Signalized Intersection Delay Models– A Primer for the Uninitiated

V. F. HURDLE

ABSTRACT

Delay is being used increasingly as the primary indicator of level of service at signalized intersections, but for many traffic engineers delay estimation is a new task. Some of the currently available estimation techniques are introduced, and the assumptions on which they are based are examined. In general, these assumptions are unrealistic; so accurate delay estimates are not really possible. The difficulties are particularly acute when the arrival flows approach capacity. Some of the procedures avoid the worst of the problems at high flows by methods that, though they involve considerable mathematics, are based on modeling that is essentially qualitative rather than quantitative. Such methods abandon the quest for accuracy in favor of reasonableness: rather than attempting to provide right answers, they try to avoid answers that are terribly wrong. These models would seem to be useful as long as users do not expect too much from them, and the models can probably be somewhat improved. If accuracy is desired, however, a new generation of models that take more account of variations in travel demand over time is needed. The use of such models would require more information about traffic patterns than users are accustomed to providing.

There is a worldwide trend toward the use of estimated delay as the principal measure of the level of service at signalized intersections. The reasons for this trend are clear: Delay can be measured; it has obvious economic worth; and it is easily understood by both technical and nontechnical people. The price that must be paid for these advantages is that capacity analysis procedures must include methods for estimating delay. This necessarily complicates the procedures, so that they involve more computation and are likely to be somewhat harder to understand. In addition, if delay is to be a useful measure of level of service, it is clearly necessary that the methods of estimating it be reasonably accurate and that analysts have a sense of the degree of accuracy to be expected in various situations.

This review paper is intended as a primer for traffic engineers who are familiar with capacity estimation techniques but have not made much use of delay equations. Because this is a primer rather than a textbook or handbook, the emphasis will not be on the mathematical derivation or detailed use of delay equations. Instead, the discussion will be concentrated on the assumptions underlying the equations and the limitations that stem from these assumptions. Although the paper is addressed to beginners, the author hopes that engineers more experienced with delay equations will find food for thought.

BASIC MODEL

The mathematical models used to estimate intersection delay are queueing models. The modeler views the traffic on each approach as a stream of customers seeking service from a server, a system similar to the lineups at grocery store checkout counters except that the server is the intersection and the "service" provided consists simply of letting the "customer" enter the intersection. The server also differs from most grocery store clerks in that it serves very quickly when it serves at all, but every minute or so it gets a red face and insists on resting.

Assume that the customers arrive at some average rate v (vehicles per unit time) and can enter the intersection at a maximum rate s when the light is green; s is called the saturation flow and, in North American literature, v is usually called the volume but will be called the arrival flow in this paper. If the cycle length is C and the length of the green interval is G, it follows that the approach can handle a maximum of $c = (G/C)s$ vehicles per unit time.

The fraction of time the signal is green, G/C, is one of the important parameters in all delay equations. Usually G will not appear alone but only as a part of the G/C ratio.

A second important parameter is the approach capacity,

$$c = (G/C)s$$

but it will also usually appear in a ratio: v/c, the ratio of the average arrival flow to the capacity. This ratio is often called the degree of saturation and in British and Australian literature is usually denoted x. In queueing theory it is usually denoted ρ and called the traffic intensity.

The operation of the intersection approach can be modeled (1) as shown in Figure 1. The curve labeled A(t) shows the cumulative number of arrivals by time t, and D(t) the cumulative number of departures. The arrival curve is an abstraction because it does not indicate the number of actual arrivals at the stop line, but the number of vehicles that would have arrived had the light always remained green. Such a curve could be obtained by counting some distance upstream, plotting the counts, then moving the curve to the right a distance equal to the time required to drive from the count station to the stop line when there is no queue of traffic waiting to get through the signal.

The departure curve D(t) shows the actual departures from the stop line. When the light is red, there are no departures, so the curve is horizontal. In reality it would begin to curve upward as cars began to move after the start of green, then after a few seconds it would become nearly straight with a slope equal to the saturation flow s.

In Figure 1 the D(t) curve is somewhat simpler than this because the figure is a simplified mathematical model. The major simplification is that the outflow does not gradually increase to s but changes instantly from zero to s at the effective start of

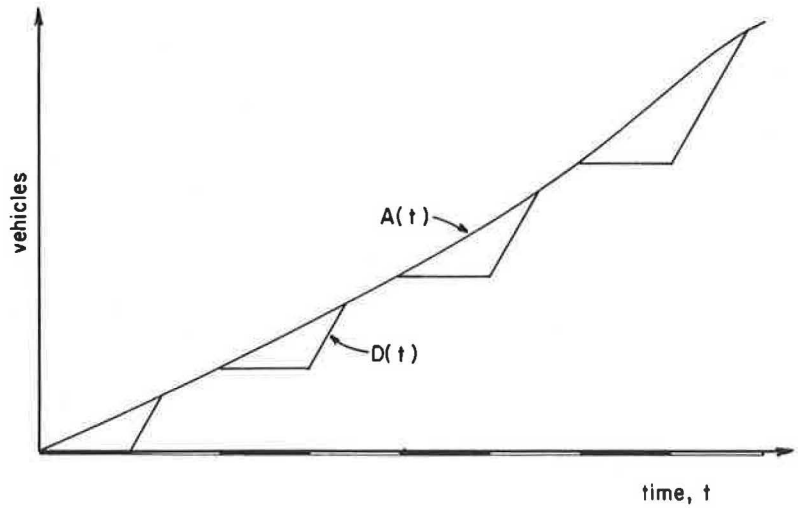

FIGURE 1 Arrival and departure curves for a signalized intersection approach.

green. To keep this simplification from being troublesome, it is, of course, important to be somewhat careful in distinguishing between the real green interval and the effective green interval used in the model, or at least to make a conscious decision that "effective green equals real green" is a good enough assumption for one's purposes. In what follows, the green interval length G will always be the effective length. (Readers wanting to know more about effective green times can find discussions in most traffic engineering textbooks.)

Figure 1 is handy because from it the queue length at any time,

$$Q(t) = A(t) - D(t)$$

can easily be read (Figure 2). Note that, because A(t) is a shifted count curve, Q(t) is an abstract queue length equal to the difference between the number of vehicles that would have crossed the stop line if the signal were not there and the number that actually did cross. Another way of thinking about the model is to say that, in the model, vehicles do not line up along the street but form a vertical stack at the stop line. The real queue is always somewhat longer than the model predicts because the queue engulfs some vehicles that the model assumes are still driving to the vertical stack at the stop line.

If the system works on at least approximately a first in-first out (FIFO) basis, one can also easily determine the waiting time of any individual vehicle as shown in Figure 2 for customer or vehicle number i, which waits w_i between arrival and departure. This is vehicle i's delay, the extra time it has to spend because the signal is there. Unlike the queue length, the abstract delay in the model's vertical stack is equal to the real delay. Vehicle i actually spends more time than w_i in queue, but the extra time, if not spent in queue, would have to be spent driving from the end of the real queue to the stop line.

If one considers vehicle i's delay as the area of a band w_i wide and one vehicle high, it is easy to see that the area between the A(t) and D(t) curves is equal to the total delay suffered by all of the customers who use the intersection approach. With a few more assumptions, a formula for the average delay can be derived by evaluating this area. If the slope of the A(t) curve is always v, the area between the A(t) and D(t) curves is made up of a

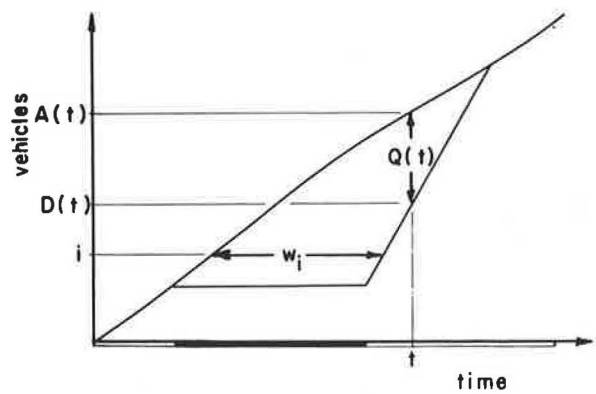

FIGURE 2 Queue length and waiting time.

series of identical triangles and the total delay per cycle can be estimated as the area of any single triangle. Dividing this area by the number of arrivals per cycle, vC, yields the average delay. This takes some algebraic manipulation because one must first locate the upper right-hand corner of the triangle. The answer is

$$UD = \{C[1-(G/C)]^2\}/\{2[1-(v/s)]\} \quad (1)$$

where

UD = average uniform delay per vehicle,
C = cycle length,
G/C = fraction of time the signal is (effectively) green, and
v/s = flow ratio (2,3) = (G/C)(v/c).

This average delay is denoted UD, which stands for uniform delay, because it was derived under the assumption that vehicles arrive at a uniform rate v throughout the cycle. In making this assumption, any random effects and any pattern imposed on the arrival stream by upstream intersections are ignored.

EFFECTS OF RANDOMNESS

The problem with Equation 1, of course, is that vehicles do not arrive at an isolated intersection uniformly, but in a random manner. At low flow

levels, this is not important, but when the arrival flow v gets close to the approach capacity, c = (G/C)s, the actual delay will be considerably larger than the uniform delay predicted by Equation 1. The reason is obvious in Figure 3, where the number of arrivals is the same during every cycle except the second, when some extra vehicles happen to arrive. Because the approach is very nearly saturated (i.e., very near capacity), it takes a long time for the extra queue that builds up in the second cycle to dissipate, and the area between the arrival and departure curves is a good deal larger than if the extra arrivals had not occurred. This extra delay is often called random delay or overflow delay. The latter term reflects the fact that the main effect of random arrivals is to cause occasional overflows of traffic from one cycle to the next (4).

To predict the random or overflow delay, it is necessary to construct a stochastic model using the methods of probability theory or computer simulation. The usual stochastic model depends on four basic assumptions:

1. The number of arrivals in a given time interval has a known distribution, often Poisson, and this distribution does not change with the time of day (i.e., the distribution is stationary) or because of the number of arrivals in any other time interval;
2. The headways between departures from the stop line either have some known distribution, again with a constant mean, or are all the same;
3. $v < c = (G/C)s$ (i.e., the system is not saturated); and
4. The system has been running long enough to have settled into a steady state.

Under these assumptions and some further assumptions about how the system operates, the queue length and waiting time should have a probability distribution that, once a steady state has been reached, does not vary with time but remains fixed. This steady-state waiting time distribution, or at least its mean and possibly its variance, can be determined in simple cases by analytic modeling and in more complicated cases by computer simulation.

Such models are known as steady-state queueing models because of the very important assumption that the system has operated for a sufficiently long time with the same average values of v and s to have settled into a steady state. For small v/c ratios, up to about 0.5, what happens in each cycle is nearly independent of what happened in the previous cycle so the time to reach a steady state is very short. As the v/c ratio increases, however, this relaxation time increases rapidly and approaches infinity as the v/c ratio approaches unity. The reason for this will be explained hereafter.

There are many steady-state models of varying complexity available; Allsop (4) gives an excellent summary, and Katsuhisa (5) gives numerical comparisons. By far the best known of these is the one constructed by Webster (6,7) in the late 1950s, assuming random (Poisson) arrivals and uniform departure headways. Webster's formula for the average delay, d, is

$$d = \{C[1-(G/C)]^2/2[1-(v/s)]\} + [(v/c)^2/2v(1-v/c)] - 0.65(C/v^2)^{1/3}(v/c)^{(2+5G/C)} \quad (2)$$

The first term of this expression is the uniform delay of Equation 1, the second can be predicted theoretically (4), and the third is a correction term obtained empirically from simulation results (6). This third term reduces the estimate by 5 to 15 percent, so a 10 percent reduction can be used as a rough approximation (7). In this latter approximation, any consistent time units may be used (i.e., if d is to be in minutes, then C must be in minutes and v and c in vehicles per minute), but the third term works properly only if all time units are seconds and all flows are in vehicles per second.

Figure 4 shows the delay predicted by Equations 1 and 2 as a function of the v/c ratio for a 60-sec cycle, G/C ratio of 0.5, and saturation flow of 3,600 vehicles/hr. The random delay component is small compared with the uniform delay when the degree of saturation is small, but it increases very rapidly at the higher v/c ratios and approaches infinity at the right edge of the figure. More complicated models with different assumptions would

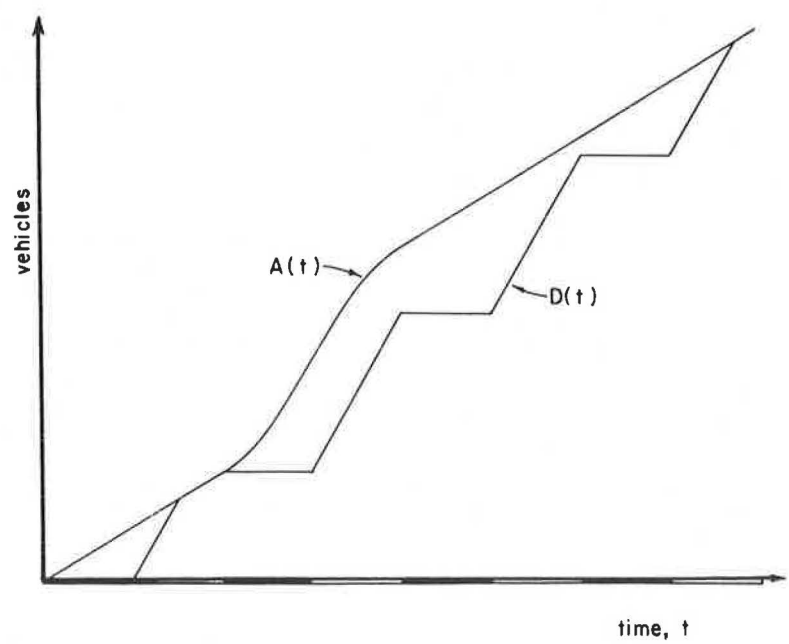

FIGURE 3 Overflow due to random variations in arrival flow.

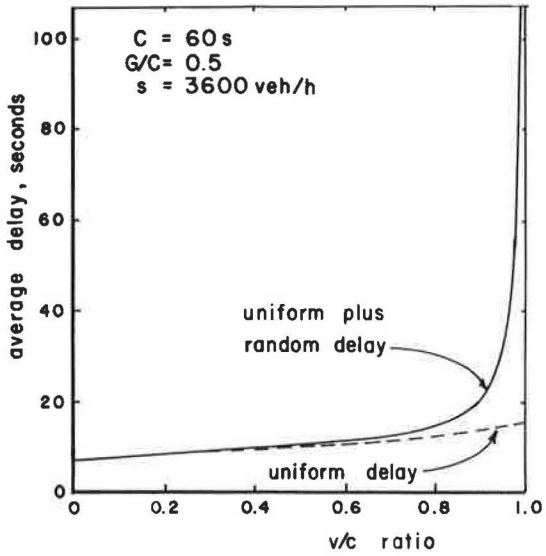

FIGURE 4 Delay predicted by a steady-state, stochastic model.

produce slightly different curves, but all steady-state stochastic models give curves with the general shape shown in Figure 4. In particular, any steady-state model that does not assume completely uniform arrivals will predict that the queue length, and therefore the delay, approach infinity as the v/c ratio approaches unity.

This is, of course, the reason that systems with a high v/c ratio take a long time to settle into a steady state; it simply takes a long time for such long queues to form, particularly since vehicles keep leaking away through the signal. As a result, one seldom sees real delays as large as those predicted for high v/c ratios by Equation 2. This discrepancy is not a result of faulty mathematics but of the unrealistic assumption that the system is in a steady state. If vehicles continued to arrive at a rate v nearly equal to the capacity c, the giant queues really would form, but in reality the peak period ends and v decreases long before a steady state is reached.

As a result, steady-state models are useful for predicting delays only at lightly loaded intersections. Some alternative models are available for the analysis of heavily loaded intersections, but they also have their shortcomings. The next model examined is at the opposite extreme from the steady-state models; it works well only if the intersection is solidly oversaturated (i.e., only if the arrival flow v is considerably greater than the capacity c for a significant length of time).

OVERSATURATION

Oversaturated systems are very easy to model (1). Suppose that a time varying arrival flow v(t) is either known or estimated. Then the cumulative arrival curve,

$$A(t) = \int_0^t v(s)ds$$

can be calculated and plotted as shown in Figure 5. [Traffic counters give A(t) directly, not v.] The dashed line in Figure 5 has slope (G/C)s and has been drawn tangent to A(t) at time zero, the time when the approach first becomes saturated. The departure curve D(t) has been drawn below, so that the dashed line divides the total area representing delay into two components: uniform delay corresponding to Equation 1 below the dashed line and overflow delay above.

The average uniform delay in this case is just half of the red interval, (1-G/C)C/2, as can be seen either by examination of Figure 5 or by substituting v = c = (g/C)s into equation 1. (In this substitution, v is not the arrival flow but the slope of the tops of the triangles.)

To obtain the average overflow delay, simply measure the area between A(t) and the dashed line--with a planimeter, by square-counting, or by any convenient method--to obtain the total delay, then divide by [A(T)-A(0)] to find the average. In practice, it is not necessary to draw the little triangles and it is usual to think of the dashed line as the departure curve, D(t), even though the real D(t) is the stairstep curve outlining the triangles.

This model ignores the effects of random variations. That is no problem when the queue is very large, because the error due to random effects will be small compared with the estimated delay, but this error can be a problem if the intersection is only slightly oversaturated. Thus, there is one group of models, the steady-state queueing models, that work well when v/c is considerably less than one and another type, the deterministic queueing model of Figure 5, that works well when v/c is considerably more than one. In between, there are problems.

Unfortunately, "in-between" is an area where good estimates could be very useful. At the extremes, it may be sufficient to say that operation is excellent or dreadful; in between, some numbers would be nice. (In truth, the ability to assign numerical values to different degrees of dreadfulness is important, but that is not the subject of this paper.)

IN-BETWEEN

As a first step toward the understanding of one model often used for v/c near unity (4,8-11), consider the special case of oversaturation delay shown in Figure 6. Figure 6a shows a very simple (and highly unrealistic) peak-period arrival pattern. Before time zero, the arrival flow is small enough that the overflow delay is negligible; during the peak from time zero to time T, it is equal to v; and after time T, there is some smaller, off-peak flow. Figure 6b shows the cumulative arrival and departure curves. The departure curve shown is not the true, saw-toothed departure curve but a straight line with slope c = (G/C)s like the dashed line in Figure 5; so the area of the triangle is the overflow delay not the total delay.

This area could be calculated with no great difficulty, but there is an easier way to obtain the average overflow delay. Begin by noting that a vehicle that arrives at time zero or τ experiences no overflow delay, but one that arrives at T must wait a time equal to the length of the horizontal dashed line near the center of Figure 6b. Between time zero and time T, the horizontal distance between the A(t) and D(t) curves varies linearly, so the average wait for those that arrive between zero and T is clearly half the length of the dashed line. The same is true for those that arrive between T and τ, so the average delay is half the length of the dashed line for both groups, hence for everyone who is caught in the overflow queue that exists from time zero until τ.

To calculate this average delay, first note that the length of the vertical dashed line is vT-cT =

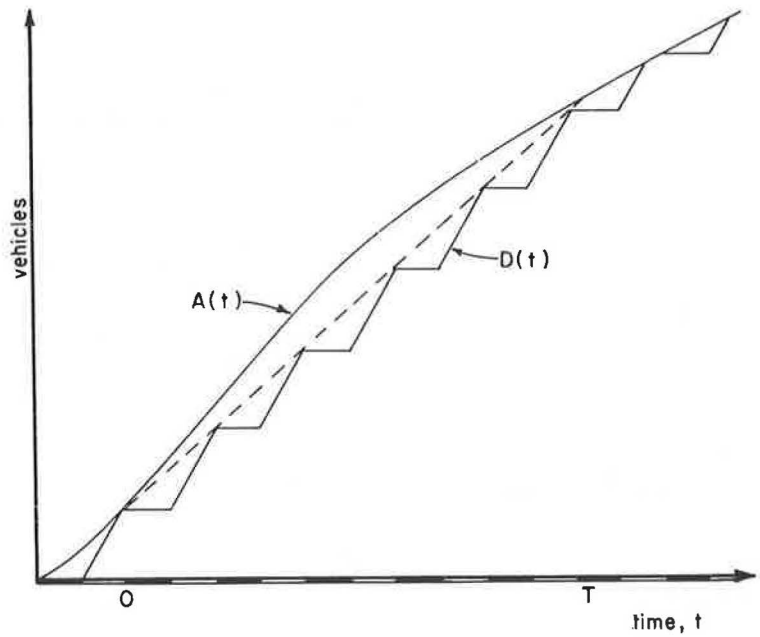

FIGURE 5 Arrival and departure curves for an oversaturated approach.

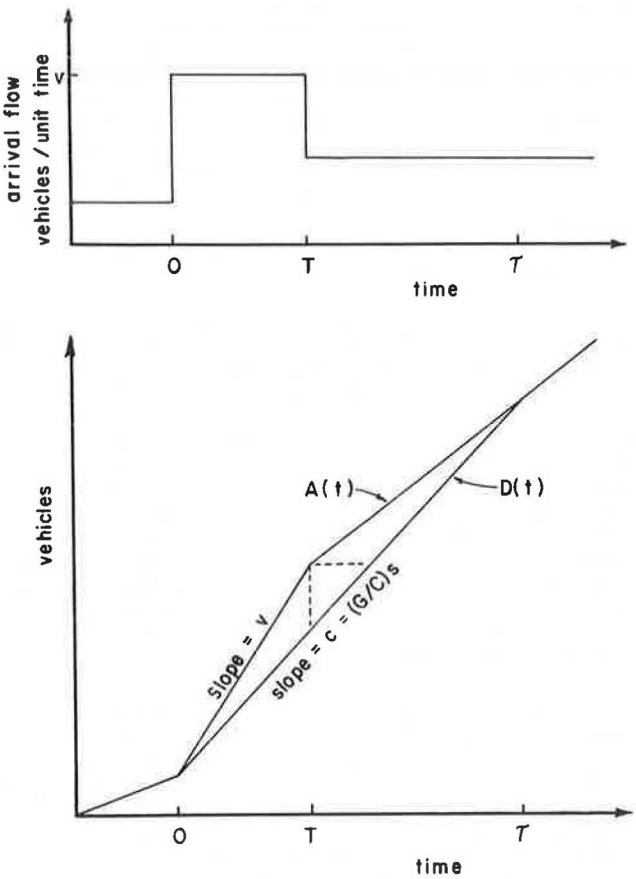

FIGURE 6 Overflow delay in a highly idealized peak period.

T(v-c). Then, because the length of the vertical dashed line is equal to the length of the horizontal dashed line times the slope of the D(t) curve, c, the horizontal dashed line must be T(v-c)/c long and the average overflow delay is

$$OD = (T/2)[(v-c)/c] = (T/2)[(v/c) - 1] \quad (3)$$

Note that the overflow delay is directly proportional to T, the length of time during which the arrival flow exceeds the capacity, and to [(v/c)-1], which is a measure of oversaturation.

Kimber and Hollis (10) and Akcelik (12) have used diagrams like Figure 7 to summarize what has been discussed so far. The solid curve on the left side, for v/c < 1, is the average overflow delay after a steady state is reached, as predicted by Equation 2. Because steady-state models predict infinitely long queues for saturated systems, this curve is asymptotic to the vertical line v/c = 1. As already noted, the enormous queues that cause these large delays do not develop at real intersections.

The solid curve on the right-hand side shows the average delay during time interval (0,τ) as predicted by the deterministic model of Figure 6 when T = 1 hr. Equation 3 indicates that this curve is a straight line with slope directly proportional to T. As was already noted, the average delay experienced by people who arrive during time interval (0,T) is also given by Equation 3 and the curve in Figure 7, but people who arrive right at time T have a delay equal to double this average.

It is reasonable to expect that both of the solid curves in Figure 7 would be good predictors of the overflow delay in the proper circumstances: the right-side, oversaturation model if the traffic really arrived as shown in Figure 6a with T of moderate length and v substantially larger than c; the left-side, steady-state model if v were substantially smaller than c or if T were very long. It is also reasonable to expect both models to deteriorate as the v/c ratio approaches unity.

Figure 7 shows clearly, however, that the problem when the v/c ratio is near unity is not just that neither model is at its best in this range, nor that one model predicts the average delay over a time interval and the other the average delay at the end of the interval. The problem is much more serious than that: the two models are utterly incompatible for v/c ratios near unity. One predicts that the overflow delay approaches infinity as v/c approaches

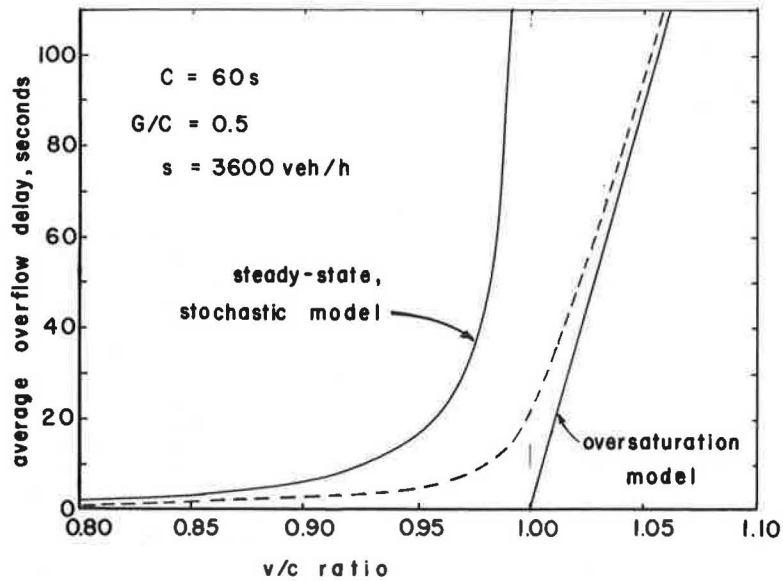

FIGURE 7 Comparison of steady-state and oversaturation models.

1 from below, but the other predicts that the delay approaches zero as v/c approaches 1 from above. A greater incompatibility is hard to imagine.

The problem, of course, lies with the assumptions. It is clear that the delay at v/c ratios slightly less than 1 will always be less than the steady-state models predict because the peak period never lasts long enough for the enormous steady-state queues to develop.

For v/c ratios greater than unity, it is not really possible to say that the real queues will always be longer than the deterministic model predicts, but surely they almost always will be. In this case, the reason is that real arrival flows do not suddenly jump from a low value to v as shown in Figure 6a but increase gradually. Therefore, because of random variations in the arrival flow and capacity, there will usually already be an overflow queue (i.e., a queue remaining at the end of green) at time zero (when the approach first becomes saturated). If this overflow queues is Q_o long at time zero, then the queue length in Figure 6b will be Q_o longer throughout time interval $(0,\tau)$ and the actual average delay will be about Q_o/c greater than the deterministic model predicts. For large v/c ratios, this error is a small fraction of the average delay, so it can be ignored, but for v/c ratios near 1, this error can be important.

Thus, the real delays are likely to be more as shown by the dashed curve in Figure 7, which agrees with the two solid curves at the edges of the figure but not in the middle. It is tempting to think that one could produce such a curve by fixing T and varying v in Figure 6a, but the assumptions of the two solid-line models are too incompatible for this to work. There would be a discontinuity at v/c = 1 because the two models average the delay in different ways. This problem could, of course, be overcome by changing the assumptions, but the effort would scarcely be worthwhile unless the shape of Figure 6a were also changed to something more realistic. It is better not to think of the dashed curve as a model of a queueing system but simply as a curve sketched quickly to illustrate some intuitive ideas.

In fact, however, the curve was not just sketched; it was calculated by the method developed by P.D. Whiting for version 6 of the Transport and Road Research Laboratory's popular computer program TRANSYT (8). This copyrighted computation method is perhaps better described as an algorithm, rather than a formula, because it requires only seven quite simple FORTRAN statements but spreads over most of a page when written as an equation. According to Robertson (9), the formula can be approximated by

$$OD = \{15T/c\}\ \{(v-c) + [(v-c)^2 + 240\ v/T]^{1/2}\} \qquad (4)$$

where OD is in seconds, T in minutes, and the flows v and c in vehicles per hour. The actual TRANSYT formula (10) is of the same general form but much more complicated and yields average delay per unit time rather than per vehicle. The curve in Figure 7 was obtained by dividing this result by the average number of vehicles served per unit time [either v or c = (G/C)s, whichever is smaller], but a formula for direct estimation of delay per vehicle is given by Kimber and Hollis (10).

Another much used version of this formula appears in the capacity analysis procedures developed in Australia by Akcelik (11):

$$OD = \begin{cases} (T/4)\left\{(x-1) + \left[(x-1)^2 + 12(x-x_o)/cT\right]^{1/2}\right\} & \text{if } x > x_o \\ 0 & \text{otherwise} \end{cases} \qquad (5)$$

where

$$x = v/c$$

and

$$x_o = 0.67 + sG/600 \qquad (6)$$

can be thought of as the smallest v/c ratio for which the random or overflow delay is large enough to be worth the effort required to calculate it. Unlike equation 4, Equation 5 can be used with any convenient time units, as long as they are consistent throughout the equation. (If the times are in seconds, then s and c should be in vehicles per second.)

The theoretical basis (12) for Equation 5 is similar to that of TRANSYT's random delay algorithm, so they can be regarded as essentially the same for

the purposes of this paper. As can be seen in Figure 8, the TRANSYT version generally seems to bend more sharply, hence produce lower delay estimates. The difference, however, is not always as large as that shown in Figure 8. Another minor, but obvious, difference between the two formulas is that the overflow delay predicted by TRANSYT increases smoothly over the entire range of v/c ratios, but Equation 5 jumps abruptly from zero to some small value when $v/c = x_o$. This both saves some unnecessary calculations and avoids the embarrassment of discovering the hard way that the upper branch of Equation 5 predicts negative overflow delays when v/c is small. (Note that x-1 is a negative number unless the intersection approach is oversaturated.)

A derivation of the TRANSYT random delay equation can be found elsewhere (10). The basic idea is really nothing more than what has already been said with respect to Figure 7, that the two solid curves should be good estimators when the v/c ratio is not close to unity, but that the real delay must be less than the steady-state models predict for v/c slightly less than 1 and more than the oversaturation model predicts for v/c slightly greater than 1. This line of intuitive reasoning leads to imagining a smooth curve that looks like the solid curve near the left side of Figure 7, then somehow cuts across the middle of the diagram and is asymptotic to the right-side, solid curve.

This is exactly what the dashed curve does. It is, in fact, exactly what it was designed to do. Whiting took a steady-state model developed by Robertson and bent it around by means of algebraic manipulation in such a way that the curve is no longer asymptotic to the vertical line, v/c = 1, but instead to the line predicted by the deterministic model of Figure 6b and Equation 3. Details of the transformation can be found elsewhere (10). The interest here, however, is not in the details but in the idea. Equations 4 and 5 and the series of statements in the TRANSYT program are simply mathematical expressions that accomplish the purpose of the dashed curve in Figure 7, provision of a smooth transition between the steady-state and oversaturation models in the region in which the two are inconsistent and likely to be unrealistic.

It is important to understand that the exact form of this transition is not the result of any detailed analysis of queue behavior but of the simple ideas expressed previously. The equations have one and only one justification: they provide a smooth transition in a way that satisfies intuitive ideas of what ought to happen. Kimber and Hollis describe the reliability of such formulas as follows: "In limiting cases . . . their results are correct, and in the intermediate regions, their functional behaviour is sensible" (10,p.11). They key word here is sensible. No claim is made that the formulas are correct; rather they yield answers that do not violate elementary logic in the troublesome region of v/c near unity where neither the steady-state nor the oversaturation models can be expected to yield reasonable results.

IS THERE A BETTER WAY?

For a practitioner needing an answer, Equation 5 or its TRANSYT equivalent obviously offers a better

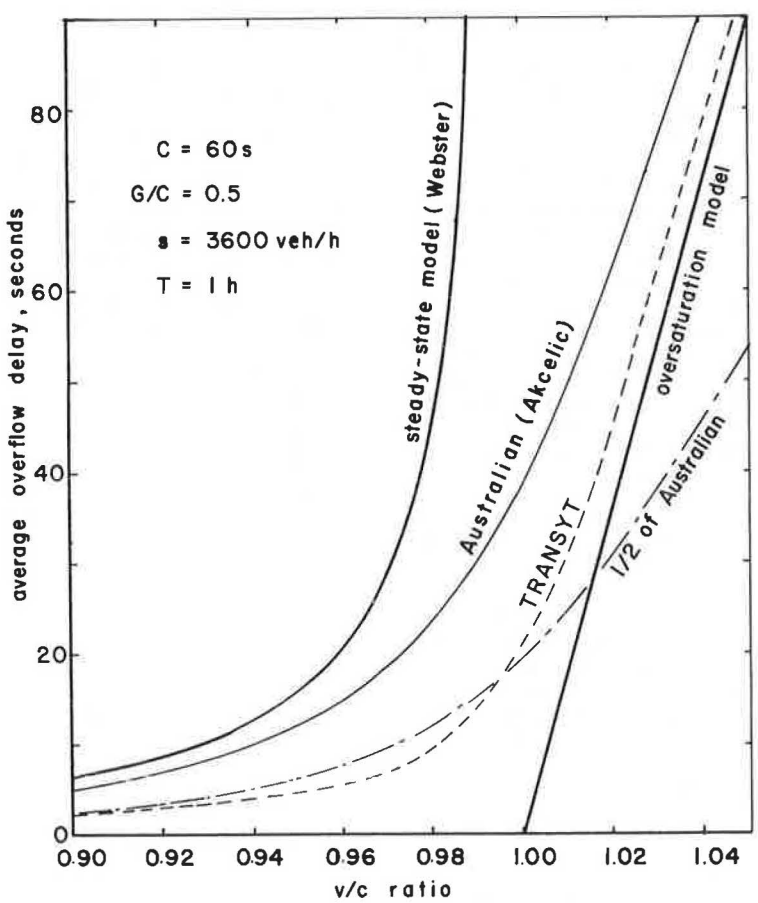

FIGURE 8 Comparison of the overflow delays predicted by five models.

solution for situations where the v/c ratio is near unity than any of the steady-state models. It will not necessarily give a good answer, but the competition offers only unreasonable answers. The choice is easy, but one really would like to be on safer ground.

The road to safer ground has two branches. The first, and obviously easier, starts by noting that the exact form of transition provided by Equations 4 and 5 is arbitrary; the dashed curve must agree with the solid curves at the edges of Figure 7, but it is clear that it could have many shapes in the middle. Therefore, an obvious thing to do is to measure delays in the field and compare then with the predictions. If there is a consistent direction of error, then maybe another transition curve can be found that is equally arbitrary but works better.

This is the approach followed by W.R. Reilly et al. (2,3). Their measurements of delays on undersaturated intersection approaches indicated that Equation 5 consistently overestimated the overflow delay, particularly at v/c ratios approaching unity (3). Therefore, they recommended a new, but very similar, overflow delay equation:

$$OD = 346.2\ T\ \{(x-1) + [(x-1)^2 + 12(x-OF)/cT]^{1/2}\} \quad (7)$$

where

$$OF = 0.67 + Cc/2{,}160{,}000 \quad (8)$$

This differs from Equation 5 in three ways, only one of which is important. The first difference is one of notation and units: OF is what has been called x_o here, and a factor of 3,600 has been introduced in both Equations 7 and 8 so that the delay and the cycle length C can be in seconds, but T is in hours and s is in vehicles per hour.

The second difference is that the equation has been divided by 1.3 to yield stopped delay. Theoretical and simulation models give estimates of the total time lost due to the presence of the signal, which is greater than the amount of time spent actually stopped. Field studies, on the other hand, often yield stopped delay, because it is easier to measure. According to Reilly et al. (13), the total delay can be estimated by multiplying the measured stopped delay by 1.3; so Equation 5 had to be divided by 1.3 to yield stopped delay.

The third difference is that the delay has also been divided by 2 to give results more nearly in agreement with the field measurements. Thus, the 346.2 T in Equation 7 comes from the T/4 of Equation 5 as follows:

$$(T/4) \times (1/1.3) \times (3{,}600\ \text{sec/hr}) \times (1/2) = 346.2\ T$$

The only important difference is the factor of 2.

The effect of this halving is to produce a new curve--shown by a broken line in Figure 8--that is lower than the one based on Equation 5 but no longer asymptotic to the straight line predicted by the oversaturation model. Reilly et al. (2) recommend that Equation 7 be used only for $v/c \leq 1.10$ and the oversaturation model for larger v/c ratios. Thus, there is a discontinuity in the delay prediction at v/c = 1.10 and a tendency to underestimate the delay when $1.00 < v/c < 1.10$, but outside of this range the overall pattern of prediction is likely to be rather similar to that of the TRANSYT model.

The hope is that Equation 7 gives better predictions than Equation 5. The price paid for this increase in accuracy includes the discontinuity, the somewhat illogical behavior for $1.00 < v/c < 1.10$ where the broken curve of Equation 7 may lie below the lower bound set by the oversaturation model, and the loss of the theoretical basis for the form of the equation--a form dictated by the abandoned requirement that the prediction equation be asymptotic to the solid line on the right side of Figure 8.

As an alternative to dividing Equation 5 by 2 or some other factor, one could try to change its shape to improve the predictions. There would seem to be two obvious goals for such an attempt: either to find a new curve that is once more asymptotic to the solid line or, if this goal is abandoned, to find a curve with a simpler formula. Researchers taking the latter path, however, will have to grapple with the problem of finding a new functional form that responds to changes in v/c, c, and T in a reasonable way. For those who would prefer to stay with the basic logic that led to Equation 5 but somehow change the curve's shape, two approaches offer hope of quick and easy results. One of these is to return to the TRANSYT formulation, which seems to agree fairly well with Equation 7 for $v/c \leq 1$. A second is to change the value of T; Equation 5 with T = 15 min gives delay estimates quite similar to Equation 7 with T = 1 hr. For $v/c < 1$, these two estimates are very similar to the TRANSYT estimate, but for v/c much greater than unity, they predict only about half as much delay, a worrisome difference. As will be discussed later, T is a very important parameter, but it is difficult to evaluate.

QUEUEING THEORY APPROACH

The second road to improvement is a modeling road; rather than simply trying to produce a reasonable curve, one can try to model actual queue behavior. This road offers more potential for good answers, but it also presents many difficulties. To be an improvement over what is now available, the model cannot assume a steady-state situation but must account for the way the queue length varies over time--its transient behavior.

Unfortunately, the modeling of transient queue behavior is not a very well-developed science and the models that are available tend to be complicated. On the other hand, more can be done than usually is. Two possible approaches are discussed by Kimber and Hollis (10) and in the last few chapters of Newell's Applications of Queueing Theory (1). Many readers of this paper will probably find both tough going, but they really can be read by graduate engineers armed with only one or two probability courses--a statement that cannot be made about most of the available literature on this subject.

Kimber and Hollis (10) propose a model in which fairly short time intervals are considered sequentially, the results from one forming the starting condition for the next. Newell (1), on the other hand, develops partial differential equations to describe the approximate behavior of the queue length distribution as it changes with time. What both modeling approaches make very clear is that the development of the queue is very dependent on the details of the arrival pattern. For example, if the numbers of arrivals in each 15 min of the peak hour at a saturated or nearly saturated approach are 200, 500, 300, and 200, the peak-hour volume and peak-hour factor will be just the same as with volumes of 200, 300, 500, and 200, but the average delay will be different because of the dynamics of queue growth and decay. If this difference is to be reflected in a model, it is clear that more information about arrival patterns must be provided than is now customary.

Because this paper is directed more to users of models than to builders of models, this is the really important lesson to be learned from the examination of non-steady-state models. If good estimates of signalized intersection delay are desired, better input to the models must be provided. Until that is done and until practical models are developed to use that input, it is not reasonable to expect delay estimation to be a very accurate process.

DIFFICULTIES IN APPLICATION

Returning to Equations 5 and 7, it should by now be clear that the most difficult problem facing an analyst willing to accept their reasonableness is the choice of the two variables describing the traffic arrival pattern, v and T. The definitions of these are given in Figure 6a: vehicles arrive at flow rate v during a period exactly T long. This, of course, is not the way real traffic arrives, and it is not true that all intervals T long in which vT vehicles arrive will have the same average delay. In fact, the delay will depend very much on the flow variations within the time interval. How, then, does one choose v and T?

The available literature seems to indicate that T should usually be 1 hr and v either the average flow during that hour (6,11) or the average within the busiest 15 min of that hour (2). Nobody, however, gives much in the way of reasons. [A partial exception is Reilly et al. (2). Equation 7 is based in part on data and its originators suggest, logically enough, that users select v and T the same way they did when they chose the equation.] Reading between the lines, it seems that nobody is very sure about this matter.

This is not surprising. Theoretical considerations indicate that there is no good answer, so those who must advise are quite correct--if not very helpful--in sounding none too confident about their advice. Nonetheless, the users are stuck with the fact that even if they cannot find a good answer, it may be possible to find a particularly bad one.

SIGNAL COORDINATION

Everything said so far has been based on the implicit assumption that the signal approach is isolated, so that the arrival pattern is at least more or less uniform. In many cases, however, the arriving vehicles are the output stream from an upstream signal. In that case, the cumulative arrival curve will not be smooth, as in Figure 1, but saw-toothed like the departure curve. In such cases, the area between the arrival and departure curves will obviously be very different if the signals are well coordinated--in which case, the area should almost vanish--than if each platoon of arrivals neatly hits a red light.

Two of the procedures discussed previously try to account for this, but in different ways. The TRANSYT model's main purpose is to achieve coordination, so it takes a very detailed approach in which the actual arrival curve is predicted. The estimate of the nonrandom delay is based on this predicted arrival curve rather than on Equation 1. The random or overflow delay estimate does not depend on the coordination.

Reilly et al. (2), on the other hand, recommend that the sum of the uniform and overflow delays predicted by Equations 1 and 7 be multiplied by an adjustment factor based on the degree of coordination (5 categories from poor to very good) and on whether the signal is pretimed, semiactuated, or fully actuated. Adjustment factors range from 0.65 to 1.50.

Obviously the detailed analysis in TRANSYT is the better approach, and just as obviously it is not always a practical one. Its advantages are likely to be realized only when good data are available, and it is really oriented to signal timing and large system design rather than to the usual design situation in which only one or a few intersections are to be designed and the upstream intersections may not even be planned, let alone timed. It is also a computer model and its methods are not always practical for hand calculation. Another difference worthy of note is that the two methods make very different assumptions about the effect of coordination on overflow delay.

SUMMARY

The primary purpose of this paper has been to introduce new users of delay models to the background of some of the models and the assumptions on which they are based. This seemed necessary because the information is widely scattered, some of it in material not readily available to the average traffic engineer. Along the way, it has been noted that the methods currently used either ignore the way in which delays vary with time or try to cope with the variation in ways that are more mathematical applications of common sense than mathematical models of traffic signal systems. The methods that ignore the variation are nearly useless for the most interesting traffic engineering problems, those where the signal system operates close to capacity, but the others seem reasonable enough to be useful as long as one remembers their limitations.

None of the models examined here can be expected to give really consistent and accurate results. To obtain such results, one would need not just better models but better information about traffic patterns. For some purposes it may be unrealistic to suppose that such information will be available. For detailed operational studies, however, it could be obtained. If estimated delay is to be a good indicator of level of service in such studies, the information will have to be obtained and the models developed to use it. This is not an impossible task.

ACKNOWLEDGMENT

This work was supported in part by the National Science and Engineering Research Council of Canada.

REFERENCES

1. G.F. Newell. Applications of Queueing Theory, 2nd ed. Chapman and Hall, London, England, 1982.
2. JHK & Associates and the Traffic Institute, Northwestern University. Urban Signalized Intersection Capacity. Draft Final Report, NCHRP Project 3-28(2). TRB, National Research Council, Washington, D.C., Feb. 1983.
3. JHK & Associates. Urban Signalized Intersection Capacity. Uncorrected Draft, NCHRP Project 3-83(2). TRB, National Research Council, Washington, D.C., 1982.
4. R.E. Allsop. Delay at a Fixed Time Traffic Signal--I: Theoretical Analysis. Transportation Science, Vol. 6, 1972, pp. 260-285.
5. K. Ohno. Computational Algorithm for a Fixed Cycle Traffic Signal and New Approximate Ex-

pressions for Average Delay. Transportation Science, Vol. 12, 1978, pp. 29-47.
6. F.V. Webster. Traffic Signal Settings. Road Research Technical Paper 39. Her Majesty's Stationary Office, London, England, 1958.
7. Road Research Laboratory. Research on Road Traffic. Her Majesty's Stationary Office, London, England, 1965.
8. D.I. Robertson and P. Gower. User Guide to TRANSYT Version 6. Supplementary Report 255. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1977.
9. D.I. Robertson. Traffic Models and Optimum Strategies of Control--A Review. Proc., International Symposium on Traffic Control Systems, edited by W.S. Homburger and L. Steinman, Vol. 1, Berkeley, Calif., 1979, pp. 262-288.
10. R.M. Kimber and E.M. Hollis. Traffic Queues and Delays at Road Junctions. Laboratory Report 909. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1979.
11. R. Akcelik. Traffic Signals: Capacity and Timing Analysis. Research Report 123. Australian Road Research Board, Victoria, 1981.
12. R. Akcelik. Time-Dependent Expressions for Delay, Stop Rate and Queue Length at Traffic Signals. Internal Report AIR 367-1. Australian Road Research Board, Numawading, Victoria, 1980.
13. W.R. Reilly, C.C. Gardner, and J.H. Kell. A Technique for Measurement of Delay at Intersections. Report FHWA-RD-76-137. FHWA, U.S. Department of Transportation, 1976.

Publication of this paper sponsored by Committee on Traffic Flow Theory and Characteristics.

Automated Collection of Vehicular Delay Data at Intersections

JAY F. LEGERE and A. ESSAM RADWAN

ABSTRACT

Most current methods used to estimate vehicular delay at intersections involve some form of manual data collection. These methods rely on statistical techniques (such as multiple linear regression) to improve the accuracy of the delay estimates. In addition, most require significant data collection and reduction efforts. The theory, design, operation, and evaluation of a microprocessor-based system for the collection of vehicular delay data at intersections is presented. The principle of the automated system, including definitions of pertinent variables and equations, is discussed. An overview of the system design, including a description of the vehicle detection scheme and the microprocessor's recognition of vehicle arrivals and departures, is presented. There is a discussion of the system software as well as a description of the data collection and reduction processes. The system performance was evaluated both in the laboratory and through analysis of data collected in the field. Recommendations for further development of the device are presented.

Since the Arab oil embargo of 1973, the United States has become increasingly concerned with its energy supply and with ways in which that supply can be conserved. One area of particular interest is the conservation of energy within the transportation sector. About 40 percent of this nation's petroleum consumption is attributable to passenger travel by automobile.

Much traffic engineering research has been done on delay and fuel consumption at signalized intersections simply because they are considered the locations where most delay and excess fuel consumption occur. Unfortunately, the most accurate methods of data collection and analysis have proven to be extremely time consuming and costly.

Estimates of intersection delay are used in numerous applications, some of which are validation and calibration of computer simulation models, estimates of road-user costs, before-and-after studies, comparisons of the efficiency of various types of intersection control, and comparison of specific signal timing and phasing.

The theory, design, implementation, and evaluation of a microprocessor-based system for the collection of vehicular delay data at intersections are presented here. The primary application of the system is for the collection of data at intersections that are under some form of signalized control. The system is also applicable to any intersection or, in general, to any section of highway for which values of average travel time and delay are desired. Details of the hardware specifications, the software routines, and the assembly language program for this application are fully documented elsewhere (1).