simple contour plot of occupancies in the study site contains significant noise and cannot give reliable information about blocking traffic.

IV. Effect of Spillovers in Urban Traffic

One interesting question that arises is the following: How important is the effect of intersection queue spillovers at the network level. In addition, are spillovers local system disturbances, or do they spread and decrease the overall system performance, and if yes, how? We address these questions based on a network-based approach, motivated by recent findings in the macroscopic modeling of traffic in cities.

Recent research findings indicate that we can model traffic in large urban areas at an aggregate level if the area can be partitioned into regions that are uniformly congested (see [12] and [13]). The findings that are based on a microsimulation of downtown San Francisco and a field experiment in downtown Yokohama, Japan, show the following: 1) Congested city centers approximately exhibit a well-defined macroscopic fundamental diagram (MFD) that relates the number of vehicles (accumulation) in the region to the region's average flow or speed, and 2) there is a linear relation between the region's average flow and its total output (rate vehicles reach their destinations). It was also found that the average network flow is maximum for the same value of vehicle density, independent of the origin–destination tables, and that the average trip length for the study area is about constant with time, i.e., the total output versus the density curve is a scaled-up version of the average network flow versus the density curve. Further analysis of the San Francisco simulations is presented here to investigate the effect of spillovers in network traffic.

This test network is a 2.5-mi^2 portion of the San Francisco downtown area (financial district and south of the market area), including about 100 intersections, with link lengths varying from 400 ft to 1300 ft (see Fig. 7). The number of lanes for through traffic varies from two to five lanes, and the free-flow speed is 30 mi/h. Traffic was simulated for a period of 4 h with time- and space-dependent traffic demands, starting from low flows and increasing to higher flows until the system reaches a gridlock. Several simulation runs were made with different demand profiles. Note that the total output increased with accumulation up to a critical value ($\sim$3500 vehicles) and then decreased to a gridlock state with almost zero output (each scenario is shown with different colors in Fig. 8).

Our conjecture is that the existence of spillovers is highly related to the decrease in the system output. If the system reaches a congested state in the decreasing part of the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

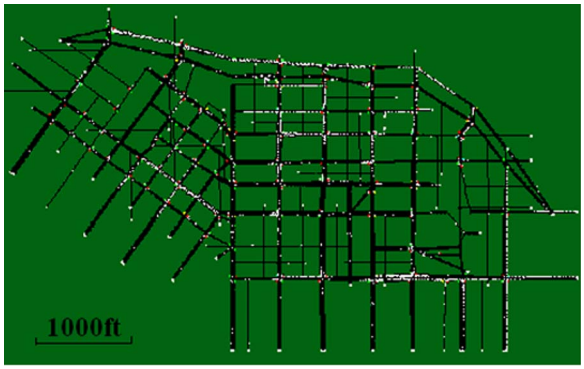

Fig. 7. View of the San Francisco network.

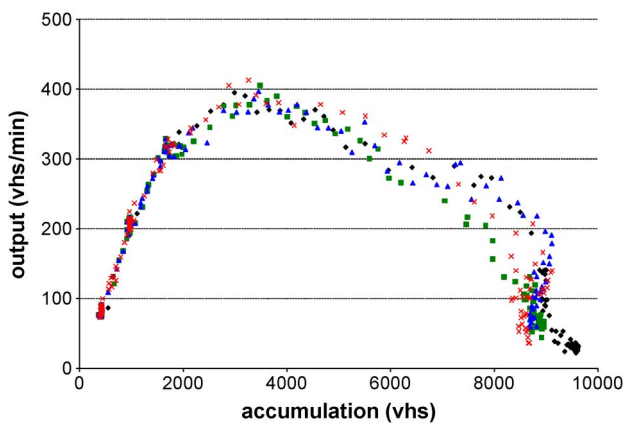

Fig. 8. Output versus accumulation pairs for different runs in the San Francisco network aggregated every two cycles [12].

output-accumulation curve and the demand continues to be high, then accumulation increases, and this case can lead the system to a gridlock. To validate this conjecture, the number of spillovers is calculated in each cycle during the simulation. To investigate the effect (if any) of the travel demand in a macroscopic relation between performance network variables and spillovers, we analyze two different scenarios, as presented in Fig. 9. This figure shows the spatial distribution of the total number of trips originated and terminated in the study region. Note that, in scenario 2, most of the trip origins are external from the southwest corner of the city, whereas in scenario 1, the trips uniformly originated in the periphery and a substantial internal portion (∼30%). With regard to destinations, in scenario 1, trip terminations are almost uniformly distributed among the city, whereas in scenario 2, most trip terminations are external, and cars exit from the boundaries of the network.

We now quantitatively describe how the number of spillover-blocked vehicles S_t in every cycle t is estimated. This value is the number of vehicles that could get served if queues from the downstream link did not spill back and block the arrivals. If c is the signal cycle duration, s is the saturation flow, q_{it} and g_{it} are the output and the duration of the green phase at link i during cycle t, respectively, and x_{it} is a binary variable with a value equal to 1 if phase failures occur (when vehicles queued at a signal are not all discharged in one cycle during the phase that serves the queue), then

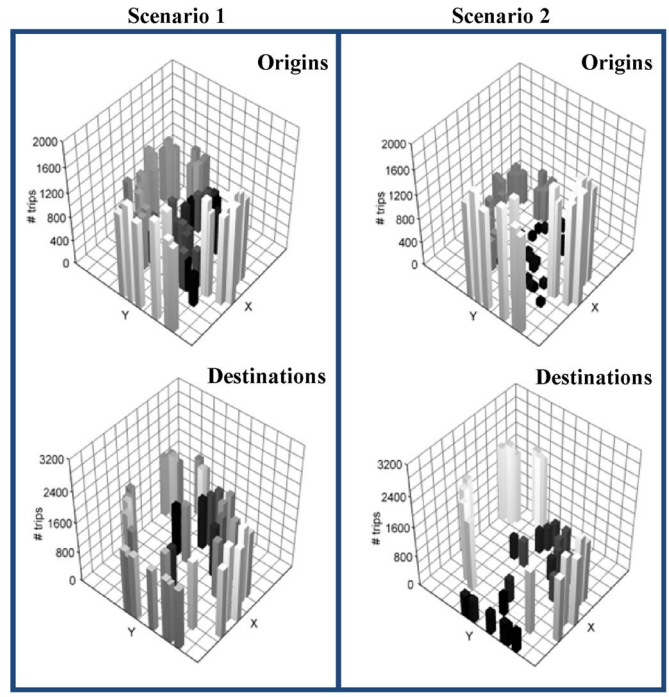

Fig. 9. Output versus spillovers pairs for two scenarios in the San Francisco network.

$$S_t = \sum_i (sg_{it} - q_{it}c) \cdot x_{it}. \tag{11}$$

Fig. 10 shows the system output (in vehicles per minute), average network speed (in kilometers per hour), and system accumulation (in number of vehicles) versus the number of spillover-blocked vehicles S_t (in vehicles per cycle) aggregated per two cycles for scenarios 1 and 2. It is clear that, when the number of spillovers increases above a threshold, the performance of the system drastically decreases, because fewer vehicles reach their destinations. Note that the output increases for small values of S_t, which might be counterintuitive at a first look. This case is because demand increases and more links operate close to capacity. Thus, there are a few instances that spillovers occur but with no significant effect in the overall network performance (note the distribution of link occupancy in Fig. 12). Although it is difficult to estimate the rate that vehicles reach their destinations (system output) without monitoring equipment in every vehicle, the space–mean speed is an easily observable quantity, even by utilizing loop detector data. There is a clear functional relationship between the space–mean speed and the number of spillovers, which is insensitive to vastly different origin–destination tables [see Fig. 10(b)]. Note that, although S_t remains small (< 250) for values of accumulation below the critical accumulation, it increases with an almost-constant rate (*one* additional spillover per cycle for every two additional cars in the network) until the system reaches a state of gridlock [see Fig. 10(c)]. These results suggest that, by keeping the number of spillover-blocked vehicles below specific thresholds, we can guarantee efficient service for the network users.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

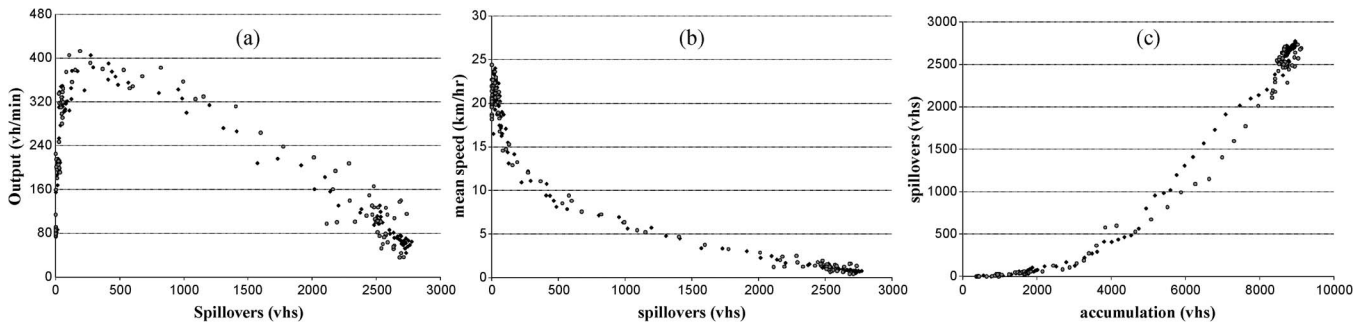

Fig. 10. Effect of spillovers in system performance measures for two different demand scenarios in the San Francisco network. (a) Output versus spillovers. (b) Average network speed versus spillovers. (c) Spillovers versus accumulation.

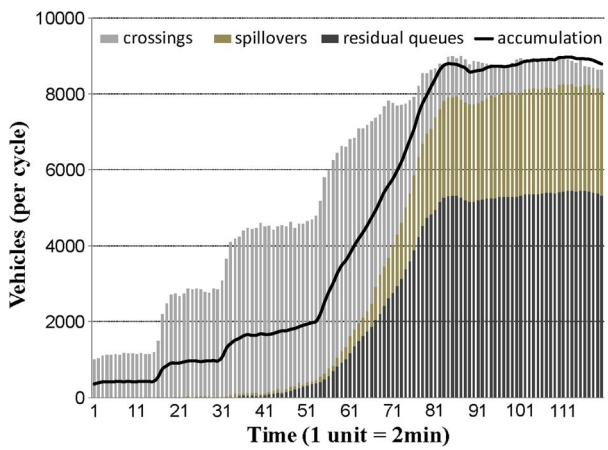

Fig. 11. Time series of the number of spillovers, the number of vehicles in residual queues, the total number of vehicles in the network (accumulation), and the total number of intersection crossings per cycle.

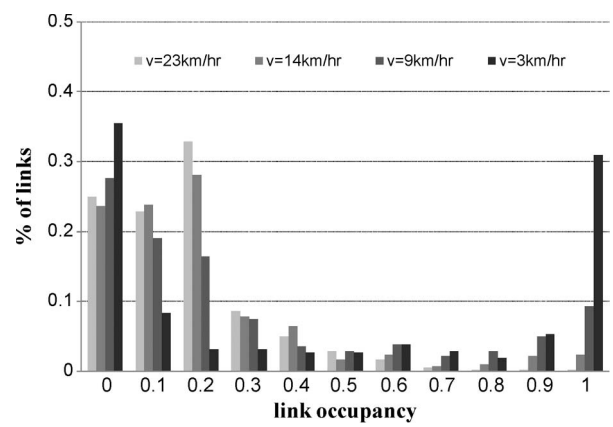

Fig. 12. Spatial distribution of link occupancy from four network snapshots at different congestion levels.

One interesting observation is that the system output is maximized for a value of spillovers higher than zero (∼250), which means that the topology of the network and the trip distribution do not allow an even and deterministic spatial distribution of vehicles in the network, i.e., it is not possible that all links perform at capacity. This condition can be verified by calculating the sum of flows from all links and noticing that this value is smaller than $\sum_i s \cdot g_{it}/c$.

To have a better understanding of congested traffic conditions at the network level and explain the significant decrease in the system output and space–mean speed, we distinguish the number of vehicles in every link i for every cycle t between the following three categories: 1) vehicles that exit the link from the downstream end by crossing the stop line; 2) vehicles that join a residual queue at the upstream end of the link and cannot get served in the same cycle at which they arrived; and 3) vehicles that could be served in the same cycle if spillovers did not occur. We then estimate in every cycle the total number of vehicles for each category. The number of intersection crossings (total circulating flow in one cycle) is $IC_t = \sum_i q_{it}$; the number of vehicles in residual queues is $RQ_t = \sum_i \max\{0, n_{it} - sg_{it}\}$, where n_{it} is the number of vehicles in link i in cycle t; and the number of spillover-blocked vehicles S_t is given by (11).

Fig. 11 shows the number of vehicles in each category with time for scenario 1. The maximum system output occurs at approximately 60 time units when $S_t = 250$ veh/cycle and the system is at the critical accumulation ($n_{cr} \approx 3500$ veh). The circulating flow is the maximum for the same value of accumulation. Note that, when S_t and RQ_t are small, the circulating flow in one cycle is larger than the number of vehicles in the system, because each vehicle travels more than one link and crosses more than one intersection per cycle. However, as queues become longer and vehicles join residual queues, some vehicles need more than one cycle to travel one link. Note that, for highly congested conditions, the sum of these three quantities is equal to the number of vehicles. When RQ further increases, residual queues spill back in the upstream links, and spillovers are created.

In general, the relationship between RQ and S depends on the length of the links and the cycle duration. For very short links with length $l < L_{eff} \cdot g \cdot s$, RQ is always zero, and spillbacks can occur very fast after a demand increase, whereas for long links, several cycles of high demand are needed.

In our experiment, we note that, when the number of vehicles n is higher than the critical accumulation, the sum of intersection crossings plus spillovers is smaller than the maximum value of circulating flow, which occurs at the critical accumulation, i.e., $IC_t + S_t < IC_{\max} \quad \forall t: \; n(t) > n_{cr}$. This case happens because the spatial distribution of vehicles in the network changes with the level of congestion. For highly oversaturated conditions, vehicles concentrate in a smaller number of links in the network, and even for links without blocking traffic from downstream, the total circulating flow decreases because of

residual queues. Fig. 12 provides a quantitative evidence for the aforementioned point. It plots histograms of link occupancy $o_{it} = n_{it}L_{eff}/l$ (each column represents a 10% range) for network snapshots with space–mean speeds of 3, 9, 14, and 23 km/hr (14 km/hr is the speed when the circulating flow is maximum). By comparing histograms for speed equal to 14 km/hr with 9 and 3 km/hr, we notice that the fraction of empty and full links increases, whereas the fraction of links with no residual queues decreases. This condition results in a significant decrease in the circulating flow, which is much higher than the number of spillovers.

These results illustrate that simple control strategies can enhance the traffic performance. These strategies should focus on reducing the variability of the vehicle densities by avoiding spillovers and a large number of vehicles in residual queues. This type of research is under way.

V. Conclusion

We have presented and validated a methodology of identifying queue spillovers in urban networks with signalized intersections and have shown the significant effect of spillovers in the system output and efficiency. The practical implications of the models, after sufficient validation and fine-tuning, can be important; for example, it is a smart approach to preempt congestion by predicting congestion locations/times and have remedial strategies in place (as opposed to the current state of the art, where traffic management strategies react to congestion). This paper has also provided some evidence that control for equalizing density throughout the network can be beneficial for network performance. Ongoing and future research involves the development of strategies for monitoring traffic in congested urban networks using different types of monitoring equipment (e.g., loop detectors, cameras, and GPS technology).

Effective monitoring is essential for developing observation-based control. New technologies of wireless sensors (for example, see the work of Kwong and others) can provide substantial improvement in accurate measures of flow and occupancy, which are needed for different performance measures, including spillovers. Avoiding spillovers in congested parts of a city can lead to better utilization of the system and an increase in mobility and accessibility. This condition is possible by prioritizing critical vehicle queues (for example, see the work of Lammer and Helbing) or restricting access to regions that exceed certain density thresholds (for example, see the work of Geroliminis and Daganzo). Recently, Helbing and Mazloumian have proposed a signal control strategy that explains the slower-is-faster effect when the utilization of a road section is very small such that it requires extra time to collect enough vehicles for an efficient "platooning" service during the green phase. Similarly, the slower-is-faster strategy would suggest restricting the inflow to congested areas to keep the number of spillovers low. At the local control level, further modeling is required, because decisions should be made before the queues reach the critical length of the link. The application of methodology to different types of signal control and networks (e.g., adaptive, pretimed, and mixed) should also be investigated.

References

[1] D. C. Gazis, "Optimum control of a system of oversaturated intersections," *Oper. Res.*, vol. 12, no. 6, pp. 815–831, Nov./Dec. 1964.

[2] P. G. Michalopoulos and G. Stephanopoulos, "Optimal control of oversaturated intersections: Theoretical and practical considerations," *Traffic Eng. Control*, vol. 19, no. 5, pp. 216–222, 1978.

[3] A. K. Rathi, "A control scheme for high-traffic-density sectors," *Transp. Res. B*, vol. 22B, no. 2, pp. 81–101, Apr. 1988.

[4] G. Abu-Lebdeh and R. Benekolah, "Development of traffic control and queue management procedures for oversaturated arterials," *Transp. Res. Rec.*, no. 1603, pp. 119–127, 1996.

[5] C. F. Daganzo, "Queue spillovers in transportation networks with a route choice," *Transp. Sci.*, vol. 32, no. 1, pp. 3–11, Feb. 1998.

[6] D. Helbing, A. Johansson, and H. Z. Al-Abideen, "The dynamics of crowd disasters: An empirical study," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 75, no. 4, p. 046 109, Apr. 2007.

[7] J. Y. Cheah and J. M. Smith, "Generalized M/G/C/C state-dependent queuing models and pedestrian traffic flows," *Queueing Syst.*, vol. 15, no. 1–4, pp. 365–386, Mar. 1994.

[8] C. Osorio and M. Bierlaire, "An analytic finite-capacity queuing network model capturing the propagation of congestion and blocking," *Eur. J. Oper. Res.*, vol. 196, no. 3, pp. 996–1007, Aug. 2009.

[9] A. Skabardonis and N. Geroliminis, "Real-time monitoring and control on signalized arterials," *J. Intell. Transp. Syst. Technol., Plan., Oper.*, vol. 12, no. 2, pp. 64–74, 2008.

[10] G. Zhang and Y. Wang, "Optimizing minimum and maximum green-time settings for traffic actuated control at isolated intersections," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 1, pp. 164–173, Mar. 2011.

[11] F. Y. Wang, "Parallel control and management for intelligent transportation systems: Concepts, architectures, and applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 3, pp. 630–638, Sep. 2010.

[12] N. Geroliminis and C. F. Daganzo, "Macroscopic modeling of traffic in cities," in *Proc. 86th Annu. Meeting Transp. Res. Board*, Washington, DC, 2007.

[13] N. Geroliminis and C. F. Daganzo, "Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings," *Transp. Res. Part B*, vol. 42, no. 9, pp. 759–770, Nov. 2008.

[14] P. Athol, "Interdependence of certain operational characteristics within a moving traffic stream," in *Highway Research Record 72*. Washington, DC: Nat. Res. Council, 1965, pp. 58–87.

[15] J. M. Ishimaru and M. E. Hallenbeck, "Flow evaluation design technical report," Washington Dept. Transp., Seattle, WA, Tech. Rep. WA-RD 466.2, 1999.

[16] B. R. Hellinga, "Improving freeway speed estimates from single-loop detectors," *ASCE J. Transp. Eng.*, vol. 128, no. 1, pp. 58–67, Jan./Feb. 2002.

[17] Y. Wang and N. L. Nihan, "Freeway traffic speed estimation using single-loop outputs," *Transp. Res. Rec.*, no. 1727, pp. 120–126, 2000.

[18] M. J. Lighthill and G. B. Whitham, "On kinematic waves I: Flood movement in long rivers," *Proc. R. Soc. Lond. A, Math. Phys. Sci.*, vol. 229, no. 1178, pp. 281–316, May 1955.

[19] M. J. Lighthill and G. B. Whitham, "On kinematic waves II: A theory of traffic flow on long crowded road," *Proc. R. Soc. Lond. A, Math. Phys. Sci.*, vol. 229, no. 1178, pp. 317–345, May 1955.

[20] P. I. Richards, "Shockwaves on the highway," *Oper. Res. B*, vol. 22, pp. 81–101, 1956.

[21] H. Yeo, A. Skabardonis, J. Halkias, J. Colyar, and V. Alexiadis, "Oversaturated freeway flow algorithm for use in next-generation simulation," *Transp. Res. Rec.*, no. 2088, pp. 68–79, 2008.

[22] A. Skabardonis and N. Geroliminis, "Real-time estimation of travel times along signalized arterials," in *Proc. 15th Int. Symp. Transp. Traffic Theory*, 2005, pp. 387–406.

[23] K. Kwong, R. Kavaler, R. Rajagopal, and P. Varaiya, "Real-time measurement of link vehicle count and travel time in a road network," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 4, pp. 814–825, Dec. 2010.

[24] S. Lammer and D. Helbing, "Self-control of traffic lights and vehicle flows in urban road networks," *J. Stat. Mech., Theory Exp.*, vol. 2008, p. P04 019, 2008.

[25] D. Helbing and A. Mazloumian, "Operation regimes and slower-is-faster effect in the control of traffic intersections," *Eur. Phys. J. B*, vol. 70, no. 2, pp. 257–274, Jul. 2009.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Nikolas Geroliminis received the Diploma in civil engineering from the National Technical University of Athens, Athens, Greece, and the M.Sc. and Ph.D. degrees in civil engineering from the University of California, Berkeley.

He is currently the Director of the Urban Transport Systems Laboratory and an Assistant Professor with the School of Architecture and Civil Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. Before joining EPFL, he was an Assistant Professor with the Department of Civil Engineering, University of Minnesota, Minneapolis, where he still has an Adjunct Professor's appointment. He serves on the Editorial Boards of *Transportation Research Part B* and *Transportation Letters*. His research interests focus primarily on urban transportation systems, traffic flow theory and control, public transportation, and logistics.

Prof. Geroliminis is a Member of the Traffic Flow Theory Committee of the Transportation Research Board. He received the University of California Transportation Student of the Year Award in 2007 and the Outstanding Graduate Student Instructor Award in 2006.

Alexander Skabardonis received the Diploma in civil engineering from the National Technical University of Athens, Athens, Greece, and the Ph.D. degree in transportation engineering from the University of Southampton, Southampton, U.K.

He is an internationally recognized expert in traffic flow theory, traffic management and control systems, the modeling and simulation of traffic operations, the design and operation of transportation facilities, intelligent transportation systems, and the energy and environmental impact of transportation facilities. He is currently the Director of the California Partners for Advanced Transportation Technology, Institute of Transportation Studies, University of California, Berkeley, where he is also a Professor of civil and environmental engineering and a Research Engineer. He has extensively worked on the development and application of models and techniques for traffic control, performance analysis of highway facilities, and applications of advanced technologies to transportation. He serves on the Editorial Board of the *Intelligent Transportation Systems Journal* and as a Reviewer for several transportation journals. He has served as the Principal Researcher for more than 60 funded contracts and grants and has published more than 200 papers and technical reports. He is a Codeveloper of the California Freeway Performance Measurement System and the Berkeley Highway Laboratory.

Prof. Skabardonis is a Member of the Traffic Flow Theory and Freeway Operations Committees of the Transportation Research Board.