Available online at www.sciencedirect.com
SCIENCE DIRECT®
Applied Mathematical Modelling 28 (2004) 445–468
APPLIED MATHEMATICAL MODELLING
www.elsevier.com/locate/apm

Stochastic modeling and real-time prediction of incident effects on surface street traffic congestion

Jiuh-Biing Sheu [a,*], Yi-Hwa Chou [b], Allen Chen [b]

[a] *Institute of Traffic and Transportation, National Chiao Tung University, 4F, 114 Chung Hsiao W. Rd., Sec. 1, Taipei 10012, Taiwan, ROC*

[b] *Department of Civil Engineering, National Taiwan University, Taipei, Taiwan, ROC*

Received 7 November 2001; received in revised form 2 September 2003; accepted 24 October 2003

Abstract

Modeling and real-time prediction of incident-induced time-varying lane traffic states, e.g., mandatory lane-changing fractions, queue lengths, and delays are vital to investigate the time-varying incident effects on traffic congestion in both the spatial and temporal domains. This paper presents a discrete-time nonlinear stochastic model to characterize the time-varying relationships of specified lane traffic states under the condition of lane-blocking incidents on surface streets. The proposed stochastic model is composed of four types of equations: (1) recursive equations, (2) measurement equations, (3) delay-aggregation equations, and (4) boundary constraints. In addition, a recursive estimation algorithm is developed for real-time prediction of the specified time-varying lane traffic states. The proposed method is tested with simulated data generated using the Paramics traffic simulator. The preliminary tests indicate the capability of the proposed method to estimate incident effects on surface street traffic congestion in real time. We also expect that this study can provide real-time incident-related traffic information with benefits both for understanding the impact of incidents on non-recurrent traffic congestion of surface streets, and for developing advanced incident-responsive traffic control and management technologies.

Keywords: Stochastic modeling; Real-time prediction; Incident-induced traffic congestion; Discrete-time

* Corresponding author. Tel.: +886-2-2349-4963; fax: +886-2-2349-4953.
E-mail address: jbsheu@mail.nctu.edu.tw (J.-B. Sheu).

0307-904X/$ - see front matter
doi:10.1016/j.apm.2003.10.004

1. Introduction

Non-recurrent traffic congestion caused by incidents is a critical traffic problem in urban areas. The increasing impacts of incidents on traffic congestion not only critically undermine mobility but also can be a major cause of bottlenecks or even secondary accidents. Earlier studies [1] have revealed that incident-induced traffic congestion continues to impose frustrating delays on road users, and remains a growing trend in the near future. Furthermore, such anomalous traffic congestion may significantly affect the system stability of lane traffic states either in the time domain or in the space domain, thus invalidating existing traffic control and management strategies [2].

Clearly, real-time estimation of incident effects on surface street traffic congestion remains a critical issue in related fields. Although over the past few decades significant advances have been made in investigating the problems of incident-induced traffic congestion, studies on real-time estimation of surface street incident effects are rare. That is, the early research related to modeling of traffic congestion appears to be limited for use either in the past-event off-line analysis [3–11] or in the model-based estimation of freeway flow variables for non-incident cases [12–14]. More importantly, most previous literature has focused on the scope of freeway incidents rather than surface street incidents, which may be more complicated and difficult to address. Details on these limitations are discussed below.

An attempt was made to use the kinematic wave theory in predicting individual travel times during freeway incidents by Messer et al. [3]. However, their test results revealed that inaccuracy in the estimation of wave speeds may cause serious misinterpretation in the prediction of incident-induced travel time. Chow [4] compared the methods of shock wave analysis and queuing analysis in order to assess the performance of incident delay calculation. That study implied that the use of a time-varying flow-density relationship may lead to more realistic results in calculating the total incident delay on a freeway section. The deterministic approach proposed by Morales [5]was based on the assumption that the demand and the capacity are constant in small time intervals, so that the cumulative delay can then be estimated using the linear arrival and departure curves according to this method. Similar attempts can be found in [6,7], where specific empirical models were proposed to estimate aggregated incident delays on freeways using a real incident database. Despite the potential advantage of the aforementioned deterministic queuing (DQ) models in estimating incident-induced delay, several issues, e.g., underestimation of total delay and the stochastic nature of incident duration, can make these DQ-based models unrealistic, as pointed out in [8]. To estimate freeway incident congestion, Al-Deek et al. [9] proposed a macroscopic method based on shock wave analysis. Although cases of single and multiple incident delays are explored in their study, the use of homogeneous traffic data for estimating wave speeds and delays may cause the congestion estimates to be unrealistic in response to incident occurrence in real time. Using multiple regression analytical techniques, two regression models were proposed in [10] to predict cumulative incident delay. Although factors affecting non-recurring congestion are considered in the model development, their models appear to be incapable of characterizing the time-varying nature of incident-induced lane traffic maneuvers. In [11], a probability-based model was proposed for incident-induced average delay estimation in the case that incident and induced traffic characteristics, e.g., incident-induced capacity, traffic arrival and departure rates, are known. Nevertheless, both the specifications of appropriate probability distribution functions in terms of incident duration and delay, and the stochastic modeling of time-varying traffic characteristics may remain as major issues in their approach for real-time applications.

In the other aforementioned field, model-based estimation of freeway flow variables, numerical mathematical models and algorithms have all been proposed to estimate the system dynamics of freeway traffic flows [12–14]. It should be noted that much effort made in the early research is focused on the development of macroscopic dynamic models to estimate basic traffic variables such as flow, speed and density for non-incident cases. Therefore, these studies seem to lack solid evidence to show the applicability of the published dynamic estimation approaches to incident-induced traffic congestion cases.

The methods published in the traditional field of automatic incident detection (AID) also fail to characterize the effects of surface street incidents on traffic congestion. It can be found that the functionality provided by traditional AID technologies seems to be restricted to recognizing incident occurrence [15–18]. Such binary output information, incident or non-incident, appears to be inadequate for further use in formulating incident impacts.

Apparently, there has been little research which address the aforementioned issues for further real-time applications, except for two previous studies: one was conducted for the real-time estimation of incident effects, including time-varying delays and queue lengths, on freeway traffic congestion [2]; and the other was the real-time estimation of mandatory lane-changing fractions in the presence of lane-blocking incidents on surface streets [19]. Although these published methods appeared to permit characterizing incident-induced traffic congestion in real time using raw traffic data collected from point detectors, it was implied that considerable effort remained necessary to in re-formulate the incident-induced traffic congestion problems for the cases of surface streets incidents. It is also worth mentioning that in contrast to freeway incident cases, traffic signal control can be another significant factor that should be considered in formulating the incident-induced traffic congestion for surface street incident cases. One typical example is the estimation of lane traffic states under the condition of signal transition steps where traffic signal turns either from RED to GREEN or from GREEN to RED at a given time step. Correspondingly, traffic signal control may significantly affect the continuity of system state prediction in case of intersection incidents.

To efficiently mitigate arterial incident-induced impacts, and even to avoid the formation of bottlenecks on surface streets during incidents, the development of advanced approaches to real-time estimation of incident effects on arterial traffic congestion appears vital. It is generally agreed that incident-induced traffic congestion can be mitigated utilizing two advanced strategies: (1) providing real-time incident-induced traffic information via advanced traveler information systems (ATIS) to travelers so they can alter their decisions on route choices, and (2) estimating time-varying incident-induced traffic variables which are used as the inputs of advanced traffic management systems (ATMS) in response to anomalous changes in incident impacts.

This study therefore investigates a new methodology, which can be used for real-time estimation of incident effects on surface street traffic congestion. The proposed method is developed on the basis of stochastic modeling approach. To estimate time-varying incident-induced lane traffic variables and impacts, only the upstream and downstream point detector data are needed in the proposed method. Correspondingly, the upstream and downstream detector data, e.g., lane traffic counts, are used as the input of the proposed model to estimate all the specified incident-induced lane traffic states. Time-based and space-based incident impacts on traffic congestion, including delays and queue lengths, respectively, are then calculated in real time using the estimates of the time-varying lane traffic states. Herein, the estimates of incident-induced lane traffic states and impacts are very useful for real-time characterization of incident-induced traffic congestion on surface streets in both the spatial and temporal domains. Details of the methodology development are described in the following sections.

2. Specification of system states

The system investigated in the study is bounded by any given pair of upstream and downstream detector stations on surface streets, where the area within the upstream and downstream detector stations is referred to as a detection zone. In view of the difference in traffic patterns between arterials and intersections, incidents investigated in this study are further classified into two categories: (1) arterial lane-blocking incidents (i.e., incidents occurring on the roadway between two adjoining intersections), and (2) intersection incidents (i.e., incidents occurring within a given intersection). Fig. 1 depicts the detector layouts which are proposed particularly for estimating

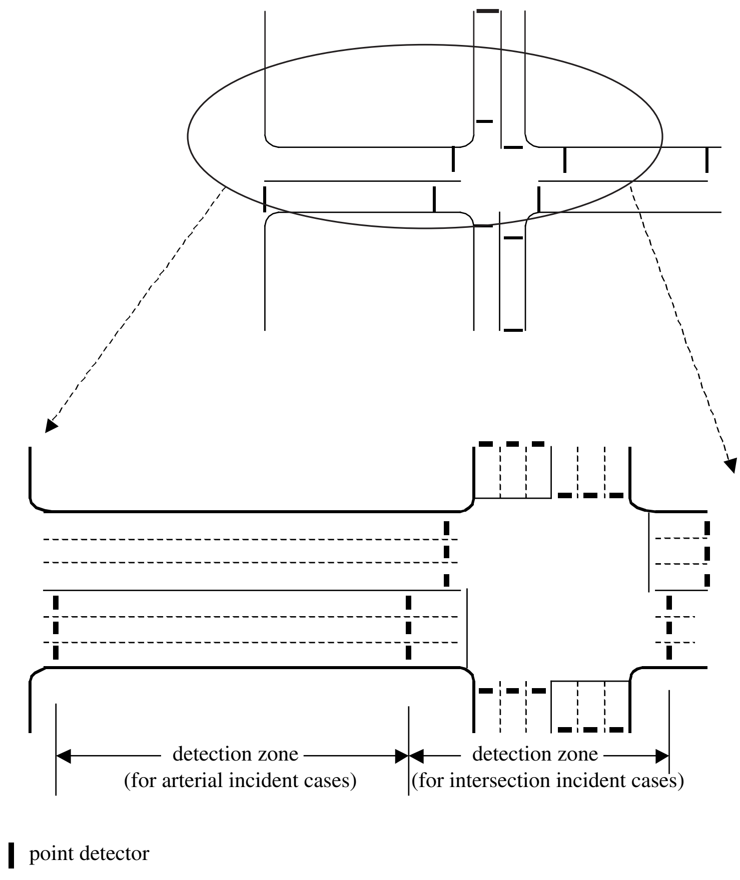

Fig. 1. Detector layouts for arterial and intersection incident cases on surface streets.

surface street incident impacts in real time if the proposed model is utilized. Such detector layouts are the basic requirement for implementation of the proposed model, and they are feasible from a technical point of view. However, our focus in this study is formulating real-time incident impact prediction problems with a stochastic modeling approach, rather than economically evaluating the feasibility of system implementation. The tasks of system implementation and corresponding evaluation are thus omitted in this study. Once a given type of incident is identified as being either an intersection incident or arterial incident, the corresponding upstream and downstream detector data will be used for real-time state estimation, all within the corresponding detection zone. Note that state estimation for sequential detection zones may be needed only in case of queue overflowing beyond the given downstream detection zone. Under this condition, the issues of multi-zone state estimation may be another major concern. However, as mentioned above, in this study we focus on the system bounded by any given pair of upstream and downstream detector stations on surface streets, and thus the aforementioned multi-zone state estimation issues are not addressed in this study.

In order to model two different types of lane-changing maneuvers which are potentially conducted during arterial incidents, we further specified two subsystems for arterial incident cases. Subsystem 1 geographically represents the area upstream from the incident site, and is specified to formulate the incident-induced mandatory lane changing from blocked lanes to adjacent lanes. In contrast, Subsystem 2, which is situated downstream from the incident site, is specified to depict the maneuvers of discretionary lane changing from adjacent lanes to blocked lanes in that area.

Once a lane is blocked within a given detection zone on an arterial, two types of lane traffic maneuvers are significant in the system: lane changing and queuing, which are also two sources of traffic congestion during incidents. In the presence of a lane-blocking arterial incident, vehicles in blocked lanes are compelled to make lane changes, and in this study such lane-changing behavior is referred to as incident-induced mandatory lane changing. However, it seems too idealistic to assume that vehicles present in the blocked lane are able to complete lane-changing maneuvers at will because of a variety of traffic conditions in adjacent lanes, and thus vehicular queuing may form in the blocked lane. On the other hand, the behavior of lane changing from any adjacent lane to the blocked lane may occur downstream from the incident site once the vehicles have passed the incident site by the adjacent lane. Such return-lane-changing behavior is viewed as a type of discretionary lane-changing behavior in this study.

Note that here we capture the downstream discretionary lane-changing fractions for two purposes. First, such lane-changing behavior may affect the time-varying lane traffic characteristics, e.g., approaching delays, lane densities and queue lengths, downstream from the incident side, all within the incident link; and thus this behavior must be considered for real-time incident impact prediction. Second, both the downstream lane traffic counts (referring to the measurements to update prior predictions of lane traffic states of the proposed model) and discretionary lane changes are mutually affected. If the downstream lane changing fractions are not considered in the proposed model, the real-time estimation of all the other incident-induced lane traffic states may be biased.

It is induced that the inter-lane and intra-lane traffic maneuvers aforementioned may significantly affect the stability of traffic flow, either in the blocked lane or in the adjacent lane. Therefore, the various delays are induced during incidents, as illustrated in Fig. 2.

In contrast with the case of arterial incidents, as mentioned above, further specification of two subsystems is not necessary for the cases of intersection incidents. Note that the detector

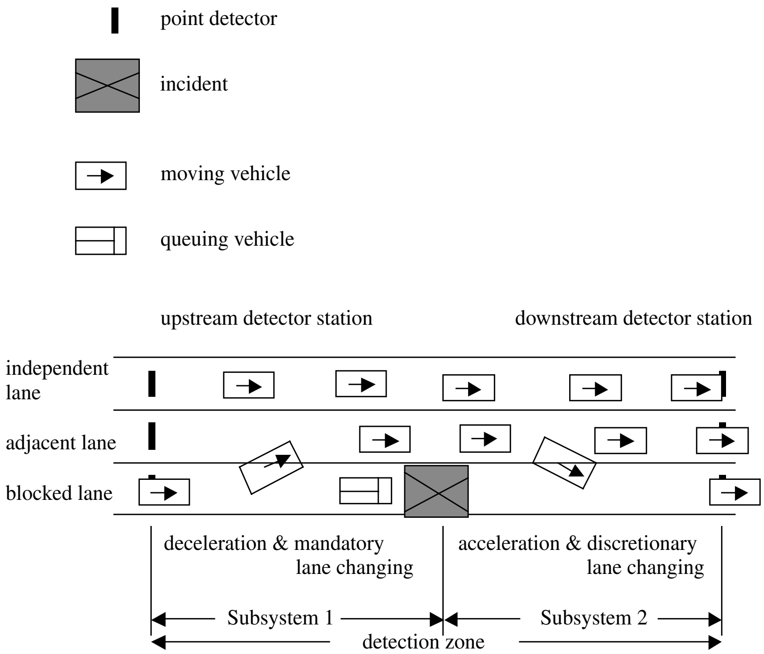

Fig. 2. Potential traffic maneuvers during lane-blocking arterial incidents.

spacing in the case of intersection incidents may not be long enough to observe the aforementioned return-lane-changing behavior completely within the detection zone. Therefore, traffic characteristics possessed by Subsystem 2 are ignored in the cases of intersection incidents.

Based on the above hypotheses, the following three groups of time-varying traffic variables are proposed to characterize incident-induced traffic congestion on surface streets: (1) basic lane traffic states, (2) space-based incident impacts, and (3) time-based incident impacts.

Basic lane traffic states refer to the elements used to derive the other groups of lane traffic variables, and their time-varying relationships are formulated in the recursive equations of the proposed stochastic model, as detailed in the following section. In the system, eight types of basic lane traffic states are specified as follows:

(1) $p_{i,j}^{m,1}(k)$ represents the time-varying mandatory lane-changing fraction from blocked lane i to adjacent lane j in Subsystem 1 of link m at time step k;
(2) $r_{j,j}^{m,1}(k)$ corresponds to the proportion of the vehicles present in adjacent lane j of Subsystem 1 which can leave from Subsystem 1 of link m at time step k;
(3) $r_{i,j}^{m,1}(k)$ represents the proportion of the vehicles conducting lane-changing maneuvers from blocked lane i to adjacent lane j in Subsystem 1 of link m at time step k which can leave from Subsystem 1 at time step k;
(4) $r_{l,l}^{m,1}(k)$ is the proportion of the vehicles present in independent lane l of Subsystem 1 which can leave from Subsystem 1 of link m at time step k;
(5) $p_{j,i}^{m,2}(k)$ is defined as the discretionary lane-changing fraction from adjacent lane j to blocked lane i in Subsystem 2 of link m at time step k;

(6) $r_{j,j}^{m,2}(k)$ represents the proportion of the vehicles present in adjacent lane j of Subsystem 2 which can leave from Subsystem 2 of link m at time step k;
(7) $r_{l,l}^{m,2}(k)$ represents the proportion of the vehicles present in independent lane l of Subsystem 2 which can leave from Subsystem 2 of link m at time step k;
(8) $r_{j,i}^{m,2}(k)$ corresponds to the proportion of the vehicles conducting discretionary lane changing from adjacent lane j to blocked lane i in Subsystem 2 of link m which can leave from Subsystem 2 of link m at time step k.

In the aforementioned lane traffic states, subscriptions i, j, and l represent the lane codes associated with the blocked lane, the lane adjacent to the blocked lane, and the independent lane which herein corresponds to any lane, excluding the blocked and adjacent lanes. Furthermore, all the aforementioned eight types of basic lane traffic states apply to arterial incident cases; whereas, the first half of the states apply to intersection incident cases, which exhibit only the traffic characteristics of Subsystem 1.

Space-based incident impacts refer to the time-varying section-wide lane traffic variables which are used to signify the changes in the severity of incident effects on traffic congestion in the spatial domain. The space-based incident impacts specified in the system are primarily time-varying lane traffic loads and queue lengths. Lane traffic loads correspond to the numbers of vehicles moving in given lanes; in contrast, incident-induced queue lengths represent the numbers of vehicles stopped in blocked lanes during incidents. The definitions of these two space-based incident impacts are given below.

Time-varying lane traffic loads specified in the system can be further classified into five types of variables as follows:

(1) $\delta_j^{m,1}(k)$ represents the number of vehicles present in adjacent lane j of Subsystem 1 on link m at time step k;
(2) $\delta_l^{m,1}(k)$ represents the number of vehicles present in independent lane l of Subsystem 1 on link m at time step k;
(3) $\delta_i^{m,2}(k)$ corresponds to the number of vehicles present in blocked lane i of Subsystem 2 on link m at time step k;
(4) $\delta_j^{m,2}(k)$ represents the number of vehicles present in adjacent lane j of Subsystem 2 on link m at time step k; and
(5) $\delta_l^{m,2}(k)$ represents the number of vehicles present in independent lane l of Subsystem 2 on link m at time step k.

Utilizing the pre-specified basic lane traffic states, the aforementioned lane traffic load variables can be mathematically expressed as:

$$\delta_j^{m,1}(k) = \left[a_i^{m,1}(k) + q_i^{m,1}(k-1)\right] \times \left[1 - r_{i,j}^{m,1}(k)\right] \times p_{i,j}^{m,1}(k) + \left[a_j^{m,1}(k) + \delta_j^{m,1}(k-1)\right] \times \left[1 - r_{j,j}^{m,1}(k)\right], \quad (1)$$

$$\delta_l^{m,1}(k) = \left[a_l^{m,1}(k) + \delta_l^{m,1}(k-1)\right] \times \left[1 - r_{l,l}^{m,1}(k)\right], \quad (2)$$

$$\delta_i^{m,2}(k) = \sum_{\forall j \in J} \left\{ \left[\left(a_j^{m,1}(k) + \delta_j^{m,1}(k-1) \right) \times r_{j,j}^{m,1}(k) + \delta_j^{m,2}(k-1) \right] \times p_{j,i}^{m,2}(k) + \delta_i^{m,2}(k-1) \right\} \times \left[1 - r_{j,i}^{m,2}(k) \right], \tag{3}$$

$$\delta_j^{m,2}(k) = \left\{ \left[a_i^{m,1}(k) + q_i^{m,1}(k-1) \right] \times p_{i,j}^{m,1}(k) \times r_{i,j}^{m,1}(k) + \left[\left(a_j^{m,1}(k) + \delta_j^{m,1}(k-1) \right) \times r_{j,j}^{m,1}(k) + \delta_j^{m,2}(k-1) \right] \times \left[1 - p_{j,i}^{m,2}(k) \right] \right\} \times \left[1 - r_{j,j}^{m,2}(k) \right], \tag{4}$$

$$\delta_l^{m,2}(k) = \left\{ \left[a_l^{m,1}(k) + \delta_l^{m,1}(k-1) \right] \times r_{l,l}^{m,1}(k) + \delta_l^{m,2}(k-1) \right\} \times \left[1 - r_{l,l}^{m,2}(k) \right], \tag{5}$$

where $a_i^{m,1}(k)$, $a_j^{m,1}(k)$ and $a_l^{m,1}(k)$ are the lane traffic counts collected from the upstream detectors in blocked lane i, adjacent lane j, and independent lane l, respectively at time step k; $q_i^{m,1}(k-1)$ corresponds to the time-varying queue length in blocked lane i of Subsystem 1 on link m at time step $k-1$; and J represents the group of the lanes adjacent to the blocked lane.

Compared to lane traffic loads, vehicular queuing can be regarded as an extreme case of lane traffic loads which signify the difficulty of intra-lane traffic movements. In this study, two groups of time-varying queuing variables are specified to characterize the intra-lane static traffic states during incidents on surface streets. These two groups of variables are the queue length upstream from the incident site in a given blocked lane, and the number of vehicles stopping in either an adjacent lane or an independent lane during red intervals in the case of intersection incidents. These time-varying queue lengths can be formulated as

$$q_i^{m,1}(k) = \left[a_i^{m,1}(k) + q_i^{m,1}(k-1) \right] \times \left[1 - \sum_{\forall j \in J} p_{i,j}^{m,1}(k) \right], \tag{6}$$

$$q_j^{m,1}(k) = \left[a_j^{m,1}(k) + q_j^{m,1}(k-1) \right] + \left[a_i^{m,1}(k) + q_i^{m,1}(k-1) \right] \times p_{i,j}^{m,1}(k), \tag{7}$$

$$q_l^{m,1}(k) = \left[a_l^{m,1}(k) + q_l^{m,1}(k-1) \right], \tag{8}$$

where $q_i^{m,1}(k)$ is the time-varying queue length in blocked lane i of Subsystem 1 on link m at time step k; $q_j^{m,1}(k)$ and $q_l^{m,1}(k)$ represent the time-varying queue lengths in adjacent lane j and independent lane l, respectively on link m at time step k during red intervals in case of intersection incidents.

Note that using the corresponding mathematical forms, all the aforementioned space-based incident impacts are updated for each time step in the proposed recursive estimation algorithm, as described in Section 4.

In contrast to the space-based incident impacts specified above, delay is regarded as a significant variable indicating the severity of incident effects on traffic congestion in the temporal domain. In this study, the specified time-varying delays are categorized into three groups: (1) the stopping delays caused by either the vehicles queuing in blocked lanes or red intervals, (2) the acceleration-or-deceleration delays caused either by vehicular acceleration usually present in Subsystem 2 or by vehicular deceleration occurring frequently in Subsystem 1, and (3) lane-changing delays. The details of their notations are given below.

The incident-induced stopping delay, for the most part, occurs upstream from the incident site in blocked lanes. In this study, the time-varying stopping delay in blocked lane i of Subsystem 1 on link m at a time step k can be expressed as

$$d_i^{m,1}(k) = t, \tag{9}$$

where t represents the length of any given time step.

In addition, stopping delay caused by red intervals exists in the case of intersection incidents. To characterize the signal effects on stopping delay in the presence of an intersection incident, the stopping delays in blocked lane i, adjacent lane j and independent lane l, respectively on link m at time step k during red intervals ($\Psi_i^{m,1}(k)$, $\Psi_j^{m,1}(k)$, $\Psi_l^{m,1}(k)$) can be specified as

$$\Psi_i^{m,1}(k) = R_i^m(k), \tag{10}$$

$$\Psi_j^{m,1}(k) = R_j^m(k), \tag{11}$$

$$\Psi_l^{m,1}(k) = R_l^m(k), \tag{12}$$

where $R_i^m(k)$, $R_j^m(k)$, and $R_l^m(k)$ represent the lengths of the red intervals associated with blocked lane i, adjacent lane j, and independent lane l, respectively on link m at time step k.

The acceleration or deceleration delays may exist either in Subsystem 1 when traffic arrivals are approaching the platoon present in a given lane, or in Subsystem 2 once the vehicles have passed by the incident site. Therefore, five types of acceleration-or-deceleration delays are specified:

(1) $d_{i,i}^{m,1}(k)$ corresponds to the deceleration delay caused by an unit vehicle approaching from blocked lane i to the end of the vehicles queuing in blocked lane i of Subsystem 1 on link m at time step k, and is given by

$$d_{i,i}^{m,1}(k) = \text{Min}\left\{ \left[e^{m,1} - s \times q_i^{m,1}(k-1)\right] \times \left[\frac{1}{u_i^{m,1}(k)} - \frac{1}{u_i^{m,1}(0)}\right], t \right\}, \tag{13}$$

where $e^{m,1}$ corresponds to the length of Subsystem 1 on link m (i.e., the distance between the upstream detector station and the incident site); s is defined as the average vehicle length; $u_i^{m,1}(k)$ represents the speed detected in blocked lane i on link m at time step k via the upstream detector station; and $u_i^{m,1}(0)$ corresponds to the highest speed detected in lane i of link m in incident-free cases. Herein, it is assumed that the lengths of Subsystems 1 and 2 are derivable once the incident location is determined from other data/information sources, e.g., incident detection technologies and manual reports from on-site drivers. Because this study scope is limited to real-time incident impact prediction rather than incident detection, the availability of the aforementioned information is not discussed in this paper.

(2) $d_{j,j}^{m,1}(k)$ is defined as the deceleration delay caused by an unit vehicle which approaches from adjacent lane j to the incident site on link m at time step k, and is given by

$$d_{j,j}^{m,1}(k) = \text{Min}\left\{ \frac{\text{Max}\{u_j^{m,2}(k), u_j^{m,1}(k)\} \times t}{\text{Min}\{u_j^{m,2}(k), u_j^{m,1}(k)\}} - t, t \right\}, \tag{14}$$

where $u_j^{m,1}(k)$ corresponds to the average speed in adjacent lane j on link m measured from the upstream detector station at time step k; and similarly $u_j^{m,2}(k)$ is referred to as the speed estimated in adjacent lane j right adjacent to the incident site on link m at time step k. Note that by considering the relationships among traffic flow, density and speed, $u_j^{m,2}(k)$ can be estimated by

$$u_j^{m,2}(k) = \frac{\frac{\rho_j^{m,2}(k)}{t}}{\frac{\delta_j^{m,1}(k) + \delta_j^{m,2}(k)}{e^{m,1} + e^{m,2}}}, \tag{15}$$

where $e^{m,2}$ corresponds to the length of Subsystem 2 on link m; and $\rho_j^{m,2}(k)$ is the estimated traffic count in lane j right adjacent to the incident site on link m at time step k. Using the time-varying state variables, we have

$$\rho_j^{m,2}(k) = \left[a_j^{m,1}(k) + \delta_j^{m,1}(k-1)\right] \times r_{j,j}^{m,1}(k) + \left[a_i^{m,1}(k) + q_i^{m,1}(k-1)\right] \times r_{i,j}^{m,1}(k) \times p_{i,j}^{m,1}(k). \tag{16}$$

(3) $d_{l,l}^{m,1}(k)$ represents the deceleration delay caused by an unit vehicle which moves in independent lane l of Subsystem 1 on link m at time step k, and is given by

$$d_{l,l}^{m,1}(k) = \mathrm{Min}\left\{\left[\frac{1}{u_l^{m,1}(k)} - \frac{1}{u_l^{m,1}(0)}\right] \times e_l^{m,1}, t\right\}, \tag{17}$$

where $e_l^{m,1}$ is the length of either Subsystem 1 in independent lane l on link m with a lane blockage or the detector spacing in independent lane l on incident-free link m; $u_l^{m,1}(k)$ corresponds to the speed detected in independent lane l on link m via the upstream detector station at time step k, and $u_l^{m,1}(0)$ represents the observed highest speed associated with $u_l^{m,1}(k)$ under incident-free conditions.

(4) $d_{j,j}^{m,2}(k)$ represents the acceleration-or-deceleration delay which may occur in adjacent lane j of Subsystem 2 on link m at time step k, and is given by

$$d_{j,j}^{m,2}(k) = \mathrm{Min}\left\{\frac{\mathrm{Max}\{u_j^{m,2}(k), \upsilon_j^{m,2}(k)\} \times t}{\mathrm{Min}\{u_j^{m,2}(k), \upsilon_j^{m,2}(k)\}} - t, t\right\}, \tag{18}$$

where $\upsilon_j^{m,2}(k)$ is the speed in adjacent lane j on link m at time step k measured from the downstream detector station.

(5) $d_{l,l}^{m,2}(k)$ represents the deceleration delay caused by an unit vehicle which moves in independent lane l of Subsystem 2 on link m at time step k, and is given by

$$d_{l,l}^{m,2}(k) = \mathrm{Min}\left\{\left[\frac{1}{\upsilon_l^{m,2}(k)} - \frac{1}{\upsilon_l^{m,2}(0)}\right] \times e_l^{m,2}, t\right\}, \tag{19}$$

where $e_l^{m,2}$ is the length of either Subsystem 2 in independent lane l on link m with a lane blockage or the detector spacing in independent lane l on incident-free link m; $\upsilon_l^{m,2}(k)$ is defined as the speed in independent lane l on link m measured by the downstream detector station at time step k, and $\upsilon_l^{m,2}(0)$ represents the observed highest speed associated with $\upsilon_l^{m,2}(k)$ under incident-free conditions.

Lane-changing delays denote the magnitude of the time-based incident impacts on inter-lane traffic movements, and in this study, two types of time-varying lane-changing delays are specified for the maneuvers of incident-induced mandatory lane changing in Subsystem 1 and discretionary lane changing in Subsystem 2, respectively. Their notations can be expressed as follows:

(1) $d_{i,j}^{m,1}(k)$ represents the time-varying lane-changing delay caused by an unit vehicle in an effort to complete mandatory lane changing from blocked lane i to adjacent lane j of Subsystem 1 on link m at time step k, and is given by

$$d_{i,j}^{m,1}(k) = \mathrm{Min}\left\{ \frac{\mathrm{Max}\{u_j^{m,2}(k), u_i^{m,1}(k)\} \times t}{\mathrm{Min}\{u_j^{m,2}(k), u_i^{m,1}(k)\}} - t + d_{\mathrm{mc}}, t \right\}, \tag{20}$$

where d_{mc} means the average time spent in conducting mandatory lane-changing behavior, which is predetermined in this study. In this study, d_{mc} is set to be 3 s, according to the calibration results of our previous research [20]. Nevertheless, it is suggested that such a value be treated as time-varying, a topic which warrants more investigations in future research.

(2) $d_{j,i}^{m,2}(k)$ represents the time-varying lane-changing delay caused by an unit vehicle in an effort to complete discretionary lane changing from adjacent lane j to blocked lane i of Subsystem 2 on link m at time step k, and is given by

$$d_{j,i}^{m,2}(k) = \mathrm{Min}\left\{ \frac{\mathrm{Max}\{\upsilon_i^{m,2}(k), u_j^{m,2}(k)\} \times t}{\mathrm{Min}\{\upsilon_i^{m,2}(k), u_j^{m,2}(k)\}} - t + d_{\mathrm{dc}}, t \right\}, \tag{21}$$

where d_{dc} represents the average time spent in conducting discretionary lane changing, and is predetermined in this study; and $\upsilon_i^{m,2}(k)$ corresponds to the speed detected in blocked lane i on link m via the downstream detector station at time step k.

Note that the specified disaggregated delay variables are used in the delay-aggregation equations of the proposed stochastic model (see Section 3) to update the time-based incident impacts for each time step in the proposed recursive estimation algorithm.

3. Stochastic modeling

Using the time-varying traffic states specified above, a discrete-time nonlinear stochastic model is proposed to formulate the incident-induced traffic congestion problems on surface streets. The proposed stochastic model is composed of four types of equations: (1) recursive equations, (2) measurement equations, (3) delay-aggregation equations, and (4) boundary constraints. Their mathematical forms are expressed below.

3.1. Recursive equations

The recursive equations indicate the relationships between the time-varying basic lane traffic states which are assumed to follow Gaussian–Markov processes in the discrete-time stochastic system. In this study, the generalized form of the recursive equations can be expressed as

$$X(k+1) = F[x(k-\tau), c(k), k-\tau] + L[x(k-\tau), c(k), k-\tau]W(k), \tag{22}$$

where $X(k+1)$ is a $\left[\sum_{m=1}^{M} \sum_{s=1}^{S} (3n_j^{m,s} + n_l^{m,s})\right] \times 1$ time-varying vector of basic lane traffic states at time step $k+1$; $n_j^{m,s}$ and $n_l^{m,s}$ correspond to the number of lanes adjacent to blocked lane i and the number of independent lanes in Subsystem s on link m, respectively; M is the total number of links connecting to the targeted intersection; S is the number of Subsystems (i.e., $S = 1$ in case of intersection incidents and $S = 2$ in case of arterial incidents); $F[x(k-\tau), c(k), k-\tau]$ represents a $\left[\sum_{m=1}^{M}\sum_{s=1}^{S}(3n_j^{m,s} + n_l^{m,s})\right] \times 1$ time-varying vector of basic lane traffic states at time step k; $c(k)$ represents a signal-effect variable for the case of intersections, and is constant in case of arterial incidents; τ is referred to as a time-lag index used to estimate basic lane traffic states during the periods of phase switching in case of intersection incidents; $L[x(k-\tau), c(k), k-\tau]$ corresponds to a $\left[\sum_{m=1}^{M}\sum_{s=1}^{S}(3n_j^{m,s} + n_l^{m,s})\right] \times \left[\sum_{m=1}^{M}\sum_{s=1}^{S}(3n_j^{m,s} + n_l^{m,s})\right]$ state-dependent noise matrix; and $W(k)$ corresponds to a $\left[\sum_{m=1}^{M}\sum_{s=1}^{S}(3n_j^{m,s} + n_l^{m,s})\right] \times 1$ state-independent noise vector following Gaussian Processes. The notations of $X(k+1)$, $F[x(k-\tau), c(k), k-\tau]$, $L[x(k-\tau), c(k), k-\tau]$, and $W(k)$ are explained as follows.

In Eq. (22), $X(k+1)$ and $F[x(k-\tau), c(k), k-\tau]$ indicate the relationships between the next-time-step and current-time-step basic lane traffic states in a deterministic system, which is exclusive of noise terms, and can be expressed as

$$X(k+1) = \begin{bmatrix} p_{i,j}^{m,1}(k+1) \\ r_{j,j}^{m,1}(k+1) \\ r_{l,l}^{m,1}(k+1) \\ r_{i,j}^{m,1}(k+1) \\ ------- \\ p_{j,i}^{m,2}(k+1) \\ r_{j,j}^{m,2}(k+1) \\ r_{l,l}^{m,2}(k+1) \\ r_{j,i}^{m,2}(k+1) \end{bmatrix}_{m=1,2,\ldots,M}, \tag{23}$$

$$F[x(k-\tau), c(k), k-\tau] = \begin{bmatrix} p_{i,j}^{m,1}(k-\tau) \\ c_j^{\lambda}(k) r_{j,j}^{m,1}(k-\tau) \\ c_l^{\lambda}(k) r_{l,l}^{m,1}(k-\tau) \\ c_j^{\lambda}(k) r_{i,j}^{m,1}(k-\tau) \\ ---------- \\ p_{j,i}^{m,2}(k-\tau) \\ c_j^{\lambda}(k) r_{j,j}^{m,2}(k-\tau) \\ c_l^{\lambda}(k) r_{l,l}^{m,2}(k-\tau) \\ c_j^{\lambda}(k) r_{j,i}^{m,2}(k-\tau) \end{bmatrix}_{m=1,2,\ldots,M}, \tag{24}$$

where $c_j^{\lambda}(k)$ and $c_l^{\lambda}(k)$ represent the signal-effect variables associated with adjacent lane j and independent l at time step k under the condition of phase λ, respectively. Note that the aforementioned signal-effect variables $(c(k))$ and time-lag index τ are specified particularly to deal with the issue of discontinuity of system state estimation in a signal transition step where the traffic control signal turns either from GREEN to RED or from RED to GREEN in the case of intersection incidents. In $F[x(k-\tau), c(k), k-\tau]$, the time-varying values of $c(k)$ and τ are determined in one of the following four conditions:

Condition 1. If time step $k+1$ is a signal transition step during which the traffic signal turns from GREEN to RED (see Fig. 3), then $\tau = 0$ and $c(k) = \frac{G(k+1)}{t}$, where $G(k+1)$ is the length of a green interval at time step $k+1$.

Condition 2. If time step $k+1$ is a signal transition step during which the traffic signal turns from RED to GREEN (see Fig. 4), then $\tau = \tau_1 - 1$ and $c(k) = \frac{G(k+1)}{t}$, where τ_1 is the time lag corresponding to the number of time steps between the transition step (RED to GREEN) and the full-green time step one step prior to the last transition step (GREEN to RED), counting this previous full-green time step (e.g., $\tau_1 = 5$ in Fig. 4).

Condition 3. If time step $k+1$ is the first full-green time step of the current green interval (see Fig. 5), then $\tau = \tau_1$ and $c(k) = 1$.

Condition 4. For all other time steps, $\tau = 0$ and $c(k) = 1$.

In addition, $L[x(k-\tau), c(k), k-\tau]$ and $W(k)$ depicted in Eq. (22) form the state-dependent and state-independent noise terms of the stochastic model, respectively, and their notations are given, respectively, by

$$L[x(k-\tau), c(k), k-\tau] = \begin{bmatrix} \zeta_{11}(k) & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & \zeta_{22}(k) & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & \zeta_{33}(k) & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & \zeta_{44}(k) & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & \zeta_{55}(k) & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & \zeta_{66}(k) & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \zeta_{77}(k) & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \zeta_{88}(k) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}_{m=1,2,\ldots,M}, \quad (25)$$

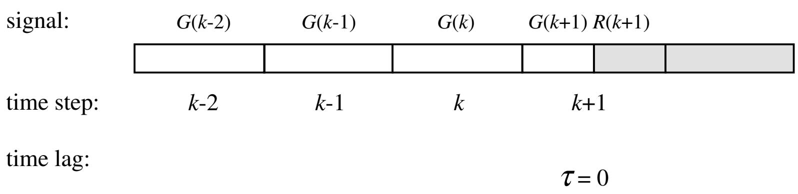

Fig. 3. Signal condition (1).

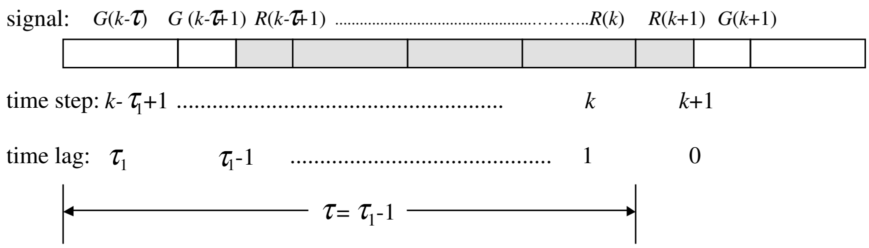

Fig. 4. Signal condition (2).

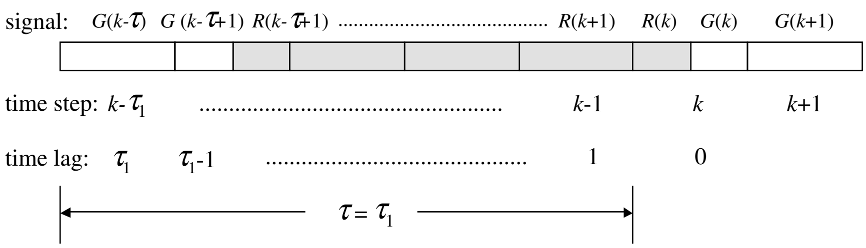

Fig. 5. Signal condition (3).

$$W(k) = \begin{bmatrix} w_{p_{i,j}^{m,1}}(k) \\ w_{r_{j,j}^{m,1}}(k) \\ w_{r_{l,l}^{m,1}}(k) \\ w_{r_{i,j}^{m,1}}(k) \\ w_{p_{j,i}^{m,2}}(k) \\ w_{r_{j,j}^{m,2}}(k) \\ w_{r_{l,l}^{m,2}}(k) \\ w_{r_{j,i}^{m,2}}(k) \\ \vdots \end{bmatrix}_{m=1,2,\ldots,M}, \tag{26}$$

where the diagonal elements of $L[x(k-\tau), c(k), k-\tau]$ can be further expressed as:

$$\zeta_{11}(k) = \left[1 - \sum_{\forall j \in J} p_{i,j}^{m,1}(k-\tau)\right] \times c_j^{\lambda}(k) \times r_{i,j}^{m,1}(k-\tau), \tag{27}$$

$$\zeta_{22}(k) = \left[1 - \sum_{\forall j \in J} p_{i,j}^{m,1}(k-\tau)\right] \times p_{i,j}^{m,1}(k-\tau) + \left[1 - c_j^{\lambda}(k) \times r_{j,j}^{m,1}(k-\tau)\right], \tag{28}$$

$$\zeta_{33}(k) = 1 - c_l^{\lambda}(k) \times r_{l,l}^{m,1}(k-\tau), \tag{29}$$

$$\zeta_{44}(k) = \left[1 - \sum_{\forall j \in J} p_{i,j}^{m,1}(k-\tau)\right] \times p_{i,j}^{m,1}(k-\tau) + \left[1 - c_j^{\lambda}(k) \times r_{i,j}^{m,1}(k-\tau)\right], \tag{30}$$

$$\zeta_{55}(k) = \left[1 - p_{j,i}^{m,2}(k-\tau)\right] \times c_i^{\lambda}(k) \times r_{j,i}^{m,2}(k-\tau), \tag{31}$$

$$\zeta_{66}(k) = \left[1 - p_{j,i}^{m,2}(k-\tau)\right] + \left[1 - c_j^{\lambda}(k) \times r_{j,j}^{m,2}(k-\tau)\right], \tag{32}$$

$$\zeta_{77}(k) = \left[1 - c_l^{\lambda}(k) \times r_{l,l}^{m,2}(k-\tau)\right], \tag{33}$$

$$\zeta_{88}(k) = \sum_{\forall j \in J} \left[1 - p_{j,i}^{m,2}(k-\tau)\right] \times p_{j,i}^{m,2}(k-\tau). \tag{34}$$

The state-dependent noise terms represented by $L[x(k-\tau), c(k), k-\tau]$ herein are specified to quantify the effects of the current-time-step lane traffic phenomena, such as lane changing and queuing, on the stability of the next-time-step system states. The magnitude of these state-dependent noise terms, on the other hand, relies on the changes in lane traffic arrivals, which are assumed to follow Gaussian processes, and thus, the state-independent noise terms $W(k)$ are included.

3.2. Measurement equations

The measurement equations serve primarily to update in real time the pre-predictions of the basic lane traffic states, which are generated merely on the basis of the aforementioned recursive equations. In the stochastic model, these measurement equations denote the time-varying relationships between the lane traffic counts measured from the detector stations and the basic lane traffic states. Their generalized form can be expressed as

$$Z(k+1) = H[x(k+1), k+1] + v(k+1), \tag{35}$$

where $Z(k+1)$, $H[x(k+1), k+1]$, and $v(k+1)$ are $\left[\sum_{m=1}^{M} n_i^m + n_j^m + n_l^m\right] \times 1$ time-varying vectors; n_i^m, n_j^m, and n_l^m represent the numbers of blocked lanes, adjacent lanes and independent lanes on link m, respectively. In Eq. (35), the elements of $Z(k+1)$ correspond to the time-varying lane traffic counts in blocked lanes, adjacent lanes, and independent lanes collected from the downstream detector stations on each link. In the stochastic model, they can be further characterized by the elements of $H[x(k+1), k+1]$, which decompose the measured downstream lane traffic counts into the basic lane traffic states and the collected lane traffic arrivals; and $v(k+1)$ represents the error terms of the time-varying lane traffic counts, which are assumed to follow Gaussian processes. $Z(k+1)$, $H[x(k+1), k+1]$, and $v(k+1)$ can be further expressed as follows

$$Z(k+1) = \begin{bmatrix} z_i^m(k+1) \\ z_j^m(k+1) \\ z_l^m(k+1) \\ \vdots \end{bmatrix}_{m=1,2,\ldots,M}, \tag{36}$$

$$H[x(k+1), k+1] = \begin{bmatrix} h_i^m(k+1) \\ h_j^m(k+1) \\ h_l^m(k+1) \\ \vdots \end{bmatrix}_{m=1,2,\ldots,M}, \tag{37}$$

$$v(k+1) = \begin{bmatrix} v_i^m(k+1) \\ v_j^m(k+1) \\ v_l^m(k+1) \\ \vdots \end{bmatrix}_{m=1,2,\ldots,M}, \tag{38}$$

where $z_i^m(k+1)$, $z_j^m(k+1)$, and $z_l^m(k+1)$ are referred to as the downstream lane traffic counts in blocked lane i, adjacent lane j, and independent lane l on link m at time step $k+1$, respectively; $v_i^m(k+1)$, $v_j^m(k+1)$, and $v_l^m(k+1)$ correspond to the Gaussian error terms associated with the collected lane traffic counts in blocked lane i, adjacent lane j, and independent lane l on link m at time step $k+1$, respectively; $h_i^m(k+1)$, $h_j^m(k+1)$, and $h_l^m(k+1)$ are the components of $z_i^m(k+1)$, $z_j^m(k+1)$, and $z_l^m(k+1)$, respectively. Here, $h_i^m(k+1)$ is set to be zero for the cases of intersection incidents; otherwise, $h_i^m(k+1)$, $h_j^m(k+1)$, and $h_l^m(k+1)$ are given by

$$h_i^m(k+1) = \sum_{\forall j \in J} \left\{ \left[\left(a_j^{m,1}(k+1) + \delta_j^{m,1}(k) \right) \times r_{j,j}^{m,1}(k+1) + \delta_j^{m,2}(k) \right] \times p_{j,i}^{m,2}(k+1) + \frac{\delta_i^{m,2}(k)}{n_j^{m,2}} \right\} \times r_{j,i}^{m,2}(k+1), \tag{39}$$

$$h_j^m(k+1) = \left\{ \left[\left(a_i^{m,1}(k+1) + q_i^{m,1}(k)\right) \times p_{i,j}^{m,1}(k+1) \times r_{i,j}^{m,1}(k+1) + \left(a_j^{m,1}(k+1) + \delta_j^{m,1}(k) \right) \times r_{j,j}^{m,1}(k+1) \right] \times (1 - p_{j,i}^{m,2}(k+1)) \right\} \times r_{j,j}^{m,2}(k+1), \tag{40}$$

$$h_l^m(k+1) = \left[\left(a_l^{m,1}(k+1) + \delta_l^{m,1}(k)\right) \times r_{l,l}^{m,1}(k+1) + \delta_l^{m,2}(k) \right] \times r_{l,l}^{m,2}(k+1). \tag{41}$$

3.3. Delay-aggregation equations

The delay-aggregation equations are used to calculate the diverse time-varying delays during incidents employing the estimates of basic lane traffic states. The generalized form of the delay-aggregation equations is given by

$$D(k+1) = G[x(k+1), k+1]Y(k+1). \tag{42}$$

In Eq. (42), $D(k+1)$ and $Y(k+1)$ are either $\left[\sum_{m=1}^{M} 3n_i^{m,1} + 3n_j^{m,1} + 2n_l^{m,1} + 2n_j^{m,2} + n_l^{m,2}\right] \times 1$ vectors for the cases of arterial incidents, or $\left[\sum_{m=1}^{M} 3n_i^{m,1} + 3n_j^{m,1} + 2n_l^{m,1}\right] \times 1$ vectors for the cases of intersection incidents where each element of $D(k+1)$ indicates the aggregate amount associated with a specific time-varying delay variable shown in Vector $Y(k+1)$; and similarly, $G[x(k+1),\ k+1]$ is referred to as a $\left[\sum_{m=1}^{M} 3n_i^{m,1} + 3n_j^{m,1} + 2n_l^{m,1} + 2n_j^{m,2} + n_l^{m,2}\right] \times \left[\sum_{m=1}^{M} 3n_i^{m,1} + 3n_j^{m,1} + 2n_l^{m,1} + 2n_j^{m,2} + n_l^{m,2}\right]$ time-varying traffic matrix in case of arterial incidents; otherwise, a $\left[\sum_{m=1}^{M} 3n_i^{m,1} + 3n_j^{m,1} + 2n_l^{m,1}\right] \times \left[\sum_{m=1}^{M} 3n_i^{m,1} + 3n_j^{m,1} + 2n_l^{m,1}\right]$ matrix where each element of $G[x(k+1), k+1]$ represents the number of vehicles associated with a specific type of delay. $D(k+1)$, $G[x(k+1), k+1]$ and $Y(k+1)$ can be further denoted as

$$D(k+1)=\begin{bmatrix} \widetilde{D}_i^{m,1}(k+1) \\ \widetilde{\Psi}_i^{m,1}(k+1) \\ \widetilde{D}_{i,i}^{m,1}(k+1) \\ \widetilde{\Psi}_j^{m,1}(k+1) \\ \widetilde{D}_{j,j}^{m,1}(k+1) \\ \widetilde{D}_{i,j}^{m,1}(k+1) \\ \widetilde{\Psi}_l^{m,1}(k+1) \\ \widetilde{D}_{l,l}^{m,1}(k+1) \\ ----- \\ \widetilde{D}_{j,j}^{m,2}(k+1) \\ \widetilde{D}_{j,i}^{m,2}(k+1) \\ \widetilde{D}_{l,l}^{m,2}(k+1) \\ \vdots \end{bmatrix}_{m=1,2,\ldots,M}, \tag{43}$$

$$G[x(k+1),k+1]=\begin{bmatrix} g_{11}^m(k+1) & 0 & 0 & 0 & \cdots \\ 0 & g_{22}^m(k+1) & 0 & 0 & \cdots \\ 0 & 0 & \ddots & 0 & \cdots \\ 0 & 0 & 0 & g_{11,11}^m(k+1) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}_{m=1,2,\ldots,M}, \tag{44}$$

$$Y(k+1)=\begin{bmatrix} c_i^{\lambda}(k)d_i^{m,1}(k+1) \\ \Psi_i^{m,1}(k+1) \\ c_i^{\lambda}(k)d_{i,i}^{m,1}(k+1) \\ \Psi_j^{m,1}(k+1) \\ c_j^{\lambda}(k)d_{j,j}^{m,1}(k+1) \\ c_j^{\lambda}(k)d_{i,j}^{m,1}(k+1) \\ \Psi_l^{m,1}(k+1) \\ c_l^{\lambda}(k)d_{l,l}^{m,1}(k+1) \\ ---------- \\ d_{j,j}^{m,2}(k+1) \\ d_{j,i}^{m,2}(k+1) \\ d_{l,l}^{m,2}(k+1) \\ \vdots \end{bmatrix}_{m=1,2,\ldots,M}, \tag{45}$$

where the diagonal elements of $G[x(k+1), k+1]$ shown in Eq. (44) are given by

$$g_{11}^m(k+1) = \left[1 - \sum_{\forall j \in J} p_{i,j}^{m,1}(k+1)\right] \times q_i^{m,1}(k), \tag{46}$$

$$g_{22}^m(k+1) = \left[a_i^{m,1}(k+1) + q_i^{m,1}(k)\right] \times \left[1 - \sum_{\forall j \in J} p_{i,j}^{m,1}(k+1)\right], \tag{47}$$

$$g_{33}^m(k+1) = \left[1 - \sum_{\forall j \in J} p_{i,j}^{m,1}(k+1)\right] \times a_i^{m,1}(k+1), \tag{48}$$

$$g_{44}^m(k+1) = \left[a_j^{m,1}(k+1) + q_j^{m,1}(k)\right] + \left[a_i^{m,1}(k+1) + q_i^{m,1}(k)\right] \times p_{i,j}^{m,1}(k+1), \tag{49}$$

$$g_{55}^m(k+1) = \left[a_j^{m,1}(k+1) + \delta_j^{m,1}(k)\right] \times \left[1 - r_{j,j}^{m,1}(k+1)\right], \tag{50}$$

$$g_{66}^m(k+1) = \left[a_i^{m,1}(k+1) + q_i^{m,1}(k)\right] \times p_{i,j}^{m,1}(k+1), \tag{51}$$

$$g_{77}^m(k+1) = a_l^{m,1}(k+1) + q_l^{m,1}(k), \tag{52}$$

$$g_{88}^m(k+1) = \left[a_l^{m,1}(k+1) + \delta_l^{m,1}(k)\right] \times \left[1 - r_{l,l}^{m,1}(k+1)\right], \tag{53}$$

$$g_{99}^m(k+1) = \Big\{\left[a_i^{m,1}(k+1) + q_i^{m,1}(k)\right] \times p_{i,j}^{m,1}(k+1) \times r_{i,j}^{m,1}(k+1) + \Big[\Big(a_j^{m,1}(k+1) + \delta_j^{m,1}(k)\Big) \times r_{j,j}^{m,1}(k+1) + \delta_j^{m,2}(k)\Big] \times \left[1 - p_{j,i}^{m,2}(k+1)\right]\Big\} \times \left[1 - r_{j,j}^{m,2}(k+1)\right], \tag{54}$$

$$g_{10,10}^m(k+1) = \Big[\Big(a_j^{m,1}(k+1) + \delta_j^{m,1}(k)\Big) \times r_{j,j}^{m,1}(k+1) + \delta_j^{m,2}(k)\Big] \times p_{j,i}^{m,2}(k+1), \tag{55}$$

$$g_{11,11}^m(k+1) = \Big\{\left[a_l^{m,1}(k+1) + \delta_l^{m,1}(k)\right] \times r_{l,l}^{m,1}(k+1) + \delta_l^{m,2}(k)\Big\} \times \left[1 - r_{l,l}^{m,2}(k+1)\right]. \tag{56}$$

3.4. Boundary constraints

The boundary constraints are specified in consideration of the lower and upper bounds of the time-varying basic lane traffic states $(x(k+1))$ in the recursive estimation procedure, so we have

$$0 \leqslant \forall x(k+1) \leqslant 1. \tag{57}$$

4. Recursive estimation

To estimate in real time the time-varying traffic variables of the model, we propose a recursive estimation algorithm. In addition to fundamentals of an extended Kalman filter, the proposed algorithm involves several primary computational steps including the procedures of time-based and space-based incident impact prediction. Furthermore, the effect of signal control is considered in processing the estimation of lane traffic states for the cases of intersection incidents. The sequence of the major computational steps is summarized below:

Step 0. Given $k = 0$, initialize state variables and the covariance matrix of the state estimation error.
Step 1. Check the status of the signal phase at the current time step to determine the signal-effect variable $c(k)$ for the cases of intersection incidents. Note that in cases of arterial incidents, the basic lane traffic states are assumed to be unaffected by the signal control, so Step 1 can be ignored.
Step 2. Compute pre-predictions of basic lane traffic states $(x(k+1|k))$ and the covariance matrix of the state estimation error utilizing the recursive equations (see Eq. (22)).
Step 3. Calculate the time-varying Kalman gain.
Step 4. Update the prior estimates of basic lane traffic states $(x(k+1|k+1))$ using the measurement equations (see Eq. (35)) together with the current-time-step raw traffic data.
Step 5. Truncate and normalize the estimates of basic lane states with boundary constraints (see Eq. (57)).
Step 6. Update the covariance matrix of the state estimation error.
Step 7. Estimate time-varying lane traffic loads and queue lengths using the updated basic lane traffic states to characterize the space-based incident impacts on traffic congestion.
Step 8. Compute aggregated delays via the delay-aggregation equations (see Eq. (42)) to characterize the time-based incident impacts on traffic congestion.
Step 9. Check incident status to determine whether or not the recursive estimation continues. If the incident is removed, then stop the estimation. Otherwise, let $k = k + 1$; input the next-time-step raw traffic data, and then go back either to Step 1 in the case of intersection incidents or to Step 2 in the case of arterial incidents to continue to the next-time-step state estimation.

5. Numerical results

The preliminary tests primarily serve to demonstrate the capability of the proposed method in terms of real-time estimation of time-varying mandatory lane-changing fractions and queue lengths which are regarded as two key elements in characterizing incident-induced inter-lane and intra-lane traffic states for further use to derive incident effects on traffic congestion in the study. Data acquisition procedures and preliminary tests, as well as generalizations of the test results are summarized below.

Due to the difficulty in collection of enough real incident-related traffic data for diverse incident cases on surface streets, simulation data generated from Paramics, which is a microscopic traffic simulator, was used in the preliminary tests. Paramics was calibrated prior to this study. Preliminary tasks conducted for qualitatively and quantitatively evaluating the Paramics simulator can also be found in our previous research [21], which details the reasons for using the Paramics simulator.

To simulate diverse lane-blocking incidents on surface streets, a small traffic network consisting of five intersections was built using Paramics. Fig. 6 illustrates the scheme of the study network, where each intersection represented by a specific node was coded with an integer value for its identification. Arterial lane-blocking incidents were strategically generated on the 3-lane link between nodes 1 and 3, and lane blockages close to the stop lines were generated to simulate intersection incidents.

Thirty-six types of lane-blocking incidents were simulated in total, and each event was associated with different incident position on the link (e.g., upstream, middle-stream, and downstream), the lane blocked (e.g., inside lane, central lane, and outside lane), and traffic flow conditions (e.g., high-volume, medium-volume, and low-volume). Out of the 36 lane-blocking incidents, 27 simulated incidents were located on the main segment of the link, and the rest situated at the approach. These simulated events were set to be 30 min for each one: the first 5 min for warming up, the next 20 min for incident duration, and the rest for incident removing. The output data simulated from Paramics, including lane traffic counts, lane-changing fractions, and queue lengths were collected at each 10-s time step during any given 20-min incident event, and thus, there were a total of 4320 data samples associated with each time-varying lane traffic variable used in the tests.

The model tests compare the time-varying estimates of incident-induced lane traffic states, including mandatory lane-changing fractions and queue lengths upstream to the incident site to the simulation data under different incident circumstances. Under different conditions of lane-blocking incidents, the estimates of time-varying lane-changing fractions and queue lengths were generated using the proposed method, and then compared with the simulation data via two measures: (1) mean absolute percentage error (MAPE), and (2) patterns of root mean square estimation errors.

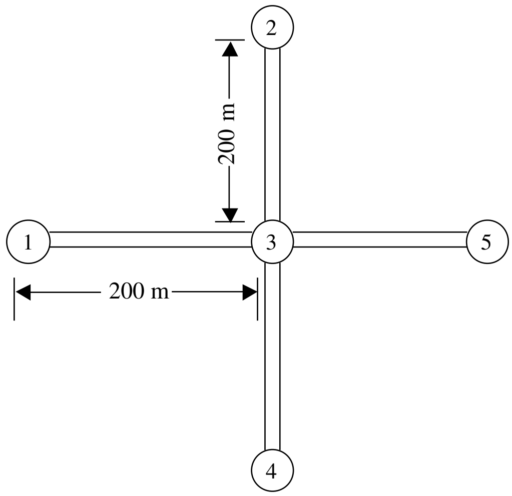

Fig. 6. The scheme of the simulation network.

The statistics of MAPE are given by

$$\mathrm{MAPE} = \begin{cases} \dfrac{\sum_{k=1}^{K} \left| \frac{x_\theta(k) - \tilde{x}_\theta(k)}{x_\theta(k)} \right|}{N_\theta} \times 100\%, & \text{given } x_\theta(k) \neq 0, \\ \dfrac{\sum_{k=1}^{K} \left| \frac{x_\theta(k) - \tilde{x}_\theta(k)}{1 - x_\theta(k)} \right|}{N_\theta} \times 100\%, & \text{otherwise,} \end{cases} \tag{58}$$

where N_θ is the data sample size for a specific incident case θ; $x_\theta(k)$ and $\tilde{x}_\theta(k)$ represent the samples of a time-varying state estimate computed from the proposed method and its simulated value generated via Paramics, respectively; K is the incident duration represented with the time-step unit. The results of the MAPE tests with respect to the estimates of lane changing and queuing are summarized in Table 1.

Table 1 illustrates the results of case-by-case MAPE-value tests in terms of the estimates of lane-changing fractions and queue lengths in blocked lanes with several generalizations, as follows. First, the estimates of mandatory lane-changing fractions and queue lengths, generally, fit with the simulation data. Out of the 72 MAPE estimates, 11 ones are less than 10%, 42 ones fall within the range of 10% and 20%, and the rest are greater than 20%. Correspondingly, 73.6% of the MAPE measures with respect to the state estimates satisfy the threshold of 0.2, which is frequently used as a reasonable criterion for the evaluation of MAPE-based tests. Secondly, the proposed method may suffer from slightly higher prediction bias under medium-volume incident cases compared with the other incident cases. According to our observations from the simulated incident events, this may result from the unstable changing patterns of lane-changing fractions under medium-volume incident cases. It is noteworthy that the lane-changing fractions tend to remain steadily low under high-volume conditions, and consistently high under low-volume

Table 1
Results of the MAPE-value (%) tests

Spatial characteristics of incidents	Traffic flow conditions					
	Low volume ($\leqslant$ 700 vph)		Medium volume (700 vph < $\leqslant$ 1100 vph)		High volume (1100 vph<)	
	Lane changing	Queue	Lane changing	Queue	Lane changing	Queue
Arterial incidents						
Inside-lane, upstream	26.6	24.1	27.1	19.8	24.3	16.3
Inside-lane, mid-stream	15.2	13.2	24.3	15.1	26.5	13.9
Inside-lane, downstream	25.9	17.8	24.1	23.5	20.2	19.1
Central-lane, upstream	16.7	12.6	26.8	21.4	27.5	16.4
Central-lane, mid-stream	9.8	13.3	16.3	18.2	17.3	19.8
Central-lane, downstream	14.9	11.5	18.7	23.6	19.5	18.3
Outside-lane, upstream	12.8	20.8	15.9	16.8	21.6	11.7
Outside-lane, mid-stream	13.0	18.0	26.0	18.0	15.1	20.4
Outside-lane, downstream	14.2	12.6	19.6	17.9	13.2	19.5
Intersection incidents						
Inside-lane	12.8	20.6	15.7	18.4	19.7	18.8
Central-lane	8.2	6.5	6.3	8.9	5.8	9.0
Outside-lane	15.4	8.4	7.9	7.2	6.0	10.5

conditions. Therefore, utilization of the proposed stochastic model makes it easier to capture the patterns of the time-varying lane-changing fractions under either low-volume or high-volume incident conditions than under medium-volume conditions in the tests.

The second type of model testing examines the prediction stability of the proposed method in the temporal domain utilizing the time-varying index of root mean square estimation error ($\lambda(k)$). In the tests, $\lambda(k)$ is given by

$$\lambda(k) = \sqrt{\frac{\sum_{\iota=1}^{N_k}(x_\iota(k) - \tilde{x}_\iota(k))^2}{N_k}}, \tag{59}$$

where N_k is the data sample size at time step k; $x_\iota(k)$ and $\tilde{x}_\iota(k)$ represent the samples of state estimates computed from the proposed method and generated via Paramics, respectively, at time step k. The time-varying values of $\lambda(k)$ associated with lane-changing fractions and queue lengths are plotted in Figs. 7 and 8, respectively.

Figs. 7 and 8 indicate that the problem of estimation divergence may not exist in the tests primarily because the $\lambda(k)$ values shown in either Fig. 7 or Fig. 8 do not exhibit a tendency to increase in the temporal domain. The $\lambda(k)$ values associated with lane changing fractions shown in Fig. 7 are mostly lower than 0.4, and oscillate around the value of 0.3. Similarly, the $\lambda(k)$ values associated with queue lengths shown in Fig. 8 are lower than 4, and tend to be convergent. Therefore, the test results imply that the proposed method may not be overwhelmed by the problem of estimation divergence.

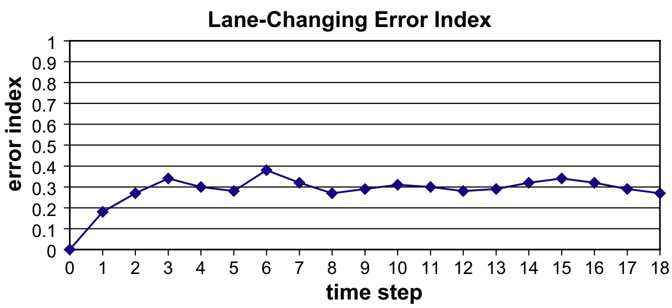

Fig. 7. Root mean square estimation error (time-varying lane-changing fractions).

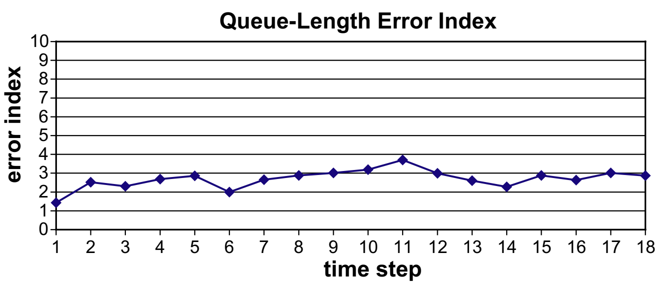

Fig. 8. Root mean square estimation error (time-varying queue lengths).

Overall, the preliminary test results in terms of real-time estimation of incident-induced inter-lane and intra-lane traffic states have demonstrated the applicability of the proposed approach for both arterial and intersection lane-blocking incident cases to further characterize incident effects on traffic congestion in real time.

6. Concluding remarks

A new approach proposed for formulating the incident-induced traffic congestion problems on surface streets, and for real-time estimation of the incident effects on lane traffic states has been presented. In order to characterize the inter-lane and intra-lane traffic maneuvers in the presence of surface street incidents, three groups of time-varying lane traffic variables were specified. After conducting this, we proposed a discrete-time nonlinear stochastic model and a recursive estimation algorithm to estimate the incident-induced lane traffic states as well as incident impacts including delays and queue lengths in real time. In the real-time estimation procedure, raw traffic data collected from pairs of point detectors were necessary to update the estimates of lane traffic variables.

Our preliminary test results indicated the acceptability of the estimated lane traffic variables generated from the proposed method, and also showed the applicability of the proposed method in generalizing one-lane-blocking incident cases on surface streets. Furthermore, the potential of the proposed approach in the development of related technologies such as real-time incident-responsive traffic management systems and automatic road congestion warning systems for further use in ATMIS is highly suggested on the basis of our generalizations. We should also mention that one of our related studies in terms of real-time incident-responsive traffic control at isolated intersections is underway, and this study can be regarded as a striking example of an application of the proposed method.

Nevertheless, tasks conducted to improve the robustness of the proposed method warrant further research. More complicated cases such as multi-lane-blocking incidents, queues spilling back to the upstream detectors, incidents occurring within an intersection etc. can be considered in model extension as well as in numerical experiment. In addition, efforts on either revising the published traffic simulators or developing specific simulation models for dealing appropriately with incident-induced traffic maneuvers under diverse incident conditions need immediate attention. According to our observations from the preliminary tests, it is inferred that a portion of the estimation errors may be related to the simulation data used in the tests. It is noteworthy that modeling of incident-induced lane-changing maneuvers may remain ambiguous in most of existing traffic simulators, according to our previous literature review [19].

References

[1] J.A. Lindley, Urban freeway congestion: quantification of the problem and effectiveness of potential solutions, J. ITE 57 (1987) 27–32.

[2] J.-B. Sheu, Y.-H. Chou, L.-J. Shen, A stochastic estimation approach to real-time prediction of incident effects on freeway traffic congestion, Transport. Res. B 35B (6) (2001) 575–592.

[3] C.J. Messer, C.L. Dudek, J.D. Friebele, Method for predicting travel time and other operational measures in real-time during freeway incident conditions, Highway Res. Rec. 461 (1973) 1–16.

[4] W.A. Chow, A study of traffic performance models under incident conditions, Highway Res. Rec. 567 (1974) 31–36.
[5] J.M. Morales, Analytical procedures for estimating freeway traffic congestion, Public Road 50 (2) (1986) 55–61.
[6] A. Skabardonis, K. Petty, H. Noeimi, D. Pydzewski, P.P. Varaiya, I-880 field experiment: data-base development and incident delay estimation procedures, Transport. Res. Rec. 1554 (1996) 204–212.
[7] E.C. Sullivan, New model for predicting freeway incidents and incident delays, J. Traffic Eng. 123 (4) (1997) 267–275.
[8] T. Olmstead, Pitfall to avoid when estimating incident-induced delay by using deterministic queuing model, Transport. Res. Rec. 1683 (1999) 38–46.
[9] H. Al-Deek, A. Garib, A.E. Radwan, Methods for estimating freeway incident congestion, in: Proceedings of the 74th Annual Meeting of the Transportation Research Board, Washington, DC, 1995.
[10] A. Garib, A.E. Radwan, H. Al-Deek, Estimating magnitude and duration of incident delays, J. Traffic Eng. 123 (6) (1997) 459–466.
[11] L. Fu, L. Rilett, Real-time estimation of incident delay in dynamic and stochastic networks, Transport. Res. Rec. 1603 (1997) 99–105.
[12] M. Szeto, D.C. Gazis, Application of Kalman filtering to the surveillance and control of traffic systems, Transport. Sci. 6 (1972) 419–439.
[13] H.J. Payne. Analysis and evaluation of estimators of traffic parameters: final report, part II. USCEE Report 463, University of California, Electronic Sciences Laboratory, 1973.
[14] M. Cremer, H. Schutt, A comprehensive concept for simultaneous state observation, parameter estimation and incident detection, in: M. Koshi (Ed.), Proceedings of the 11th International Symposium on Transportation and Traffic Theory, Elsevier, Amsterdam, 1990, pp. 95–111.
[15] P.G. Michalopoulos, R.D. Jacobson, Automatic incident detection through video image processing, Traffic Eng. Control 34 (2) (1993) 66–75.
[16] S. Sellam, A. Boulmakoul, Intelligent intersection: artificial intelligence and computer vision techniques for automatic incident detection, Artif. Intell. Appl. Traffic Eng. (1994) 189–200.
[17] J.N. Ivan, Neural network representations for arterial street incident detection data fusion, Transport. Res. C 5C (3/4) (1997) 245–254.
[18] S.I. Khan, S.G. Ritchie, Statistical and neural classifiers to detect traffic operational problems on urban arterials, Transport. Res. C 6C (5–6) (1998) 291–314.
[19] J.-B. Sheu, S.G. Ritchie, Stochastic modeling and real-time prediction of vehicular lane-changing behavior, Transport. Res. B 35B (7) (2000) 695–716.
[20] J.-B. Sheu, A prototype of incident-responsive signal control for isolated intersections, National Science Council Research Report Taiwan, NSC 90-2211-E-321-027, 2001.
[21] B. Abdulhai, J.-B. Sheu, W.W. Recker, Simulation of ITS on the Irvine FOT area using the 'Paramics 1.5' scalable microscopic traffic simulator: phase I: model calibration and validation, California PATH Research Report UCB-ITS-PRR-99-12, 1999, pp. 1055–1425.