Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections

Francois Dion a,*, Hesham Rakha b, Youn-Soo Kang c

a Virginia Tech Transportation Institute, 3500 Transportation Research Plaza (0536), Blacksburg, VA 24061, USA
b Charles Via Department of Civil and Environmental Engineering, 3500 Transportation Research Plaza, Blacksburg, VA 24061, USA
c Korean Transportation Institute, South Korea

Received 20 March 2002; received in revised form 28 April 2002; accepted 30 October 2002

Abstract

Delay is an important parameter that is used in the optimization of traffic signal timings and the estimation of the level of service at signalized intersection approaches. However, delay is also a parameter that is difficult to estimate. While many methods are currently available to estimate the delays incurred at intersection approaches, very little research has been conducted to assess the consistency of these estimates. This paper addresses this issue by comparing the delays that are estimated by a number of existing delay models for a signalized intersection approach controlled in fixed-time and operated in a range of conditions extending from under-saturated to highly saturated. Specifically, the paper compares the delay estimates from a deterministic queuing model, a model based on shock wave theory, the steady-state Webster model, the queue-based models defined in the 1981 Australian Capacity Guide, the 1995 Canadian Capacity Guide for Signalized Intersections, and the 1994 and 1997 versions of the Highway Capacity Manual (HCM), in addition to the delays estimated from the INTEGRATION microscopic traffic simulation software. The results of the comparisons indicate that all delay models produce similar results for signalized intersections with low traffic demand, but that increasing differences occur as the traffic demand approaches saturation. In particular, it is found that the delay estimates from the INTEGRATION microscopic simulation model generally follow the delay estimates from the time-dependent models defined in the 1997 HCM, 1995 Canadian Capacity Guide, and 1981 Australian Capacity Guide over the entire range of traffic conditions considered.

© 2003 Published by Elsevier Science Ltd.

*Corresponding author. Tel.: +1-540-231-9619; fax: +1-540-231-5214.
E-mail addresses: fdion@vtti.vt.edu, fdion@ctr.vt.edu (F. Dion), hrakha@vt.edu (H. Rakha).

0191-2615/03/$ - see front matter © 2003 Published by Elsevier Science Ltd.
doi:10.1016/S0191-2615(03)00003-1
**Nomenclature**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_i$</td>
<td>area of region $i$ in time–space diagram (kmh)</td>
</tr>
<tr>
<td>$c$</td>
<td>$s \cdot g_e/C = $ capacity of intersection approach (veh/h)</td>
</tr>
<tr>
<td>$C$</td>
<td>traffic signal cycle length (s)</td>
</tr>
<tr>
<td>$d$</td>
<td>average delay per vehicle (s/veh)</td>
</tr>
<tr>
<td>$D$</td>
<td>total delay incurred on intersection approach (veh s)</td>
</tr>
<tr>
<td>$d_o$</td>
<td>deterministic overflow delay (s/veh)</td>
</tr>
<tr>
<td>$d_1$</td>
<td>uniform delay (s/veh)</td>
</tr>
<tr>
<td>$d_2$</td>
<td>incremental delay accounting for randomness of vehicle arrivals and over-saturation delay (s/veh)</td>
</tr>
<tr>
<td>$d_3$</td>
<td>residual delay for over-saturation queues that may have existed before the analysis period (s/veh)</td>
</tr>
<tr>
<td>$f_{PF}$</td>
<td>adjustment factor accounting for the quality of progression in coordinated systems</td>
</tr>
<tr>
<td>$f_r$</td>
<td>adjustment factor for residual delay component</td>
</tr>
<tr>
<td>$f_p$</td>
<td>adjustment factor for situations in which the platoon arrives during the green interval (0.9–1.2)</td>
</tr>
<tr>
<td>$g_e$</td>
<td>effective green interval duration (s)</td>
</tr>
<tr>
<td>$k$</td>
<td>incremental delay factor accounting for pre-timed or actuated signal controller settings</td>
</tr>
<tr>
<td>$k_a$</td>
<td>density of approaching traffic (veh/km)</td>
</tr>
<tr>
<td>$k_d$</td>
<td>density of discharging traffic (veh/km)</td>
</tr>
<tr>
<td>$k_i$</td>
<td>traffic density in zone $i$ (veh/km)</td>
</tr>
<tr>
<td>$k_j$</td>
<td>jam density (veh/km)</td>
</tr>
<tr>
<td>$I$</td>
<td>adjustment factor for upstream filtering/metering</td>
</tr>
<tr>
<td>$n, m$</td>
<td>capacity guide model parameters</td>
</tr>
<tr>
<td>$N$</td>
<td>number of cycles over which calculations are performed</td>
</tr>
<tr>
<td>$P$</td>
<td>proportion of vehicles arriving during effective green interval</td>
</tr>
<tr>
<td>$r_e$</td>
<td>effective red interval duration (s)</td>
</tr>
<tr>
<td>$s$</td>
<td>saturation flow rate (veh/h)</td>
</tr>
<tr>
<td>$SW_{ij}$</td>
<td>speed of shock wave between traffic zones $i$ and $j$ (km/h)</td>
</tr>
<tr>
<td>$t_c(u)$</td>
<td>time to clear the queue of vehicles in under-saturated conditions (s)</td>
</tr>
<tr>
<td>$t_{m(u)}$</td>
<td>time to maximum extent of queue in under-saturated conditions (s)</td>
</tr>
<tr>
<td>$T$</td>
<td>evaluation period (h)</td>
</tr>
<tr>
<td>$TT$</td>
<td>travel time of vehicle during trip</td>
</tr>
<tr>
<td>$TT_o$</td>
<td>total travel time on intersection approach during evaluation period, over-saturated conditions (veh s)</td>
</tr>
<tr>
<td>$TT_u$</td>
<td>total travel time on intersection approach during signal cycle, under-saturated conditions (veh s)</td>
</tr>
<tr>
<td>$u_f$</td>
<td>vehicle speed under free-flow conditions</td>
</tr>
<tr>
<td>$u_i$</td>
<td>traffic speed in zone $i$ (km/h)</td>
</tr>
<tr>
<td>$u(t)$</td>
<td>instantaneous speed at time $t$</td>
</tr>
</tbody>
</table>
1. Introduction

Vehicle delay is perhaps the most important parameter used by transportation professionals to evaluate the performance of signalized intersections. This importance of vehicle delay is reflected in the use of this parameter in both design and evaluation practices. For example, delay minimization is frequently used as a primary optimization criterion when determining the operating parameters of traffic signals at isolated and coordinated intersections. The Highway Capacity Manual (HCM) further uses the average control delay incurred by vehicles at intersection approaches as a base for determining the level of service provided by the traffic signals located at the downstream end of these approaches (TRB, 1997).

The popularity of delay as an optimization and evaluation criterion is attributed to its direct relation to what motorists experience while attempting to cross an intersection. However, delay is also a parameter that is not easily determined. Teply (1989), for instance, indicated that a perfect match between field-measured delay and analytical formulas could not be expected. The difficulty in estimating vehicle delay at signalized intersections is also demonstrated by the variety of delay models for signalized intersections that have been proposed over the years.

Despite differences between the proposed delay models, very little research has been concerned with the consistency of delay estimates from one model to the other. This paper addresses this problem by comparing the delays that are estimated by a number of analytical delay models, including deterministic queuing, shock wave, steady-state stochastic, and time-dependent stochastic delay models, and by further comparing these estimates to the delays that are produced by a microscopic traffic simulation model. To achieve this goal, the paper first presents some background material on vehicle delays at signalized intersections, followed by a description of the various delay models that are being compared. Evaluations of the consistency of delay estimates from these models are conducted by using them to evaluate delays on both under-saturated and over-saturated signalized intersection approaches.

2. Delay at signalized intersections

Delay at signalized intersections is computed as the difference between the travel time that is actually experienced by a vehicle while going across the intersection and the travel time this ve-
Vehicle would have experienced in the absence of traffic signal control. The diagram of Fig. 1 further indicates that the total delay experienced by a vehicle can be categorized into deceleration delay, stopped delay and acceleration delay. Typically, transportation professionals define stopped delay as the delay incurred when a vehicle is fully immobilized, while the delay incurred by a decelerating or accelerating vehicle is categorized as deceleration and acceleration delay, respectively. In some cases, stopped delay may also include the delay incurred while moving at an extremely low speed. For example, the 1995 Canadian Capacity Guide for Signalized Intersections (ITE, 1995) defines stopped delay as any delay incurred while moving at a speed that is less than the average speed of a pedestrian (1.2 m/s).

Fig. 2 illustrates in more detail the distinction between deceleration, stopped and acceleration delay. The figure illustrates the simulated trajectory and speed profiles of a number of vehicles arriving at a signalized intersection during a single cycle. These simulated trajectories were obtained using the INTEGRATION microscopic traffic simulation model (Van Aerde and Associates, 2001). In the figure, it is first observed that only the first eight vehicles reaching the intersection come to a complete stop. These vehicles need to stop either as a consequence of their arrival during the red interval or during the green interval when the queue of vehicles that had formed during the previous red interval has not yet fully dissipated. It is further observed that the following three vehicles only experience deceleration and acceleration delay, as these vehicles reach the intersection when all previously queued vehicles have already started to move and therefore only need to slow down to maintain a safe distance with the vehicles ahead of them.

While most of the delay incurred at signalized intersections is directly caused by the traffic signal operation, a fraction of the total delay is attributable to the time required by individual drivers to react to changes in the signal display at the beginning of the green interval, to me-
mechanical constraints, and to individual driver behavior. In ideal situations, vehicles queued at an intersection would start moving at their ideal speed immediately following the display of a green signal. However, the first few drivers usually hesitate a few seconds before starting to accelerate, thus causing additional delays to all queued vehicles. This delay in the start of the queue dissipation process may even cause additional vehicles to join the queue before its complete dissipation. When acceleration occurs, the rate at which vehicles accelerate also depends on mechanical constraints dictating the maximum feasible acceleration rate and on the rate at which individual driver chose to accelerate.

As an example, Fig. 3 illustrates the simulated headways between successive stop line cross times for the 12 vehicle arrivals of Fig. 2. As can be observed, the first vehicle crosses the stop line 4.3 s after the green initiation, while the second, third, and fourth vehicles follow with respective headways of 3.0, 2.7 and 2.4 s. In this case, the larger headways observed at the beginning of the interval are entirely caused by acceleration constraints. In the absence of such constraints, all vehicles would have crossed the intersection with average 2-s headway. This would have resulted in a 24-s reduction in the total delay incurred during the simulated signal cycle.

To account for the additional delays due to driver reaction time and vehicle acceleration constraints, the operation of a signalized intersection is usually defined in terms of effective signal intervals instead of actual intervals in delay estimation models, as shown in Fig. 4. Instead of
explicitly considering green, yellow and amber intervals and attempting to model variable departure rates, delay calculations are typically performed by dividing the signal cycle into effective periods of stopped and moving traffic within which constant traffic characteristics can be assumed. The amount of difference between the actual and effective timings will thus depend on the assumptions regarding driver reaction time at the beginning of the green interval and vehicle accelerations.

A final element that may affect the delays incurred at intersection approaches is the randomness in vehicle arrivals. If vehicles were to arrive at uniform intervals, the delays incurred by vehicles within successive signal cycles would be identical, as there would then be an exact replication of the arrival and departure patterns. However, under random arrival patterns, the number of arrivals may fluctuate from one cycle to the other, thus resulting in different queue lengths. This may in turn result in arrival demands that occasionally exceed the approach capacity, and therefore, in higher delays. Finally, platooned arrivals may also occur in coordinated traffic signal systems. In this case, the delay incurred by vehicles will depend on the degree to which the signals at successive intersections are timed to provide a green indication during the periods of high arrival flow rate.

Fig. 3. Simulated headway distribution at beginning of green interval.

Fig. 4. Queue modeling under deterministic queuing analysis.
3. Delay models for signalized intersections

3.1. Deterministic queuing model

Classic, deterministic queuing models can predict delay at signalized intersections for which the number of vehicles that can be served during a green interval is greater than the number of arrivals per cycle. These models view traffic on each intersection approach as a uniform stream of arriving vehicles seeking service from a control device that provides a high service rate, but that also periodically stops servicing vehicles to accommodate traffic on a conflicting movement.

To illustrate how deterministic queuing models predict delays, consider the upper diagram of Fig. 5, which illustrate cumulative arrivals and departures at an under-saturated intersection.

Fig. 5. Idealized cumulative arrivals and departures for under- and over-saturated conditions.
From the figure, it can be determined that the area between the arrival and departure curves represents the total uniform delay incurred by all vehicles attempting to cross an intersection within a signal cycle. Assuming uniform arrivals and service times, i.e., a queuing system of the type D/D/1, Eqs. (1) and (2) can then be derived to calculate the average uniform delay incurred every signal cycle by vehicles attempting to cross the intersection. Eq. (2) is identical to what is used in the HCM and the Canadian Capacity Guide for Signalized Intersections, as will be demonstrated in the following sections:

\[
\begin{align*}
    d &= \frac{r^2_c}{2 \cdot C} \cdot \left(\frac{s}{s - v}\right) \\
    d &= \frac{C(1 - \frac{g_e}{C})^2}{2 \left(1 - X\frac{C}{g_c}\right)}
\end{align*}
\]  

(1)  

(2)

The model of Eq. (1) was generated by first assuming that vehicles arrive at a uniform and constant rate. A consequence of this assumption is that the queue of vehicles that form at an intersection operating in under-saturated conditions can always be cleared before the return of the red signal. In reality, the randomness of traffic may cause some vehicles to remain queued at the end of the green interval, especially at intersections operating near saturation. A second assumption is that vehicles decelerate and accelerate instantaneously. As was illustrated in Fig. 1, this assumption converts all deceleration and acceleration delays into equivalent stopped delay, and thus allows a direct estimation of the total delay incurred by vehicles attempting to cross an intersection. This assumption also implies that all drivers follow average driving patterns, in addition to assigning all incurred delays to the intersection approach, even though some delay occur in reality on the exit link when vehicles are accelerating. A final assumption is that vehicles queue vertically at the intersection stop line. While this assumption does not represent a normal queuing behavior and may not accurately represent the exact number of queued vehicle at a given instant, it does not bias the delay estimation process over an entire queue formation and dissipation process and is therefore a valid simplification when only considering delay estimations.

In over-saturation conditions, the number of vehicles reaching the intersection exceeds the number of vehicles that can be served by the traffic signal. This causes a growing residual queue to occur, as illustrated in the lower diagram of Fig. 5. The overflow delay associated with this situation corresponds to the area between the line representing the arrivals that can be served at capacity and the line representing the actual arrivals. In this case, Eqs. (3) and (4) can be derived to express the average delay over the number of vehicles discharged during the evaluation period \(T\).

\[
\begin{align*}
    d_o &= \frac{3600T}{2} \left(\frac{v}{C} - 1\right) \\
    d_o &= 900T \left((X - 1) + \sqrt{(X - 1)^2}\right)
\end{align*}
\]  

(3)  

(4)

This model is time dependent, as the overflow delay increases with any increase in the evaluation period. This is logical, as the residual queue keeps growing through the period. However, while the formation of Eq. (4) expresses the total delay incurred during the time period \(T\), it does
not include the delay that is incurred by the vehicles that remain in queue after the conclusion of \( T \). It should be noted that Eq. (4) provides a general relationship that is valid for volume-to-capacity ratios that are less than and/or greater than 1.0.

3.2. Shock wave delay model

Traffic flow can be characterized using flow, density and speed through an analogy with fluid dynamics. Lighthill and Whitham (1955), as well as Richards (1956), made the first successful attempts at such a description. They both demonstrated the existence of traffic shock waves and proposed a first theory of one-dimensional waves that could be applied to the prediction of highway traffic flow behavior. Eqs. (5) and (6) represent their model. The first equation defines the relation between volume, density, and speed that has been developed from the application of fluid dynamics theory. Using Eq. (5), Eq. (6) was then developed to describe the speed at which a change in traffic characteristics, or shock wave, propagates along a roadway.

\[
\begin{align*}
  v_i &= k_i \cdot u_i \\
  \text{SW}_{ij} &= \frac{v_j - v_i}{k_j - k_i}
\end{align*}
\]

Using the model of Eqs. (5) and (6), Rorbech (1968) investigated the queue formation at intersection approaches at the beginning of red intervals. Stephanopoulos and Michalopoulos (1979) further investigated the dynamics of queue formation and dissipation at isolated intersections using the flow conversation principle of Eq. (5). In another study, Michalopoulos et al. (1980) analyzed traffic dynamics between signalized intersections and demonstrated the existence of shock waves propagating downstream of an intersection caused by the periodic operations of traffic signals. Michalopoulos et al. (1981) and Michalopoulos and Pisharody (1981) further developed a real-time control algorithm based on shock wave theory that minimizes total delay at isolated intersections subject to constraints regarding maximum queue lengths on individual approaches.

The main difference between shock wave and deterministic queuing models is in the way vehicles are assumed to queue at the intersection. While queuing analysis assumes vertical queuing, shock wave analysis considers that vehicles queue horizontally. As illustrated in Fig. 6, the consideration of the horizontal extent of a queue enables the capturing of more realistic queuing behavior and the determination of the maximum queue reach. This is not possible with deterministic queuing models, as these models only track the number of queued vehicles, not their spatial location.

In Fig. 6, the total travel time spent by all vehicles going through the intersection can be estimated using the density and flow rate associated with each region. Since delay represents the added travel time caused by the traffic signal operation, the total delay incurred by the traffic within one signal cycle can be estimated by calculating the difference between the total travel time with traffic signals and the total overall travel time without traffic signals, as demonstrated in Eq. (7).

\[
D = TT_{u(\text{signals})} - TT_{u(\text{no signals})} = 3600 \left[ \sum_{i=A,B,C} A_i k_i - \sum_{i=A,B,C} A_i k_C \right]
\]
Finally, Eq. (8) can be derived to compute the average total delay incurred by individual vehicles due to the operation of traffic signals.

\[
d = 3600 \frac{x_{m(u)}}{2 \cdot v \cdot C} \cdot \left[ r_c \cdot (k_j - k_a) + (t_{m(u)} + t_{c(u)}) \cdot (k_d - k_a) \right]
\]  

(8)

with:

\[
x_{m(u)} = \frac{1}{3600} \cdot \left[ -v \cdot r_c \cdot s \right] \frac{s(k_j - k_a) - v(k_j - k_d)}{s(k_j - k_a) - v(k_j - k_d)}
\]  

(9)

\[
t_{m(u)} = \frac{v \cdot r_c \cdot (k_j - k_d)}{s(k_j - k_a) - v(k_j - k_d)}
\]  

(10)

\[
t_{c(u)} = 3600 \left[ \frac{x_{m(u)}}{SW_N} \right] = 3600x_{m(u)} \cdot \frac{k_a - k_d}{v - s}
\]  

(11)

Similar to the deterministic queuing models, the shock wave delay model of Eq. (8) assumes that vehicles follow a non-random and consistent path and that all vehicles accelerate and decelerate instantaneously. These two elements are represented in Fig. 6 by the constant interval between the lines representing the trajectories of individual vehicles and by the sharp angles in the trajectories when a vehicle passes from one traffic zone to another. Similar to Eq. (2), the consequences of these two assumptions are that Eq. (8) only estimates uniform delay and assumes that all delays are incurred on the approach side of an intersection. Finally, another common element with Eq. (2) is the use of the effective signal interval durations to account for the start loss and end gain.

Shock wave theory can also be used to estimate approach delays in over-saturated conditions. In this case, delay estimation is similar to the under-saturated case. As an example, Fig. 7 illustrates the spatial and temporal evolution of a queue in the first two cycles of operation of an over-saturated signalized approach as seen from the shock wave analysis. In the figure, the shock waves labeled SW_i, SW_R and SW_N have the same origin as the similar shock waves shown in Fig. 6. In this case, however, the over-saturation creates a new shock wave, SW_S, locating the back of the area containing fully stopped vehicles after the return of the red indication.

![Shock wave analysis for under-saturated approach.](image-url)
In Fig. 7, the areas 1, 3 and 4 all contain vehicles queued at the approach jam density, while the areas 2, 5 and 6 are characterized by vehicles moving at saturation flow and with a density corresponding to the discharge density. Based on these observations, it can be determined that the area of regions 3 and 5 is identical to the area of regions 1 and 2. If the area of regions 4 and 6 is known, the estimation of the delay incurred by vehicles on the approach is then only a matter of knowing the number of cycles over which the estimation is performed.

Similar to the under-saturated scenario, Eqs. (12) and (13) can be derived from Fig. 7 to estimate the total travel time incurred by vehicles traveling on the over-saturated intersection approach with and without the traffic signals.

$$T_{T_{0(s)}} = N \left\{ \left( \frac{x_{m(o)}}{2} \right) r_e \cdot k_j + \left( \frac{x_{m(o)} + x_{c(o)}}{2} \right) g_e \cdot k_d \right\} + \sum_{i=1}^{N-1} i \cdot x_{c(o)} \cdot \left( r_e \cdot k_j + g_e \cdot k_d \right)$$  \hspace{1cm} (12)

$$T_{T_{0(no \ signals)}} = N \cdot k_a \left\{ \frac{x_{m(o)}}{2} r_e + \frac{x_{m(o)} + x_{c(o)}}{2} g_e \right\} + \sum_{i=1}^{N-1} i \cdot x_{c(o)} \cdot \left( r_e + g_e \right) \cdot k_a$$  \hspace{1cm} (13)

with:

$$x_{c(o)} = \frac{C \cdot \left( v - s \right)}{3600 k_a}$$  \hspace{1cm} (14)

Finally, the average over-saturation delay is computed as the difference between Eqs. (12) and (13) divided by the number of vehicle departures within the analysis period $T$, as indicated in Eq. (15).

$$d = \frac{N \left\{ \frac{x_{m(o)}}{2} \cdot r_e \cdot (k_j - k_a) + \left( \frac{x_{m(o)} + x_{c(o)}}{2} \right) g_e \cdot (k_d - k_a) \right\}}{s \cdot g_e \cdot N/3600} + \sum_{i=1}^{N-1} i \cdot x_{c(o)} \cdot \left( r_e \cdot (k_j - k_a) + g_e \cdot (k_d - k_a) \right)$$  \hspace{1cm} (15)

### 3.3. Steady-state stochastic delay models

While deterministic queuing and shock wave delay models both assume uniform arrivals, stochastic delay models attempt to account for the randomness of vehicle arrivals. One of the

![Fig. 7. Shock wave analysis for over-saturated approach.](image-url)
fundamental and most often quoted models in the literature is the Webster model (Webster, 1958). This model, which is expressed by Eq. (16), is comprised of three terms. The first term estimates the average approach delay assuming uniform arrivals, which is consistent with Eq. (2) that was derived earlier in the paper. The second term considers the additional delays attributed to the randomness of vehicle arrivals. The third term is an empirical correction factors that reduces the estimated delay by 5–15%, to be consistent with simulation results.

\[
d = \frac{C(1 - \frac{c}{g})^2}{2(1 - X \frac{c}{g})} + \frac{X^2}{2v(1 - X)} - 0.65\left(\frac{c}{v^2}\right)^{1/3}X^{2+\frac{\pi}{6}}
\] (16)

Following Webster’s work, other stochastic models were proposed. These include the models by Miller (1963), Newell (1960, 1965), McNeil (1968), and Heidemann (1994). These models all share the same general basic assumptions. First, they all consider that the number of arrivals in a given time interval follows a known distribution, typically a Poisson distribution, and that this distribution does not change over time. This implies that these models could not be applied to estimate delays at intersections within a coordinated system, where arrivals are platooned as a result of upstream traffic signals. Second, they all assume that the headways between departures from the stop line follow a known distribution with a constant mean, or are identical. Third, while it is recognized that temporary over-saturation may occur due to the randomness of arrivals, it is assumed that the system remains under-saturated over the analysis period. Fourth, the system is assumed to have been running long enough to allow it to have settled into a steady state. Fifth, all these models still consider that vehicles decelerate and accelerate instantaneously and thus, that all drivers behave similarly.

3.4. Time-dependent stochastic delay models for under-saturated and over-saturated conditions

A main consequence of the stochastic delay modeling described in the previous section is that the estimated delays tend to infinity as traffic demand approaches saturation (\(v/c\) ratio of 1.0). This was considered as a weakness by many researchers (Akcelik, 1988; McShane and Roess, 1990; Fambro and Roupahil, 1997). Fig. 8 illustrates how delay estimation models should theoretically behave under varying demand levels. For low \(v/c\) ratios, the models should produce delay estimates that are similar to those produced by deterministic queuing delay models assuming constant uniform arrivals. As the load increases, a larger proportion of the delay is caused by the randomness of vehicle arrivals and attributed to the inability to clear all queued vehicles in some cycles. As the \(v/c\) ratio approaches 1.0, the models should finally not tend to infinity, but should instead produce estimates that become tangent to the deterministic over-saturation model of Eq. (2).

The concept of a general time-dependent delay model was originally conceived by Robertson (1979) and further enhanced by Kimber and Hollis (1979) using the coordinate transformation technique illustrated in Fig. 8. This technique transforms the equation defining a steady-state stochastic delay model so that it becomes asymptotic to the deterministic over-saturation model of Eq. (4). Although there is no rigorous theoretical basis for this approach (Hurdle, 1984), empirical evidence indicates that these models yield reasonable results. This explains why numerous time-dependent delay formulas based on the coordinate transformation technique have been proposed
over the years (Brilon and Wu, 1990; Akcelik, 1981, 1988; Akcelik and Rouphail, 1993; Fambro and Rouphail, 1997) and have been incorporated into a number of capacity guides, such as those from the United States (TRB, 1994, 1997), Australia (Akcelik, 1981) and Canada (ITE, 1995).

The capacity guide delay models currently used in the United State, Australia and Canada are all similar. The general form of these models is expressed by Eqs. (17)–(20), with Table 1 indicating the specific values assigned to the parameters in each model.

\[
d = d_1 \cdot f_{PF} + d_2 + d_3 \cdot f_t
\]

with:

Table 1
Capacity guide delay model parameters

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameters</th>
<th>(f_t)</th>
<th>(N)</th>
<th>(m)</th>
<th>(k)</th>
<th>(I)</th>
<th>(T)</th>
<th>(X_o)</th>
<th>(f_{PF})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australian (1981)</td>
<td>0 0 6 or 12(^a)</td>
<td>(n/a)</td>
<td>(n/a)</td>
<td>Variable</td>
<td>0.67 + (s_{g_e}/600)</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Canadian (1995)</td>
<td>0 0 4</td>
<td>(n/a)</td>
<td>(n/a)</td>
<td>Variable</td>
<td>0</td>
<td>1.0 or Eq. (20)(^b)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCM (1994)</td>
<td>0 2 4 to 16(^c)</td>
<td>(n/a)</td>
<td>(n/a)</td>
<td>15 min</td>
<td>0</td>
<td>1.0, 0.85, or Eq. (20)(^d)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCM (1997)</td>
<td>1 0 8</td>
<td>0.04–0.50(^e)</td>
<td>1.0(^f)</td>
<td>Variable</td>
<td>0</td>
<td>Eq. (20)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) 12 for random arrivals; 6 when platooning occurs.
\(^b\) 1.0 for isolated intersections; Eq. (20) in other cases.
\(^c\) Function of arrival type (16 for random arrivals, 12 for favorable or non-favorable progression, 8 for very poor or highly favorable progression, 4 for very unfavorable progression).
\(^d\) 1.0 for pre-timed, non-coordinated signals; 0.85 for actuated, non-coordinated systems; Eq. (20) for coordinated systems.
\(^e\) 0.50 for pre-timed signals; 0.04–0.50 for actuated controllers.
\(^f\) 1.0 for isolated intersection only.
\[ d_1 = 0.50C \frac{(1 - \frac{g_c}{c})^2}{(1 - \frac{c}{g_c} \cdot \min(X, 1.0))} \]  

(18)

\[ d_2 = 900X^aT \left( X - 1 \right) + \sqrt{(X - 1)^2 + \frac{mkI}{cT} \cdot (X - X_o)} \]  

(19)

\[ f_{PF} = \frac{(1 - P)f_P}{1 - \frac{g_c}{c}} \]  

(20)

The models defined in Table 1 all assume steady-state traffic conditions. They estimate delays under stochastic equilibrium conditions, when the arrival and departure flow rates have been stationary for an indefinite period of time. They also assume that the number of arrivals in a given interval follow a Poisson distribution that remains constant over time, and that the headways between departures have a known distribution with a constant mean value. Finally, these models do not include all of the delay incurred by either arriving or discharged vehicles during the control period considered. These models only include the delays incurred up to the end of the control period and do not consider the additional delay that is incurred as the queue is being served. However, it can be demonstrated that accounting for the additional delay that is incurred during the decay of the queue and averaging over all the vehicle departures (in this case the arrival rate), the average delay is the identical. Specifically, if it is assumed that there are no vehicle arrivals at the conclusion of the analysis period, then the average delay can be computed using Eq. (21), which is identical to Eq. (4).

\[ d = \frac{0.5(v - c) \cdot 3600T \cdot \left[ T + \frac{(v-c)T}{c} \right]}{T \cdot v} = \frac{1800(v - c)T}{c} = 900T \left( X - 1 \right) + \sqrt{(X - 1)^2} \]  

(21)

Similar to the models presented in the previous sections, the models of Table 1 were developed assuming instantaneous accelerations and decelerations to simplify the delay estimation. In this case, however, a number of relations have been proposed to estimate the proportion of stopped and acceleration/deceleration delays. For instance, the 1994 HCM estimates the stopped delay at 76% of the total estimated delay. Both Teply (1989) and Olszewski (1993) agreed that this factor was incorrect for very low and very high signal delays and proposed alternative evaluation methods. Teply first recommended a multiplicative adjustment factor that varies 0.36–0.83 and is a function of the duration of the red interval, which was later adopted in the 1995 Canadian Capacity Guide for Signalized Intersections, while Olszewski recommended a subtractive adjustment factor varying with the approach speed.

The main differences between the various capacity guide models are in the way they attempt to consider non-Poisson arrivals:

(a) In the Australian model, adjustments for the type of arrivals are considered by simply altering the value assigned to the parameter \( m \). A value of 12 is used for random arrivals, while a value of 6 is used for platooned arrivals.

(b) In the 1994 HCM model, both the parameter \( m \) and the progression factor \( f_{PF} \) are modified. In this case, values ranging from 4 to 16 are used for the parameter \( m \), while alternative suggested values for the progression factor \( f_{PF} \) consider not only varying arrival patterns, but also the use of pre-timed or actuated controllers at coordinated and non-coordinated intersections.
(c) In the 1995 Canadian model, the type of arrivals only affects the progression factor \( f_{PF} \) as the parameter \( m \) is assigned a fixed value of 4.

(d) The 1997 HCM model fixes the parameter \( m \) at 8, but introduces the parameters \( k \) and \( I \) to explicitly consider the effect of signal controller types and of arrival patterns affected by upstream signalized intersections. As a result, the product \( mkl \), which is always equal to 4.0 in the Canadian model, could vary in this case from 0.32 to 4.0.

In addition to the above differences, the 1994 HCM model is the only one to assign a non-zero value to the parameter \( n \). This parameter was added to compensate for the model’s assumptions of a zero initial queue and fixed 15-min analysis period. While the 1997 HCM model reassigns a value of zero to the parameter \( n \), this model is also the only one to define a non-zero value to the parameter \( f_t \) and explicitly considers the residual delays for over-saturation queues that may have existed before the analysis period.

### 3.5. Microscopic simulation delay models

Microscopic traffic simulation models have the ability to track individual vehicle movements within simulated street networks. Vehicle tracking is usually done using car-following, lane-changing, and gap-acceptance logic. This allows such models, among other things, to consider virtually any traffic conditions, ranging from highly under-saturated to highly over-saturated conditions.

Because of their ability to track the movements of individual vehicles, microscopic simulation models can determine the delay incurred by an individual vehicle while traveling a network of links with different characteristics by comparing simulated and ideal travel times. No specific formulas are therefore required to evaluate uniform and overflow delay, or delays in under-saturated and over-saturated traffic conditions, thus allowing for the evaluation of complex traffic situations. In addition, the ability to record vehicle speed and position on a second-by-second basis further allows the recording of speed profiles and the direct estimation of deceleration, stopped and acceleration delays.

For this study, the INTEGRATION microscopic traffic simulation software is used to obtain simulated delay estimates. This simulation model features an integrated, dynamic traffic simulation and traffic assignment model. Within the model, delay is estimated for each individual vehicle by calculating, for each traveled link, the difference between the vehicle's simulated travel time and the travel time that the vehicle would have experienced on the link at free speed, as expressed by Eq. (22) (Rakha et al., in press).

\[
D = \int_{t=0}^{T} \left[ u_t - u(t) \right] \frac{u_t}{u_f} dt
\]

### 4. Test scenario

To evaluate the consistency of delay estimates from the various models presented in this paper, delay evaluations were carried out for the sample network of Fig. 9 using the INTEGRATION
microscopic traffic simulation software and the deterministic, shock wave, Webster and capacity
guide delay models defined by Eqs. (2), (4), (8), and (15)–(17), respectively. The example of Fig. 9
features a single-lane intersection approach at a fixed-timed traffic signal operating with a 60-s
cycle length and a 30-s effective green interval. A 1-km exit link is also included to allow the
INTEGRATION model to capture the delays incurred by vehicles accelerating as they leave the
intersection and to compile total approach delay, as is done in the analytical models.

For each model, delay evaluations were specifically carried out for \( v/c \) ratios varying between
0.1 and 1.4. This allowed evaluations to be conducted for a range of traffic conditions extending
from highly under-saturated to highly over-saturated conditions. For each scenario, vehicle ar-
rivals were further assumed to follow a random process with a constant average arrival rate. For
the analytical models, this only required a single application of the equations defining each model.
For the INTEGRATION model, however, 10 replications were made of each scenario to account
for the stochastic nature of the model’s simulation process. Thus, unless otherwise noted, the
results from the INTEGRATION simulation model are an average of 10 simulations, while the
results from the analytical models are the delays reported by Eqs. (2), (4), (8), and (15)–(17).

Finally, to ensure that appropriate comparisons are made for all \( v/c \) scenarios considered,
particularly in the over-saturation range, simulation were only carried out for a 15-min control
period. This 15-min period was imposed by the fact that some of the time-dependent delay models
considered, namely the 1994 HCM model, require a 15-min analysis period.

5. Test results

Figs. 10 and 11 illustrate the results of the delay estimations that were carried out for the
example of Fig. 9. Fig. 10 illustrates the delays that were estimated over the entire range of \( v/c \)
ratios considered, while Fig. 11 provides a more detailed look at the delay estimates for \( v/c \) ratios
below 1.0.
Fig. 10. Delay estimates for under- and over-saturated conditions.

Fig. 11. Delay estimates for under-saturated conditions only.
5.1. Consistency of delay estimates for under-saturated conditions

In Fig. 11, it is first observed that there is a general agreement between all the analytical delay models considered when these are applied to the analysis of signalized intersections with very low \( v/c \) ratios. However, it is also observed that this agreement tends to decrease with increasing \( v/c \) ratios. For example, the difference between the minimum and maximum delay estimates from the various analytical models does not exceed 6.0% for \( v/c \) ratios below 0.2. For a \( v/c \) ratio of 0.4, the maximum difference increases to 14.4%, while differences exceeding 30% are observed for \( v/c \) ratios above 0.6.

Fig. 11 further illustrates that the deterministic queuing and shock wave models always produce the lowest estimates. This is due to the fact that these two models consider only uniform arrivals. As a result, these models cannot consider the potential for additional delays that arise from the probability of having temporary over-saturation delays due to surges of arriving vehicles. It is also observed that the deterministic and stochastic models produce relatively similar delay estimates at very low \( v/c \) ratios. This is an indication that the randomness of vehicle arrivals can be neglected when estimating delays for highly under-saturated conditions and that the use of either deterministic or stochastic models is valid in such conditions.

A more detailed analysis of Fig. 11 further reveals that the delays predicted by the Webster, 1995 Canadian Capacity Guide and 1997 HCM models are virtually identical for \( v/c \) ratios below 0.8. Above this ratio, the delays from the Webster model tend towards infinity for \( v/c \) ratios approaching 1.0. This is a well-known behavior of the model and other similar steady-state stochastic models that invalidates their use when the \( v/c \) ratio tends to 1.0. In comparison, the delays estimated at a \( v/c \) ratio of 1.0 from the various capacity guide models oscillate between 43.7 and 45.0 s per vehicle, while the average delays estimated by the INTEGRATION model approach 38.8 s per vehicle.

It should be noted at this point that although the vehicles are generated randomly at their origin within the INTEGRATION software, the level of randomness during a simulation is reduced as a result of the car-following behavior that is applied while vehicles are traveling along a link. On one hand, the car-following logic allows a vehicle with a large distance headway with the vehicle in front of it to speed up to reduce the headway. On the other hand, the logic forces a vehicle that has been generated close to its predecessor to slow down to maintain a safe car-following distance. Consequently, it is expected that the simulation delays would be slightly less than the random delay estimates.

While a general agreement exists between the various capacity guide delay models in the trend of increasing delays with increased \( v/c \) ratios, differences are also observed. The main differences are more particularly found with the Australian Capacity Guide and 1994 HCM models, which both consistently produce lower delay estimates than the 1997 HCM and 1995 Canadian Capacity Guide models. These differences are primarily due to the coordinate transformation technique that was applied in each case to produce a model providing delay estimates that are asymptotic to the deterministic over-saturation delay model of Eq. (2). The 1995 Canadian Capacity Guide and 1997 HCM models both produce identical results, as these two models are similar in this case. As explained earlier, the main differences between these two models are in the values assigned to the parameters \( f_r, m, k, \) and \( l \) in Eqs. (17) and (19). In this case, however, the parameter \( f_r \) has no impact since it is assumed that no over-saturation queue exists before the start of the analysis.
Both models also exhibit the same value for the $mkl$ product in Eq. (19). While the Canadian model fixes the value of $m$ at 4 and ignores the parameters $k$ and $l$, the HCM model assigns a value of 8 to the parameter $m$, a value of 0.5 to the parameter $k$ to account for pre-timed signal operation, and ignores the parameter $l$.

The analysis of Fig. 11 further reveals that the average delay estimates from the INTEGRATION simulation model are in general agreement with the estimates from the various capacity guide models. This agreement is particularly evident in Fig. 12, which superimposes the delay estimates from the individual INTEGRATION simulation runs to the corresponding capacity guide delay estimates. As can be observed, the estimates from the four capacity guide delay models all fall within the range of delays obtained from the individual simulation runs. This result thus either validates the under-saturated delay models defined in the various capacity guides, or the INTEGRATION simulation process, depending on the point of view taken.

In Figs. 11 and 12, while the estimates from the various analytical models generally increase monotonically with higher $v/c$ ratios, it is observed that there is more variability in the delay estimates from the INTEGRATION simulation model than from the analytical models, particularly at low $v/c$ ratios. A portion of this variability is attributed to the stochastic nature of the INTEGRATION simulation process, while another portion is attributed to the discrete nature of the simulation process, which only considers complete vehicles and not fractions of vehicles.

A main effect of considering discrete vehicle departures is to increase the sensitivity of delay estimates to specific vehicle arrival times. As an example, Table 2 computes the delay associated with three arriving flows for the example of Fig. 9. In all cases, the vehicles arrive at the intersection with a constant 5-s headway and depart at a constant 2-s saturation flow headway. The
only difference between the arrival patterns is in the time at which the first vehicle arrives relative to the start of the effective green. In the second cycle, vehicle arrivals are all offset by 1 s relative to the first cycle, while arrivals in the third cycle are offset by 4 s. As can be observed, a 1-s offset is sufficient in this case to decrease the estimated average delay per cycle from 13.8 to 12.9 s, while a four-second offset further results in a 10.5 s average delay. If delay calculations were made using the deterministic queuing or shock wave models of Eqs. (2) and (8), the estimated average delay would then be 12.5 s for all three cycles. In the case of the Webster and capacity guide models, similar estimates would also be obtained from all cycles, since these models only consider average flow rates and traffic patterns.

Fig. 13 provides a more detailed look at the variability of simulated delay estimates. The figure plots for each v/c ratio the standard deviation and coefficient of variation (COV) (standard deviation divided by mean) of delay estimates that are obtained from the ten repetitions per v/c ratio. As can be observed, significant variability as measured by the COV in delay estimates is observed both at low v/c ratios and around a v/c ratio of 1.0. While the literature documents the higher variability in delay as the v/c ratio approaches 1.0, it does not describe the high variability that is observed for extremely low traffic demands.

The two observed peaks of high COV that are observed in Fig. 13 can be attributed to two factors. First, at very low v/c ratios there are larger temporal headways between successive vehicles and thus the standard deviation is higher than the case of higher v/c ratios. This provides greater opportunities for vehicle spacing to be varied, as was demonstrated earlier in Table 2. As traffic demand increases and the average temporal headway between successive vehicles decreases, the impact of the vehicle departure randomness on delay estimates decreases, thus explaining the reduced variability that is observed for v/c ratio less than 0.7. In addition, given that the mean delay at low v/c ratios is very small and that the COV is computed as the ratio of the standard deviation to the mean, it is not surprising to observe higher coefficients of variation at very low v/c ratios.
Another element to consider when interpreting the above results is the fact that the INTEGRATION car-following logic reduces the level of randomness of vehicle arrivals. As was explained earlier, the logic reduces differences in vehicle spacing. On one hand, the logic allows a vehicle with large distance headway with the vehicle in front of it to speed up to reduce the headway. On the other hand, the logic forces a vehicle that has been generated close to its predecessor to slow down to maintain a safe distance.

For $v/c$ ratios above 0.7, signal cycle over-saturation is the main factor responsible for the increase in delay variability. As demand approaches saturation, there is an increased probability that surges in vehicle arrivals may cause temporary over-saturation of a signal cycle. Since higher delays typically result from over-saturated cycles, an increased variability in delay estimates thus results as the probability of having over-saturated cycles increases. This variability reaches its peak at a $v/c$ ratio of 1.0, as demand fluctuations then create equal probabilities for over-saturated and under-saturated cycles. Finally, the variability decreases again as the $v/c$ ratio increases past 1.0 as the increasing congestion and number of over-saturated cycles constrain the freedom of movements of vehicles and lead to more uniform flow patterns.

5.2. Consistency of delay estimates for over-saturated conditions

For the over-saturated domain, the results of Fig. 10 indicate that there is a general agreement between the INTEGRATION simulation model, the 1981 Australian Capacity Guide, the 1995 Canadian Capacity Guide and the 1997 HCM delay models. For these four models, the delay estimates gradually approach the delays predicted by the deterministic over-saturation model of Eq. (2) as the $v/c$ ratio increases. The only exception to this trend is for the 1994 HCM model,
which does not produce a delay curve that is asymptotic to the deterministic over-saturation model. This behavior is explained by the fact that this model is the only one to include the $X^2$ adjustment parameter in Eq. (19), as summarized in Table 1. As explained earlier, this parameter, which was later removed in the 1997 model, was introduced to compensate for the assumption of zero initial queue and fixed 15-min analysis period.

In this case, Fig. 12 again emphasizes the consistency of delay estimates from the INTEGRATION, the 1981 Australian Capacity Guide, the 1995 Canadian Capacity Guide and the 1997 HCM delay models. Similar to the under-saturation case, it is observed that the delays estimated from these three capacity guide models fall within the range of delays estimated by the INTEGRATION simulation model. In this case, it is interesting to note that the INTEGRATION model always produces the lowest estimates and that the difference between the various models dissipates as the $v/c$ ratio increases. The lower delay estimates from the INTEGRATION software can be attributed to the lower level of randomness that results from travel from the entry node to the traffic signal stop line. More consistent results between the various models are observed as the $v/c$ ratio increases since the randomness in vehicle arrivals is then gradually absorbed by the queue formed upstream of the traffic signal, thus creating more uniform traffic conditions across the various models.

6. Conclusions

This paper compared the delays predicted by the INTEGRATION microscopic traffic simulation software and a number analytical delay models on a one-lane approach to a pre-timed signalized intersection approach for traffic conditions ranging from under-saturation to over-saturation. The analytical models that were compared were representative of deterministic queuing, shock wave, steady-state stochastic and time-dependent stochastic delay models. For the steady-state stochastic models, the Webster model was used as an example. For the time-dependent stochastic models, the models defined in the 1981 Australian Capacity Guide, the 1995 Canadian Capacity Guide and the 1994 and 1997 HCM were considered.

The delay estimates predicted by each model were compared over a range of $v/c$ ratios extending from 0.1 to 1.4 to assess their consistency. Over this range, the delay models from the 1981 Australian Capacity Guide, the 1995 Canadian Capacity Guide, the 1997 HCM, and the INTEGRATION microscopic traffic simulation model produced delay estimates that generally agree with each other. Depending on the point of view considered, it can be concluded that these results either validate the delay models currently used in the capacity guides from the United States, Canada and Australia, or validates the INTEGRATION microscopic simulation software. In addition, it was further determined that all the delay models considered in this paper produce relatively consistent delay estimates when applied to the analysis of under-saturated signalized intersections with $v/c$ ratios below 0.6, thus indicating the validity of the use of all these models in such conditions.

While the study indicates a strong consistency between the time-dependent stochastic delay models used in the more recent capacity guide for signalized intersections and the delays estimated by the INTEGRATION microscopic traffic simulation model for the case considered, efforts should be made to evaluate this consistency for more complex situations. In particular, the
consistency of delay models should be evaluated for multi-lane approaches, intersections controlled by actuated controllers, and intersections where non-random vehicle arrivals occurs as a result of signal coordination with upstream intersections. The impact of varying driver behavior should also be investigated, as this may impact the saturation flows used in the various delay models, as well as the approach speeds and amount of time lost every cycle due to driver reaction time.

Acknowledgements

The authors would like to acknowledge the effort of the late Dr. Michel Van Aerde in the realization of this study. Dr. Van Aerde acted as a supervisor for Dr. Youn-Soo Kang during the initial stages of his doctoral studies at Virginia Tech.

References


Rakha, H., Dion, F., Sin, H., in press. Field evaluation of energy and emission impacts of traffic flow improvement projects using GPS data: issues and proposed solutions. Transportation Research Record.


