Estimation of Independence of Vehicle Arrivals at Signalized Intersections: A Modelling Methodology

GANG-LEN CHANG AND JAMES C. WILLIAMS

The assumption of independent vehicle arrivals at traffic signals, such as that in the Poisson distribution, has been widely used for modelling delay at urban intersections. The degree of correlation among vehicles determines whether this convenient assumption of independence is realistic. Collection of vehicle headways seems to be the only well-known method by which an estimate of their autocorrelation can be found. This, however, is a tedious and time-consuming task. To cope with this problem, an effective methodology using a discrete model for estimating the degree of interaction among vehicles under given traffic and geometric conditions is proposed in the present study. As an example of this technique's usefulness, a model is proposed and estimated, using information collected from 40 locations. Preliminary results appear to confirm the strength and applicability of the proposed method even though the developed model is constrained by the limited available data.

One of the prerequisites for the effective design and evaluation of the operation of traffic signals is to accurately estimate the delay incurred by traffic passing through the signalized intersection. The use of delay in determining intersection level of service in the 1985 *Highway Capacity Manual* (*1*) highlights the need to accurately estimate delay. Clayton was one of the first to make an attempt in this respect (*2*); he proposed a model to calculate the delay at fixed-time signals, where vehicle arrivals at and departures from a signalized intersection were presumed to be at strictly regular intervals. Winsten (*3*) revised this model, using a more realistic distribution (binomial) to simulate the pattern of arrivals. Webster (*4*) and Newell (*5*) have contributed delay formulas using a similar methodology but a different distribution (Poisson) for the arrival of vehicles. Subsequent models by Miller (*6, 7*) and Newell (*8*) have incorporated the variance-to-mean ratio of the vehicle arrivals to accommodate arrival distributions other than the Poisson, and in Hutchinson's numerical comparisons of various delay equations this term was added to Webster's model (*9*). Because little significant progress has occurred in this fundamental aspect over the past two decades, the delay formulas developed by Webster, Miller, and Newell still predominate in practice (particularly Webster's delay equation), and are incorporated in many popular traffic computer packages, such as SOAP (*10*) and PASSER-II (*11*).

A common feature of the above-mentioned delay formulas is that the arrival pattern of vehicles at intersections is presumed to be an independent Markov process. The average delay and queue length are then estimated based on the given flow rate and distribution, such as the Poisson or binomial. The explicit assumption of independence among arrivals of vehicles, as embedded in the basic properties of the Poisson distribution, is convenient for deriving the desired performance indicators such as the average delay, maximum delay, and average queue, and indeed provides a reasonable approximation of reality as long as the traffic is light. It is, however, obviously inconsistent with what can be observed at highly congested intersections where vehicles significantly interact with others in the arriving flows.

As is well recognized, to characterize the complex traffic patterns at the desired level of accuracy is a difficult yet essential task that enables the model to possess realistic features. The uniform distribution, as used in these well-accepted delay models, generally provides a good representation of departure distributions, since queued vehicles at the beginning of green time are usually discharged at a more or less constant rate. The Poisson or binomial distributions, however, cannot capture the possible interaction between arriving vehicles that may vary with the degree of congestion, driver behavior, physical features between adjacent intersections, speed limit, and so on. In very light traffic, it seems reasonable to expect that vehicles will arrive independently and follow a Poisson process at intersections. The degree of independence decreases as the degree of interference among consecutively arriving vehicles, resulting from congestion and other factors, increases. A well-known phenomenon is the car-following relationship (*12*) that describes the action-response effect between the leading and following vehicles. As long as such interrelationships are developed in the traffic flows, the assumption of independence among arrivals is obviously no longer realistic, and will inevitably lead to a biased estimate of the degree of delay and the other performance measures, such as queue length.

The determination of the distribution of vehicle arrivals, however, is a tedious and time-consuming process, necessitating the collection of vehicle headways or, at the very least, vehicle arrivals in consecutive time periods (of 30 sec or less). Hutchinson's numerical work (*9*) shows that the commonly used delay equations show little difference when vehicle arrivals are Poisson distributed; otherwise, significant differences exist among models. In addition, preliminary results of a data collection effort by the authors indicate significant differences between estimated delay from independent and nonindependent arrivals (*13*).

Civil Engineering Department, University of Texas at Arlington, Arlington, Tex. 76019.

While independent observations are not necessarily Poisson distributed, this distribution is typically assumed in most traffic-related studies if vehicle arrivals are independent. Therefore, it would be useful to have a simple test by which the independence of vehicle arrivals for specific traffic and geometric conditions could be determined without collecting vehicle headways. Of course, if the test indicates nonindependent arrivals, assumption of the Poisson distribution would not be valid, and further field studies and statistical tests would be necessary to select an appropriate distribution for the proper application of existing delay models.

A general model is presented in the next section that can be used to estimate the probability of independent vehicle arrivals to determine whether the commonly used Poisson assumption is applicable. Geometric features and traffic factors likely to affect the degree of interaction are included in this conceptual model. Techniques used to determine an appropriate specific model are also briefly discussed. An example of an application of this model is given in the third section; it is based on data collected from 40 intersections. In the final section, some conclusions and directions for further research are presented.

MODELLING CONCEPT AND ESTIMATION METHODOLOGY

A conceptual modelling system is presented in this section. The system relates the key features of a traffic system to its service load from which the degree of interaction among vehicles in the system can be estimated. This problem considers a traffic system consisting of an urban road section with N traffic lanes connecting two signalized intersections that are not interconnected, shown in Figure 1. Interconnected and actuated signals are not taken into account in this model, as each would require additional factors than those considered here. The goal, given the geometric factors and traffic conditions of such a road section, is to estimate the probability that vehicle arrivals at the downstream location are independent.

A single direction of a traffic system, such as that in Figure 1, can be analogized with a one-way channel with a unique entrance and exit at each end. Every vehicle entering the system, from the microscopic perspective, can then be viewed as a particle following a predetermined path (the available lanes) to pass through the channel. As such, whether the interference among particles in the channel is significant or not apparently depends on the channel's key physical features (e.g., length, number of paths), the number of particles (or the flow of particles), and their characteristics. Similarly, interaction among vehicles in the sort of traffic system shown in Figure 1 may vary with factors associated with the road section's physical features and the traffic flow characteristics.

More specifically, the system's key features primarily determine the available space for the flows. This space can be represented by

$$SA_i = f_1 (L_i, N_i, G_i, O_i) + \zeta_i \quad (1)$$

where

SA_i = the amount of space available for traffic flows;
L_i = the road section length, a major factor in the degree of platoon dispersion from the upstream intersection;
N_i = the number of available lanes;
G_i = the road section grade;
O_i = other associated factors; and
ζ_i = a random variable used to capture the effect of unobserved factors.

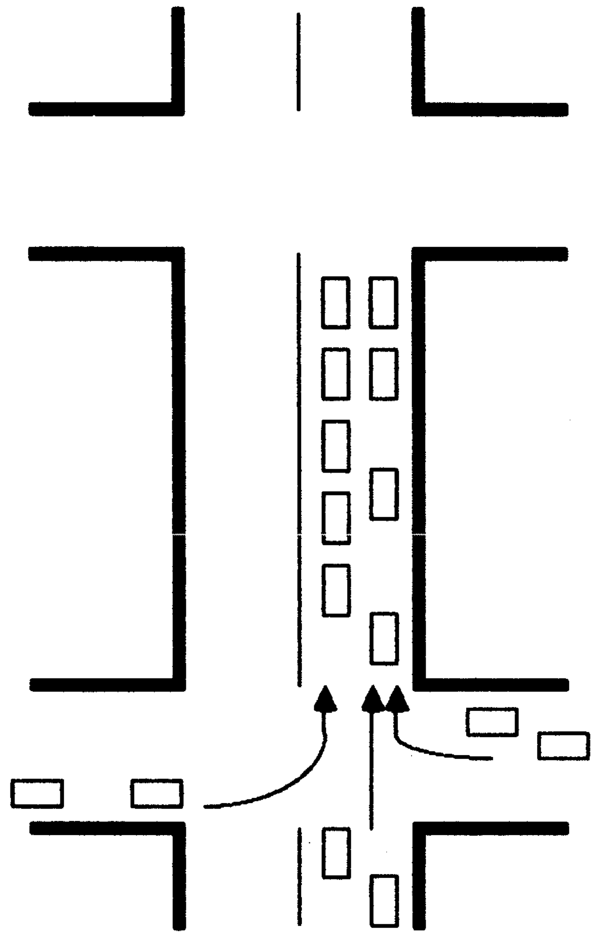

FIGURE 1 Graphical representation of two noninterconnected traffic signals.

The subscript i identifies a particular direction of a road section.

On the other hand, volume, average speed, concentration, and driver behavior can characterize the roadway space needed to provide the independent-arrival environment. This can be stated as

$$CS_i = f_2 (S_i, Q_i, K_i, DB_i, OF_i) + \varepsilon_i \quad (2)$$

where

CS_i = the critical amount of space needed for independent arrivals at the downstream intersection;
S_i = average vehicle speed;
Q_i = the flow (or volume);
K_i = the average concentration;
DB_i = driver behavior, often characterized by various indicators;
OF_i = other associated factors; and

ε_i = random variable used to capture unobserved associated factors.

As before, the subscript i identifies a particular direction of a road section.

As such, it can be expected that the interaction among arriving vehicles in road section i may occur if CS_i is greater than SA_i. Considering the uncertainties arising from unobservable factors, this statement can further be elaborated as a probabilistic formulation. If X_i and Y_i denote the vectors of the observable explanatory variables for SA_i and CS_i, respectively, as shown in Equations 1 and 2, and if A_i and B_i represent vectors of parameters associated with the variables in X_i and Y_i, respectively, this probabilistic formulation is as follows:

Prob [having independent vehicle arrivals in road section $i|X_i, Y_i$]

$$= \text{Prob}\,[SA_i > CS_i] = \text{Prob}\,[f_1(A_i, X_i) + \zeta_i > f_2(B_i, Y_i) + \varepsilon_i]$$

$$= \text{Prob}\,[f_1(A_i, X_i) - f_2(B_i, Y_i) > \varepsilon_i - \zeta_i] \quad (3)$$

Accordingly, given adequate observations (varying with the number of unknown parameters) and presumed properties of the error terms, estimation of parameters can be carried out by using the maximum likelihood method. Note that the specifications for SA_i and CS_i (equations 1 and 2) vary with the available information, measurable key factors, and their interrelationships. Many formal statistical procedures, such as the likelihood ratio test, the Lagrangian multiplier test, and tests of non-nested hypotheses are available for specification testing (*14–16*).

Prediction of the Probability of Vehicle Interaction Leading to Independent Arrivals

As is well recognized, the methodology for estimating parameters of a specified model varies with the presumed properties of the error terms. A description follows of two commonly used econometric approaches (binary logit and probit models) that can be applied for the estimation of equation 3. A detailed discussion of their statistical features is, however, not within the scope of this paper and is available elsewhere (*17*).

Binary Probit Model

Assume ε_i and ζ_i follow normal distributions with zero means, a covariance of σ_{12} and variances of $\sigma^2{}_1$ and $\sigma^2{}_2$, respectively. Accordingly, $\varepsilon_i - \zeta_i$ is also normally distributed with zero mean, but with a variance σ^2 $(= \sigma^2{}_1 + \sigma^2{}_2 - 2\sigma_{12})$. The probability of independent vehicle arrivals can then be solved as follows:

Prob [independent vehicle arrivals]

$$= \text{Prob}\,[f_1(A_i, X_i) - f_2(B_i, Y_i) < \zeta_i - \varepsilon_i]$$

$$= \Phi\,[f_1(A_i, X_i) - f_2(B_i, Y_i)/\sigma] \qquad \sigma > 0 \quad (4)$$

where Φ denotes the the standardized cumulative normal distribution.

Binary Logit Model

Another commonly used technique is to assume that $\varepsilon^*{}_i$ $(= \zeta_i - \varepsilon_i)$ is logistically distributed (is Gumbel distributed), with a cumulative distribution:

$$F(\varepsilon^*{}_i) = 1/(1 + \exp(-u \cdot \varepsilon^*{}_i)) \quad (5)$$

$$u > 0, \; -\infty < \varepsilon^*{}_i < \infty$$

where u is a positive scale parameter. This distribution approximates the normal distribution (as used in the probit model) quite well, but is much more convenient in terms of analytical computation. Under this assumption, the probability of independent vehicle arrivals is given by the following expression:

Prob [independent vehicle arrivals]

$$= \exp(u \cdot f_1(A_i, X_i))\, u$$

$$\div\, [\exp(u \cdot f_1(A_i, X_i)) + \exp(u \cdot f_2(B_i, Y_i))] \quad (6)$$

Note that for convenience, but without loss of accuracy, the scale parameter, u, is generally assumed to equal 1.

Estimation of Model Parameters

Let each road section i, with associated key attributes as described previously, be viewed as one observation, then, given N observations, the likelihood function of the parameters in vectors A and B (Equation 3) can be constructed as follows (*18*):

$$L^*(A_i, B_i) = \prod_{i=1}^{N} \left[P_i(a)^{\delta_i} \cdot (1 - P_i(a))^{1-\delta_i} \right] \quad (7)$$

where

$P_i(a)$ = Prob [independent vehicle arrivals in road section i]
δ_i = 1 if the independent vehicle arrivals were observed in road section i
δ_i = 0 otherwise

To facilitate computation, Equation 7 is often rewritten in the following logarithmic form, denoted as L:

$$L(\beta_1, \ldots \beta_k) = \sum_{i=1}^{N} [\delta_i \ln P_i(a) + (1 - \delta_i) \ln (1 - P_i(a))] \quad (8)$$

By differentiating L with respect to each of the βs (parameters in vectors A_i and B_i) and setting the partial derivatives equal to zero, parameters satisfying max L $(\beta_1, \beta_2, \ldots, \beta_k)$ can be obtained. In many cases of practical interest it has been proven that the likelihood function (Equation 7) is globally concave and is, therefore, unique if a solution to the first-order conditions exists.

ILLUSTRATIVE EXAMPLE

This section presents an example that illustrates the modelling procedures and methods used for specification testing. The data set for parameter estimation consists of 40 noninterconnected signalized intersections, all located in the Salt Lake City metropolitan area in Utah. Due to limited resources, our data collection focused on those variables relatively inexpensive to collect, yet are critical to the model development. The available data and proposed formulation for a model that will evaluate the degree of correlation, or interaction, among arriving vehicles are stated below. Each variable corresponds to a particular road section i and the traffic approaching the particular intersection in question. The available key variables are as follows:

1. Length of road section i (L_i), measured from the stop line of the selected intersection back to the nearest signalized intersection upstream;
2. Number of lanes, omitting exclusive turn lanes (N_i);
3. Average vehicle speed (S_i);
4. Flow rate to the selected intersection (Q_i); and
5. Vehicle headways, taken separately for each lane (if $N_i > 1$), between each successive vehicle.

The vehicle arrivals at the downstream intersection are assumed to be independent if the vehicle headways are not significantly correlated. Estimation of the correlation was carried out by analyzing the vehicle headway time series collected over a 30-min interval in the selected road section with a general ARIMA model (*19*).

Model Specification

Given the available information above, several plausible specifications for Equation 1 and 2 were proposed and examined based on the estimated results. Primary criteria used to carry out the comparisons are the t-statistic, likelihood ratio index, and physical implications of the estimated parameters. Of the specifications tested, the one providing the best fit (highest likelihood ratio index) and having a reasonable physical meaning reflected by the proper parameter signs is

$$\text{Amount of space available} = SA_i = N_i \cdot L_i \tag{9}$$

Critical amount of space needed =

$$CS_i = a_1(Q_i)^{a_2} + a_3(S_i)^{a_4} + \varepsilon_i \tag{10}$$

where a_1, a_2, a_3, and a_4 are model parameters. The notion embedded in this formulation is that under the given environment (as characterized by SA_i) the flow rate and speed are critical factors that contribute to the formation and the degree of interaction among vehicles. More specifically, the vehicle arrivals along road section i may be significantly correlated if SA_i is less than CS_i. The probability of existing dependent arrivals of vehicles is as follows:

$$\begin{aligned} \text{Prob}\,[SA_i < CS_i] &= \text{Prob}\,[N_i \cdot L_i < a_1(Q_i)^{a_2} + a_3(S_i)^{a_4} + \varepsilon_i] \\ &= \text{Prob}\,[(N_i \cdot L_i) - a_1(Q_i)^{a_2} - a_3(S_i)^{a_4} < \varepsilon_i] \end{aligned} \tag{11}$$

This proposed formulation cannot be considered the standard model for predicting the independence among vehicle arrivals because of limitations of the collected data; it simply serves as an illustrative example. For instance, the value of SA_i, in practice, depends on the number of lanes and the length of the road section and varies with the system's other geometric characteristics, such as grade. In addition, information regarding drivers' behavior or risk attitude that could be one of the critical factors was not collected due to the prohibitive cost of data acquisition and classification. Also, the arrival pattern depends, to some extent, on the departure pattern from the upstream signalized intersection; and for the same flow rate, many possible departure patterns exist. Although this is not directly modelled, the degree of variation of the arrival flow patterns is constrained by the distance between the signalized intersections, which has been incorporated in the model.

Parameter Estimation and Implications

Because the variable CS_i is not directly observable, the parameters in Equations 9 and 10 should be estimated using discrete methods, such as either the logit or probit model, as presented in the second section. Justification of a proper specification for a discrete model, as for regression, is often carried out by examination of (1) the exhibited sign of estimated parameters, (2) the asymptotic standard error (or t-statistic), and (3) a goodness-of-fit index such as ρ^2, the likelihood ratio index. This index measures the fraction of an initial log likelihood value explained by the presented model and is defined as $[1 - L(\beta)/L(0)]$, where $L(0)$ and $L(\beta)$ denote the initial value (when all the parameters are zero) and the maximum value (at convergence) of the log likelihood function, respectively. The ρ^2 is analogous to R^2 used in the regression models, and must lie between zero and one for a binary discrete model. It is particularly useful in comparing alternative specifications developed on the same data set. A more in-depth discussion regarding the statistical properties of the likelihood ratio index is available elsewhere (*18*).

Table 1 summarizes the results of the parameter estimation for the illustrative model, as presented in Equations 9 and 10 using the discrete logit model. As expected, all parameters exhibit positive signs, except a_3. The implications are that, within a given road section, as partly characterized by SA_i, an increase in the flow will significantly increase dependent vehicle arrivals. On the other hand, a facility with better geometrics generally allows a higher speed under a given traffic volume, and thus results in less interaction among vehicles as reflected by the negative sign of parameter a_3.

Turning to the t-statistics as shown in Table 1, it can be noted, using a 0.95 statistical confidence interval, that both parameters a_2 and a_4 are not significantly different from one. As such, the model represented by Equations 9 and 10 was re-estimated with a simple linear specification presuming that parameters a_2 and a_4 are known to equal one. Estimated results of the simplified specification present similar information and are reported in Table 2.

In addition to the commonly used t-statistics, Tables 1 and 2 present the result of the log likelihood ratio test, defined as $-2\,[L(0) - L(\beta)]$, a statistic used to test the null hypothesis that all parameters are zero. This statistic is asymptotically distributed with an χ^2 distribution with K degrees of freedom, where K is the number of parameters. In this example the value of the log likelihood ratio is 37.74 (see Table 2), which indicates the null hypothesis can be rejected (the χ^2 value at the 0.05 level of significance is 9.49).

TABLE 1 ESTIMATION RESULTS OF THE ILLUSTRATIVE MODEL

Parameter	Associated Variable	Estimated Value	Standard Error	t Statistic
a_1	Total approach volume: Q_i	0.00865	0.0027	3.20
a_2	Total approach volume: Q_i	0.96377	0.0357	1.01[a]
a_3	Average speed: S_i	−0.17642	0.0618	2.85
a_4	Average speed: S_i	0.91425	0.0733	1.17[a]

NOTE: Number of observations = 40; $L(0) = -27.726$; $L(\beta) = -8.114$; $-2(L(0) - L(\beta)) = 31.224$; $\rho^2 = 1 - L(\beta)/L(0) = 0.707$.
[a]Not significant at 95 percent confidence level ($a_i = 1$ tested).

TABLE 2 REESTIMATED RESULTS OF THE ILLUSTRATIVE MODEL

Parameter	Associated Variable	Estimated Value	Standard Error	t Statistic
a_1	Total approach volume: Q_i	0.00911	0.0038	2.372
a_3	Average speed: S_i	−0.17518	0.0649	2.698

Number of observations = 40; $L(0) = -27.726$; $L(\beta) = -8.856$; $-2(L(0) - L(\beta)) = 37.74$; $\rho^2 = 1 - L(\beta)/L(0) = 0.681$.

With respect to the goodness of fit, the likelihood ratio index, ρ^2, generally serves as a good indicator for discrete models, as described above, although other rigorous, yet complex, statistical tests are available (*14–16*). Owing to the inevitable involvement of various behavioral or unobservable factors, the likelihood ratio index in most discrete models, which may be considered to be well specified, often cannot achieve as high a value as R^2 in successful regressions. In fact, the illustrative example with ρ^2 near 0.7, as shown in Tables 1 and 2, will generally have a reasonably good specification.

Model Application

Instead of spending time and money in collecting vehicle headways, a traffic engineer could use a model such as this (Equations 9 and 10) to predict the probability of dependent vehicle arrivals under the given traffic conditions and geometric characteristics. The following example illustrates the procedure for estimating this probability. If L_i = 400 ft = 0.0758 mi, S_i = 15 mi per hour, Q_i = 150 vehicles per hr per lane, and N_i = 3 lanes, then

$$SA_i = N_i \cdot L_i = (3)\ (0.0758) = 0.2273$$

$$CS_i = a_1\ (Q_i) + a_3\ (S_i)$$
$$= (0.00911)\ (150)\ (3) - (0.17518)\ (15)$$
$$= 1.4718$$

Accordingly, the probability of incurring correlated arriving vehicles under the above traffic conditions is

$$\text{Prob [nonindependent vehicle arrivals} \mid L_i, S_i, N_i, Q_i]$$
$$= \exp\ (CS_i)/[\exp\ (CS_i) + \exp\ (SA_i)]$$
$$= 4.3571\ /\ (4.3571 + 1.4718) = 0.77$$

In other words, one can conclude that the traffic flows in this road section are highly correlated about 77 percent of the time; thus, the commonly used Poisson distribution is not valid under this scenario. This result should not be unexpected, given the volume and the close signal spacing used for this example. Other distributions that can provide for dependent and independent arrivals should be considered.

CONCLUSION

The present study has introduced an effective yet economic approach to estimate the degree of correlation among arriving vehicles under given conditions and geometric characteristics. With the proposed technique, traffic professionals can easily determine if the existing delay formulas and other traffic models based on the Poisson distribution are applicable. In particular, the results of this test can indicate when the Poisson assumption may be used.

As the primary focus of the paper is to introduce the modelling methodology and its application, only the simplest case comprising two noninterconnected intersections is considered. This model can be extended to more complex traffic systems if appropriate model variables are included in the specification, e.g., a variable indicating whether a platoon arrives during the red or green time would be necessary when formulating a similar model for a progressive signal system.

The formulation proposed in the example serves only as an illustration. More complete information, as described in previous sections, must be collected, and rigorous statistical procedures applied for the testing of the model specification to determine the most appropriate formulation.

ACKNOWLEDGMENT

The data used in the illustrative example was collected by Tsai Ying-Chieh.

REFERENCES

1. *Special Report 209: Highway Capacity Manual*. TRB, National Research Council, Washington, D.C., 1985.
2. A. J. H. Clayton. Road Traffic Calculations. *Journal of the Institution of Civil Engineers*, Vol. 16, 1940.
3. M. Beckman, C. B. McGuire, and C. B. Winsten. *Studies in the Economics of Transportation*, Yale University Press, 1956.
4. F. V. Webster. Traffic Signal Settings. *Road Research Technical Paper Nr. 39*, 1958.
5. G. F. Newell. Queues for a Fixed-Cycle Traffic Light. *Annual Mathematical Statistics*, Vol. 31, 1960.
6. A. J. Miller. Settings for Fixed-Cycle Traffic Signals. *Operational Research Quarterly*, Vol. 14, 1963.
7. A. J. Miller. The Capacity of Signalized Intersections in Australia. *Australian Road Research Board Bulletin Nr. 3*, 1968.
8. G. F. Newell. Approximation Methods for Queues with Application to the Fixed-Cycle Traffic Light. *SIAM Review*, Vol. 7, 1965.
9. T. P. Hutchinson. Delay at a Fixed Time Traffic Signal, II: Numerical Comparisons of Some Theoretical Expressions. *Transportation Science*, Vol. 6, 1972.
10. *Signal Operations Analysis Package (SOAP)*. Implementation Package IP-79-9, five volumes. FHWA, U.S. Department of Transportation, 1979.
11. *Analysis of Reduced-Delay Optimization and Other Enhancements to PASSER II-80—PASSER II-84—Final Report*. Report No. TTI-2-18-83-375-1F. Texas Transportation Institute, 1984.
12. R. Herman and R. B. Potts. Single-Lane Traffic Theory and Experiment. *Proc., 1st Symposium on the Theory of Traffic Flow*, ed. R. Herman, Elsevier, 1961.
13. G.-L. Chang and J. C. Williams. Empirical Investigation of Theoretical Delay Models, working paper.
14. J. L. Horowitz. Testing Probabilistic Discrete Choice Models of Travel Demand by Comparing Predicted and Observed Aggregate Choice Shares. *Transportation Research B*, Vol. 19B, 1985.
15. T. S. Breusch and A. R. Pagan. A Simple Test for Heteroskedasticity and Random Coefficient Variation. *Econometrica*, Vol. 47, 1979.
16. J. L. Horowitz. Statistical Comparison of Non-Nested Probabilistic Discrete Choice Models. *Transportation Science*, Vol. 17, 1983.
17. C. F. Daganzo and L. Schoenfeld. *CHOMP User's Manual*. ITS Research Report UCB-ITS-RR-78-7. Institute of Transportation Studies, University of California, Berkeley, 1978.
18. M. Ben-Akiva and S. R. Lerman. *Discrete Choice Analysis*. MIT Press, 1985.
19. G. E. P. Box and G. M. Jenkins. *Time Series Analysis: Forecasting and Control*. Holden-Day, 1976.

The authors are solely responsible for the content of this paper.

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