Event-Based Short-Term Traffic Flow Prediction Model

K. Larry Head

The problem of predicting traffic flow for the purpose of real-time traffic-adaptive signal control in an urban street network is explored. A prediction model is described that combines data from traditional vehicle loop detectors and known relationships from traffic flow theory. The model is demonstrated using a microscopic traffic simulation model. Results of the simulation demonstrate that the model can provide the information required to develop truly proactive real-time traffic-adaptive signal control.

One of the greatest challenges to the development of real-time traffic-adaptive signal control is the prediction of traffic flows on the network and the relationship of these flows to the traffic control signal settings.

The need for prediction was recognized in the development of the UTCS system in the early 1970s. The development of second-generation (UTCS-2) and third-generation (UTCS-3) control logic included prediction as a primary system component (7). UTCS-2 based its signal timing decisions on predictions of demand for the next 5 to 15 min. UTCS-3 based its signal timing on predictions of demand over much shorter periods of approximately a cycle length, although UTCS-3 logic was not based on fixed cycle length.

For real-time traffic-adaptive signal control logic to be effective, it must have an accurate view of the state of traffic conditions on the network and be able to predict, at least over short periods, how the current network conditions will evolve. The importance of the temporal distribution of information in the prediction can be understood by considering the signal timing problem given two possible arrival patterns during the planning horizon as depicted in Figure 1.

Each arrival pattern represents a flow profile where the magnitude of the profile represents the number of vehicles to arrive at an intersection in a fixed time interval. (For the purpose of this discussion, the time intervals should be considered to be 1 or 2 sec in length.) Both arrival patterns are identical until time \( t_0 \), when the signal control logic is required to decide whether to serve this or another approach. There is significantly more demand immediately following \( t_0 \) in the upper flow profile than in the lower during the same time interval. In each case the number of vehicle arrivals over the time horizon shown is equal, but the control decision should be different. It is of fundamental importance to know the temporal arrival distribution to build a truly real-time traffic-adaptive signal control logic.

This paper explores the issues and problems of generating the necessary traffic flow predictions to allow the development of proactive real-time traffic-adaptive signal control logic. In the following section, the flow prediction problem is addressed and several relevant issues are discussed. Then an event-based short-term traffic flow prediction model is presented and followed by a simulation-based demonstration of the model's capabilities.

FLOW PREDICTION

Three issues are important to predicting traffic flow: (a) length of the prediction time horizon; (b) number of prediction points per time horizon, called the prediction frequency; and (c) number and location of information sources used in making the prediction. The prediction time horizon provides the real-time traffic-adaptive signal control logic with the ability to plan future signal timing decisions. If the prediction horizon is short, perhaps several seconds, then the signal timing decisions are restricted. For example, if the predictions are made over a 10-sec horizon, the signal timing logic can only make timing decisions that extend or shorten the current phase. Actuated signal control logic operates in this mode. If the predictions are made over a longer horizon, the signal timing decisions can include decisions on phase termination times and phase sequencing. For example, if the prediction horizon is 30 to 40 sec, the signal timing logic might schedule the next two or three phases and their durations on the basis of predicted demand.

The prediction frequency provides information about the distribution of vehicle arrivals over time. If the predictions are made at a frequency of only one prediction for the decision time horizon, then the signal timing logic must assume that the vehicles are distributed uniformly over that time. If the predictions are made more frequently—say, 10 to 30 times over the prediction horizon—then the signal timing logic will have a more accurate representation of the distribution of vehicle arrivals over time. Figure 2 depicts the information content of predictions at a frequency of once (dashed) and 10 times (solid) per horizon.

Traffic flow is, in general, a time-space phenomenon. The number and location of information sources determine the ability of any prediction algorithm to predict conditions on the basis of current conditions at related spatial locations. For example, if a detector is located 10-sec upstream of the desired prediction point, then prediction will be easier but only for a 10-sec horizon. The farther away the location of other information sources, the longer the potential prediction horizon. But the temporal information may become more distorted (e.g., platoon dispersion) and thus less valuable for prediction. In addition, the farther away the information sources, the greater the effects of exogenous factors, such as traffic signals and traffic sources/sinks. There is a trade-off between the distance between information sources and prediction accuracy. A system with many well-placed detectors will give the best prediction information, but the cost of such a system may be prohibitive.

Stephanedes et al. (2) conducted a critical review of the UTCS predictors and three other demand predictors. They com-
pared the prediction accuracy of UTCS-2, UTCS-3, historical averages, current measurement, and a new algorithm proposed by the authors. The proposed predictor had a parametric form similar to a proportional-integral-differential (PID) controller.

Each predictor was compared using mean squared error (MSE) and mean absolute error (MAE) for 5-min predictions and cycle-based predictions. It was concluded that for 5-min predictions, the historical average performed better than UTCS-2, and that both predictors were superior to the others. For cycle-by-cycle comparisons, the UTCS-2 and the historical average predictors were not applicable since synchronization of cycles over historical periods was impossible. A moving average version of the PID predictor was superior to the UTCS-3 and current measurements. Some versions of the proposed PID predictor performed better than the moving average version, but the performance was sensitive to the selection of the model parameters.

Each algorithm that Stephanedes et al. studied addresses the prediction problem on the basis of a fixed time horizon, either five min or one cycle, and updates the prediction at a frequency of only once per horizon. Table 1 gives a summary of each of these algorithms in terms of its characteristics: prediction horizon, prediction frequency, number and location of information sources, and performance.

Okatani and Stephanedes (3) used a Kalman filter model structure to consider information from multiple sources (i.e., detectors on a number of links). They made predictions at a frequency of once per 15-sec time horizon. Their results indicate a substantial improvement over the UTCS-2 prediction algorithm but fail to address the need for higher-frequency predictions as required for real-time traffic-adaptive signal control logic.

In a discussion of the prediction problem, Gartner (4) concluded that the deficiency in providing good temporally distributed predictions could be addressed by relying on actual flows rather than average volumes. A possible method for obtaining actual flows would be to place detectors on the links upstream from the intersection and use the flows at these points to provide predictions. This approach has been adopted by several real-time signal control systems, including SCOOT (5), OPAC (6), and UTOPIA (7,8). A major limitation of this approach is that the distance between the intersection and the upstream detector could constrain the prediction time horizon.

Another approach, one used in SCATS (9), is to locate the detectors at the stop bars of the upstream intersection and use the departure profiles together with a dispersion factor to predict the downstream arrivals. This approach allows the effect of the upstream signal to be included in the prediction.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Characteristic</th>
<th>Horizon</th>
<th>Frequency</th>
<th>Sources</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTCS-2</td>
<td></td>
<td>5-15 min</td>
<td>1</td>
<td>Single</td>
<td>Best for 5 min</td>
</tr>
<tr>
<td>UTCS-3</td>
<td></td>
<td>5-15 min, cycle</td>
<td>1, 1</td>
<td>Single</td>
<td>Poor Overall</td>
</tr>
<tr>
<td>Historical Average</td>
<td></td>
<td>5-15 min</td>
<td>1</td>
<td>Single</td>
<td>Best for cycle</td>
</tr>
<tr>
<td>Current Measurement</td>
<td></td>
<td>5-15 min, cycle</td>
<td>1, 1</td>
<td>Single</td>
<td>Poor due to time delay</td>
</tr>
<tr>
<td>PID</td>
<td></td>
<td>5-15 min, cycle</td>
<td>1, 1</td>
<td>Single</td>
<td>Sensitive to Parameters</td>
</tr>
</tbody>
</table>

*The notation 1,1 refers to the frequency based on the horizon, e.g. the UTCS-3 algorithm was evaluated with a 5-15 min horizon and a cycle based horizon. The frequency of each was 1 prediction per horizon.*
PREDICTION MODEL

The prediction model presented here is based on the use of detectors on the approach of each upstream intersection, together with the traffic state (arrivals and queues), and the control plan for the upstream signals to predict future arrivals. The model is data-driven and combines actual traffic detector data with traffic flow theory.

The prediction scenario geometry is depicted in Figure 3. It is desired to predict the flow approaching intersection A at detector $d_A$, where the actual flow can be measured; hence, the quality of the prediction can be assessed in real-time. The prediction of each arrival at the downstream intersection depends on the event of a vehicle crossing one of the upstream detectors and not (directly) on the traditional detection parameters of count and occupancy at a single detector.

Consider the process of arrivals at an intersection, as observed at a detector, as a sequence of observations $\{n(t)\}_{t=1}^{\infty}$ with $n(t)$ representing the number of vehicle arrivals during time interval $t$. It is assumed that at any time $t$, $n(t) = 0, 1, 2, \ldots$ and depends on the number of lanes and length of the time interval. The prediction model assumes that this arrival process can be divided into two parts—a predictable part and an unpredictable part—hence,

$$n(t) = n_p(t) + n_u(t)$$

where $n_p(t)$ represents the predictable part and $n_u(t)$, the unpredictable part. From a traffic engineering perspective, the unpredictable part of the arrival process may result from sources and sinks such as parking lots, garages, shops, and on-street parking.

If several sources or sinks affect the arrival process—that is, if the contribution of $n_u(t)$ is significant—the control strategy at the intersection probably will be different than if the arrival process is highly predictable. For example, if the process is highly predictable, the control strategy could be to allow platoon progression (assuming platoons exist in the flow). If the process is highly unpredictable, then the control strategy could be to gather arrivals into platoons that can be predicted or accommodated at downstream intersections.

A possible measure of predictability may be defined similarly to the signal-to-noise-ratio (SNR) familiar to signal processing and communication engineers:

$$SNR = \frac{\int_{t=1}^{t=T} n_p(t) dt}{\int_{t=1}^{t=T} n_p(t) + n_u(t) dt}$$

where $T$ is the prediction horizon. Intuitively, if $SNR \approx 1$, the predictable part of the process is dominant; if $SNR << 1$, the unpredictable part of the process is dominant. When $SNR = 0.5$, there are approximately equal volumes from each process.

The rest of this paper is concentrated on the predictable part of the arrival process. However, it is noted that if the unpredictable arrival process contributes significantly to the actual arrival process, the model presented here may not be the best choice for prediction. In cases in which the unpredictable process dominates, additional roadway detectorization may be required to observe the traffic flows that contribute to the unpredictable component of the process.

Traffic contributing to the predictable arrival process, or traffic flow, at $d_A$ originates from the approaches to intersection $B$ and can be measured at detectors $d_1$, $d_2$, and $d_3$, which represent the flows that will turn left, pass through, and turn right, respectively. Consider the event of a vehicle crossing a detector, say, $d$, where $i \in \{1, 2, r\}$, at time $t_c$. Let this event be denoted $e_i(t_c)$.

Several factors affect when and if the vehicle will arrive at $d_A$, including:

- Travel time from $d$ to the stop bar at intersection $B$.
- Delay due to an existing queue at $B$.
- Delay due to the traffic signal at $B$.
- Travel time between $B$ and $d_A$ and
- Probability the vehicle will travel along a route that includes location $d_A$.

$\text{FIGURE 3 Geometric layout of prediction scenario.}$
In Figure 4c, the arrival at \( d_i \) encounters delay for the signal as well as a standing queue and must travel from \( d_i \) to the stop bar at \( B \) and from the stop bar to \( d_i \). The travel time is defined as

\[
t_a = t_{d_i} + \max\{T_{d_i, s_B}, T_{q_i}\} + T_{s_B, d_i}
\]

The delay due to the standing queue, \( T_{q_i} \), can be estimated using a relationship of the form

\[
T_{q_i} = a_0 + a_1 N_{q_i}
\]

where \( a_0 \) and \( a_1 \) are parameters that can be selected on the basis of the particular intersection and \( N_{q_i} \) is the number of vehicles in the queue \((10)\). Equation 6 has the form of the Greenshields equation and has been used to estimate the amount of time required to clear a queue.

Equation 6 assumes some knowledge of the number of vehicles in the queue, \( N_{q_i} \). Since current detection technology does not provide this as a direct traffic measurement, it must be estimated. Baras et al. \((11)\) investigated a point process–based estimation/prediction filter that provides this information. In this paper the authors have compared the Baras et al. filter with a simple counting estimator and have found that a simple counting estimator provides reasonably accurate information for prediction and requires considerably less computational effort in the process.

Figure 4d depicts the case when the arrival at \( d_i \) occurs after the signal has begun serving the desired phase, but a standing queue is present. In this case the prediction time is

\[
t_a = t_{d_i} + \max\{T_{d_i, s_B}, T_{q_i}\} + T_{s_B, d_i}
\]

This case is similar to Equation 5, except that the delay due to the standing queue must be adjusted using the amount of time that has elapsed between the onset of the signal and the arrival of the vehicle at \( d_i \) and the travel time to the back of the queue. Equation 5 captures this relationship accurately.

Recall that the prediction of the downstream arrival was initiated on the basis of the event, \( e(t_d) \), of a vehicle crossing an upstream detector. Given this estimate of the predicted arrival time, an arrival event at intersection \( A \), at detector \( d_A \), can be anticipated with probability \( p_{BA} \). This probability reflects the uncertainty that vehicle crossing the upstream detector will actually travel on a route that will cross the detector at \( d_A \).

This uncertainty, along with the possibility of multiple lanes or time intervals in which more than a single vehicle may cross one of the upstream detectors, can be incorporated into the model by predicting the expected number of arrivals at \( d_A \) instead of a single arrival event. If \( n_i(t_d) \) vehicles cross detector \( d_i \) in time interval \( t_d \), then using Equations 3–7, the expected number of arrivals at \( d_A \) at time \( t_{d_A} \) can be predicted to be

\[
\tilde{n}_A(t_{d_A}) = \sum_{i \in \{L,R\}} \sum_{t_p = t_{d_i}} p_{BA} n_i(t_p)
\]

The inner summation estimates that the expected number of arrivals at \( d_i \) predict that these arrivals will occur at future time \( t_p = t_{d_i} \) for movement \( i \). The outer summation is over each of the movements feeding link \( BA \).

From an operational algorithmic point of view, the model can be implemented by maintaining a data base table that is updated each time an event occurs on one of the upstream approaches. In this manner the prediction at \( d_A \) evolves as the information becomes available.

Several operational issues, such as right turn on red and permitted left turns, have not been addressed directly. These factors can greatly affect the predictions and can be incorporated into the
model. For example, right turn on red can be incorporated easily by conditioning the probability of a vehicle making a right turn on the signal state and the opposing volume. If the signal is in a red state, there are vehicles queued that may make a right turn, and a gap is observed in the opposing flow, then a right turn can be predicted. In this case the queue size estimate can be adjusted accordingly and an arrival at the downstream intersection can be predicted. A similar enhancement can be made for permitted left turns. These factors have been included in the simulation study discussed in the following section.

EXAMPLE

The prediction model just presented was implemented as part of a research effort to develop real-time traffic-adaptive signal control logic. The signal control logic is a hierarchical-distributed logic called RHODES (12). The prediction model was implemented as part of the intersection control logic within the RHODES hierarchy and was used to evaluate the performance of the RHODES intersection control logic at a single intersection using computer simulation.

A computer simulation was developed using a modified version of the TRAF-NETSIM traffic simulation model developed by FHWA (13). The simulation model was modified to support external real-time traffic-adaptive signal control logic by passing surveillance data from the simulation and accepting signal-state control decision inputs on a second-by-second basis.

The simulated traffic network was based on an actual network in Tucson, Arizona. Actual signal timing plans, detector locations, traffic volumes, and turning percentages were used as the basis for the simulation. The simulated network consisted of 28 intersections, although the prediction and control algorithms were applied to a single intersection. It was necessary to simulate the area surrounding the intersection of interest to ensure realistic traffic flows since the current version of TRAF-NETSIM has limited traffic generation capabilities. The simulation model did not include any interlink sources and sinks, which provides the most desirable environment for the prediction model to be successful.

The geometric scenario for collecting the following data is as shown in Figure 3. Nodes A and B are located 716.5 m (2,350 ft) apart. Each of the upstream detectors is located 39.65 m (130 ft) from Node B. The through approach is three lanes plus a left-turn pocket. Each side street approach consists of two lanes plus a shared turning lane. Detector $d_A$ is located 152.5 m (500 ft) upstream of Intersection A. For the purposes of this study, the simulation was run for 1,170 sec, of which 400 sec were used to allow the network to reach equilibrium. Since the prediction model does not require steady-state conditions, all data, both transient and steady-state, are collected for analysis.

Figure 5 shows a plot of actual versus predicted travel time for vehicles traveling along a route that crosses detector $d_A$. The plot shows the ability of the prediction model to estimate actual travel time. The general trend in the plot is along the line $y = x$, which is the perfect prediction line.

The scatter in the plot is due to several factors, including the natural stochastic variations in travel times. The model produces significant errors in two areas. The first is when the actual travel time is long but the predicted travel time is short, the other is when the actual travel time is short and the predicted travel time is long. Further investigation of these errors shows that each occurs at the end or beginning of a signal phase. If the phase ends before a vehicle crosses the stop bar, but the prediction model expected the vehicle to clear the intersection, the predicted travel time will be much shorter than the actual travel time. Similarly, if a vehicle passes through the intersection but the prediction model expected the vehicle to stop, a significant error will occur.

Figures 6a–c show a plot of the actual and predicted travel times as a function of the time when the prediction was made. Each figure also includes the signal control state when vehicle movements are permitted for the associated approach. It is important to note that there are more predictions than actual travel because of the probabilistic nature of vehicles traveling along a route that crosses the downstream detector. This is most apparent in Figure 6a, where there are relatively few left-turning vehicles. There are no actual travel times reported from approximately Time 1100 to Time 1170, since the simulation terminated before the generating vehicles completed their trips.

Close examination of Figure 6 shows that the prediction model exhibits the same temporal behavior as the actual travel process. This is especially evident in Figure 6b. During the period when the signal is red, the travel times are long. As the green phase nears, the travel times become shorter until eventually the queue has dispersed and vehicles flow freely through the intersection.

Figure 6c shows the highly variable behavior of the right-turning vehicles. This variability is due primarily to right turn on red behavior.

Figure 7 shows actual and predicted flow profiles, $\hat{n}_s(t_{d_A})$, at $d_A$ as a function of time. To capture the "flow profile" characteristic, the cumulative number of arrivals in a 5-sec time interval are shown. Only a portion of the total 1,170 sec of simulation time is shown.

The performance of the prediction model can be assessed quantitatively by examining the prediction error statistics. Standard forecasting/prediction measures include mean error (ME), sum of the squared errors (SSE), MSE, MAE, and mean absolute relative error (MARE) (14). Since there are time instances where no actual arrivals occur, the MARE measure was modified to include only the
absolute error and not the infinite relative error that would result from dividing by the 0-valued number of observed arrivals. Table 2 gives these measures for the simulation experiment that generated the flow profile shown in Figure 7. In addition to these descriptive statistics, a Durbin-Watson (D-W) statistic is reported in Table 2. The D-W statistic measures the existence of any pattern in the prediction errors. If the D-W statistic is near 2, as it is in this experiment, the errors are essentially random.

Figure 8 shows a histogram of the errors with each error cell taken to cover a range of 0.5 and the center point shown as the cell label. Note that all 0-valued errors are included in the range from 0 to 0.5. In addition, the cumulative frequency is shown on Figure 8. It is interesting that most of the second-by-second prediction errors are within a single vehicle.

These descriptive statistics are valuable primarily when one or more models are to be compared and have limited value alone. Since this type of high-frequency prediction is new, no other models can be used for comparison. In fact, it is not known how sensitive real-time traffic-adaptive signal control will be to informational errors. As Gartner noted (4), the true test for a prediction model is its ability to work with signal control logic to improve traffic performance. The prediction model presented in this paper has been coupled with a dynamic programming-based traffic-adaptive intersection control optimization algorithm (15) for the purpose of evaluation.

As described previously, a network of 28 intersections was simulated using TRAF-NETSIM. The load (demand) on the network was varied across 30 simulation runs by increasing the vehicle input rate at each source node over a range of ±20 percent. The optimization logic was instructed to minimize total delay. Figure 9 shows the average delay per vehicle using the combined (optimization and prediction) intersection control logic over the range of observed loads at a single intersection. For the purpose of comparison, the average delay per vehicle using well-timed semiactuated control logic is also given in Figure 9.
DISCUSSION OF RESULTS

Although the simulation study presented in this paper is limited, it appears that the prediction model provides valuable information for the development of real-time traffic-adaptive signal control logic. Further study and evaluation are required before the limitations and properties will be fully understood; however, the current results are extremely promising.

It is the author's belief that different prediction algorithms will be required for different situations. In some cases the use of upstream detectors will be sufficient to provide the desired level of performance. In others, more complex algorithms will be required. In still others, prediction will not be possible.

The prediction model presented here was based on several considerations. One was that the predictions should be based on the actual observations of traffic on the network—that is, it should be data-driven. Another was that operating agencies (cities, counties, states) have already made significant investments in detector systems. For real-time traffic-adaptive signal control systems to be economically feasible, they must use as much of the existing surveillance system as possible. The model does require communication between adjacent intersections to provide signal timing and upstream detector information. This additional communication requirement may or may not be possible in modern traffic signal control systems, but it will most certainly be required in future real-time traffic-adaptive signal control systems such as the RT-TRACS under development by Farradyne Systems, Inc., for FHWA (16).

The primary limitation of the prediction model is its dependence on turning percentages, link travel times, and queueing delay. However, almost all existing traffic signal systems and signal optimization software require similar information and, when properly calibrated, work relatively well. In the simulation study this information has been included as input parameters. It has been the author's experience, using this simulation study, that the prediction model is not highly sensitive to some of these parameters. However, if these factors were to change significantly, it is expected that the quality of the predictions would be compromised.

### TABLE 2 Descriptive Statistics of Predicted Traffic Arrival Process

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ME</td>
<td>0.09</td>
</tr>
<tr>
<td>SSE</td>
<td>358.13</td>
</tr>
<tr>
<td>MSE</td>
<td>0.31</td>
</tr>
<tr>
<td>MAE</td>
<td>0.29</td>
</tr>
<tr>
<td>MARE</td>
<td>0.27</td>
</tr>
<tr>
<td>D-W</td>
<td>1.94</td>
</tr>
</tbody>
</table>

FIGURE 7 Flow profile showing actual and predicted number of arrivals during time intervals of 5 sec.

FIGURE 8 Histogram of prediction errors.
It is hoped that some of these types of information will become available through advanced technologies as part of the intelligent transportation system (ITS). Perhaps, more important, developments such as this prediction model can identify the types of information that ITS developers should be attempting to provide.

Within the model itself are several possible improvements. A better prediction of left- and right-turn permitted movements should be included. The model tested in the simulation used a heuristic rule for this behavior. Another possible improvement would be to treat link travel times as random variables and to allow the predictions to be distributed over time with some probability distribution.

Despite these limitations, the prediction model appears to have many promising characteristics, including the fact that it is data-driven and combines these data with traffic flow knowledge. The ability to work with the traffic-adaptive signal control logic to improve the performance of a single intersection is the best evidence that this type of prediction model is feasible and, more important, valuable.

ACKNOWLEDGMENTS

The work presented in this paper has been partly supported by a contract with the Arizona Department of Transportation and FHWA. The author wishes to acknowledge several colleagues who have contributed to this effort, including Pitu B. Mirchandani, Suvrajeet Sen, Dennis Sheppard, Doug Tarico, Mike Whalen, Doug Gettman, and Vadim Parfenov.

REFERENCES