Real-Time Estimation of Incident Delay in Dynamic and Stochastic Networks

LIPING FU AND LAURENCE R. RILETT

The ability to predict the link travel times is a necessary requirement for most intelligent transportation systems (ITS) applications such as route guidance systems. In an urban traffic environment, these travel times are dynamic and stochastic and should be modeled as such, especially during incident conditions. In contrast to traditional deterministic incident delay models, the model presented explicitly considers the stochastic attributes of incident duration. This new model can be used for predicting the delay that a vehicle would experience as it travels through nonrecurring congestion brought about by an incident. The model is operational in the sense that it does not require significant data and computational abilities beyond that which is traditionally used and can be used within traffic models or within actual ITS implementations. A mixed discrete and continuous vehicle-delay model is first derived and estimators of the mean and variance of vehicle delay are identified. A sensitivity analysis subsequently is performed, and a method for updating the estimated delay as new information becomes available is provided.

The estimation and prediction of link travel times in a road traffic network are critical for many intelligent transportation systems (ITS) applications such as route guidance systems (RGS) and freeway traffic management systems. The common objective of these systems is to provide information necessary to help individual drivers identify optimal routes based on real-time information on current traffic conditions. To identify these optimal routes, the travel times on links for the future time periods when the vehicle is expected to traverse the link are required. In an urban traffic environment these link travel times should be modeled as dynamic and stochastic, particularly during incident conditions.

Traditionally, traffic delay due to incidents, referred to in this paper as incident delay, is estimated using a deterministic queueing model that assumes that the traffic arrival rate, capacity reduction, and incident duration can be identified exactly. This approach may be adequate for after-incident evaluation, for which information on the traffic volume and incident situation is readily available; however, it is inappropriate for prediction of incident delay in real-time applications such as dynamic RGS because the only information typically available is the time when an incident occurs (or is detected), the current status of the incident (removed or not, reduced capacity), and the link demand volume. In real-time situations, the incident duration is unknown. Therefore, incident delay is best modeled by a random variable that represents the stochastic characteristics associated with the incident rather than by using a deterministic value. Another potential drawback to using a deterministic model of incident duration in real-time situations is that the variance of incident delay is ignored, although it is clearly significant for many ITS applications (1). The objective of this paper is to develop an incident delay estimation model that will consider exactly the randomness of incident duration.

An overview of existing methods for estimation and prediction of incident delay is presented first. A stochastic model is then developed to estimate the probability distribution of incident delay, from which the mean and variance of incident delay are derived. An example incident is created and used to demonstrate the performance of the new model. Sensitivity analyses of the estimation error of estimated incident delay and the variance of incident delay as a function of the incident duration variance also are performed. Finally, how real-time information can be incorporated into the estimation of incident delay is examined.

OVERVIEW OF INCIDENT DELAY ESTIMATION METHODS

Most of the existing incident delay estimation methods focus on total incident delay caused by incidents, for example, the methods of Chow (2), the Highway Capacity Manual (3), Wirasinghe (4), Messer et al. (5), and Al-Deek et al. (6). These methods are intended only for after-incident evaluation, and therefore information on the traffic volume and incident situation is assumed known.

Messer et al. (7) developed a method for predicting the travel time required to traverse a freeway segment that is experiencing incident congestion. The model was developed on the basis of shock-wave theory for use in the operation and control of variable message signs. In their research it was assumed that all of the inputs were known a priori, and consequently the models may be considered deterministic.

The rapid development of the ITS field in the last decade has spurred research in this area with various link travel-time estimation and prediction methods having been proposed for demonstration ITS projects and simulation studies. Hoffman and Janko (8) developed a link travel-time estimation and prediction method that has been used in the ALI-SCOUT system. In their approach the link travel time is predicted by scaling the historical travel time on the basis of current detected link travel time. Koutsopoulos and Xu (9) presented an approach based on information-discounting theory as an attempt to improve on Hoffman’s approach. In the ADVANCE project (10), the proposed link travel-time estimation method treats the incident-absent and incident-present situations separately and historical link travel-time is used for calculating vehicle routes.

It is important to note that all of these methods are fundamentally heuristic in that they tend to disregard the patterns that develop during incident congestion. For example, incident delay tends to have a build-up period and a decline period. Therefore, techniques....
that simply use a scaling factor may result in overestimation or underestimation of the actual link travel time, and consequently the corresponding RGS may provide suboptimal routes to their guided vehicles during the incident period.

Conversely, numerous simulation studies have been conducted to evaluate the potential benefits of RGS. One of the most essential components of these simulation models is the link travel-time model, which is used to estimate the link travel time for use in route calculation. Koutsopoulos and Yablonski (11) presented a theoretical link travel-time estimation model in which incident delay is estimated by a deterministic model. Although in their study the incident and its attributes (reduced capacity and incident duration) are randomly generated, this information is assumed to be known and is used directly for routing. Al-Deek and Kanafani (12) evaluated the benefits of an RGS specifically in the case of incident congestion. In their model, the incident situation also is assumed to be deterministic, and a deterministic queueing model is then used to estimate the queueing delay. All these simulation applications assume an advanced knowledge about the incident situation and then guide the vehicles on the basis of this information. However, in a real-time operational situation the evolution of an incident situation cannot be predicted exactly and the use of such extra information may result in overestimation of the benefits of an RGS. This research examines the situation in which an incident has been detected and develops a real-time estimation of the incident delay distribution pattern that a vehicle would experience if it were to travel through the incident location at some future time.

DYNAMIC AND STOCHASTIC INCIDENT DELAY MODEL

Assumptions and Notation

Many factors affect the delay a vehicle experiences as a result of an incident. These include incident severity (capacity reduction), incident duration, arrival pattern, traffic volume, and the future time when the vehicle arrives at the incident location. In a real-time environment all of these factors are random variables, which makes incident delay modeling a complex process.

The approach adopted in this paper is to model incident duration as a random variable within a traditional deterministic queueing model approach. A queueing diagram is presented in Figure 1 that shows the cumulative vehicle arrivals and vehicle departures before and after the incident is cleared. The incident occurs at time $T^\ast$ and lasts for $D^\ast$ amount of time. It is assumed that the traffic arrival rate, denoted as $q$ [passenger car units/hour (pcu/hr)], is constant and known. This value is represented by the slope of the cumulative arrival function. The nonincident capacity, denoted as $c$ (pcu/hr), and incident capacity, denoted as $c^\ast$ (pcu/hr), also are assumed known and constant. These are represented by the slope of the cumulative departure curve before the incident is cleared ($c^\ast$) and after the incident is cleared ($c$).

The incident duration, $D^\ast$, is modeled as a random variable with a known probability density function (PDF) and is denoted by $f_{D^\ast}(x)$. It is anticipated that PDF will be developed on the basis of historical data. For example, it has been found that incident duration follows a lognormal distribution (13, 14), and this fact will be explored further in a later section. It also is assumed that when the incident is removed and the link capacity returns to the nonincident value ($c$), the link travel-time prediction problem reverts to one that may be handled using traditional, deterministic methods. The variable $T_0$ presented in Figure 1 represents the current time $T_o$ or the time the estimation is made. For the present it will be assumed that $T_0$ and $T^\ast$ are equivalent. $T_1$ is the estimated time of arrival on the link and is assumed fixed in this analysis. The parameter $d_a$ represents the incident delay experienced by a vehicle that arrives at the incident spot at time $T_a$. It is a random variable and its derivation will be the focus of the following sections. Note that there are essentially two time periods of interest. The first is when the incident occurs and vehicles begin to form a queue at the rate of $q - c^\ast$. The second is when the incident clears and the queued vehicles dissipate at a rate of $c$.

Probability Distribution of Incident Delay

The probability distribution of the incident delay of a given vehicle ($d_a$) depends on the probability distribution pattern of the incident duration and the time the vehicle arrives at the link. The relationship between the two variables can be established by using standard
incident delay formulas. Essentially, when a vehicle arrives at the link at time $T_a$, three delay regimes may apply: The vehicle can proceed through without delay, it can experience a maximum delay, or it can experience a delay somewhere between these two extremes. These three regimes and their associated mathematical formulations will be discussed in the following sections.

No-Delay Regime

In the no-delay regime the vehicle arrives after the incident has been cleared and the associated queue has been dissipated. Consequently, the vehicle does not experience any delay. As indicated in Figure 1, $T_2$ is a random variable that represents the time when the incident is cleared and the associated queue has dissipated. Its value is a function of the incident duration, the arrival rate, the capacity, and the adjusted capacity. $T_2$ may be calculated using standard queuing theory as presented in Equation 1.

$$T_2 = T^* + \left( \frac{c - c^*}{c - q} \right) D^*$$

Note that $T_2$ is a random variable because the incident duration also is a random variable. The probability that the vehicle delay equals zero is simply the probability that $T_2$ is equal to or greater than $T_2$. The incident duration that makes the arrival time of an individual vehicle coincident with the time when the incident is cleared (i.e., $T_a = T_2$) is denoted by $D_1$. It can be derived from Equation 1 as follows:

$$D_1 = \frac{c - q}{c - c^*} (T_2 - T^*)$$

It is assumed that the arrival time is fixed and consequently that $D_1$ is a deterministic value as well. $D_1$ represents the longest incident duration that will result in zero delay to a vehicle that arrives at time $T_a$. The probability that the vehicle will have no delay is equivalent to the incident duration’s being less than $D_1$:

$$P(D_a = 0) = P(D^* \leq \frac{c - q}{c - c^*} (T_2 - T^*))$$

$$= P(D^* \leq D_1)$$

$$= P_1$$

$P_1$ may be calculated by calculating the appropriate area of the lower tail of the incident duration PDF:

$$P_1 = P(D^* < D_1) = \int_0^{D_1} f_{D^*}(x)dx$$

This is illustrated in Figure 2.

Fixed-Delay Regime

The second regime occurs when the vehicle arrives at the incident location at time $T_a$ where the incident queue (a) has not dissipated, and (b) will not be dissipated until after the vehicle traverses the link. In this situation the queue dissipates at a rate of $c^*$ and consequently the vehicle experiences the maximum delay (for a vehicle arriving at time $T_a$). $T_1$ represents the minimum time at which the maximum delay occurs. It can be represented as a function of the reduced capacity, arrival rate, and incident duration $D^*$:

$$T_1 = T^* + \left( \frac{c - c^*}{c - q} \right) D^*$$

$$D_2$$

represents the incident duration that makes the arrival time of an individual vehicle ($T_a$) coincident with the time when a maximum delay occurs (i.e., $T_a = T_1$). That is, if the incident duration is larger than $D_2$, then $T_1$ will be greater than $T_a$, and the vehicle will experience the maximum delay. The functional form of $D_2$ is

$$D_2 = \frac{q}{c^*} (T_a - T^*)$$

The parameter $d_a$ represents the maximum possible incident delay and is a function of (a) the reduced capacity, (b) the difference in time between the start of the incident and the time of arrival at the link, and (c) the queue dissipation rate. It may be calculated as follows:

$$d_a = \frac{q - c^*}{c^*} \cdot (T_a - T^*)$$

For a given $T_a$, if the incident duration is greater than $D_2$, the vehicle delay $D_a$ is equivalent to $d_a$. The probability that delay will be equal to $d_a$ is expressed mathematically as

$$P(D_a = d_a) = P(D^* \geq \frac{q}{c^*} (T_a - T^*))$$

$$= P(D^* \geq D_2)$$

$$= P_2$$

$P_2$ (Figure 2) is the probability that the incident duration is greater than $D_2$, and is represented by the upper tail of the incident duration PDF as follows:

$$P_2 = P(D^* > D_2) = \int_{D_2}^{\infty} f_{D^*}(x)dx$$

Variable-Delay Regime

The third regime occurs when the vehicle arrives at the link and either (a) the incident has been cleared but some portion of the queue remains or (b) the incident has not been cleared but will be cleared before the vehicle exits the link. In this situation, the dissipation rate
of the standing queue is some combination of \(c\) and \(c^*\), and consequently the queue delay will lie between the previous two cases—that is, between zero and the maximum delay \(d_m\). The delay also will be dependent on the incident duration and may be expressed mathematically as

\[
d_a = \left(\frac{c - c^*}{c}\right)D_a - \left(\frac{c - q}{c}\right)(T_a - T^*)
\]  

(10)

Equation 10 is valid for \(D_1 < D^* < D_2\) and may be written as the following inequality by using Equations 2 and 6:

\[
\left(\frac{c - q}{c - c^*}\right)(T_a - T^*) < D^* < \left(\frac{q}{c^*}\right)(T_a - T^*)
\]

(11)

The relationship between \(d_a\) and \(D^*\) is given by Equation 10 and therefore the PDF of \(d_a\) over the range zero to \(d_m\) may be calculated as

\[
f_{d_a}(x) = \left(\frac{c}{c - c^*}\right) f_{D_a}\left[\left(\frac{c}{c - c^*}\right)x + D_a\right]
\]

(12)

Note that to identify the probability of a given vehicle delay, the function in Equation 12 would have to be integrated over the appropriate range.

\(P_1\) is the probability that the incident duration is greater than \(D_1\) and less than \(D_2\); and may be calculated by subtracting the results of Equations 4 and 10 from 1.0 as follows:

\[
P_1 = 1 - P_2 - P_3
\]

(13)

Note that Equation 13 simply represents the probability that the delay experienced will be neither zero nor \(d_m\). Because the PDF of the incident duration \((D^*)\) is known, \(P_1\) also could have been calculated by integrating Equation 12 over the limits \(D_1\) and \(D_2\).

Mean and Variance of Incident Delay

It was demonstrated in the previous section that incident delay \(d_a\) may be modeled as a mixed discrete and continuous random variable. The two extreme mass points are calculated by using Equations 4 and 9, and the PDF is calculated by using Equation 12. Figure 3 schematically illustrates the distribution pattern of the incident delay. Obviously, the PDF of incident duration for a given application will be dependent on the PDF of the incident delay.

Although this derivation is useful in and of itself, for the majority of real applications it is the first moment, or mean, of the incident delay distribution that is more commonly used. The vehicle-routing algorithms in RGS tend to use mean values exclusively, although the use of a measure of dispersion (i.e., variance) has been advocated (15,16). To date, there have been no large-scale tests that have attempted to use the PDF distribution within the route-selection optimization process.

The general formula for the mathematical expectation of the discrete and continuous function \(d_a\) is

\[
E(d_a) = 0 \cdot P(d_a = 0) + \int_0^{d_m} f_{d_a}(x)dx + d_m \cdot P(d_a = d_m)
\]

(14)

Equation 14 may be transformed into Equation 15 where \(D_{12}\) is the conditional expectation of the incident duration and is defined in Equation 16. \(D_{12}\) is simply a truncated expectation of the incident duration between \(D_1\) and \(D_2\):

\[
E(d_a) = \frac{c - c^*}{c} (D_{12} - D_1P_1) + P_1d_m
\]

(15)

where

\[
D_{12} = \int_{D_1}^{D_2} x f_{D_a}(x)dx
\]

(16)

The variance may be calculated by

\[
\text{VAR}(d_a) = E(d^2_a) - [E(d_a)]^2
\]

(17)

By using Equations 3, 8, and 13, Equation 18 can be rewritten as

\[
E(d^2_a) = \left(\frac{c - c^*}{c}\right)^2 \left(V_{12} + D_1^2P_1 - 2D_1D_{12}\right) + P_1d_m^2
\]

(19)

\(V_{12}\) is the conditional expectation of the squared incident duration, which is defined as follows:

\[
V_{12} = \int_{D_1}^{D_2} x^2 f_{D_a}(x)dx
\]

(20)

EXAMPLE PROBLEM

Expected Incident Delay

Incident delay traditionally is estimated by using a deterministic model that assumes that the attributes of an incident (capacity reduction, duration) are known or can be estimated exactly. If the average incident duration used is \(\mu^*\), the incident delay \((D_a)\) is calculated by using Equation 21, which is simply the mathematical relationships developed in an earlier section with the random variable \(D^*\) replaced by its mean \(\mu^*\).

\[
D_a = \begin{cases} 
0 & \text{if } T_a \geq T_2 \\
\frac{q - c^*}{c^*} (T_a - T^*) & \text{if } T_1 < T_a < T_2 \\
\frac{c - c^*}{c} \mu^* - \frac{c - q}{c} (T_a - T^*) & \text{if } T_1 < T_a < T \end{cases}
\]

(21)
Apart from the fact that the deterministic model does not provide information on the incident variation, it also will generate a biased estimate of the mean incident delay as in Equations 14 and 19. This section provides a numerical example to demonstrate the estimation bias of the deterministic model and its sensitivity to the variation of the incident duration.

Assume a unidirectional, two-lane highway with capacity equal to 3,600 pcu/hr. An accident is detected on the road, which reduces the highway capacity to 1,800 pcu/hr. The average traffic volume under normal traffic-flow conditions during this time period is estimated to be 3,000 pcu/hr, among which approximately 500 pcu/hr are assumed to divert to other routes because of the incident. Furthermore, it is assumed that the incident duration is lognormally distributed with a mean incident duration of 30 min and that the standard deviation can range from 0 to 30 min.

The delay caused by the incident, calculated by using Equation 15, as a function of the arrival time at the link is illustrated in Figure 4 for different standard deviations. From Figure 4, it can be observed that, as expected, the mean delays estimated by both models are exactly the same if there is no variation in the incident duration. However, as the variation of the incident duration increases, the deterministic model may overestimate or underestimate the incident delay. For example, if the standard deviation of the incident duration is 30 min [Figure 4(c)], the deterministic model would overestimate the expected incident delay by approximately 50 percent for a vehicle that arrives at the link 20 min after the incident occurs. Conversely, if the vehicle arrives at the link 40 min after the incident, the deterministic model would underestimate the expected incident delay by approximately 50 percent. In addition, the deterministic model would predict no delay for a vehicle arriving at 50 min. This result clearly would be undesirable for an actual RGS.

Figure 5 illustrates the relationship between the maximum overestimation error and underestimation error as a function of the standard deviation of the incident duration. The estimation error is defined as the ratio of the difference in the expected delays estimated by the deterministic and stochastic models to the expected delay by the stochastic model. As given in Figure 5, the estimation error is proportional (and approximately linear) to the standard deviation of the incident duration.

**Variance of Incident Delay**

As with the expected length of incident delay, the variance of incident delay is dependent on the standard deviation of the incident duration. Figure 6 illustrates the standard deviation of the incident delay as a function of the arrival time at the incident location under different standard deviations of the incident duration. The expected length of the incident delay also is given in Figure 6. As expected, the larger the variation of the incident duration, the larger the variance of the incident delay.

From Figure 6 it can be seen that there is a large amount of variation in the average incident delay and that this variation is more significant for the trips arriving after the expected incident duration time. For example, when the standard deviation of the incident duration is greater than 15 min, the coefficient of variation of the incident delay for trips arriving after the expected incident duration (30 min) has values larger than 2.0.

Another fact that can be observed from Figure 5 is that, although the expected incident delay is small when a vehicle arrives around the expected incident-clearance period, the variation of the incident delay could be very large. As illustrated in Figure 6(c), when the arrival time is 80 min after the occurrence of the incident, the expected delay is approximately 2 min and the standard deviation about this estimate is approximately 6 min. One implication of this is that a routing decision based on the average travel time would provide a route with a higher order of risk of being delayed.

**INCIDENT DELAY ESTIMATION WITH REAL-TIME INFORMATION**

As discussed in previous sections, one of the major pieces of information required for estimating the incident delay is the incident duration distribution. It can be expected that the incident duration
will be a function of the incident managing capability of the local authority, the incident location, and the incident severity among other factors. However, it still is feasible to establish location-specific distribution functions based on historical data \((13,9)\). In addition, the information on the incident status (i.e., whether it has been removed) also may be available. In an ITS context this information may be managed by a traffic management center (TMC). The following sections focus on how to update the probability distribution of the incident duration and how to apply this updated PDF to improve the incident-delay estimation model developed earlier.

### Prior Probability Distribution of Incident Duration

Previous theoretical and empirical work \((13,9)\) has shown that the incident duration typically has a lognormal distribution. This research therefore assumes that the incident duration is lognormally distributed and its distribution can be established and categorized if necessary. These incident-duration distributions can be considered as prior knowledge on the incident duration. If the mean of the natural logarithmic of the incident duration \([\ln(D^*)]\) is \(\lambda\) and the standard deviation of \(\ln(D^*)\) is \(\xi\), then \(\ln(D^*)\) is \(N(\lambda, \xi)\) with density function noted as \(f_{\ln(D^*)}(x)\).

### Posterior Probability Distribution of Incident Duration

Assume that the TMC at the current time \((T_0)\) receives new information showing that an incident still has not been cleared since its occurrence at time \(T^*\). The implication of this information is that the incident duration must be longer than \((T_0 - T^*)\). Therefore, the probability distribution of the incident duration should be modified to take into account this new information. The modified PDF of the incident duration, posterior PDF \([f_{D^*}(x)]\), can be obtained by applying Bayesian theory:

\[
f_{D^*}(x) = k \cdot L(x) \cdot f_{\ln(D^*)}(x)
\]

where

\[
L(x) = \text{the likelihood function of the observed output, which is}
\]

\[
k = \begin{cases} 0 & \text{if } x \leq T_0 \\ 1 & \text{if } x > T_0 \end{cases}
\]

\[
k = \left[ \frac{L(x) f_{\ln(D^*)}(x) dx}{\int_{0}^{\infty} f_{\ln(D^*)}(x) dx} \right]^{-1} = \left[ \frac{\int_{0}^{\infty} f_{\ln(D^*)}(x) dx}{L(x) f_{\ln(D^*)}(x) dx} \right]^{-1}
\]

\[
k = \left[ 1 - \phi \left( \frac{\ln(T_0) - \lambda}{\xi} \right) \right]^{-1}
\]

Figure 7 schematically illustrates the relationship between the prior PDF, posterior PDF, and the likelihood function. It should be noted that the above method also can be extended to incorporate other types of information on the incident situation such as an estimation of incident duration from an emergency vehicle operator or police officer.
a priori then the problem simplifies to one of identifying the parameters of the PDF. Regardless, some type of statistical analysis of the incident data currently being collected by transportation management centers will be required.

ACKNOWLEDGMENT

Funding for this research was supplied by the Natural Science and Engineering Research Council of Canada and is gratefully acknowledged.

REFERENCES


Publication of this paper sponsored by Committee on Freeway Operations.