TRANSPORTATION RESEARCH

RECORD

No. 1510

*Highway Operations, Capacity,
and Traffic Control*

Traffic Flow Theory and Characteristics with Applications for Intelligent Transportation System Technologies

A peer-reviewed publication of the Transportation Research Board

TRANSPORTATION RESEARCH BOARD
NATIONAL RESEARCH COUNCIL

NATIONAL ACADEMY PRESS
WASHINGTON, D.C. 1995

Transportation Research Record 1510
ISSN 0361-1981
ISBN 0-309-06203-9
Price: $ 23.00

Subscriber Category
IVA highway operations, capacity, and traffic control

Printed in the United States of America

Sponsorship of Transportation Research Record 1510

GROUP 3—OPERATION, SAFETY, AND MAINTENANCE OF TRANSPORTATION FACILITIES
Chairman: Jerome W. Hall, University of New Mexico

Facilities and Operation Section
Chairman: Jack L. Kay, JHK & Associates

Committee on Traffic Flow Theory and Characteristics
Chairman: Hani S. Mahmassani, University of Texas at Austin
Secretary: James C. Williams, University of Texas
Siamak (Sia) A. Ardekani, Rahim F. Benekohal, Gang-Len Chang, Gary A. Davis, Richard W. Denney, Jr., Nathan H. Gartner, Fred L. Hall, Benjamin G. Heydecker, R. Jayakrishnan, Masao Kuwahara, Michael Kyte, John Leonard, Henry C. Lieu, David Mahalel, Carroll J. Messer, Abbas Mohaddes, Kyriacos C. Mouskos, Markos Papageorgiou, Ajay K. Rathi, Nagui M. Rouphail, Michael A. P. Taylor, Michel Van Aerde

Transportation Research Board Staff
Robert E. Spicher, Director, Technical Activities
Richard A. Cunard, Engineer of Traffic and Operations
Nancy A. Ackerman, Director, Reports and Editorial Services

Sponsorship is indicated by a footnote at the end of each paper. The organizational units, officers, and members are as of December 31, 1994.

Transportation Research Record 1510

Contents

Foreword

The papers in this volume are from the 1995 Annual Meeting of the Transportation Research Board and were peer reviewed by the TRB Committee on Traffic Flow Theory and Characteristics.

The study of the relationships among traffic flow characteristics and flow models is fundamental to traffic operations and is the subject of considerable interest as a result of the development of intelligent transportation systems.

Readers with an interest in traffic flow models and characteristics will find papers pertaining to relationships among traffic speed, flow, and concentration; neural network modeling of the macroscopic relationships between traffic flow variables; microscopic models for traffic operations at the individual vehicle level; the analysis of day-to-day variations in real-time traffic flow data for providing in-vehicle real-time information on traffic conditions; and the validation of traffic simulation networks.

The use of traffic flow theory for control applications in real-time advanced traffic management systems is the focus of the last series of papers. Included are papers on an event-based traffic flow prediction model for real-time traffic-adaptive signal control, estimation of intersection turning movements, arterial incident detection, and deceleration behavior and prediction models.

Another Look at A Priori Relationships Among Traffic Flow Characteristics

JAMES H. BANKS

Past derivations of a priori relationships among speed, flow, and concentration (such as the fundamental relationship and the speed-flow-occupancy relationship) have involved unrealistic assumptions of uniformity in at least one traffic flow characteristic. Several relationships are derived for which these assumptions of uniformity are relaxed. For relationships involving both time- and distance-based variables, this requires that the relationship be understood in probabilistic terms; where all variables are time-based, deterministic relationships are also possible. The fundamental relationship can be shown to be strictly true in the limit where the time and distance intervals over which measurements are taken approach 0. Where the order of arrival of vehicles with particular speeds is random, the fundamental relationship is found to hold for average values of the variables in question; this is also true if the section over which density is measured is empty at the beginning and end of the time interval used for averaging. Relationships derived under various other assumptions involve covariance terms, so that if particular variables are not correlated, simple relationships continue to hold. Where these variables are correlated, biases may be expected. Under certain conditions, these biases may be quite serious. Comparison of the relationships derived here with those of past empirical studies results in good agreement for the relationship between density, as estimated from the fundamental relationship, and occupancy. On the other hand, previously reported discrepancies between measured speeds and speeds calculated from flows and occupancies cannot be explained fully by the covariance terms in the relationships derived here.

The study of relationships among the traffic flow characteristics speed, flow, and concentration (either density or occupancy) has long been a fundamental part of traffic research. In general, two types of relationships among these variables are possible: a priori relationships, which proceed from the definitions of the various measures, and empirical relationships, which can be discovered only by observing actual traffic flow.

Not surprisingly, the bulk of the literature focuses on empirical relationships. The major a priori relationships were worked out early in the history of traffic flow research and have been little examined since. They have often been taken for granted and used freely to transform data from one form to another or to move back and forth among the three possible bivariate relationships involving speed, flow, and concentration. Nevertheless, several recent studies by Hall et al. have raised questions about the accuracy and applicability of these relationships (*1–4*). These studies have also presented data that appear to contradict them to some extent and have suggested using three-dimensional empirical models that are independent of them.

The relationships in question include the so-called *fundamental relationship:*

$$q = uk \tag{1}$$

where

q = flow,
u = speed, and
k = density.

A similar relationship exists among speed, flow, vehicle length, and occupancy:

$$u = \frac{qL}{H} \tag{2}$$

where L represents vehicle length and H occupancy, defined as the fraction of time that vehicles are present at a point. The classical derivation of Equation 1 is that of Wardrop (*5*). Equation 2, which most commonly has been used to estimate speeds from flow and occupancy data, was proposed by Athol (*6*).

Since the use of these relationships (especially Equation 1) has been pervasive in traffic flow theory, confirming major inaccuracies in them could have far-reaching consequences. The purpose of this paper is to reexamine the validity of these relationships, extend their derivations to address certain oversimplifications, and consider the possible reasons for apparent discrepancies between them and actual data, particularly those reported by Hall and Persaud (*1*).

FUNDAMENTAL RELATIONSHIP

Theoretical objections to the fundamental relationship (Equation 1) arise from the distinction between relationships that hold true for uniform traffic streams (those with constant, identical speed and spacing for all vehicles) and those that hold for averages of the characteristics of nonuniform traffic streams. In addition, the fundamental relationship involves both time- and distance-based variables, which may be incompatible with one another in nonuniform traffic streams.

Point Relationships Among Traffic Flow Characteristics

The version of Equation 1 presented here implicitly assumes a uniform traffic stream and under that assumption can be derived easily by means of dimensional analysis. It can also be shown to be true at a point, if all measures are regarded as continuous variables. This approach to its derivation makes use of a three-dimensional surface proposed by Makigami et al. (*7*). If vehicle trajectories are plotted and numbered (with some adjustments in cases in which vehicles pass one another), the trajectories may be considered as the contour lines of a surface of cumulative flow versus time and distance. For

Department of Civil Engineering, San Diego State University, San Diego, Calif. 92182-1324.

such a surface to exist, it must be possible to smooth out the discrete steps in the actual cumulative vehicle function, so as to treat it as continuous. Where this is a reasonable simplification, the partial derivative of the cumulative vehicle function $A(x,t)$ with respect to time represents flow, that of A with respect to distance represents density, and that of distance with respect to time represents the speed of a vehicle at an instant of time. That is,

$$q = \frac{\partial A}{\partial t} \tag{3}$$

$$k = \frac{\partial A}{\partial x} \tag{4}$$

and

$$u = \frac{\partial x}{\partial t} \tag{5}$$

Then, since

$$\frac{\partial A}{\partial t} = \frac{\partial x}{\partial t}\frac{\partial A}{\partial x} \tag{6}$$

it follows that

$$q = uk$$

Nonuniform Flow over Long Time Intervals

Obvious problems with these derivations are that (*a*) real traffic streams are never truly uniform and, under some conditions (congested flow, for instance), are far from being even approximately uniform; and (*b*) point measures of traffic characteristics such as flow and density have theoretical meaning only. To apply to more-realistic models of the traffic stream, derivations of the fundamental relationship must be able to relate average values of the traffic flow variables, measured over more extended times and distances.

A possible way of doing so is to further consider the traffic flow surface proposed by Makigami et al. (*7*). Consider two locations x_1 and x_2 such that there are no entrances or exits between them. Plots of the cumulative numbers of vehicles passing these two points result in the functions $A(x_1, t)$ and $A(x_2, t)$. Now consider a time interval T that begins when the section between x_1 and x_2 is empty, continues so long as vehicles are present in the section, and ends as soon as the section is empty again. Figure 1 shows plots of $A(x_1, t)$ and $A(x_2, t)$ for time interval T. At any given time $N(t)$ vehicles are present in the section. The average number of vehicles in the section at any time is

$$\overline{N} = \frac{\int_0^T N(t)dt}{T} \tag{7}$$

and the average density is

$$\overline{k} = \frac{\int_0^T N(t)dt}{(x_2 - x_1)T} \tag{8}$$

The total flow exiting the section during time T is $A(x_2, T)$, which, under the preceding assumptions, is also the total flow entering, or $A(x_1, T)$. The average flow, then, is

$$\overline{q} = \frac{A(x_2, T)}{T} \tag{9}$$

Finally, the total time consumed by all vehicles in the section is

$$\int_0^T N(t)dt$$

so that the average time required for a vehicle to cross the section is

$$\overline{t} = \frac{\int_0^T N(t)dt}{A(x_2, T)} \tag{10}$$

and the harmonic mean or space mean speed is

$$\overline{u}_s = \frac{x_2 - x_1}{\overline{t}} = \frac{A(x_2, T)(x_2 - x_1)}{\int_0^T N(t)dt} \tag{11}$$

Substituting $\overline{k}T$ for

$$\frac{\int_0^T N(t)dt}{x_2 - x_1}$$

and $\overline{q}T$ for $A(x_2, T)$ results in

$$\overline{u}_s = \frac{\overline{q}T}{\overline{k}T} = \frac{\overline{q}}{\overline{k}} \tag{12}$$

Note that the assumption that the section is empty at both the beginning and end of period T is necessary for this relationship to be strictly true. In the absence of this assumption, the total travel time does not represent the sum of the travel times across the entire section for any particular group of vehicles; consequently, $\overline{u}_s$ would be at best an approximation. In cases in which T is long relative to $\overline{t}$, this may not be important, but in the extreme case in which T approaches 0, the average speed used in Equation 11 would be meaningless. Also, $A(x_1,T)$ would not equal $A(x_2,T)$, so that $\overline{q}$ would be defined ambiguously. Again, if $A(x_1,T)$ is approximately equal to $A(x_2,T)$, this may not be important, but over short time intervals the difference is apt to be fairly large, especially in congested flow.

Wardrop's Derivation

A second approach to deriving the fundamental relationship for nonuniform flow is that of Wardrop. This classical derivation relaxes the assumptions of uniform speeds and densities by assuming instead that the traffic stream is composed of a set of subsidiary traffic streams. Within each stream, speeds of all vehicles are identical and constant with respect to time and distance, but vehicle spacings are random. Wardrop's subsidiary streams thus represent a discrete approximation of the speed distribution. From these assumptions, Wardrop shows that if each traffic stream i has speed u_i and flow q_i, the characteristics of the traffic streams may be combined to give

$$q = \overline{u}_s k \tag{13}$$

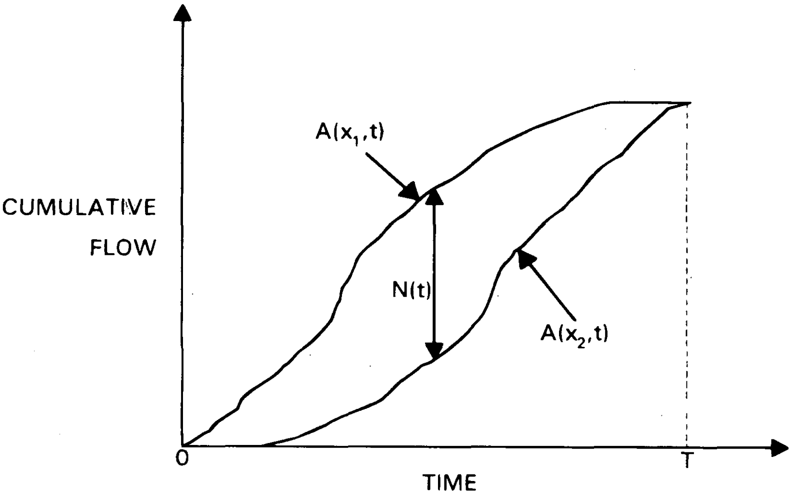


FIGURE 1 Cumulative flow versus time at two locations.

where q and k are the overall flow and density of the traffic stream and $\bar{u}_s$ is the harmonic mean or space mean speed.

Uncongested Flow: Speed Independent of Flow

A close look at Equation 13 shows that it is not exactly correct, since q and k cannot be uniform quantities if vehicles are spaced randomly. Instead, q and k must be intended to be averages or expected values. More important, the assumption of a discrete speed distribution is unrealistic. To relax this assumption, however, it is necessary to confront a fundamental difficulty that arises because of the combination of time and distance-based variables in the fundamental relationship.

Density is most properly measured by counting the number of vehicles present in a section of known length at an instant of time. The presence of any given vehicle in the section at this instant, however, is dependent on the exact time that it entered the section as well as its average speed across it. There can be no deterministic relationship among speed, flow, and density for nonuniform traffic streams because flow conveys only the average time between the arrival of successive vehicles at a point, not the exact time that each vehicle arrived.

Under certain assumptions of randomness, this difficulty can be circumvented by formulating the relationship in probabilistic terms. The resulting relationships must be understood to hold only among the average or expected values of the variables when repeated samples are taken, not for the measured values of any given sample. In the case of uncongested flow, it may be reasonable to assume a continuous speed distribution that is nearly independent of flow, especially over short periods of time, and for which the order of arrival of vehicles with given speeds is random.

Consider a traffic stream with an average flow rate of $\bar{q}$ over some period of time. Flow during this same period is also characterized by a speed distribution with a probability distribution function $p(u)$. If vehicles arrive at random, and the speeds of individual vehicles are independent of one another, the probability that a vehicle with speed u' is present in a section of length X at any instant is proportional to X/u', the amount of time that the vehicle spends in the section. Note that, to be strictly correct, the speed in question should be the average speed of the vehicle across the section, not its spot speed at any point. Thus the probability of detecting a vehicle with any given speed is

$$p = \bar{q}\left(\frac{X}{u}\right)p(u)du \qquad (14)$$

For the entire traffic stream, the expected number of vehicles present in a section of length X is

$$E(N) = \bar{q}X\int_0^\infty \frac{p(u)}{u}du \qquad (15)$$

The lower bound of the integral is shown as 0 on the assumption that negative speeds do not occur; this is not important, however, as long as it is understood that the integral of $p(u)$ from 0 to infinity is 1.0. Expected density, in turn, is the expected number of vehicles in the section divided by the length of the section, or

$$\bar{k} = \frac{E(N)}{X} = \bar{q}\int_0^\infty \frac{p(u)}{u}du \qquad (16)$$

For a continuous speed distribution, however, space mean speed is defined as

$$\bar{u}_s = \frac{1}{\int_0^\infty \frac{p(u)}{u}du} \qquad (17)$$

so that Equation 6 becomes

$$\bar{k} = \frac{\bar{q}}{\bar{u}_s} \qquad (18)$$

Consequently, it can be shown that the fundamental relationship also holds between space mean speed and the arithmetic means of flow and density in cases in which speed distributions are continuous, so long as the order of arrival of vehicles with given speeds is random.

Congested Flow: Cyclic Speed-Flow Variation

In congested flow, the assumption of random arrivals is unlikely to be valid. Instead, what is usually observed is a pattern of waves moving upstream. Flow moving through these waves is characterized by alternating periods of acceleration and deceleration and is often called stop-and-go traffic. The behavior of such waves is not very well understood, although there is literature, both theoretical and empirical, related to them (*8–12*). In any case, however, speeds, flows, and densities in congested flow appear to be strongly correlated with one another.

At one extreme, the wave pattern might be assumed to consist of a series of identical waves with periods T and wave lengths X. Under that assumption, the relationship among speed, flow, and density is affected by the length of the section over which density is measured. Suppose, for instance, density is defined over the waves' length X and flow over their period T. In this case, both the number of vehicles present in the section and the amount of time it takes each vehicle to cross the section are constants, since flow is always identical at both boundaries of the section. The density in this case is $k = \bar{q}T/X$ and the speed is $u = X/T$, so that

$$\bar{k} = \frac{\bar{q}}{u} \tag{19}$$

Note, however, that k and u are constants only if measured over X (or some integral multiple thereof). For distances less than X, they vary, and for distances much less than X, they fluctuate widely. Under these conditions, Equation 19 is no longer valid.

Consider a traffic stream consisting of a series of identical waves with period T. At any instant t, there is an instantaneous flow rate $q(t)$ arriving at some point in a section of length X (which is less than the wave period) and an average speed across the section of $u(t)$. The speed $u(t)$ (as measured over distance X) varies less than the spot speed measured at any point in the section, but it still varies and is correlated with $q(t)$, the flow passing the point at time t. The probability of detecting a particular vehicle in a sample taken at a random instant is proportional to the amount of time that the vehicle spends in the section. In this case, this time is given by

$$\frac{X}{u(t)} \tag{20}$$

and, since the cycle repeats itself over T, the probability of detecting the vehicle that passed the point at t is

$$\frac{X}{Tu(t)} \tag{21}$$

Since the number of vehicles passing the point at t is given by $q(t)dt$, the expected number of such vehicles to be detected is

$$\frac{X}{T}\frac{q(t)dt}{u(t)} \tag{22}$$

Note that in any given sample, the $q(t)dt$ vehicles passing the point t are either in the section or not; Expression 22 gives the average number of such vehicles that would be detected in repeated samples. The expected *total* number of vehicles detected in any given sample may be found by integrating over t:

$$E(N) = \frac{X}{T}\int_0^T \frac{q(t)}{u(t)}dt \tag{23}$$

Once again, the expected density is the expected number of vehicles in the section divided by X, or

$$\bar{k} = \frac{1}{T}\int_0^T \frac{q(t)}{u(t)}dt \tag{24}$$

Now let Λ represent the reciprocal of speed, so that $\Lambda(t) = 1/u(t)$ and $\bar{\Lambda} = 1/\bar{u}_s$. Also, let $q(t)$ be replaced by $\bar{q} + [q(t) - \bar{q}]$ and $\Lambda(t)$ by $\bar{\Lambda} + [\Lambda(t) - \bar{\Lambda}]$, where $\bar{q}$ and $\bar{\Lambda}$ are the mean values of q and Λ. Equation 24 may now be rewritten as

$$\bar{k} = \frac{1}{T}\int_0^T \{\bar{q} + [q(t) - \bar{q}]\}\{\bar{\Lambda} + [\Lambda(t) - \bar{\Lambda}]\}dt \tag{25}$$

Expanding Equation 25 results in

$$\bar{k} = \frac{\int_0^T \bar{q}\bar{\Lambda}dt}{T} + \frac{\bar{\Lambda}\int_0^T [q(t) - \bar{q}]dt}{T} + \frac{\bar{q}\int_0^T [\Lambda(t) - \bar{\Lambda}]dt}{T} + \frac{\int_0^T [q(t) - \bar{q}][\Lambda(t) - \bar{\Lambda}]dt}{T} \tag{26}$$

By definition, however,

$$\int_0^T [q(t) - \bar{q}]dt = 0 \tag{27}$$

$$\int_0^T [\Lambda(t) - \bar{\Lambda}]dt = 0 \tag{28}$$

and

$$\sigma_{q\Lambda} = \frac{1}{T}\int_0^T [q(t) - \bar{q}][\Lambda(t) - \bar{\Lambda}]dt \tag{29}$$

where $\sigma_{q\Lambda}$ is the covariance of flow and the reciprocal of speed. Equation 26 may now be rewritten as

$$\bar{k} = \frac{\bar{q}\bar{\Lambda}T}{T} + \sigma_{q\Lambda} = \frac{\bar{q}}{\bar{u}_s} + \sigma_{q\Lambda} \tag{30}$$

Let the estimated density that would be calculated by dividing flow by space mean speed be represented by $\hat{k} = \bar{q}/\bar{u}_s$. Then, from Equation 30,

$$\bar{k} = \hat{k} + \sigma_{q\Lambda} \tag{31}$$

In congested flow, speeds and flows tend to be correlated positively; consequently, the covariance of flow and the reciprocal of speed should be negative. This means that in congested flow characterized by a uniform wave pattern, the actual expected density should be less than that estimated by dividing average flow by space mean speed, where $\bar{u}_s$ is defined over a distance less than the wave length.

The assumption of identical waves is, of course, not very realistic [see, for instance, the wave plots by Koshi et al. (*10*)]. It is far more likely that wave periods, wave lengths, and amplitudes (in terms of speed, flow, and density) vary in some irregular pattern. This affects the preceding derivation primarily in that it is no longer sufficient to integrate over the period of a single wave, as in Equation 23; instead, if the traffic stream is assumed to exhibit long-term averages of speed and flow, the period of integration should be long enough to allow these averages to be approached. The fundamental relationship thus appears to hold approximately for nonidentical waves, provided traffic flow characteristics tend toward long-term averages and are averaged over sufficiently long periods.

Density Calculated from Measured Speed and Flow

Equation 31 gives a relationship between the expected value of density, as measured over an extended section of roadway, and density calculated as the ratio of flow to space mean speed. Because true density data are hard to obtain, one common use of the fundamental relationship has been to calculate density from measured speeds and flows (*13*). In a number of empirical studies of speed-density or flow-density relationships, the "density" data were actually the estimate $\hat{k}$ rather than measured densities.

It is clear such densities are not based on conditions measured over extended sections of roadway; rather, this type of density may more nearly represent the reciprocal of the average distance headway between successive vehicles in the vicinity of a point. This quantity will be referred to as "inverse-spacing density" and the symbol k_s used to designate it.

The inverse-spacing density k_s, unlike k, is a time-based variable. That is, it is measured over time at a point in space or, more literally, over a comparatively short distance. Consequently, relationships involving speed, flow, and inverse-spacing density avoid the difficulties that arise from combining time-based and distance-based variables. As a result, it is possible to derive relationships that hold for the measured values of the variables for particular samples, rather than just for expected values obtained in repeated samples.

The relationship between $\hat{k}$ and inverse-spacing density may be derived as follows. Let x_i be the distance that vehicle i has traveled from some point at the instant vehicle $i + 1$ reaches the point, and t_i be the time elapsed between the time vehicle i passes the point and the time vehicle $i + 1$ passes it. Time t_i for vehicle i is given by

$$t_i = \frac{x_i}{u_i} = x_i \Lambda_i \tag{32}$$

For a total of N vehicles passing the point, the average flow is defined as

$$\bar{q} = \frac{N}{\sum t_i} \tag{33}$$

Then the estimated density $\hat{k}$ is given by

$$\hat{k} = \bar{q}\overline{\Lambda} = \frac{N\overline{\Lambda}}{\sum x_i \Lambda_i} \tag{34}$$

Now, in a procedure similar to that used to derive Equation 20, let x_i be replaced by $\bar{x} + (x_i - \bar{x})$ and Λ_i be replaced by $\overline{\Lambda} + (\Lambda_i - \overline{\Lambda})$. Then

$$\begin{aligned}\hat{k} &= \frac{N\overline{\Lambda}}{\sum\left[\bar{x} + (x_i - \bar{x})\right]\left[\overline{\Lambda} + (\Lambda_i - \overline{\Lambda})\right]} \\ &= \frac{N\overline{\Lambda}}{\sum\left[\bar{x}\overline{\Lambda} + \bar{x}(\Lambda_i - \overline{\Lambda}) + \overline{\Lambda}(x_i - \bar{x}) + (x_i - \bar{x})(\Lambda_i - \overline{\Lambda})\right]} \\ &= \frac{N\overline{\Lambda}}{N\bar{x}\overline{\Lambda} + \bar{x}\sum(\Lambda_i - \overline{\Lambda}) + \overline{\Lambda}\sum(x_i - \bar{x}) + \sum(x_i - \bar{x})(\Lambda_i - \overline{\Lambda})}\end{aligned} \tag{35}$$

In this case, $\Sigma\,(\Lambda - \overline{\Lambda}_i) = 0$, $\Sigma\,(x_i - \bar{x}) = 0$, and $\sigma_{x\Lambda} = \Sigma[(x_i - \bar{x}) \times (\Lambda_i - \overline{\Lambda})]/N$, so Equation 35 may be rewritten as

$$\hat{k} = \frac{\overline{\Lambda}}{\bar{x}\overline{\Lambda} + \sigma_{x\Lambda}} \tag{36}$$

If $\sigma_{x\Lambda} = 0$,

$$\hat{k} = \frac{1}{\bar{x}} \tag{37}$$

This indicates that if there is no correlation between the vehicle spacing and the speed of the individual vehicles (so that the covariance of x and Λ is 0), the estimated density is indeed the reciprocal of the distance spacing of the vehicles in the vicinity of the point of measurement. In reality, however, vehicle spacing and the reciprocal of speed are expected to have a negative correlation, especially in congested flow; consequently, $\hat{k}$ tends to overestimate inverse-spacing density as well as density measured over a section.

This tendency is somewhat counteracted in cases in which the average speed used in the calculation is the time mean speed rather that the space mean speed. In that case

$$\hat{k} = \left(\frac{\bar{u}_s}{\bar{u}_t}\right)\frac{\overline{\Lambda}}{\bar{x}\overline{\Lambda} + \sigma_{x\Lambda}} \tag{38}$$

Since space mean speed is always less than time mean speed, the first term is always less than 1.0; however, the extent to which this counteracts the negative covariance term in the denominator is uncertain.

OCCUPANCY-BASED RELATIONSHIPS

The relationship among speed, flow, occupancy, and vehicle length given by Equation 2 most commonly has been used to calculate estimated speeds from flows and occupancies. For uniform traffic streams, Equation 2 (like Equation 1) can easily be derived by dimensional analysis. In his classical derivation of the speed-flow-occupancy relationship, Athol assumes a uniform vehicle length $\overline{L}$ and proceeds to show that under this assumption,

$$H = \frac{q\overline{L}}{\bar{u}_s} \tag{39}$$

or, in the more familiar form used in Equation 2,

$$\bar{u}_s = \frac{q\overline{L}}{H} \tag{40}$$

Hall and Persaud (*1*) question the validity of this relationship for the realistic case in which vehicle lengths vary within the traffic stream; they also present data that are incompatible with it, although the speeds in the data in question are time mean speeds rather than space mean speeds.

The effect of nonuniform vehicle lengths may be incorporated in the derivation of the speed-flow-occupancy relationship as follows.

Let the speed estimate calculated from flow and occupancy be $\hat{u}$. Then Equation 40 becomes

$$\hat{u} = \frac{q\bar{L}}{H} \tag{41}$$

where $\bar{L}$ now represents the average effective vehicle length, consisting of the sum of the detector length and the vehicle's electrical length, which is related, but not identical, to its physical length. It is assumed that the effective length of an individual vehicle L_i is independent of the speed of the vehicle as it passes the detector, although it may, of course, vary from vehicle to vehicle. Under this assumption τ_i, the time that vehicle i "occupies" the detector, is given by

$$\tau_i = \frac{L_i}{u_i} = L_i\Lambda_i \tag{42}$$

and occupancy, if measured over time interval T, by

$$H = \frac{\sum \tau_i}{T} = \frac{\sum L_i\Lambda_i}{T} \tag{43}$$

Flow, meanwhile, is defined as

$$q \equiv \frac{N}{T} \tag{44}$$

where N is the total number of vehicles passing the detector during time T. Equation 41 may now be written as

$$\hat{u} = \frac{N\bar{L}}{HT} = \frac{N\bar{L}}{\sum \tau_i} = \frac{N\bar{L}}{\sum L_i\Lambda_i} \tag{45}$$

By a derivation similar to that of Equation 36, the denominator of Equation 45 can be shown to equal $N(\bar{L}\bar{\Lambda} + \sigma_{L\Lambda})$, so that it may be rewritten as

$$\hat{u} = \frac{1}{\bar{\Lambda} + \frac{\sigma_{L\Lambda}}{\bar{L}}} \tag{46}$$

If effective vehicle lengths are not correlated with the reciprocal of vehicle speeds, the covariance term is 0 and Equation 46 reduces to

$$\hat{u} = \frac{1}{\bar{\Lambda}} = \bar{u}_s \tag{47}$$

In cases in which speeds and vehicle lengths are correlated, however, the speed estimate calculated from flow and occupancy is not the same as the space mean speed. If there is a correlation between L and Λ, it should be positive, since larger vehicles normally would be assumed to have smaller speeds and hence larger values of Λ. Consequently, $\hat{u}$ may be an underestimate of the space mean speed.

A rough idea of the bias resulting from the covariance term may be gained by assuming a traffic stream composed of two distinct types of vehicles, one large and slow and the other small and fast. By this means it can be shown that for conditions typical of noncongested urban rush hour traffic on relatively flat roads (small percentage of trucks, relatively small difference in speed between trucks and other vehicles) the bias should be small, but that on steep grades with substantial truck traffic it should be quite significant. For instance, for a traffic stream consisting of 95 percent passenger cars with effective lengths of 7 m and speeds of 85 km/hr and 5 percent trucks with effective lengths of 22 m and speeds of 70 km/hr, $\bar{u}_t$ = 84.25 km/hr, $\bar{u}_s$ = 84.09 km/hr, and $\hat{u}$ = 82.45 km/hr. On the other hand, for a traffic stream consisting of 80 percent passenger cars with effective lengths of 7 m and speeds of 85 km/hr and 20 percent trucks with effective lengths of 22 m and speeds of 40 km/hr, $\bar{u}_t$ = 76.00 km/hr, $\bar{u}_s$ = 69.39 km/hr, and $\hat{u}$ = 56.86 km/hr.

DENSITY-OCCUPANCY RELATIONSHIP

Another relationship of interest is that between density and occupancy. Athol (*6*) shows that if it is assumed that vehicles are of uniform length and that the fundamental relationship holds, this relationship is

$$H = k\bar{L} \tag{48}$$

If Athol's assumptions are relaxed, the relationship may be derived as follows for occupancy and point density. From Equation 43,

$$H = \frac{\sum L_i\Lambda_i}{T} \tag{49}$$

Meanwhile,

$$T = \sum \tau_i = \sum x_i\Lambda_i \tag{50}$$

Substituting Equation 50 into Equation 49,

$$H = \frac{\sum L_i\Lambda_i}{\sum x_i\Lambda_i} \tag{51}$$

By logic similar to that used in deriving Equations 36 and 46, it can be shown that

$$H = \frac{\bar{L}\,\bar{\Lambda} + \sigma_{L\Lambda}}{\bar{x}\bar{\Lambda} + \sigma_{x\Lambda}} = \frac{k_s(\bar{L}\,\bar{\Lambda} + \sigma_{L\Lambda}}{\bar{\Lambda} + k_s\sigma_{L\Lambda}} \tag{52}$$

If both $\sigma_{L\Lambda}$ and $\sigma_{x\Lambda}$ equal 0,

$$H = \frac{\bar{L}}{\bar{x}} = k_s\bar{L} \tag{53}$$

which is identical to Athol's result. As argued previously, if the covariances are not 0, $\sigma_{L\Lambda}$ should be positive and $\sigma_{x\Lambda}$ should be negative; consequently, where either covariance is not 0, H should be greater than $k_s\bar{L}$, and the relationship between H and k_s should be nonlinear.

Equation 52 gives the relationship between occupancy and inverse-spacing density. A more interesting comparison may be that between density estimate $\hat{k}$ and occupancy, since these have commonly been the concentration measures used in empirical studies of speed-concentration and flow-concentration relationships. By definition, the estimated density is $\hat{k} = q\bar{\Lambda}$. Meanwhile, combining Equations 41 and 46 leads to

$$\frac{q\bar{L}}{H} = \frac{1}{\bar{\Lambda} + \frac{\sigma_{L\Lambda}}{\bar{L}}} \tag{54}$$

Cross-multiplying,

$$H = q\bar{L}\left(\bar{\Lambda} + \frac{\sigma_{L\Lambda}}{\bar{L}}\right) = \hat{k}\bar{L} + q\sigma_{L\Lambda} \tag{55}$$

Since $\sigma_{L\Lambda}$ is assumed to be positive, H should also normally be greater than $\hat{k}\bar{L}$. Of the two covariances, however, $\sigma_{L\Lambda}$ is more likely to be negligible (except in certain obvious situations such as steep upgrades) than is $\sigma_{x\Lambda}$; hence, the relationship between H and $\hat{k}$ is more likely to be nearly linear than is that between H and k_x.

EMPIRICAL EVIDENCE

A few studies have attempted to verify some of the a priori relationships among speed, flow, and concentration. These include comparisons of occupancy with density by Koshi et al. (*10*) and Athol (*6*) and comparisons of measured speeds with speeds estimated from flows and occupancies by Hall and Persaud (*1*).

Koshi et al. compare occupancies with densities calculated from double-loop data from Tokyo expressways. These densities were thus (presumably) calculated as measured flow divided by measured average speed, which would result in what has been referred to here as $\hat{k}$, provided that the speed data were reduced as space mean speed. They found the relationship to be slightly nonlinear, with a negative second derivative of H with respect to k.

Athol compares occupancy with what he calls accumulation, which also turns out to be $\hat{k}$, and both occupancy and accumulation with what he calls aerial density (density defined over a section, k in the present notation; Athol calls this aerial density because it was measured from aerial photographs). Athol found good agreement between H and $\hat{k}$ and plotted a linear relationship between them, although the data could possibly indicate a nonlinear one. Plots of either H or $\hat{k}$ versus k were badly scattered, however, especially that of $\hat{k}$ versus k.

Hall and Persaud compare measured speed with speed calculated from flow and occupancy ($\hat{u}$ in the present notation) and find major discrepancies between the two, especially for very high and very low occupancies. In general, their data show that if reasonable values of $\bar{L}$ are assumed, $\hat{u}$ will significantly overestimate true speed at very low occupancies and underestimate most of the range representing congested flow.

The experiment actually performed by Hall and Persaud was to calculate a term g, defined as

$$g = \frac{q}{uH} \tag{56}$$

This term corresponds to $1/\bar{L}$ in Equation 41; however, since speeds in their data are measured in kilometers per hour and occupancies in percent, the equivalent average effective vehicle lengths in meters are given by $\bar{L} = 10/g$. In addition, the speeds used in the calculation were time mean rather than space mean speeds.

Hall and Persaud present plots of mean values of g versus H for various locations; Figure 2 is a reproduction of one of these plots. The plots indicate that mean values of g for the very lowest occupancies ranged from 1.8 to 2.4, with the average being about 2.2 for the four locations. Values of g for most of the rest of the uncongested regime are near 1.4 or 1.5, and those for the very highest occupancies (between 70 and 80 percent) range from about 0.3 to 0.6, with an average of perhaps 0.5. The values of g for the higher-volume uncongested regime correspond to a value of $\bar{L}$ of about 7 m, which is credible; those for heavily congested flow, however, correspond to $\bar{L}$ of 20 m, which is not.

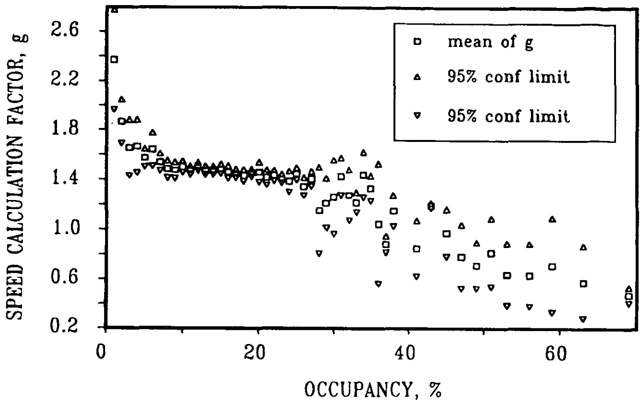


FIGURE 2 Reproduction of plot of mean values of *g* versus *H* (*1*, Figure 2).

The overestimates of speed at very low occupancies are related to data reduction practices. The traffic management system in question reported occupancies in whole percentages and truncated these to the next lowest whole percent. This practice led to the large biases at very low flows (F. L. Hall, unpublished data).

For data taken at high occupancies, biases might result from either the covariance term or the difference between time mean and space mean speed. It is unlikely that the covariance term was very large, however, because the data were taken in the median lane at sites for which trucks were excluded from this lane. This was done to reduce the variance of both speeds and vehicle lengths.

The relationship between time mean and space mean speed was shown by Wardrop (*1*) to be

$$\bar{u}_t = \bar{u}_s + \frac{\sigma_s^2}{\bar{u}_s} \tag{57}$$

where

$$\sigma_s^2 = \sum_{i=1}^{N} \frac{\Lambda_i(\bar{u}_s - u_i)^2}{\sum \Lambda_i} \tag{58}$$

Substituting for σ_s^2 in Equation 57 and simplifying gives

$$\bar{u}_t = \bar{u}_s + \frac{1}{N}\sum \frac{(\bar{u}_s - u_i)^2}{u_i} \tag{59}$$

as a general relationship between the two. From Equation 59, it may be seen that the difference between space mean and time mean speed will increase as the dispersion of the speed distribution increases and that even a very small number of very low speeds could bias the relationship significantly. From this, it may be concluded that the bias will be greatest in heavily congested traffic, but it is difficult to tell how large it might be in any given case.

One past attempt to quantify the relationship between time mean and space mean speed is that of Drake et al. (*14*), who found the following relationship:

$$\bar{u}_s = -1.88960 + 1.02619\bar{u}_t \tag{60}$$

where speeds are in miles per hour. Drake goes on to comment that the maximum difference in the two averages is about 3 km/hr (1.9

mph) at zero speed, a conclusion that proceeds directly from the regression equation. This is certainly much smaller than the bias reported by Hall and Persaud, although, since the relationship is not linear, Drake may have understated the bias at very low speeds. In any case, the discrepancies found by Hall and Persaud appear larger than can be accounted for by the known biases. This raises the possibility that there may have also been counting or data reduction errors under congested conditions, although it is not clear what these might have been.

CONCLUSIONS

Previous derivations of a priori relationships among speed, flow, and concentration variables have assumed that certain features of the traffic stream are uniform. In the case of relationships among speed, flow, and density, either speeds or vehicle spacings (sometimes both) have been assumed to be uniform. It has been shown that, for the fundamental relationship, these assumptions can be relaxed provided that the relationship is understood in probabilistic terms. For relationships involving only time-based terms, these assumptions can also be relaxed in deterministic cases.

Where the order of arrival of vehicles with particular speeds is random, it has been shown that the fundamental relationship applies to average values of the variables in question. Where there are cyclic variations in speeds and flows resulting from waves in congested flow, the expected relationship among speed, flow, and density involves the covariance of flow and the reciprocal of speed and should vary depending on the relationship between the length of the wave and of the section over which density is measured. For the relationship among speed, flow, and what has been called inverse-spacing density (the reciprocal of the average distance separations of vehicles in the vicinity of a point), relaxation of the assumptions of uniform speed or uniform spacing leads to a relationship involving the covariance of vehicle spacing and the reciprocal of speed. For the relationship among speed, flow, vehicle length, and occupancy, relaxation of the assumption of uniform vehicle length leads to a relationship involving the covariance of vehicle length and the reciprocal of speed. It has further been shown that the relationship between inverse-spacing density and occupancy contains both of these covariance terms, but that the relationship between occupancy and density estimated from flow and measured space mean speed contains only the covariance of vehicle length and the reciprocal of speed.

These findings imply that simple relationships among speed, flow, and concentration variables hold a priori not only in cases in which particular variables are uniform (which is almost always unrealistic), but also in cases in which certain variables are not significantly correlated with one another. For the relationship among speed, flow, and occupancy, this is an important advantage, since the covariance of vehicle length and the reciprocal of speed is not likely to be of practical significance except in certain easily identified circumstances such as steep grades with considerable truck traffic. For relationships including inverse-spacing density, the covariance term involving spacing and the reciprocal of speed is likely to be fairly large, especially in congested traffic. This need not be a major problem, however, since inverse-spacing density is not a very useful measure otherwise, and it can be shown that density estimated from flow and measured speed should agree closely with occupancy, except in cases in which speeds and vehicle lengths are strongly correlated.

Comparison of the relationships derived here with empirical studies of the relationship between occupancy and density estimated from flow and measured speed indicates good agreement. In the case of the empirical data in Hall and Persaud's study of speeds calculated from flows and occupancies, on the other hand, there are large discrepancies, both in cases in which occupancies were very low and in most of the congested-flow regime, where they were relatively high. Those involving very low occupancies were due to data reduction techniques. Those involving data from the congested-flow regime may to some extent be due to the difference between time mean and space mean speed or to the covariance term identified here; however, the magnitude of the discrepancy appears to be too large to be explained by the combined effects of these two sources of bias. This situation raises the possibility that there may have also been counting or data reduction errors under congested conditions, although it is not clear what these might have been.

The a priori relationships examined here have commonly been used for calculating speed estimates by traffic management systems and as a basis for studying empirical relationships among traffic flow characteristics. Given the nature of the relationships for nonuniform flow, it appears that the use of the flow-speed-occupancy relationship to estimate speeds and transform variables in empirical studies should be valid in all but a few cases—provided that flows and occupancies are measured accurately. This is certainly true for uncongested conditions. For heavily congested conditions, it should also be true provided speeds are reduced consistently as space mean and the correlation between vehicle lengths and speeds is small. Given the results of Hall and Persaud's study, however, the accuracy of the measurements should not be taken for granted.

ACKNOWLEDGMENT

The author would like to thank Fred Hall for encouraging him to pursue this topic. It was originally suggested by his published work questioning the validity of the fundamental relationship. In addition to the contribution made by Hall's published work, several points addressed in this paper were originally suggested by Hall either in personal communications or in drafts of work that is not yet published.

REFERENCES

1. Hall, F. L., and B. N. Persaud. Evaluation of Speed Estimates Made with Single-Detector Data from Freeway Management Systems. In *Transportation Research Record 1232,* TRB, National Research Council, Washington, D.C., 1989, pp. 9–16.
2. Gilchrist, R. S., and F. L. Hall. Three-Dimensional Relationships Among Flow-Theory Variables. In *Transportation Research Record 1225,* TRB, National Research Council, Washington, D.C., 1989, pp. 99–108.
3. Acha-Daza, J. A., and F. L. Hall. A Graphical Comparison of the Predictions for Speed Given by Catastrophe Theory and Some Classical Models. In *Transportation Research Record 1398,* TRB, National Research Council, Washington, D.C., 1993, pp. 119–124.
4. Pushkar, A., F. L. Hall, and J. A. Acha-Daza. Estimation of Speeds from Single-Loop Freeway Flow and Occupancy Data Using Cusp Catastrophe Theory Model. In *Transportation Research Record 1457,* TRB, National Research Council, Washington, D.C., 1994.
5. Wardrop, J. G. Some Theoretical Aspects of Road Traffic Research. *Proc., Institution of Civil Engineers II,* Vol. 1, 1952, pp. 325–378.
6. Athol, P. Interdependence of Certain Operational Characteristics Within a Moving Traffic Stream. In *Highway Research Record 72,* HRB, National Research Council, Washington D.C. , 1965, pp. 58–87.
7. Makigami, Y., G. F. Newell, and R. Rothery. Three-Dimensional Representation of Traffic Flow. *Transportation Science,* Vol. 5, No. 3, 1971, pp. 302–313.
8. Kühne, R. D. Freeway Control Using a Dynamic Traffic Flow Model and Vehicle Reidentification Techniques. In *Transportation Research*

Record 1320, TRB, National Research Council, Washington, D.C., 1991, pp. 251–259.

9. Kühne, R. D. Macroscopic Freeway Model for Dense Traffic—Stop-Start Waves and Incident Detection. *Proc., 9th International Symposium of Transportation and Traffic Theory,* NVU Press, Utrecht, The Netherlands, 1984, pp. 21–42.
10. Koshi, M, M. Iwasaki, and I. Okhura. Some Findings and an Overview on Vehicular Flow Characteristics. *Proc., 8th International Symposium on Transportation and Traffic Theory, 1981,* University of Toronto Press, Toronto, Ontario, Canada, 1983, pp. 403–426.
11. Lam, T., and R. Rothery. The Spectral Analysis of Speed Fluctuations on a Freeway. *Transportation Science,* Vol. 4, No. 3, 1970, pp. 293–310.
12. Mika, H. S., J. B. Kreer, and L. S. Yuan. Dual-Mode Behavior of Freeway Traffic. In *Highway Research Record 279,* HRB, National Research Council, Washington D.C., 1969, pp. 1–12.
13. Gerlough, D. L., and M. J. Huber *Special Report 165: Traffic Flow Theory: A Monograph.* TRB, National Research Council, Washington, D.C., 1975.
14. Drake, J. S., Shofer, J. L. and A. D. May. A Statistical Analysis of Speed Density Hypotheses. In *Highway Research Record 154,* HRB, National Research Council, Washington, D.C., 1967, pp. 53–87.

DISCUSSION

Michael Cassidy
University of California, Berkeley, 109 McLaughlin Hall, Berkeley, Calif. 94720.

This paper is based on the premise that the equation $q = vk$ is valid only for a limited range of flow conditions such as stationary conditions where all families of vehicle trajectories are parallel, equidistant, and of constant speed. Yet, Edie (*1,2*) provided definitions of flow, speed, and density so as to guarantee the validity of this equation for all traffic conditions. To resolve this apparent contradiction, we apply Edie's definitions in a manner consistent with earlier experiments performed by Hall and Persaud (*3*), which served as motivation for the present paper.

We visualize on a time-space plane a rectangular region A of spatial dimension L, the distance separating paired loop detectors (6 m), and temporal duration T, the count interval (30 sec). An instant at time t within rectangular region A is itself a region (i.e., a "slice") of spatial dimension L and elemental time duration dt. Density at time t is conventionally defined as n/L, the number of vehicles within region A at time t divided by the segment length. Equivalently, density can be expressed as $(n \cdot dt)/(L \cdot dt)$, the ratio of the total time spent by all vehicles in the slice corresponding to time t to the "area" of the slice. As our original rectangular region A is composed of elementary slices, it makes sense to define density in region A as $t(A)/(L \cdot T)$, where the numerator is the total time spent by all vehicles in A. This is Edie's generalized definition of density. Given that region A is a rectangle composed of elementary slices of fixed "area" $L \cdot dt$, this generalized definition is simply the average density over time.

We exchange the roles of space and time and next visualize a single point at location x within rectangular region A. This point defines a region of duration T and elemental spatial dimension dx so that flow can be expressed as $(n \cdot dx/T \cdot dx)$, the total distance traveled by all vehicles crossing point x divided by the "area" of the slice corresponding to this point. As region A is composed of elementary slices of "area" $T \cdot dx$, it makes sense to define flow in A as $d(A)/(L \cdot T)$, where the numerator is the distance traveled by all vehicles in A. Again, this is Edie's generalized definition of flow; it reduces to the conventional definition when A is taken to be a slice of spatial dimension dx and temporal duration T.

As we have now defined these measures, dividing flow by density results in $d(A)/t(A)$, which can be taken to be a definition of an average velocity in region A. This was proposed by Edie, and with his definitions, the equation $q = vk$ is always valid.

The expression evaluated by Hall and Persaud (*3*) is

$$\text{occupancy}/\overline{\ell}_r = \text{flow}/\overline{v}$$

where $\overline{\ell}_r$ is mean effective vehicle length and $\overline{v}$ is mean speed. This can likewise be shown to be true by definition provided measurements of $\overline{v}$ are the generalized ones because the left-hand side can be shown to be Edie's generalized definition of density and the conventional definition of flow appearing in the right-hand side coincides with the generalized one. [When correlations exist between vehicle length and speed, $\overline{\ell}_e$ must be an average value occurring over space. When traffic is not stationary, occupancy measured at a point in space over time (i.e., by a detector) is not equivalent to a spatial measure of road occupancy. We anticipate addressing these issues at some future date.] Reported discrepancies can be explained by methods used for averaging observations. Edie's definition of average speed can be computed using the arithmetic mean of trip times between two points (e.g., paired detectors) or as the harmonic mean speed of vehicles passing a single point (e.g., detector) when conditions are stationary. There is no reason to expect a $\overline{v}$ calculated in a different manner to satisfy the relation.

REFERENCES

1. Edie, L. C. Discussion of Traffic Stream Measurements and Definitions. *Proc., Second International Symposium on the Theory of Traffic Flow,* OECD, Paris, 1965, pp. 139–154.
2. Edie, L. C. *Traffic Science* (D.C. Gazis, ed.). Wiley and Sons, New York, 1974, pp. 8–20.
3. Hall, F. L., and B. N. Persaud. Evaluation of Speed Estimates Made with Single-Detector Data from Freeway Traffic Management Systems. In *Transportation Research Record 1232,* TRB, National Research Council, Washington, D.C., 1989, pp. 9–16.

AUTHOR'S CLOSURE

I wish to thank Mr. Cassidy for calling attention to work by Edie, which I probably should have mentioned. It is, however, just one more in a long series of efforts to make the fundamental relationship work by imposing special conditions or (in this case) by adopting unnatural definitions of the variables. Strictly speaking, none of Edie's variables can actually be measured, although they can be fairly closely approximated under the conditions Cassidy outlines. Certainly, they are not the conventional definitions of the variables in question. The thrust of my paper was to acknowledge (and in some cases elaborate on) these special cases while at the same time determining the nature of the discrepancies that result when conventional definitions of the variables are used.

In the case of Hall and Persaud, the special case proof is interesting, but it does not address the practical concern underlying their work. The occupancies they were concerned with were measured over time, and the traffic flow was not stationary. Also, given that the flows were not stationary, none of the possible measures of speed available to them really conformed to Edie's definition. Given that situation, I believe that it makes sense to ask whether the biases introduced by the nonstationary traffic stream account for the discrepancies they observed. I am not so sure that it makes sense to try to define a relationship that is a priori true but involves variables that are unlikely ever to be measured in practice.

DISCUSSION

Matti Pursula
Helsinki University of Technology, Rakentajanaukio 4A, FIN-02150 Espoo, Finland.

The paper deals with the fundamental traffic flow relationships and the problems that basically local measurement techniques cause in the analysis of these relationships with some sectionwide variables.

I have three main comments on the paper. The first is that it is hard for me to understand why the writer, who clearly wants to challenge some basic traffic flow theory paradigms, fails to give precise definitions for his variables in each case of analysis. This vagueness, I believe, is also the reason for some errors that can be found in the equations that he derives. In addition to that, the vague definitions can cause misinterpretations of correct results.

My second comment is that the writer does not at all describe the most general definitions of the three basic traffic flow variables (*1,2*), which guarantee that the fundamental flow relationship is valid for any kind of traffic (congested, uncongested, random, uniform, etc.) in any time-space domain XT (X being the space axis and T being the time axis).

These definitions are as follows:

- Traffic flow q = the amount of vehicle-kilometers of travel (S) in the domain divided by the area of the domain, that is,

$$q = \frac{S}{X \times T}$$

- Traffic density k = total vehicle-hours of travel (T_{Tot}) in the domain divided by the area of the domain, that is,

$$k = \frac{T_{\text{Tot}}}{X \times T}$$

- Space mean speed of traffic u = total vehicle-kilometers of travel in the domain divided by the total vehicle-hours of travel in the domain, that is,

$$u = \frac{S}{T_{\text{Tot}}}$$

From these definitions it can easily be seen that the fundamental flow relationship $q = u \times k$ is valid for the time-space domain in question. The derivation of these relationships is given for example by Leutzbach (*2*).

So, in theory, traffic flow variables can be measured in a way that is in accordance with the fundamental flow relationship. The problems arise from our inability to measure the variables simultaneously in time and space.

My third comment is related to Equations 20 through 31 in the paper. In this part of the paper the writer develops equations for a time-space domain TX in a cyclic flow situation. His derivations leading to Equation 24 are correct. This equation gives the definition of traffic density averaged over the time-space domain. It can easily be seen that his result is in accordance with the above given general definition, that traffic density equals the amount of vehicle hours in the time-space domain divided by the area of the domain. In Equation 24 only division by time is needed, because the derivation of the equation already averaged the value over the space axis.

In Equations 25 and 26 the writer makes some basically correct mathematical manipulations of Equation 24 to develop it further. But then he makes a major error in the definition of space mean speed, given in the form of the reciprocal of speed, in Equation 28. According to that equation the mean travel time in the domain is the average value of travel time $\Lambda(t)$ over the time axis without consideration of the number of vehicles traveling within the travel time in question. The error can be seen from the following equations, the first being the one given in the paper and the second the correct one (in two equivalent forms).

$$\int_0^T [\Lambda(t) - \overline{\Lambda}]dt = 0 \qquad \text{(Equation 28)}$$

$$\overline{\Lambda} = \frac{\int_0^T q(t)\Lambda(t)dt}{\int_0^T q(t)dt} \quad \text{or} \quad \frac{1}{T}\int_0^T [q(t)\Lambda(t) - \overline{q}\overline{\Lambda}]dt = 0 \qquad \text{(correct form)}$$

When Equation 28 is replaced with the correct one, the calculations based on Equation 26 result in a simple identity (i.e., Equation 25).

On the basis of Equations 25 and 27 and the correct equation for mean travel time given above, one can easily see that the fundamental flow relationship holds for this situation and no correction term is needed in the calculation of density.

REFERENCES

1. Edie, L. Flow Theories. In *Traffic Science* (D. Gazis, ed.), John Wiley and Sons, New York, 1974, pp. 2–108.
2. Leutzbach. Introduction to the Theory of Traffic Flow. Springer-Verlag, Berlin, 1988.

AUTHOR'S CLOSURE

I wish to thank Pursula for pointing out the mistake in Equation 28 and calling attention to work by Edie and Leutzbach, which I probably should have mentioned. In the case of Equation 28, his version is the correct one, and the consequence is indeed that the correction term disappears. This leads to the somewhat more satisfying conclusion that the fundamental relationship holds without bias for cyclic flow, regardless of the relationship between the wavelength and the distance over which density is measured.

The work by Edie and Leutzbach is of historical interest and should have been cited. It is, however, just one more in a long series of efforts to make the fundamental relationship work by imposing special conditions or (in this case) by adopting unnatural definitions of the variables. Strictly speaking, none of Edie's variables can be measured. Certainly, they are not the conventional definitions of the variables in question. The thrust of my paper was to acknowledge (and in some cases elaborate on) these special cases while at the same time determining the nature of the discrepancies that result when conventional definitions of the variables are used.

I am somewhat puzzled by the assertion that my definitions of the variables are vague. In most cases the exact mathematical meaning is stated. It is true that in the "special case" formulations there are variations in the precise definitions from one formulation to another, but these are a result of attempts to make the fundamental relationship work. Finally, I must deny that it was my intent to "challenge some basic traffic flow theory paradigms." On the contrary, they had already been challenged, most notably by Hall, and my intent was to try to limit the uncertainty by determining, where possible, the nature of any biases.

Description of Macroscopic Relationships Among Traffic Flow Variables Using Neural Network Models

TAKASHI NAKATSUJI, MITSURU TANAKA, POURMOALLEM NASSER, AND TORU HAGIWARA

The relationships between traffic flow variables play important roles in traffic engineering. They are used not only in basic traffic flow analyses but also in some macroscopic traffic flow simulation models. For many decades, various mathematical formulations that describe the relationships among density, flow, and speed have been proposed, including multiregime models. Previously, the best mathematical curve was determined by trying several different formulas and applying regression analysis. In these processes, one must specify in advance which mathematical formula should be adopted and where it should be shifted to another in a multiregime model. Neural network models have some promising abilities to represent nonlinear behaviors accurately and to self-organize automatically. A procedure for describing the macroscopic relationships among traffic flow variables using some neural network models is presented. First, a Kohonen feature map model was introduced to convert original observed data points into fewer, more uniformly distributed ones. This conversion improved regression precision and computational efficiency. Next, a multilayer neural network model was introduced to describe the two-dimensional relationships. The model was effective in describing the nonlinear and discontinuous characteristics among traffic flow variables. It was unnecessary to specify the regression curves and the transition points in advance. The multiple correlation coefficients resulting from the model were better than those resulting from a conventional nonlinear equation.

The relationships among traffic flow variables play important roles in traffic engineering. They are used not only in analyses of traffic flow behavior but also in some macroscopic traffic flow simulation models. For many decades, traffic-flow analysts have studied various mathematical formulations that describe the relationships among density, flow, and speed of uninterrupted traffic flows (*1–3*). The best mathematical formerly was determined by trying several formulas and applying regression analysis techniques. In some cases, one equation may be most appropriate; another may be better in the others. Moreover, multiregime models that use a few functions have been proposed, too (*1–3*). Normally they include discontinuity points not only in original functions but also in their derivatives—that is, in applying such models, one must specify in advance which mathematical formula should be adopted in each region and where it should be shifted to another.

As a matter of fact, the authors are now engaged in the development of a traffic flow simulation model (*4*) in which a characteristic curve prescribes the average traffic states. The curve is updated using traffic detector data at each observation point. The authors used to face the aforementioned difficulties in establishing a macroscopic relationship in a computer.

Some neural network models, such as a multilayer model (*5*), have the promising ability to describe nonlinear behaviors very well. So, it is expected that when they are applied to the regression problem, they can self-adjust the curvature of characteristic curves automatically while responding to the distribution of observed data. Above all, they require no preliminary knowledge of the mathematical formulas and the transition points. Another difficulty in regression analysis lies in the trimming of excessive observed data. When traffic flows are observed, often one comes across unequally distributed traffic data—distributed densely in a few restricted regions and sparsely in the others. This unequal distribution of observed data would affect the regression results badly. Excessive observed data in a region decrease the computational efficiencies, too. One must determine in advance which data should be retained and which should be trimmed. Some statistical criteria, such as AIC (Akaike information criteria) and FPE (final prediction error) (*6*), may provide useful knowledge about how *much* data should be retained, but they provide no information about *which* should be retained.

Some neural network models, such as a Kohonen feature map (KFM) model (*7*), have the ability to convert original observed data into fewer, more representative data automatically. The KFM model does not require any preliminary knowledge about the data structure. All one must do is specify the number of data points to which the original data set should be reduced.

BACKGROUND

Characteristic Curves

There are many characteristic curves proposed so far for describing the relationship between density and speed. In this study the authors used the formula derived from the car-following theory (*3*):

$$v = v_f\left[1-\left(\frac{k}{k_j}\right)^{l-1}\right]^{\frac{1}{1-m}} \tag{1}$$

Civil Engineering Department, Hokkaido University, Kita 13, Nishi 8, Kitaku, Sapporo, 060, Japan.

where

k = density (veh/km),
v = speed (km/hr),
k_j = jam density,
v_f = free speed, and
l,m = sensitivity factors from car-following theory.

Substituting Equation 1 into the relationships $q = kv$, the other relationships among density, flow, and speed can be obtained as follows:

$$q = v_f k\left[1-\left(\frac{k}{k_j}\right)^{l-1}\right]^{\frac{1}{1-m}} \qquad (2)$$

$$q = k_j v\left[1-\left(\frac{v}{v_f}\right)^{1-m}\right]^{\frac{1}{l-1}} \qquad (3)$$

where q denotes the traffic flow rate in vehicles per hour. The unknown parameters in those equations are subject to some constraints *(3)*:

$$\begin{aligned} &l > 0 \\ &m > 1 \\ &v_f^{\min} \le v_f \le v_f^{\max} \\ &k_j^{\min} \le k_j \le k_j^{\max} \end{aligned} \qquad (4)$$

Regression Analysis

Equations 1, 2, and 3 are expressed in a general form

$$y = f(x, a_1, a_2, a_3, a_4) \qquad (5)$$

where

x = control variable,
y = state variable, and
a_j (j = 1,2,3,4) = unknown parameters of l, m, v_f, and k_j in Equations 1–3, respectively.

By obtaining sets of observed data (x_i, y_i) (i = 1,2,...,N), one can identify the parameters by a regression technique. Since Equation 5 is in nonlinear form and is subject to some constraints given by Equation 4, the problem here reduces to a nonlinear constrained least mean square problem. That is, the unknown parameters are estimated so as to minimize the objective function J as follows:

$$J = \sum_i [y_i - f(x_i, a_1, a_2, a_3, a_4)]^2$$

Subject to

$$G_j \le a_j \le H_j \qquad j = 1,2,3,4 \qquad (6)$$

The authors used Box's complex algorithm to solve this problem. A detailed discussion of this algorithm can be found elsewhere *(11)*..

Multilayer Neural Network Model

Figure 1 shows a multilayer neural network model for describing the macroscopic relationships between traffic variables. It consists of three layers: an input layer, an intermediate layer, and an output layer. The strength of the connections is called synaptic weight. The normalized control variable x_i^B was entered into the input layer, such as k/k_j in Equation 1. The input signals were transmitted in sequence from the input layer to the output layer while the neural operations were repeated. The output layer produces the normalized objective variable y_k^D, such as v/v_f in Equation 1. This is the forward signal process in Figure 1. Next, the synaptic weights were adjusted so that the error between the output signals and the target signals is minimized. The backpropagation method *(5)* produces the adjustments of synaptic weights in each layer. In actual computation the synaptic weights are adjusted by the momentum method to smooth the adjustment and urge the convergence.

Kohonen Feature Map

The KFM model is a two-layered neural network that can organize a topological map from a random starting point. It has the ability to classify input patterns into several output patterns. Figure 2 depicts the basic network structure of a KFM model. the authors used a one-dimensional structure for this analysis. It consists of two layers: an input layer and a competitive layer. The interconnections (synaptic weights) are adjusted in a self-organizing manner without any target signals. the authors briefly explain how this can be done. An input pattern to the KFM is denoted here as

$$E = [e_1, e_2, \ldots, e_{n]} \qquad (7)$$

Since the observed traffic variables are adopted as the input signals, the input layer has three neurons in it (n=3). The weights from the input neurons to a single neuron in the competitive layer are denoted as

$$W_i = [w_{1i}, w_{2i}, \ldots, w_{ni}] \qquad (8)$$

where i identifies the ith neuron in the competitive layer. The number of neurons there can be specified arbitrarily.

The first step in the adjustment of synaptic weights is to find a winning neuron c in the competitive layer whose weight vector

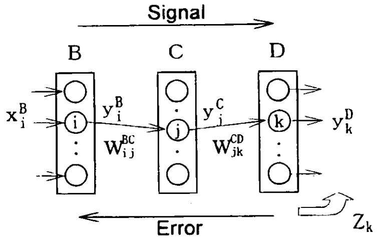


FIGURE 1 Basic structure of multilayer neural network.

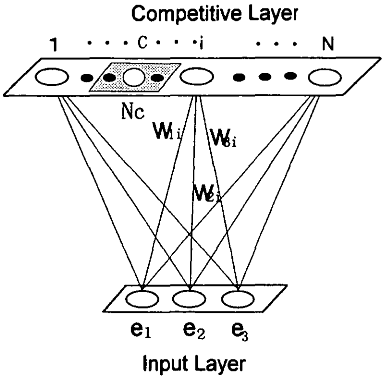


FIGURE 2 Basic structure of KFM model.

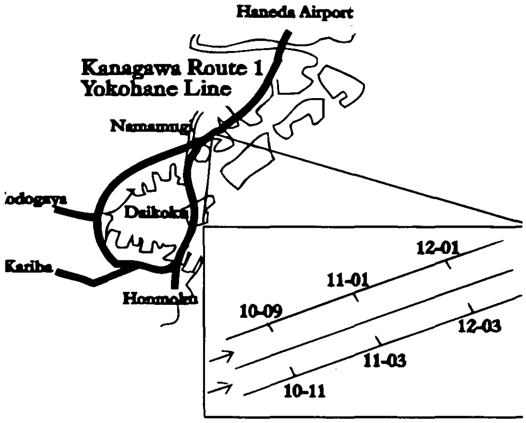


FIGURE 3 Overview of Yokohane Line and locations of traffic detectors.

matches most to each input vector E. The matching value is defined by the distance between vectors E and W_i:

$$\sqrt{\sum_j (e_j - w_{ij})^2} \tag{9}$$

The neuron with the lowest matching values wins the competition. After the winning neuron c is identified, weights are updated for all neurons that are in the neighborhood N_c of the winning neuron. The adjustment is

$$\Delta W_{ij} = \begin{cases} \beta(e_j - w_{ij}) & \text{if } i \in N_c \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where β is the learning rate, which is decreased over a span of many iterations. This adjustment results in the winning neuron becoming more likely to win the competition when the same or similar input pattern is presented subsequently. In other words, the synaptic weight vector W_i consequently represents those input patterns that resemble each other. This is what is called the integration of observed data. See work by Dayhoff (*7*) for more details.

TRAFFIC DATA

Observed Data

The observed data used here come from the Metropolitan Expressway in Tokyo. The data were collected on the Yokohane Line between Tokyo Haneda Airport and Yokohama in October 1993. Supersonic traffic detectors are installed in each of the two directional lanes every 300 m, and traffic data on flow, occupancy, and average speed are compiled every 1 min. Figure 3 depicts the schematic drawing of the freeway section and the location of the traffic detectors. Traffic data on both lanes in the eastbound direction from Yokohama to Tokyo Airport were used. This road section experiences incessant congestion in the daytime on weekdays. The authors chose such time periods that include extensive traffic situations, ranging from free-flow to congested states. In this analysis, assuming that density is proportional to time occupancy, the authors used time occupancy directly rather than converting it to density (*8*). This requires a minor change in the nonlinear equations from Equations 1 through 3. Assuming homogeneity around observation points, the authors treated the time-mean speed as identical with the space-mean speed. However, it should be noted that this assumption is not always valid. One must examine carefully what has been analyzed, in particular when traffic is congested.

Training

By using the KFM model in three-dimensional space, the original observed data were converted to fewer points of more integrated data. Figure 4 depicts the schematic drawing of the conversion. Iterative trainings by the model produce the neurons whose weights correspond to integrated data. They were projected on each two-dimensional plane for two-dimensional analysis. Next, by using a multilayer neural network model, the input-output relationships between the control and the state variables were built up. The completion of training by the backpropagation method brings a stable regression between them.

Kohonen Feature Map

To convert observed data to sets of integrated data, the authors prepared a KFM model consisting of an input layer with three neurons and a competitive layer with neurons that correspond to the number of integrated data points. Before the training, all observed data are normalized. After having given a set of observed data to the input layer in Figure 2, the authors selected a winning neuron in the competitive layer and adjusted the weights of neurons in the neighborhood of the winning neuron. This process is iterated for all input patterns consecutively. Training iteration continues until the change of synaptic weights becomes sufficiently small. Finally, a stable formation of integrated data can be obtained.

The most important problem in this process is how to determine the number of integrated data points. Generally, the appropriate number of data points depends on the use of a characteristic curve; for interpreting traffic flow behaviors, the number must be determined carefully so as to not lose the original data properties. One must determine it while checking the information statistics based on a criterion, such as AIC or FPE. On the other hand, for using a char-

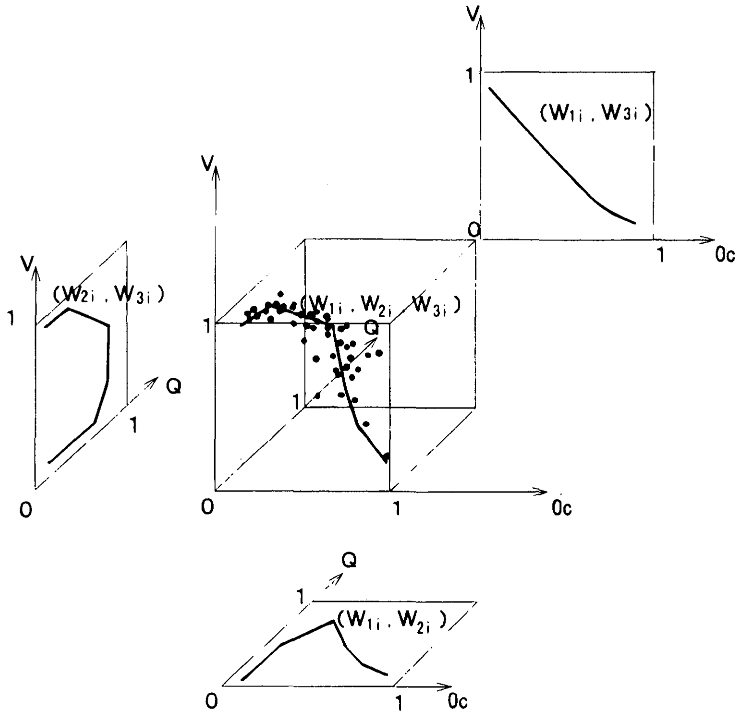


FIGURE 4 Schematic drawing of integration of observed data by Kohonen feature mapping.

acteristic curve in a simulation model, excessive data should be trimmed because such data affect the regression badly. In this case, one may be able to determine the number experimentally because only the average characteristics of traffic flow states are needed. In this paper, assuming the usage in a simulation model, the authors determined it experimentally: 20 points for each data set containing 120 points of observed data.

Figure 5 shows how original observed data are integrated as the training proceeds. For simplicity, the evolution process is projected on the occupancy-speed plane, and, for convenience, it is enlarged to the real scale. White circles in Figure 5*a* are original observed data, and black ones in the center of the graph are the initial weights that are set to the value 0.5 plus a small, within 10 percent, randomized value. Figures 5*b–d* show the distribution of the neuron weights after 50, 130, and 200 training iterations, respectively. It can be seen that the weights spread out gradually over the original space as the training proceeds. As shown in Figure 5, the KFM model requires nearly 100 to 300 iterations to complete the training.

Multilayer Neural Network

As mentioned before, the authors prepared a multilayer network with a neuron in the input layer and a neuron in the output layer for two-dimensional analysis. The synaptic weights were adjusted by the back-propagation method.

In this paper the authors adopted a training procedure (*9*) that is somewhat different from the usual one. Here, the authors adjust the weights thoroughly for an input pattern until the error between the output signal and the target signal becomes sufficiently small. The adjustment is repeated for all input patterns. The completion of adjustment for an input pattern deteriorates the synaptic weights for the other patterns, so that those training processes are iterated hundreds or thousands of times, normally 10,000 to 30,000 times. The training method adopted here was effective in avoiding entrapment into a local minimum and converged steadily to a global minimum.

RESULTS

In presenting how well the neural network models describe the nonlinear phenomena without any specific functions, the authors compare two methods: an analytical one by nonlinear equations, and one using artificial intelligence through neural network models. However, the authors refrain from interpreting the curves from the traffic flow viewpoints because there is much to do before doing so, including determining the appropriate number of integrated data points.

Occupancy-Speed Curve

First, the methods are compared using the traffic data observed at Detector Station 1201. The period is 2 hr. Figure 6 shows three regression curves: (*a*) a curve by a nonlinear equation for original observed data, (*b*) one by a neural network model without the KFM model, and (*c*) one by a neural network model with the KFM model.

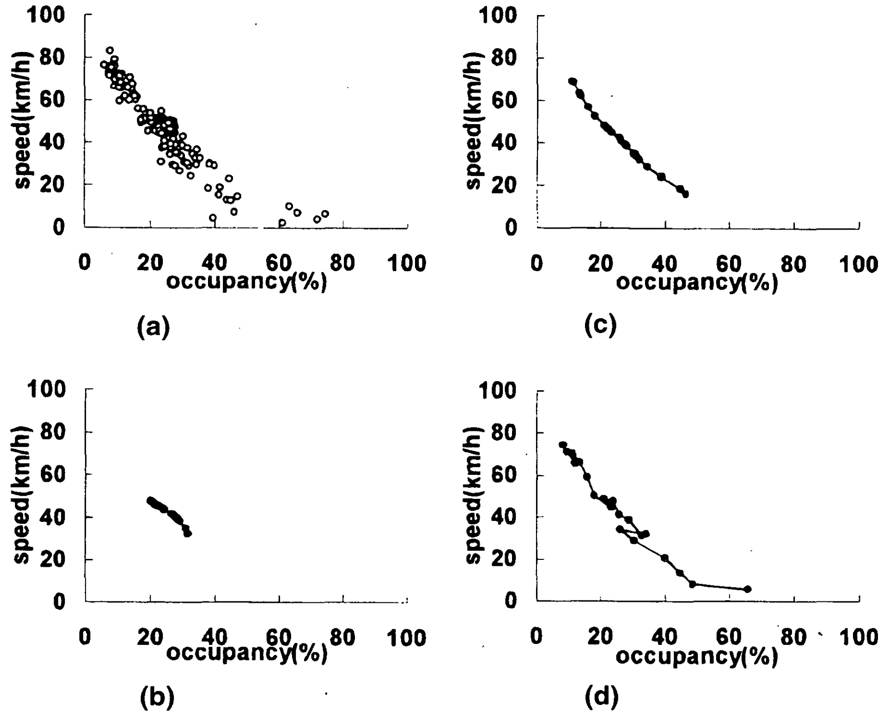


FIGURE 5 Evolution into integrated data by Kohonen feature mapping: *a,* 0 iterations; *b,* 50 iterations; *c,* 130 iterations; *d,* 200 iterations.

The white circles in Figures 6*a* and *b* present 120 points of original observed data, and the black ones in Figure 6*c* present 20 points of data integrated by the KFM model.

It is seen in Figures 6*a* and *b* that the observed data are excessively distributed in both regions where time occupancy is from 5 to 15 percent and from 20 to 40 percent. Those excessive data points affect the regression curve very badly. It should be noted that the shape of the curves is quite different in the high-occupancy region (occupancy is more than 40 percent), although there is little difference in the correlation coefficients, as presented in Table 1. This means that the densely distributed data in the low- and middle-occupancy regions almost govern the curve, and to the contrary, the data in the high-occupancy region have little effect on it.

On the other hand, Figure 6*c* shows 20 points of integrated data and the regression curve by the neural method with the KFM model. It is seen that by introducing the KFM model, the authors were able to make the original data more uniformly distributed. In particular, the five original data in the high-occupancy region in Figure 6*a* are reduced to two sets of data in Figure 6*c*. This favorably improved the regression in the region. One can see that the regression curve with the KFM is located in the middle of the original observed data in the high-occupancy region. This appears to be desirable for applying the curve in a traffic simulation model. However, for interpreting traffic flow phenomena in the region, the overtrimmed curve is not adequate. In such cases, one should increase the number of integrated data or use original raw data.

Figure 7 shows the regression curves for the other detector stations. As in Figure 6, the white circles are the original data, the black are the integrated data, and the thick line is the regression curve produced by the multilayer neural method. One realizes at a glance that the regression curves are more complicated than those of the nonlinear equation in Figure 6*a*. It is seen in the low-occupancy region that the curves have a "snake head": they are nearly flat where time occupancy is less than 15 percent. In addition, the regression curves consist of a few convex parts. In other words, they are discontinuous in their derivatives. Likewise, a small gap can be seen around the time occupancy of 20 percent in Figure 7*b*.

In this way, the neural network method has the promising ability to describe a discontinuous relationship more precisely. It needs neither to divide the whole region into several nor to introduce an individual function for each region. Unfortunately, those features of the neural network models are not easy to evaluate quantitatively. However, the correlation coefficients reflect those features indirectly. Table 1 presents the coefficients produced by both of the neural methods along with those produced by the nonlinear equation for all cases. It is seen that the neural methods are better than the nonlinear equation. Also, there is little difference between both of the neural methods. This means that the neural network models can flexibly self-adjust the curvature of regression curves according to the number of data points. Needless to say, the neural method with the KFM model is more efficient in the computation than that without the KFM model.

Occupancy-Flow Curve

Figure 8 presents the regression results by both methods for the occupancy-flow curve at Detector Station 1201. Compared with the occupancy-speed curve in Figure 6, the behaviors are a bit more

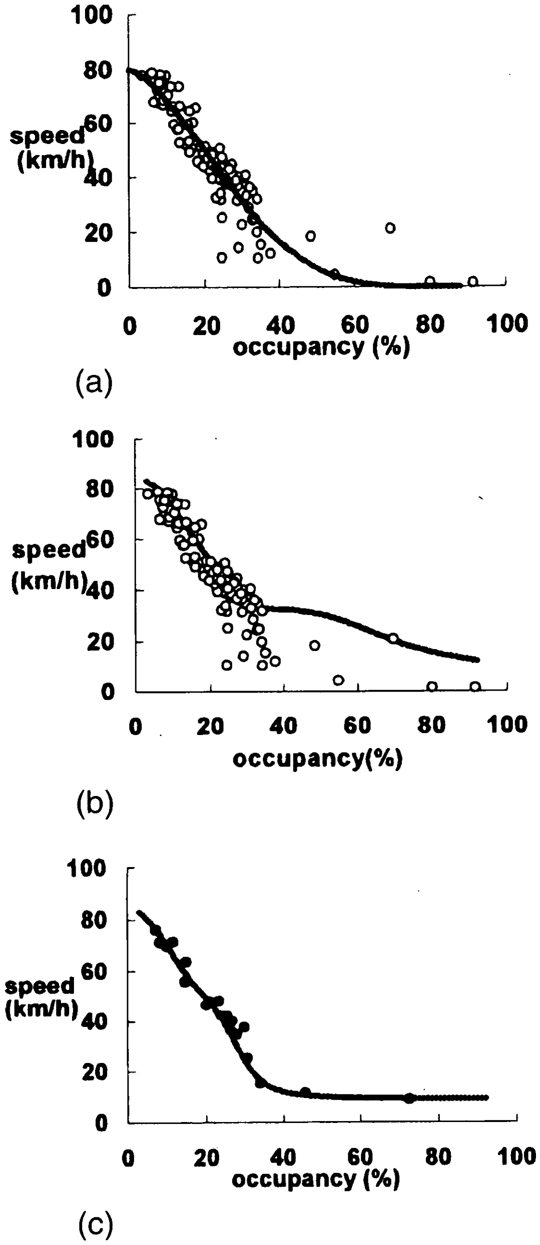


FIGURE 6 Comparison of neural network models with nonlinear equation on occupancy-speed curve: *a,* nonlinear equation; *b,* neural network without KFM; *c,* neural network with KFM.

complex. Clearly, the nonlinear equation in Figure 8*a* fails to describe the relationship in the congested region. On the contrary, as shown in Figure 8*b*, one can recognize the good regression in the region. Integration of original data in the high-occupancy region into a few data points contributed to this improvement. Of course, it must be examined carefully if the number of data points in the region is sufficient or not, according to the purpose for which the curve is used. In addition, one can see that the curve is not so well regressed in the vicinity of capacity, apparently because of the data being scattered in the region. That is, even the neural method cannot describe such data. The description for such data is the most difficult subject in the mathematical formulations.

TABLE 1 Comparison of Multiple Correlation Coefficients on Occupancy-Speed Curve

Detector Point	Non-linear Equation	Neural Network	
		without KFM	with KFM
1009	0.94	0.97	0.97
1011	0.87	0.92	0.94
1103	0.91	0.94	0.95
1201	0.88	0.91	0.92
1203	0.92	0.96	0.97

Figure 9, similar to Figure 7, shows the regression curves by the neural method for the other cases. One can see that the distribution of integrated data is more complicated than that of those in the occupancy-speed curves in Figure 7: the thin curve that connects the integrated data in sequence has two peaks. It should be noted that the regression curve (thick line) in Figure 9*a* corresponds well to the movement of the data. In this way, the neural method is able to describe such a complex relationship, too. Here also, one must care-

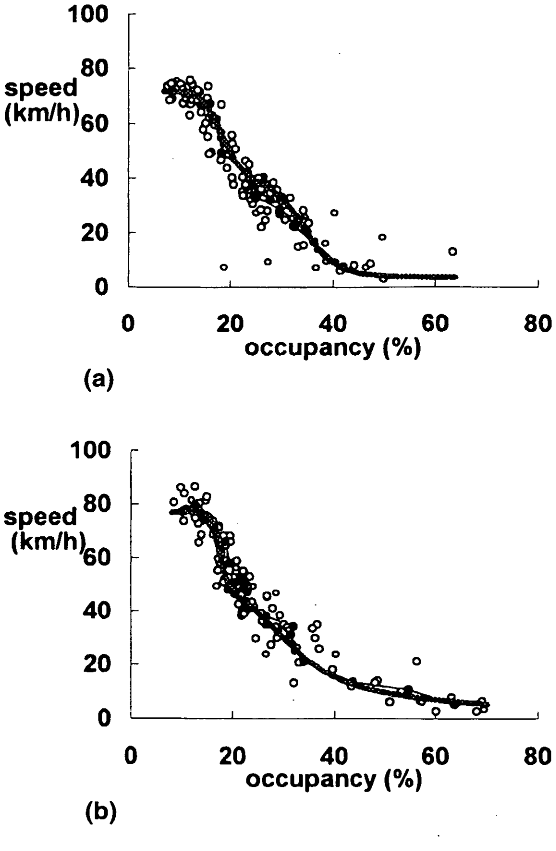


FIGURE 7 Occupancy-speed curves by neural network models: *a,* Station 1011; *b,* Station 1103.

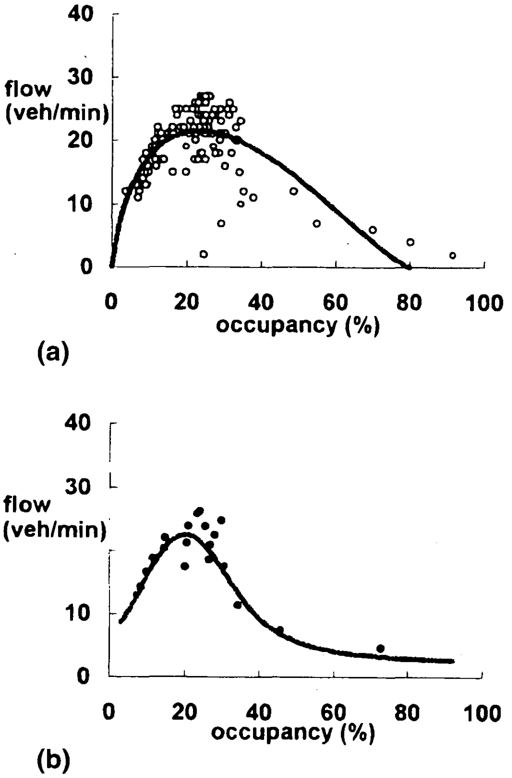


FIGURE 8 Comparison of (*a*) neural network models with (*b*) nonlinear equation on occupancy-flow curve, Station 1201.

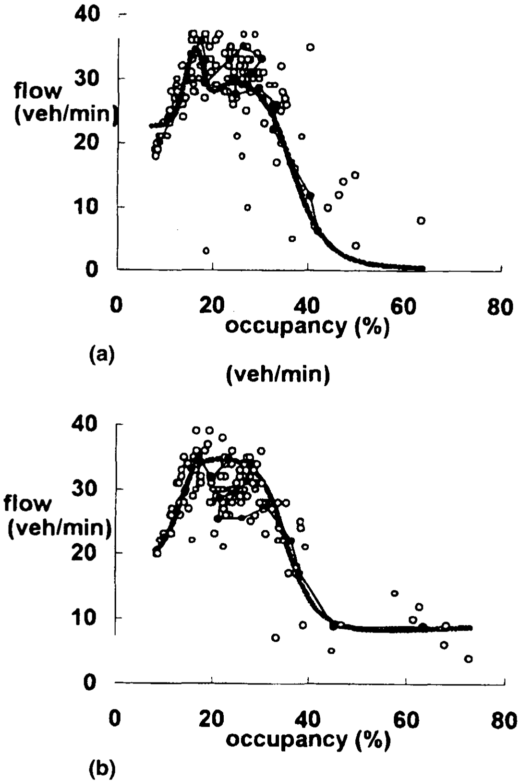


FIGURE 9 Occupancy-flow curves by neural network models: *a,* Station 1011; *b,* Station 1203.

fully examine the validity of the curve from traffic engineering viewpoints.

On the other hand, the regression curve in Figure 9*b* is relatively smooth, although the integrated data are distributed zigzag as in Figure 9*a*. This is because even the neural network model is not able to describe such a function that has two or more state values for a control value. In this case, there are two or three flow values for an occupancy value near capacity. Anyway, it should be noted that the curves are not so well regressed yet in the vicinity of capacity in both of the figures. For reference, the correlation coefficients for all cases are given in Table 2. One can see that the neural method is much better than the nonlinear equation.

Flow-Speed Curve

In general, flow-speed curves become more complicated because of the transition of traffic states (*10*). They would take a different path according to whether the traffic goes into congestion or recovers to free-flow state. However, in this paper, neglecting those dynamic behaviors, the authors treated traffic states as static ones. Figure 10, similar to Figures 7 and 9, shows regression curves for two cases, in which the authors treated speed as the control variable and flow as the state variable. Because of the lack of observed data in the free-flow state, the regression curve cannot be seen in the high-speed region. The curve in Figure 10*a* presents a somewhat poor regression with the integrated data points around capacity whereas the one in Figure 10*a* follows them somewhat better. To trace the data in Figure 10*a* more precisely, it may be necessary to change the number of data points. But this should be done only if it is meaningful from the viewpoint of traffic engineering. Here also, as indicated in Table 3, the neural method gives better correlation coefficients than the nonlinear equation.

CONCLUDING REMARKS

The relationships among traffic flow variables play important roles in traffic engineering. They are used not only in analyses of traffic flow behaviors but also in some macroscopic traffic flow simulation models. Noting that some neural network models have promising abilities to represent nonlinear behaviors and to self-organize automatically, the authors applied them to the description of the relationships. First, the authors introduced a KFM model to integrate the original observed data points into fewer, more uniformly distributed ones. Next, a multilayer neural network model was used to describe the relationships between traffic flow variables. the authors investigated the applicability of the neural network models to the regression problem and compared the results with those produced by a conventional nonlinear equation. The major findings are as follows:

1. A KFM method served to integrate original observed data points into fewer, more uniformly distributed data points. All that

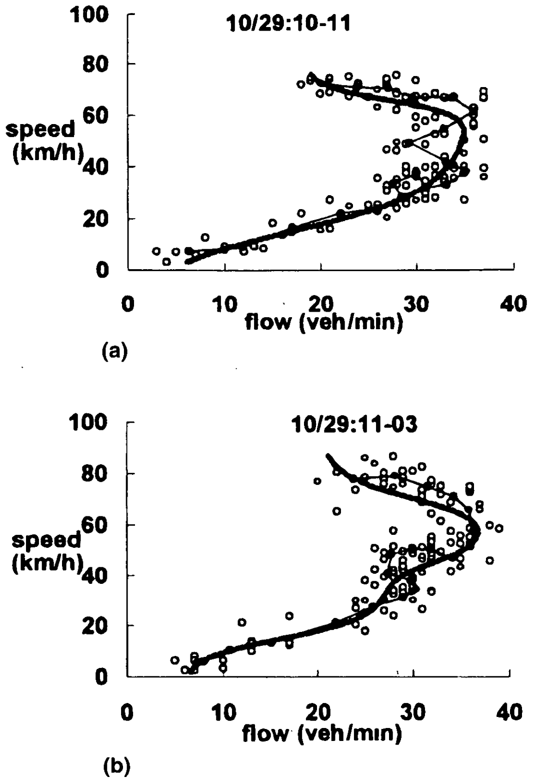


FIGURE 10 Flow-speed curves by neural network models: *a,* Station 1011; *b,* Station 1103.

TABLE 2 Comparison of Multiple Correlation Coefficients on Occupancy-Flow Curve

Detector Point	Non-linear Equation	Neural Network with KFM
1009	0.60	0.78
1011	0.47	0.58
1103	0.74	0.79
1201	0.52	0.66
1203	0.61	0.80

must be done to specify the desired number of integrated data points. This integration contributes to the improvement of regression precision and computational efficiency.

2. A multilayer neural network model was effective in describing the nonlinear and discontinuous relationships between traffic flow variables. The model made it unnecessary to specify the regression curves and the transition points in advance. In addition, the multiple correlation coefficients produced by the model were better than those produced by a nonlinear equation.

TABLE 3 Comparison of Multiple Correlation Coefficients on Flow-Speed Curve

Detector Point	Non-linear Equation	Neural Network with KFM
1009	0.70	0.83
1011	0.74	0.85
1103	0.79	0.82
1201	0.63	0.81
1203	0.75	0.81

The method proposed here still has some disadvantages: it requires a bit of burdensome work to estimate some fundamental traffic parameters, such as maximum volume, which are significant for analyzing traffic flow behavior.

In this paper, the discussion was limited to the availability of neural network models. The interpretation of traffic phenomena using them is left to future work. Moreover, the availability of other neural models that might be more effective than those used here must be examined.

ACKNOWLEDGMENTS

The authors thank Dr. Morita and his colleagues at the Metropolitan Expressway Office in Tokyo for offering them observation data. Financial support was provided by the Education Ministry of Japan.

REFERENCES

1. Gerlough, D., and M. Huber. *Special Report 165: Traffic Flow Theory.* TRB, National Research Council, Washington, DC., 1976.
2. McShane, W. R., and R. P. Roess. *Traffic Engineering.* Prentice-Hall, Englewood Cliffs, N.J., 1990.
3. May, A. D. *Traffic Flow Fundamentals.* Prentice-Hall, Englewood Cliffs, N.J., 1990.
4. Nakatsuji, T., and T. Kaku. Improvement of Traffic Flow Simulation Precision by Direct Usage of Traffic Detector Data. *Proc., Infrastructure Planning,* Vol. 16, No 1, 1993, pp. 115–120.
5. Wasserman, P. D. *Neural Computing.* Van Nostrand Reinhold, 1989.
6. Sakamoto, Y., M. Ishiguro, and G. Kitagawa. *Information Statistics.* Kyoritu, Japan, 1983.
7. Dayhoff, J. *Neural Network Architecture.* Van Nostrand Reinhold, 1990.
8. Acha-Daza, J. A., and F. L. Hall. Application of Catastrophe Theory to Traffic Flow Variables. *Transportation Research,* Vol. 28B, No. 3, 1994, pp. 235–250.
9. Nakatsuji, T., and T. Kaku. Development of a Self-Organizing Traffic Control System Using Neural Network Models. In *Transportation Research Record 1324,* TRB, National Research Council, Washington, D.C., 1991, pp. 137–145.
10. Gunter, M. A., and F. L. Hall. Transition in the Speed-Flow Relationship. In *Transportation Research Record 1091,* TRB, National Research Council, Washington, D.C., 1986, pp. 1–9.

Microscopic Modeling of Traffic Within Freeway Lanes

JONATHAN M. BUNKER AND ROD J. TROUTBECK

Microscopic models provide an understanding of traffic operations at the level of passage of individual vehicles. Roadway performance can be ascertained by understanding how vehicles interact with each other. Cowan's M3 headway distribution models were calibrated for the curb and median lanes of two-lane mainline freeway segments, using data captured at 14 sites. Calibration of the relationship among Cowan's M3 parameters, proportion of headways greater than a minimum of 1 sec, and flow rate yielded exponential decay equations for each lane. The M3 models provide a source of vehicle arrivals for gap acceptance models, which may be used to quantify the ability of drivers to change lanes, for example. It was found that the parameters calibrated for each lane are suitable for use at any mainline site, independent of site-specific conditions. The proportion of small headways was found to be higher in the median lane than the curb, for all flow rates, and for both lanes lower than their respective equivalents on arterial roads with intersections. The largest bunched headway was considered to be between 2 and 3 sec. The models predicted bunching between 85 and 93 percent of median lane vehicles, and between 75 and 90 percent of curb lane vehicles, at capacity. The lesser amount of curb lane bunching reflects its use as a slower vehicle lane with greater stream friction.

Microscopic models provide a means of modeling traffic at the level of individual vehicles passing roadside observation points by describing the headways, or times between passage of vehicles. These models can be used as inputs to gap acceptance models, so that roadway performance can be quantified with capacity and delay estimates. Because of these attributes, microscopic models provide a greater level of understanding of the processes taking place than do macroscopic models.

This paper details an analysis of within-lane traffic flow on freeway mainline segments. It discusses a method of relating the proportion of headways greater than a minimum value to the lane flow rate, for each of the curb and median lanes on a two-lane, unidirectional element.

BACKGROUND

Headways are the time intervals between passage of successive vehicles past a roadside observation point. Figure 1 illustrates a typical cumulative distribution of freeway curb lane headways, measured over 15 min. The horizontal axis represents the size of headway, and the vertical axis represents the proportion of headways less than the corresponding horizontal axis ordinate. Knowledge of the headway distribution is necessary for the application of gap acceptance theory by which the ability of a stream to absorb vehicles can be quantified.

Figure 1 shows the measured distribution and a theoretical Cowan's M3 distribution (*1*), which fits the data. This model has two components of headways: those assumed to be equal to a specified minimum, Δ, and those greater than the minimum. Those greater than the minimum are distributed exponentially. The proportion of those greater than the minimum is denoted as α. The two parameters, α and Δ, therefore are interrelated. Cowan's M3 model is given as a cumulative probability function by Equation 1:

$$F(t)=\begin{cases}1-\alpha e^{-\lambda(t-\Delta)} & t\geq\Delta\\ 0 \quad t<\Delta\end{cases} \tag{1}$$

where λ is a shape parameter, given by Equation 2:

$$\lambda=\frac{\alpha q}{1-\Delta q} \tag{2}$$

and q is the lane flow rate, equal to the reciprocal of the mean headway.

Many headways of 1 sec, and even smaller, were observed. However, only freeway gaps greater than about 1.5 sec are useful for merging, so it was important to select parameters of Cowan's M3 model that consistently facilitate the accurate modeling of these headways. Headways less than this are not particularly useful, so they do not require accurate modeling.

For a particular data set, there is a particular set of α- and Δ-values that provide the best fit. Sullivan and Troutbeck (*2*) showed that α- and Δ-values can be varied slightly, but, by maintaining a relationship between them, the resulting distribution, *F*(*t*), is not significantly affected. Consequently, the Δ-value was chosen to be a convenient constant, and α was reevaluated for each data set accordingly.

The minimum headway, Δ, was set to 2 sec for the study of arterial road operations, facilitating a maximum flow rate of 1Δ, or 1,800 veh/hr. Flow rates of up to 2,500 veh/hr were recorded in freeway lanes during this study, so a smaller value of Δ was necessary. A value of Δ equal to 1 sec was considered more appropriate for freeways, as it allows for theoretical flow rates up to 3,600 veh/hr, and accounts for a more realistic minimum headway.

Relationships between lane flow rate and proportion of headways greater than 1 sec need calibration to predict the arrival headway distributions in each lane on a two-lane unidirectional mainline freeway segment, for any lane flow rate. Headway distributions are required as input to merging and lane changing models, which quantify performance measures of capacity and delay.

Headway data were collected in 15-min observations, at mainline locations a minimum of 1 km from ramp junctions, to calibrate relationships between lane flow rate and proportion of headways greater

Physical Infrastructure Center, School of Civil Engineering, Queensland University of Technology, GPO Box 2434, Brisbane Q 4001, Australia.

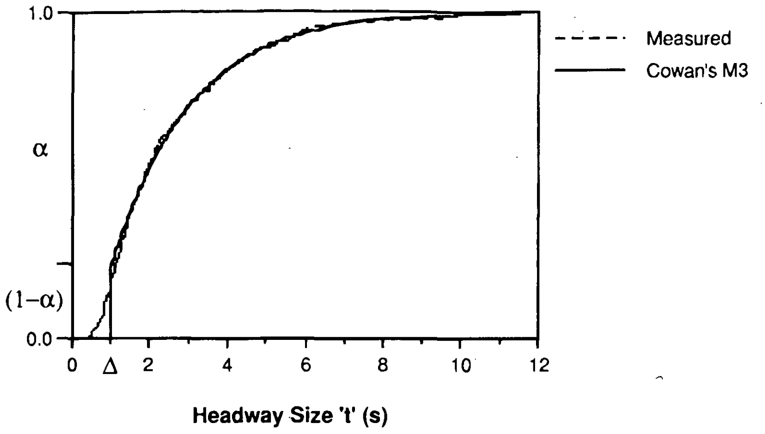


FIGURE 1 Measured distribution of headways, overlain by Cowan's M3 model representation.

than the minimum. With Δ set to 1 sec, the maximum likelihood technique (*3*) was used to find the best Cowan's M3 model fit to the measured distribution for each 15-min observation. The data acquisition led to a series of (α, *q*) data pairs for each lane at each site, for each observation period.

RELATING PROPORTION OF HEADWAYS GREATER THAN 1 SEC TO FLOW RATE

The analysis relating headway proportions to flow rate was limited to two-lane, unidirectional freeway mainline elements, where the lane nearest the edge of the road was denoted as the curb lane and the lane nearest the center of the road denoted as the median lane.

Curb Lane

Figure 2 illustrates the plots of the curb lane relationships between the proportion of headways greater than 1 sec, α, and flow rate, *q*, for all sites. For any given flow rate, there is a considerable spread of points occurring both within and between sites. The difference in environments between sites did not produce a marked effect on the overall relationship between α and *q*.

Although there is considerable spread in the data across all sites, postulating a single curve was considered reasonable, as there is a definite downward trend in α with increasing flow rate. This was expected, because more drivers would travel at small headways as flow rate increases, changing the shape of the headway distribution, so that there is a smaller proportion of the larger, exponentially distributed headways in the representative M3 model.

A model was sought between α and *q*, which would have common attributes to models representing other facilities. Current work by Sullivan and Troutbeck (*2*) showed that a negative exponential model is well suited to both lanes of an arterial road (one containing at-grade intersections). Relationships of this type were discussed and compared by Brilon (*4*) and Akçelic and Chung (*5*) and generally were found to be the best form of model.

Figure 2 shows that the data generally lie close to an α value of 1, up to a flow rate of about 0.2 veh/sec. This means that very few

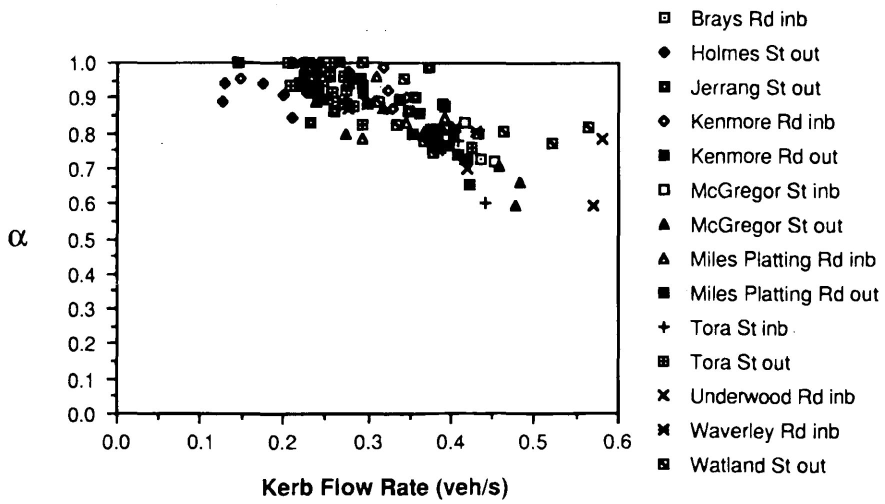


FIGURE 2 α versus *q* curb data from 14 freeway mainline segment sites.

drivers follow at the minimum headway, 1 sec, for low flow rates. Beyond this, the mass of data has a downward trend with flow rate. A constant value of α equal to 1 was considered adequate to reflect the conditions in the low flow regime. A downward trending relationship was considered for the higher flow rates.

This dichotomized relationship assumed that the turning point had the coordinates $(q_0, 1)$, where q_0 was to be found by regression analysis. The proposed relationship is given by Equation 3:

$$\alpha = \begin{cases} e^{-A(q-q_0)} & q \geq q_0 \\ 1 & q < q_0 \end{cases} \qquad (3)$$

where q is the curb lane flow rate in vehicles per second, and q_0 is the curb lane flow at the turning point between the two states of headway distribution discussed earlier.

Note that for flow rates less than q_0, where α is equal to 1, Cowan's M3 model becomes a shifted negative exponential distribution, with a minimum headway, Δ, equal to 1 sec.

The values of q_0 and A providing the minimum standard error were found to be 0.1877 veh/sec and 1.0801 sec/veh, respectively; the standard error was 0.0586. A value of q_0 equal to 0.175 veh/sec produced a standard error only 0.597 percent larger than the best-fit value. The optimum value of A was found to be 1.0027, which was then rounded to 1.0, giving a standard error of 0.059—only 0.599 percent larger than the best-fit value. Equation 4 defines the regression equation found to predict the curb lane proportion of headways greater than 1 sec, α, for a given flow rate, q:

$$\alpha = \begin{cases} e^{-1.0(q-0.175)} & q \geq 0.175 \text{ veh/sec} \\ 1 & q < 0.175 \text{ veh/sec} \end{cases} \qquad (4)$$

where Δ is equal to 1 sec.

An analysis of variance for Equation 4 yielded an F-value of 217, well exceeding the critical $F(1, 130, 0.05)$ value of 3.91, so the hypothesis that there is no relationship between α and q by Equation 4 was rejected at the 5 percent level.

Figure 3 illustrates the regression curve against the data. The dichotomized linear-exponential relationship fits the data well. A two-part linear relationship would have been equally acceptable within this range of flows. The exponential function selected for the downward trend decays very slightly, appearing almost linear anyway. The linear relationship was not chosen because it may predict negative α-values for some flow rates, which is not satisfactory. The function selected is also applicable to the median lane data, as will be discussed later. Consistency of functional form between both lanes is a positive attribute, as it is has greater flexibility in practical applications.

F-tests were used to establish whether the general relationship of Equation 4 was a suitable representation of the data for each individual site. The calculated F-values exceeding the critical $F(1, N - 2, 0.05)$ values in 10 of the 14 cases. The four sites found to bear no significant relationship by Equation 4 at the 5 percent level were Jerrang Street outbound (six points), Holmes Street outbound (six points), Underwood Road inbound (seven points), and Miles Platting Road inbound (eight points).

All of these sites had small sample sizes within narrow bands of flow rates. The data were not able to produce a strong enough trend for any relationship to be significant within each of these sites. The spread of data for each of the 4 sites, however, was not unusually high, compared with the data of all 14 sites. The generalized relationship was therefore considered to be acceptable for each of these locations.

Median Lane

As with the curb lane analysis, the median lane relationships between proportion of headways greater than 1 sec, α, and lane flow rate, q, were found to vary little between the 14 sites, for most flow rates. (Figure 4). For a given flow rate, data points lie within a band with depth of about 0.2 in terms of α, independent of site.

Two regimes of flow state can be seen in Figure 4 for flows above 0.6 veh/sec. A branch of data conforms to the trend of the lower flow data, below 0.6 veh/sec. However, there is also a branch where α-values are higher. Five or six data points that do not conform have α-values greater than 0.4 and result from the distribution's being relatively insensitive to α at these higher flow rates. For instance, if q is 0.65 veh/sec and α is 0.4 or 0.8, the pro-

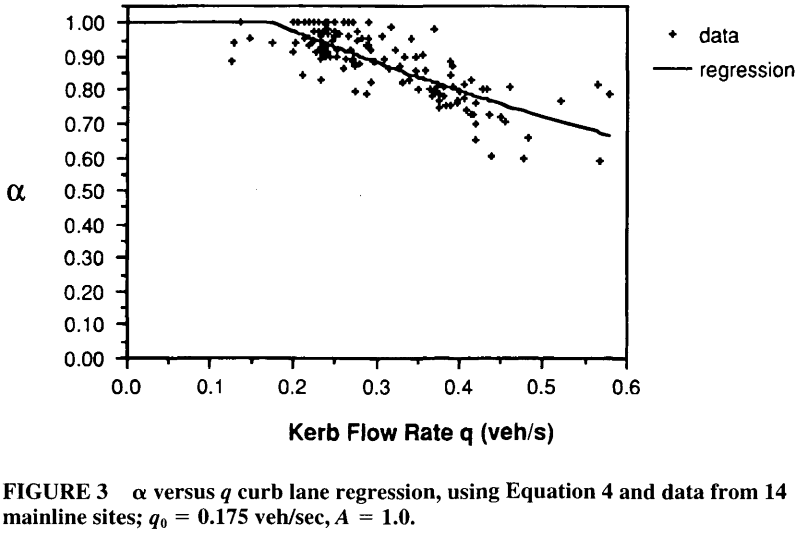


FIGURE 3 α versus q curb lane regression, using Equation 4 and data from 14 mainline sites; q_0 = 0.175 veh/sec, A = 1.0.

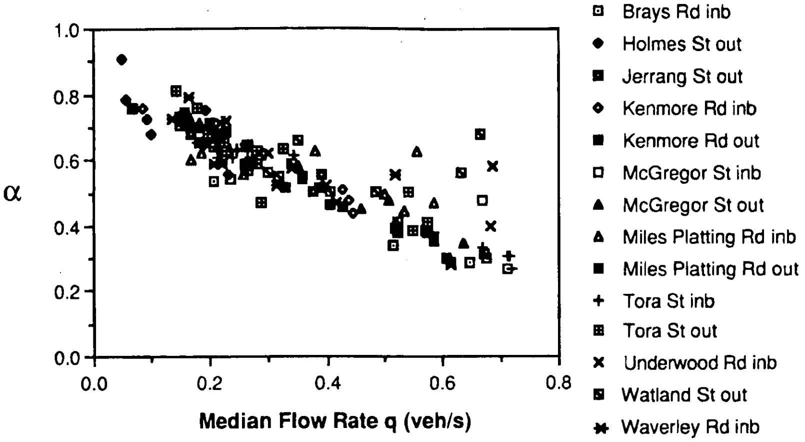


FIGURE 4 α **versus** q **median data from 14 freeway mainline sites.**

portions of headways greater than 2 sec are 19 and 18 percent, respectively. For purposes of modeling, operations were assumed to occur in the low state only, requiring only one curve for the entire flow regime.

In Equation 3, the exponential curve was shifted to the right to account for the low flow state in which practically all headways are greater than 1 sec. However, for the median lane data, a shift to the left was more appropriate, as the trend indicates that even for low flows the proportion of headways greater than 1 sec will not reach unity. When flow rate is 0, for an isolated vehicle, α must equal 1. A model incorporating these features is given by Equation 5.

$$\alpha = \begin{cases} e^{-A(q+q_0)} & q > 0 \\ 1 \quad q = 0 \end{cases} \tag{5}$$

Regression analysis using Equation 5 yielded optimum values of q_0 and A equal to 0.0869 veh/sec and 1.4070 sec/veh, respectively; the standard error in α was 0.0534. For q_0 rounded to 0.075 and A to 1.45, the standard error was 0.0535, compared with 0.0534 for the best-fit parameters. The difference is negligible. The regression curve for the relationship between α and q for the median lane is given by Equation 6:

$$\alpha = \begin{cases} e^{-1.45(q+0.075)} & q > 0 \text{ veh/sec} \\ 1 \quad q = 0 \text{ veh/sec} \end{cases} \tag{6}$$

where the minimum headway, Δ, is 1 sec.

Figure 5 illustrates the curve of Equation 6 against the field data. An analysis of variance for the equation gave an F-value of 723, compared with a critical F (1,125, 0.05) value of 3.91. The hypothesis that Equation 6 is unsuitable was rejected at the 5 percent level of significance.

F-tests were used to determine whether the common relationship of Equation 6 gives a reasonable representation to the data of each individual site. The calculated F-values exceeded the critical $F(1, N - 2, 0.05)$ values in all but 1 of the 14 cases. Again, insufficient data were available at this site to produce a strong enough trend for any relationship to hold.

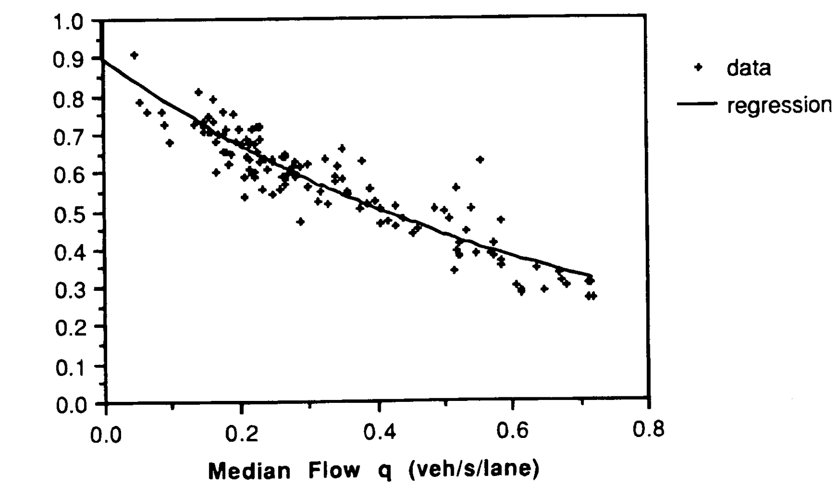


FIGURE 5 α **versus** q **median regression, using Equation 6 for data from 14 mainline sites;** $q_o = 0.075, A = 1.45$**.**

DISCUSSION OF RESULTS

Uses

The models for proportion of headways greater than 1 second, α, versus lane flow rate, *q,* have practical and theoretical applications in traffic engineering problems.

Bunker and Troutbeck (*6*) described relationships for estimating the flow rate in each of the curb and median lanes of a freeway mainline segment, given a total flow rate. Dichotomized linear models were selected to model the relationship between curb lane and total freeway flow rates. The relationships developed here may then be used to estimate the proportion of headways greater than 1 second in each lane, under the specified total demand. Using Equation 2, the decay constant of Cowan's M3 model may be calculated for each lane. All of the parameters necessary for using Cowan's M3 model for distribution of traffic within a lane, given in cumulative form as Equation 1, are then available. Thus, the amount of traffic in each lane and the distribution of headways within lanes may then be predicted for any freeway mainline segment, for any total flow demand.

This compound model has uses in prescribing the arrival of traffic on freeway mainline segments and estimating traffic inputs to gap acceptance models. Cases in which gap acceptance theory may be used for a mainline segment include merging and lane changing. Subsequent to the study described here, a gap acceptance model was established to predict delays and the distances required to perform lane changes, which are valuable performance measures. The model requires the distribution of headways in the target lane for the flow rate under consideration, as was calibrated here, and parameters for driver critical acceptance.

The models developed here may also be used to gain an understanding of the operation of a freeway and to compare it with other facilities. Comparisons are now made between the performance of freeway lanes and lanes on arterial roads with intersections.

Comparison of Curb and Median Lanes

Curb and median mainline freeway lanes do not operate in the same manner, as Figure 6 shows. For any given flow rate, the curb lane has a higher proportion of headways greater than 1 second and therefore would be expected to have fewer vehicles following at close headways.

The curb lane flow rates do not reach the high flows observed in the median lane. The maximum flow rate recorded in the curb lane at any site was approximately 0.58 veh/second, or 2,100 veh/hr, whereas in the median lane, the highest flow rate recorded was approximately 0.72 veh/sec, or 2600 veh/hr. This is consistent with the findings of Bunker and Troutbeck (*6*), who studied lane flows on freeway mainline segments. In those analyses they found that the curb lane is the dominant carrier under low total flows, and the median lane is the dominant carrier under high total flows, hence the discrepancy between maximum flows recorded in each lane.

The models for α versus *q* were not extended beyond the maximum flow rates recorded, as it is likely that they are close to capacity. The Cowan's M3 headway distribution model may not be applicable to congested operations. The relationship between α and *q* certainly would not be consistent with the model established earlier under those conditions.

The higher bunching in the median lane for any given lane flow rate relates to the apportioning of total flow between lanes (*6*). The median lane is reserved principally for overtaking on divided roads. Drivers using the median lane are likely to be more dissatisfied with their speeds than curb drivers, who tend to travel at more comfortable headways (of greater than 1 sec from the vehicle in front). Because the median lane is considered to be the fast lane, drivers might tend to be more alert and, as a result, travel closer to vehicles in front. This could be because a driver in a median platoon may intend to be in the median lane only until he or she passes the vehicle in the curb lane so is prepared to follow at a closer distance, or because the driver wishes to pressure the driver in front to speed up or move to the curb lane.

Drivers in the median lane may also be prepared to travel at close headways more often, as there is not as much stream friction created in the curb lane by merge and diverge areas.

If the proportion of vehicles following closely behind others can be considered to be a measure of the quality of service, then Figure 6 indicates that drivers in the curb lane have a better quality of service than those in the median lane. Of course, the driver elects to use a particular lane, so an improved speed that may be available,

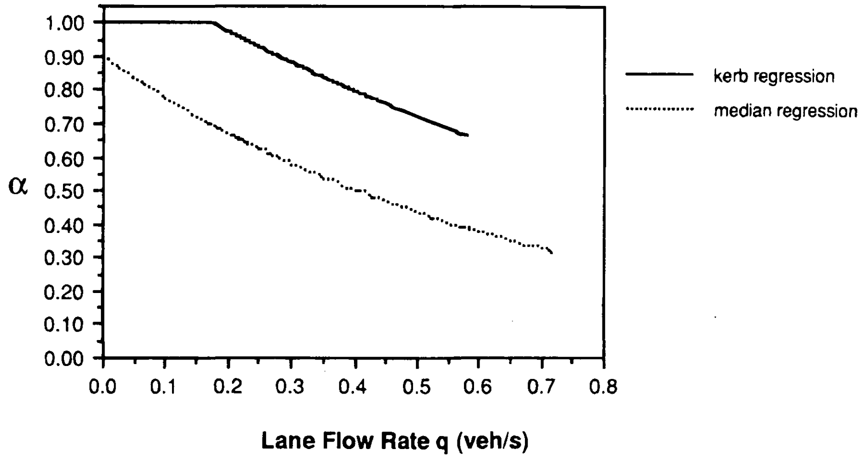


FIGURE 6 Models developed for α versus *q* in each lane, for freeway mainline segments. Differences show that each lane operates in a unique manner.

or perceived to be available, in using the median lane may be an offsetting quality-of-service measure.

Comparison of Freeway and Urban Arterial Facilities

Sullivan and Troutbeck (2) analyzed the behavior of traffic within lanes on urban arterial roads, classified as those with at-grade intersections, including traffic circles and unsignalized and signalized intersections. Cowan's M3 model was also used to model the distribution of flows within lanes for that analysis. However, the minimum headway, Δ, selected for that analysis was 2 seconds, the value recommended by road authorities for safe travel.

Sullivan and Troutbeck quantified the relationship between the proportion of "free" vehicles, having headways greater than 2 sec, and lane flow rate, *q,* for each lane on urban arterial link segments, away from intersections. Analyses were conducted for two- and three-lane segments; those for two-lane segments only are discussed here. Exponential models were selected for these relationships.

To compare freeway operations with arterial roads, it was necessary to establish relationships between the proportion of headways greater than 2 seconds and lane flow rate, *q,* for freeways. Using Equations 5 and 6 to predict α, the proportions of headways greater than 2 seconds were predicted by Equation 1. This proportion could then be compared directly with the equivalent quality for arterial roads. The results are plotted in Figure 7.

Figure 7 shows that for a given lane flow rate, a greater proportion of drivers closely follow others on an arterial road than on a freeway. This is partly due to the formation of platoons of vehicles at intersections on arterial roads. There are not as many opportunities to bunch vehicles together on freeways, where there are no interruptions in the uncongested state. The lower-speed environment of an urban arterial would also act to maintain a higher level of bunching. Vehicles on arterials are limited to the lower speeds necessary to maintain safety and order, which allows drivers to tolerate shorter headways.

Implications for Bunching

The development of the models for predicting the headway distribution in each lane, given the lane flow rate, helps in assessing the amount of bunching occurring. A bunched driver is considered to be one closely following a vehicle ahead. This assessment is most important at capacity conditions.

Figure 5 shows that the maximum flow rate recorded in the median lane at a site was approximately 0.7 veh/sec, or 2,500 veh/hr. It is postulated that this high flow rate is at, or very near to, capacity. Equation 6 gives the corresponding value of the proportion of headways greater than 1 sec, α, equal to 0.325. Equation 1 gives a proportion of headways less than or equal to 2 sec, of 85 percent. This can also be seen in Figure 7. The proportion less than or equal to 3 sec is 93 percent. If the largest headway in front of drivers who are bunched is between 2 and 3 sec, then between 85 and 93 percent of median lane vehicles are bunched at capacity, according to the model. This value appears to be reasonable.

Figure 3 shows that the maximum flow rate recorded in the curb lane was approximately 0.6 veh/sec, or 2,160 veh/hr. It is postulated that this value is also at or very near to capacity. The discrepancy between capacity flow rates in each lane is expected, as the median lane is usually the dominant carrier under such conditions. According to Equation 4, the value of α corresponding to capacity, is 0.65. The proportion of headways less than or equal to 2 sec is then 75 percent, and the proportion of headways less than or equal to 3 sec is 90 percent. If the largest bunching headway is between 2 and 3 sec, then between 75 and 90 percent of curb lane vehicles are bunched at capacity, according to the model. Again, this appears to be reasonable.

The lesser amount of bunching in the curb lane at capacity is to be expected because of the slower drivers who wish to travel at more relaxed headways, and the cautious drivers who expect vehicles to be merging into the curb lane from an on-ramp or the median lane. The result shows that it would be more difficult to move into the median lane at capacity than into the curb lane. This conclusion matches observations of capacity operations.

CONCLUSIONS

Headways are the times between passage of successive vehicles in a lane. Knowledge of the headway distribution is important when using gap acceptance theory to assess the ability of a lane to absorb merging vehicles. Cowan's M3 model was used to model the head-

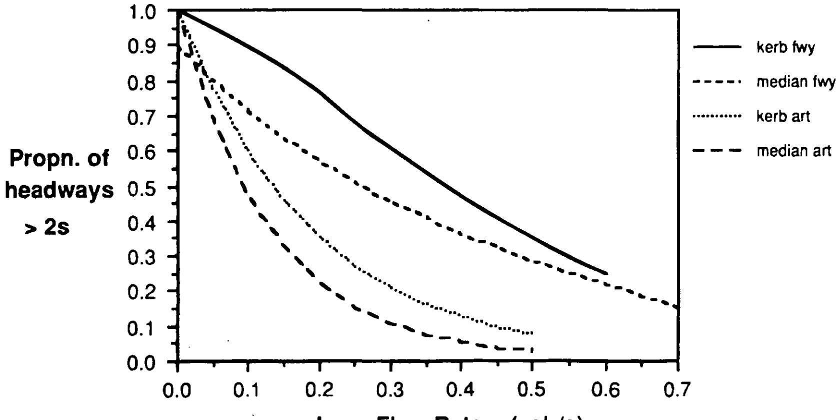


FIGURE 7 Proportion of headways greater than 2 sec versus lane flow rate, *q,* for curb and median lanes of freeway and arterial road types.

way distributions in the curb and median lanes on two-lane, unidirectional mainline segments, a minimum of 1 km from ramp terminals. Parameters of the M3 model include the minimum headway modeled, proportion of headways greater than the minimum, and a shape parameter, which is a function of the lane flow rate. Headways greater than the minimum are distributed exponentially.

A constant minimum headway of 1 sec was selected for convenience, as this value permits accurate modeling of all useful headways greater than about 1.5 sec. Small measured headways close to the minimum of 1 sec are poorly modeled; however, they are not considered useful to entering drivers so do not require specific attention.

Two relationships were found that relate the proportion of headways greater than 1 sec to flow rate, for each of the curb and median lanes. Both equations were based on exponential regression across 14 sites and were found to be significant at the 5 percent level, based on the results of F-tests. They were also shown to be suitable estimators of the relationships for individual sites. Although these empirical equations are recommended for estimating the value of α, given a lane flow rate, q, it must be emphasized that there was a considerable amount of spread in the data. The standard errors were 0.059 for the curb lane and 0.054 for the median lane, in terms of α.

The relationship may be used in conjunction with lane flow models to predict the distribution of vehicles in both lanes for any two-lane mainline location, given the total flow rate. Gap acceptance theory may then be used to predict delays, and subsequently distances required to change lanes, which are valuable performance measures.

Comparison of the relationships between proportion of headways greater than 1 sec and lane flow rate, for both lanes, shows that the proportion of small headways is greater in the median lane for any flow rate. This relates to its function as a fast or overtaking lane. Drivers in the median lane are more likely to be dissatisfied with their speeds, traveling at closer headways.

Relationships were compared with similar ones calibrated for arterial roads with at-grade intersections. The minimum headway modeled for these facilities was 2 sec. For any flow rate, and for both lanes, the proportion of headways greater than 2 sec is always greater for freeway traffic than for arterial traffic, because there are not as many opportunities to bunch traffic together on freeways.

The development of the models for prediction headway distributions in each lane enabled assessment of the amount of bunching. The largest bunched headway was expected to be between 2 and 3 sec. At capacity flow rates, between 85 and 93 percent of median vehicles and between 75 and 90 percent of curb vehicles are bunched, according to the models. These values match observations. The lesser amount of bunching is expected in the curb lane, as was the lesser amount of the very small 1-sec headways.

REFERENCES

1. Cowan, R. J. Useful Headway Models. *Transportation Research,* Vol. 9, No. 6, 1975, pp. 371–375.
2. Sullivan, D. P., and R. J. Troutbeck. The Use of Cowan's M3 Headway Distribution for Modelling Urban Traffic Flow. *Traffic Engineering and Control,* Vol. 35, No. 7/8, July–Aug. 1994, pp. 445–450.
3. Troutbeck, R. J. *Evaluating the Performance of a Roundabout.* Special Report SR45. Australian Road Research Board, Nunawading, 1989.
4. Brilon, W. Recent Developments in Calculation Methods for Unsignalized Intersections in West Germany. In *Intersections Without Traffic Signals,* Vol. 1, Springer-Verlag, Berlin, Germany 1988, pp. 111–153.
5. Akçelik, R., and E. Chung. Calibration of the Bunched Exponential Distribution of Arrival Headways. *Road and Transport Research,* Vol. 3, No. 1, March 1994, pp. 43–59.
6. Bunker, J. M., and R. J. Troutbeck. *Lane Flow Distribution on a Two Lane Unidirectional Freeway Link Segment.* PIC Report 94–6. Queensland University of Technology, Brisbane, Australia, 1994.
7. Bunker, J. M., and R. J. Troutbeck. *Modelling Traffic Distribution Within Freeway Lanes Microscopically:* PIC Report 94–9. Queensland University of Technology, Brisbane, Australia, 1994.

Statistical Analysis of Day-to-Day Variations in Real-Time Traffic Flow Data

H. RAKHA AND M. VAN AERDE

In the absence of intelligent vehicle-highway system technologies, commuters tend to select their routes through a congested network primarily on the basis of expected average link travel times. For this average to be representative of the current day, it is essential that the traffic conditions be relatively similar each day. However, if the traffic conditions vary considerably from one day to the next, the historical information will be insufficient for commuters to find the optimum routes through the network, and the provision of real-time traffic information could provide major benefits. Furthermore, simulation is becoming an important tool in evaluating different traffic control strategies. As a result it has become more and more important not only that the average typical traffic conditions be established but also that the upper and lower bounds of these average conditions be estimated. Consequently, two related issues are examined: the spatial and temporal magnitude of the variability in traffic conditions during typical nonincident conditions, and the magnitude of this variability during incident conditions. It was shown that in the absence of incidents, the temporal and spatial variations in traffic conditions were very similar for weekdays but varied considerably relative to the typical conditions during weekends. Major incidents, however, were found to alter drastically the average recurring conditions, thus creating a window of opportunity for achieving travel benefits by using dynamic data in real time.

The main objective of most advanced traveler information systems (ATIS) is to provide drivers with accurate real-time information on traffic conditions. Drivers can select optimum routes to their intended destinations based on this information. Various studies have investigated the potential benefits of ATIS (*1,2*).

In general, the benefits of such ATIS have been shown to depend on the level of market penetration and on the relative accuracy of the information provided to the equipped vehicles when compared with the accuracy of the historical data available to nonequipped vehicles.

Furthermore, as simulation becomes an important evaluation tool, it is important that one calibrates these simulation models to the existing traffic conditions.

Therefore, various questions remain. For example, how large must typical day-to-day variations in weekday traffic conditions be before they provide a sufficient window of opportunity for benefits to be accrued through the provision of real-time data to equipped vehicles? By how much do traffic conditions typically vary from day to day? By how much do incidents increase the window of opportunity for achieving benefits through the provision of real-time data?

This paper attempts to address most of these questions through a qulitative and quantitative analysis of 75 days of freeway management center (FMC) data along Interstate 4 in Orlando, Florida. The specific objectives of this paper are twofold: to investigate the variability in traffic conditions during (*a*) typical nonincident conditions and (*b*) incident conditions.

It is anticipated that the findings will be of assistance to both intelligent vehicle-highway system (IVHS) designers and to those who simulate such systems, as they will be able to perform their analysis based on tangible traffic network statistics rather than on hypothetical ones.

BACKGROUND

As part of the Advanced Driver and Vehicle Advisory Navigation Concept (ADVANCE), static prediction models were developed that could be applied to a series of traffic flow data: travel time, volume, and occupancy (*3*). In their model, Shbaklo et al. studied the effect of link type, time of day, day of week, and season on the flow and occupancy measurements for arterial and freeway links. This work was an extension of previous work on travel time analysis on links (*4*).

Shbaklo et al. (*3*), using 5-min loop detector data, for 72 days conducted analysis of variance (ANOVA) tests on freeway data in Chicago. They found the season to be an insignificant factor and the day of the week (2.5 to 9.7 percent of squared error) and time period (50 to 77 percent of squared error) to be significant factors on the flow and occupancy measurements. In their analysis, Shbaklo *et al.* did not examine whether Fridays or Mondays were statistically different from midweek days (Tuesdays, Wednesdays and Thursdays). Furthermore, they did not study the effect of incidents on these typical traffic conditions.

In this paper, the work conducted by Shbaklo *et al.* is extended to investigate variability within weekdays, spatial variability, and the effect of incidents on typical traffic conditions. Furthermore, the temporal and spatial variability in flow, speed, and occupancy measurements about a typical average temporal and spatial surface is investigated in an attempt to estimate statistical bounds to identify a typical weekday traffic conditions.

STUDY DESCRIPTION

Network Configuration

A 16-km (10-mile) portion of the I-4 freeway in Orlando, Florida, was considered in this study. I-4 is a major route that travels across the center of Florida from the southwest (Tampa) to the northeast (Daytona), passing by Disney World. The detectorized portion of the I-4 freeway is located near downtown Orlando, extending from 33rd Street to the southwest and ending downstream of Maitland Boulevard to the northeast.

Department of Civil Engineering, Queen's University, Kingston, Ontario, Canada K7L 3N6.

Twenty-four loop detector stations along I-4 were numbered from 1 to 25, with no data being provided for Station 10. The spacing of the detector stations ranged from approximately 0.40 to 0.90 km (0.25 to 0.54 mil). There were no major terrain variations along the detectorized section of the I-4 freeway, as Orlando is rather flat. However, at many interchanges with arterials, the freeway was elevated. The entire detectorized section of I-4 was composed of three lanes in each direction.

Data Collection Time Frame

The analysis period included traffic data for portions of 4 months during the winter of 1992–1993. The data included 11 days in November 1992, 29 days in January 1993, 26 days in February 1993, and 11 days in March 1993. This amounted to 75 days of 30-sec data, yielding approximately 10 days of data for each day of the week.

The FMC dual loop detectors measured and logged the flow, occupancy, and space-mean speed for each of the three lanes at 30-sec intervals. These data were aggregated into 5-min data summaries in order to reduce the level of data to be processed while still capturing most of the trends in the varying traffic conditions. Average lane flow, occupancy, and space-mean speed estimates were generated from the individual loop detector measurements for each station. In estimating the average lane speed at a specific station, loop speeds were weighted by the volume on each set of dual loops.

INITIAL ANALYSIS OF FMC DATA

An analysis of the FMC traffic data is presented in order to assess the variability in traffic conditions within weekdays. Subsequently, different weekdays are compared and the effect of incidents on the average typical traffic conditions is assessed. The analysis in this paper defines Saturdays and Sundays to constitute weekend days.

Generation of Typical Weekday Surfaces

Using the FMC data available during the 4-month period, it was possible to generate a surface that represented the average for all the days at a particular station of all the speed, flow, and occupancy measurements at a particular time of day. Equations 1 and 2 demonstrate how an estimate of each observation for the flow and occupancy was generated. In the case of the speed surface, a volume-weighted average was used. Core weekdays were considered to be Tuesday through Thursday, as it was initially not clear if Mondays or Fridays would be consistently similar to Tuesdays, Wednesdays, and Thursdays. There were 33 pure core weekdays during the analysis period. These core weekdays were checked for any abnormal traffic conditions such as vehicle detector failures (indicated as -1) or major incidents, as indicated in the incident data base that was provided by the FMC. The suspected days were removed from the estimated average.

The selection process resulted in 22 weekdays being considered in developing the average eastbound weekdays surfaces (*nd* = 22). The entire 33 weekday were used to generate the average westbound weekday surfaces (*nd* = 33). The resulting average flow surface for the eastbound direction only is presented in Figure 1; the results for the westbound direction were very similar.

$$x^n_{i,j} = \sum_{k=1}^{10} x^n_{i,j,k} \qquad \forall x^n_{i,j,k} \ge 0 \tag{1}$$

$$\bar{x}_{i,j} = \frac{\sum_{n=1}^{nd} x^n_{i,j}}{n\text{day}} \qquad \forall x^n_{i.j} \ge 0 \tag{2}$$

where

nd = total number of nonincident weekdays;
*n*day = number of good observations ($x^n_{i,j} \ge 0$);
$x^n_{i,j,k}$ = 30-sec observation on day *n* at station *i*, at 5-min time interval *j*, at 30-sec period *k* during 5-min interval;
$x^n_{i,j}$ = 5-min observation on day *n* at station *i* at time interval *j*; and
$\bar{x}_{i,j}$ = average weekday 5-min observation at station *i* at time interval *j* (flow or occupancy; speed was generated as a volume-weighted average).

The typical average spatial and temporal flow variation in the eastbound direction for an entire 24-hr period along the detectorized

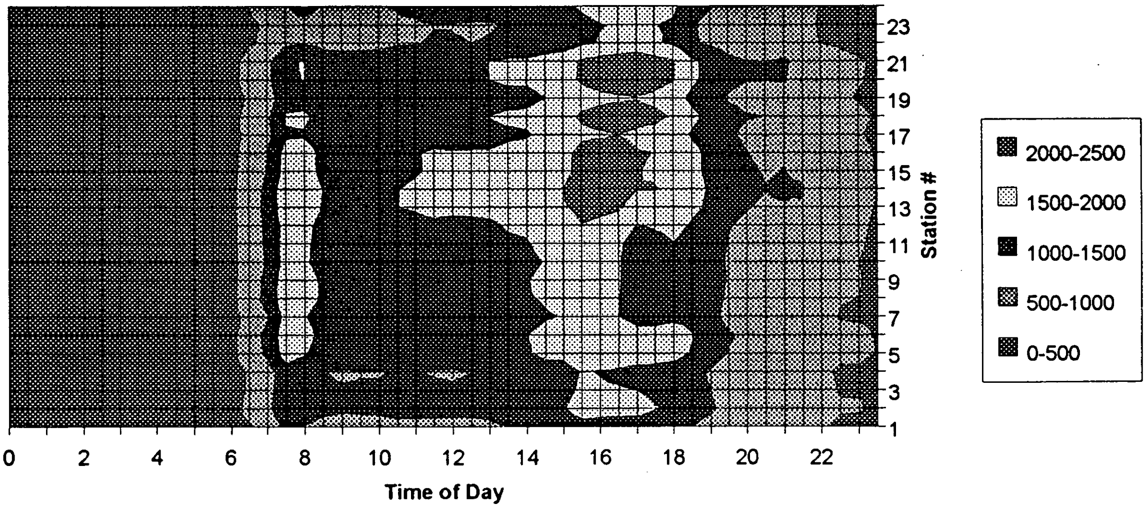


FIGURE 1 Spatial and temporal eastbound flow variation for average weekday [vehicles per hour (vph)].

I-4 section is presented in Figure 1. The *x*-axis represents the time of day, which ranges from 0 at midnight at the start of the day to 24 at midnight at the conclusion of the day, whereas the *y*-axis represents the station numbers traversed. The eastbound flow proceeds in the upward direction from Station 1 to Station 25. For each cell combination of time of day and station, the *z*-axis represents the average hourly lane flow measured.

It can be noted from Figure 1 that the flow increased gradually at 6:00 a.m. along all stations until it reached approximately 2,000 vehicles per hour per lane (vphpl) at 7:30 a.m. along most of the detector stations. The flow increased again during the p.m. peak at approximately 3:00 until 6:30 p.m. at Stations 12 through 22. It appears from Figure 1 that the flow from 5:00 to 7:00 p.m. at Stations 7 through 12 was lower (1,000 to 1,500 vphp). However, after examining Figure 2, it appears that the speed was also low, ranging from 20 to 40 km/hr. Thus the lower flow measurements were most likely due to the presence of congestion rather than to a reduction in demand. It appears from Figures 1 and 2 that a strict analysis of flow contours can be deceiving, as it is not clear whether a reduction in flow is caused by congestion or by a simple reduction in demand.

Single-Factor ANOVA of Weekday Data

To investigate whether the variability in traffic conditions between the different days of the core of the week (Tuesday, Wednesday, and Thursday) was statistically significant, a single-factor ANOVA was conducted using the SYSTAT model (*5*)-The ANOVA tested if the root mean square error (*RMSE*) associated with the different day surfaces about the typical average weekday surface was greater than the variation within the samples for each specific day of the week using Equation 3. Table 1 presents the ANOVA results for flow variations in the eastbound direction. These results, based on the 22 observations, indicate that the different days were not found to be statistically different at a level of significance of 95 percent. Similar results were obtained when comparing the speed as well as occupancy in the eastbound direction, as indicated in Table 1. Consequently, the observations in the eastbound direction for Tuesdays, Wednesdays, and Thursdays were grouped together as weekdays.

$$RMSE = \sqrt{\frac{\sum_i \sum_j (x_{i,j}^n - \bar{x}_{i,j})^2}{n\text{obs}}} \qquad \forall x_{i,j}^n, \bar{x}_{i,j} \geq 0 \tag{3}$$

where *n*obs is the number of good observations $(x_{i,j}^n, \bar{x}_{i,j} \geq 0)$.

A similar single-factor ANOVA on the different weekdays in the westbound direction was conducted as presented in Table 1. Again, the ANOVA results demonstrated that there was no statistical difference between the observations for Tuesdays, Wednesdays, and Thursdays at the 95 percent confidence level. Consequently, the data for these days were grouped together as core weekdays.

In order to examine the ANOVA assumption of homogeneity of variance, the variation in residuals as a function of the estimated values (day mean) is plotted in Figure 3. The Studentized residuals were used because it is convenient to reference them against a *t* distribution. In Figure 3 the residuals for the typical weekdays were all within two standard deviations. It appears from Figure 3 that the residuals are homogeneous as there appears to be no trend to the residuals. Similar trends were found for the residual plots generated for the eastbound speed and occupancy surfaces. Similar trends were also found for the westbound flow, speed, and occupancy surfaces but because of limited space are not presented here.

COMPARISON OF MEAN SURFACES

A typical average core weekday was compared with a typical Monday, a typical Friday, a typical Saturday, and a typical Sunday to determine if the traffic conditions are qualitatively and statistically different. An incident scenario is also compared with the typical average weekday conditions in order to demonstrate qualitatively the relative difference in flow conditions from one day to the next, versus an incident day to a nonincident day.

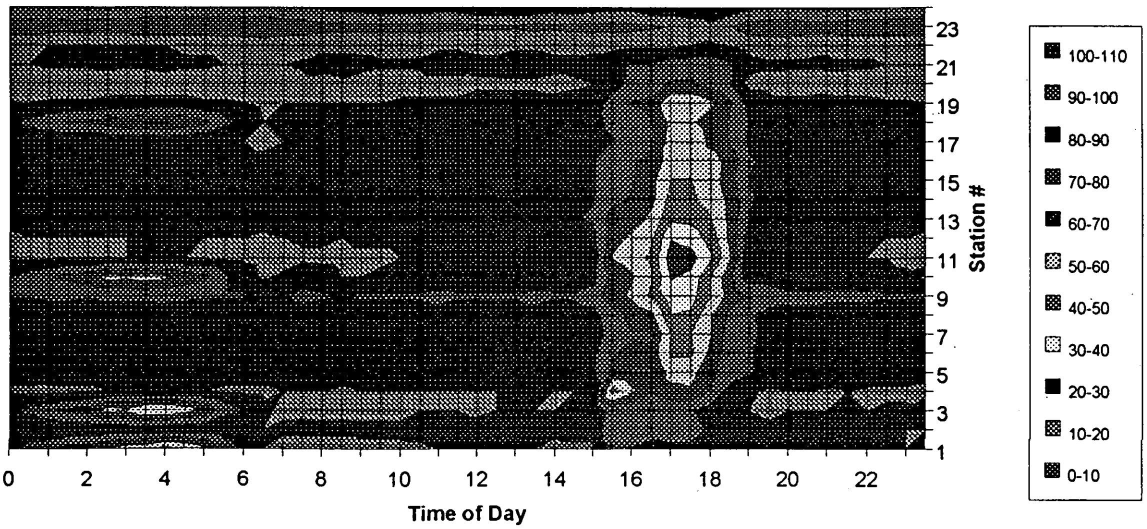


FIGURE 2 Spatial and temporal eastbound speed variation for average weekday (km/hr).

TABLE 1 Single-Factor ANOVA Results

Description	ANOVA groups	DF (within groups)	DF (total)	F	F_{crit}	Sig (95%)
Flow (EB)	Tue. vs. Wed. vs. Thur.	19	21	1.16	3.52	No
	weekday vs. Mon.	29	30	5.32	4.18	Yes
	weekday vs. Fri.	30	31	101.87	4.17	Yes
	weekday vs. Sat.	30	31	682.84	4.17	Yes
	weekday vs. Sun.	32	33	384.79	4.15	Yes
Speed (EB)	Tue. vs. Wed. vs. Thur.	19	21	2.76	3.52	No
	weekday vs. Mon.	29	30	2.40	4.18	No
	weekday vs. Fri.	30	31	101.87	4.17	Yes
	weekday vs. Sat.	30	31	682.84	4.17	Yes
	weekday vs. Sun.	32	33	384.79	4.15	Yes
Occ. (EB)	Tue. vs. Wed. vs. Thur.	19	21	1.88	3.52	No
	weekday vs. Mon.	29	30	1.20	4.18	No
	weekday vs. Fri.	30	31	17.13	4.17	Yes
	weekday vs. Sat.	30	31	16.01	4.17	Yes
	weekday vs. Sun.	32	33	47.25	4.15	Yes
Flow (WB)	Tue. vs. Wed. vs. Thur.	30	32	0.85	3.32	No
	weekday vs. Mon.	41	42	7.03	4.08	Yes
	weekday vs. Fri.	41	42	66.39	4.07	Yes
	weekday vs. Sat.	41	42	1678.67	4.08	Yes
	weekday vs. Sun.	43	44	1668.55	4.07	Yes
Speed (WB)	Tue. vs. Wed. vs. Thur.	30	32	0.55	3.32	No
	weekday vs. Mon.	41	42	0.11	4.08	No
	weekday vs. Fri.	41	42	12.15	4.07	Yes
	weekday vs. Sat.	41	42	22.34	4.08	Yes
	weekday vs. Sun.	43	44	23.54	4.07	Yes
Occ. (WB)	Tue. vs. Wed. vs. Thur.	30	32	0.62	3.32	No
	weekday vs. Mon.	41	42	0.30	4.08	No
	weekday vs. Fri.	41	42	15.98	4.07	Yes
	weekday vs. Sat.	41	42	113.03	4.08	Yes
	weekday vs. Sun.	43	44	208.02	4.07	Yes

Average Monday Surface

The average Monday flow, speed, and occupancy eastbound and westbound surfaces were generated in a similar fashion to the average core weekday surfaces. The eastbound average Monday surfaces were estimated by averaging over 9 Mondays, and the westbound average surfaces were estimated by averaging over 10 days.

The average Monday flow surface was found to be quite similar to the core weekday surface, and thus a typical Monday may qualitatively be considered to be similar to a core weekday. The same trends were found in comparing the occupancy and speed

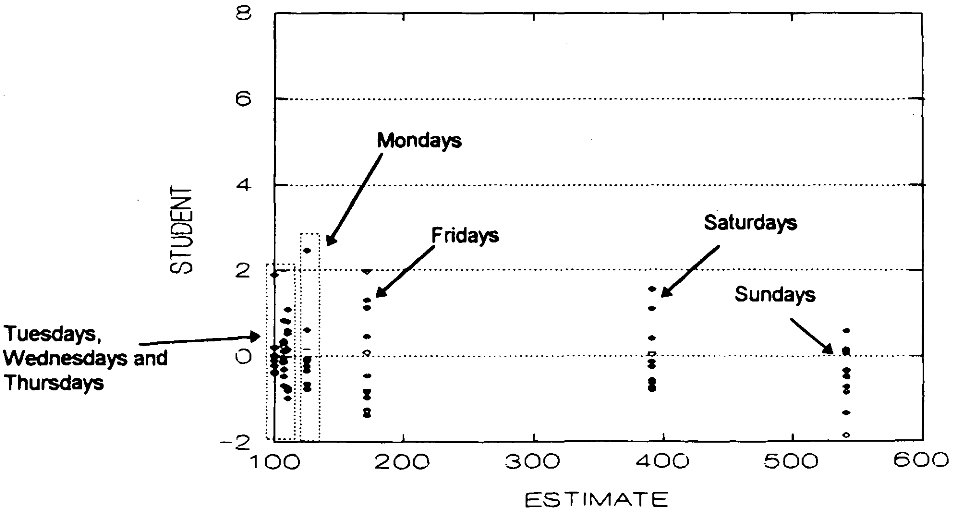


FIGURE 3 Variation in residual error as a function of *RMSE* estimate for eastbound flows.

surfaces. However, the limited space in this paper prevents their inclusion.

To verify quantitatively the similarity or variability between the Monday traffic conditions and the typical core weekday conditions, a single-factor ANOVA was conducted. The results of the ANOVA for the eastbound direction, presented in Table 1, demonstrate that the Monday flow conditions were statistically different from the typical weekday conditions at the 95 percent confidence level. However, the speed and occupancy measurements were not statistically different from the typical core weekday measurements (at the 95 percent confidence level), as given in Table 1. The same trend of results was obtained in conducting an ANOVA for the westbound direction, as indicated in Table 1.

It appears that Mondays are different from core weekdays in terms of flow but not in terms of speed or occupancy. Mondays therefore were not included in the data sample to create an average core weekday. These results were found to be consistent with the homogeneity assumption of ANOVA as illustrated by the residuals in Figure 3.

Average Friday Surface

The eastbound and westbound average Friday flow, speed, and occupancy surfaces were generated by averaging over 10 Fridays. By comparing the weekday and Friday surfaces, it was found that the p.m. peak on Friday started earlier (11:00 a.m. versus 12:00 p.m.) and extended over an extra hour (until 8:00 p.m. versus 7:00 p.m.).

The statistical results were found to verify the preceding qualitative comparison, as given in Table 1. Specifically, the ANOVA results for the eastbound direction indicated that the flows, speeds, and occupancies on a typical Friday were statistically different from the traffic conditions of typical core weekdays at the 95 percent confidence level. The results for the westbound direction were similar, as indicated in Table 1. These results, again, were found to be consistent with the homogeneity assumption of ANOVA, as illustrated by the residuals in Figure 3.

Average Saturday Surface

The eastbound and westbound average Saturday flow, speed, and occupancy surfaces were generated by averaging over 10 Saturdays; the plots are not presented because of the limited space in this paper. For the average Saturday flow surface, the traffic flows increased gradually from 7:00 a.m. until they reached a maximum flow of approximately 1,800 vphpl at noon at Station 15. The flow characteristics for a typical Saturday were very different from the traffic characteristics of a typical core weekday, as might be expected. The ANOVA results for the eastbound direction, presented in Table 1, demonstrate that the Saturday traffic conditions were statistically different from the typical weekday conditions. The results for the westbound direction, presented in Table 1, also demonstrate this trend.

It is noteworthy that in terms of eastbound flow and speed, Saturdays were much more distinct from core weekdays than Fridays. However, in terms of occupancy, Saturdays were different from core weekdays by only as much as were Fridays. In the westbound direction, flow and occupancy were much different, but speeds were not quite so different. These results, again, were found to be consistent with the homogeneity assumption of ANOVA as illustrated by the residuals in Figure 3.

Average Sunday Surface

An ANOVA of the eastbound Sunday traffic conditions and the weekday conditions, presented in Table 1, demonstrates that traffic conditions on Sundays were also statistically different from typical weekday conditions. Similar results were found for the westbound direction, as given in Table 1. As for Saturdays, the results presented in these tables indicate that the flow and speed on a typical Sunday were very different from a typical core weekday for the eastbound direction. The flow and occupancy in the westbound direction were also very different from the core weekday. These results, again, were found to be consistent with the homogeneity assumption of ANOVA as illustrated by the residuals in Figure 3. However, there appeared to be an outlier point, as illustrated in Figure 3.

Incident Effects

During the analysis of the core weekday data, a severe incident that resulted in the total closure of the eastbound direction of I-4 occurred on Thursday, November 5, 1992, as illustrated by the speed surface plot presented in Figure 4. The incident started at approximately 3:20 p.m. and lasted until approximately 5:00 p.m. The incident site was located between Stations 9 and 11 at Robinson Street, as indicated by the stationary frontal shock wave.

Following the clearance of the incident it can be noted in Figure 4 that the traffic proceeded downstream as a continuous platoon, and thus one can observe a surge of low speeds proceeding downstream up to Station 21. The forward-forming shock wave appears to be sloped steeply because the vehicles proceeded to Station 21 within one 5-min analysis period. This incident resulted in a queue that extended as far back as Station 1.

Note that a vehicle entering the system at 6:00 p.m. would experience delay at a location downstream of the incident at a point sometime after the incident was actually cleared.

Summary

In summary, based on statistical comparison of the traffic conditions for various days, the following conclusions can be made:

- Traffic flow conditions within core weekdays appear to be highly similar and consistent.
- Some traffic flow parameters on Mondays are similar to traffic conditions on core weekdays (Tuesday, Wednesday, and Thursday).
- Traffic conditions on Fridays differ from core weekday conditions in each of the three measures. Specifically, it appears that the p.m. peak on Fridays extends further in the day.
- Traffic conditions on weekends differ from traffic conditions on weekdays, and Saturdays differ in flow from Sundays.
- Major incidents can cause significant disruptions to typical weekday traffic conditions.

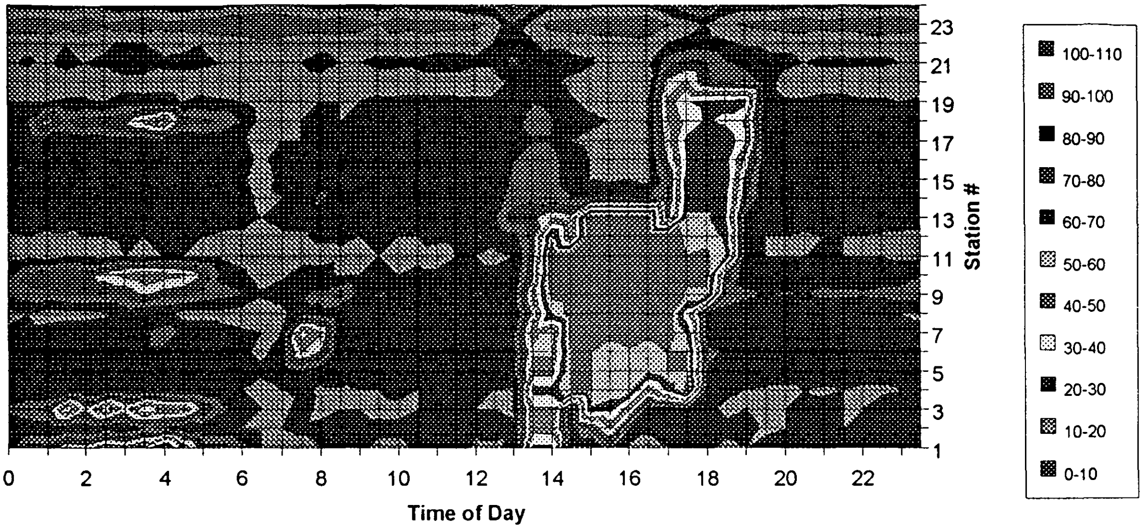


FIGURE 4 Spatial and temporal eastbound speed variation during an incident (km/hr).

OVERALL COMPARISON OF TRAFFIC CONDITIONS

The traffic conditions for each day were compared with the average weekday flow, speed, and occupancy surfaces. Two measures of comparison were estimated. The first was an estimate of the coefficient of determination (R^2) and will be labeled the regression measure. The second measure was an estimate of the number of observations within two standard deviations of the average weekday observation, assuming a normal distribution, and will be labeled the success measure. The findings for each of these measures are discussed in this section.

Regression Measure

A regression measure similar to R^2 was used to compare the traffic conditions for each day. For each day, three matrices of flow, speed, and occupancy observations were generated. These matrices were 288 rows (number of 5-min intervals in the day) by 24 columns (number of loop detector stations). A separate overall mean for the average weekday flow, speed, and occupancy measurements was also estimated, as demonstrated in Equation 4 (mean over all stations and all time periods $\bar{x}$).

For each of these surfaces, the squared error about the average core weekday surface was estimated as the difference for each station and time-of-day combination from the average core weekday surface using Equation 5 (sum of squared errors about the average surface S_1). The sum of squared errors for the flow, speed, and occupancy measurements of each day about their respective overall means was also estimated using Equation 6 (S_t). The sum of squared error, explained by each of the flow, speed, and occupancy average weekday surfaces, S_2, was estimated as the difference between S_t and S_1 using Equation 7. The R^2 measure for each of the three surfaces for each day was calculated as the ratio of S_2 to S_t (S_2/S_t). Thus, R^2 was a measure of the amount of error captured by the average weekday surface. An R^2 of 1 would mean that the average surface explained 100 percent of the squared error, whereas an R^2 of 0 would mean that the average surface did not explain any of the error.

$$\bar{x} = \frac{\sum_{i=1}^{24}\sum_{j=1}^{288}\bar{x}_{i,j}}{n\text{obs}} \qquad \forall \bar{x}_{i,j} \geq 0 \tag{4}$$

$$S_1 = \sum_{i=1}^{24}\sum_{j=1}^{288}(x^n_{i,j} - \bar{x}_{i,j})^2 \qquad \forall x^n_{i,j}, \bar{x}_{i,j} \geq 0 \tag{5}$$

$$S_t = \sum_{i=1}^{24}\sum_{j=1}^{288}(x^n_{i,j} - \bar{x})^2 \qquad \forall x^n_{i,j} \geq 0 \tag{6}$$

$$S_t = S_1 + S_2 \tag{7}$$

where

$n$obs = number of good observations ($\bar{x}_{i,j} \geq 0$; maximum = 6, 912),
x = overall average observation (flow, speed, and occupancy),
S_t = total sum of squared errors about overall mean (flow, speed, and occupancy),
S_l = sum of squared errors about average surfaces (flow, speed, and occupancy), and
S_2 = sum of squared errors explained by average surface (flow, speed, and occupancy).

The variation of R^2 over the 75-day analysis period from the average core weekday flow surface in the eastbound direction is presented in Figure 5. It appears that the R^2 for weekdays exceeded 90 percent and that an R^2 of 30 percent was estimated for the major incident day (November 5, 1992: Day 24). This low R^2 indicated that this incident had a substantial effect on the average traffic conditions. Mondays also had a relatively high R^2 (exceeded 90 percent), except for a Monday that had an incident in addition to a failure in some loop detectors. Fridays had a lower R^2, ranging from 75 to 90 percent. The Saturday and Sunday flow surfaces differed considerably from the weekday average surface (R^2 from 0 to 60 percent). The same trend was found for the westbound direction, but because of limited space, the results are not presented here.

The variation from the average weekday speed surface in the eastbound direction in R^2 during the 75-day analysis period was also analyzed but is not presented because of lack of space. Unlike the flow surface comparisons in Figure 5, the speed variation appeared to be much more scattered. The scatter in the speed variation was probably the result of shock waves proceeding along the detectorized section at different rates, even though the overall flow remained very similar. Interestingly, the major incident did not result in an R^2 worse than nonincident weekdays (Day 24).

The variation, from the average weekday occupancy surface in the eastbound direction, in R^2 during the 75-day analysis period was less scattered than the speed variation. Specifically, the R^2 ranged from 65 to 95 percent for the core weekdays, 45 to 90 percent for Mondays, 60 to 90 percent for Fridays, and 0 percent for Saturdays and Sundays. As was the case for the flow, the R^2 for the major incident day (Day 24) was much lower than the typical weekday R^2 (38 percent).

Success Measure

The original loop detector measurements, which were made at thirty 30-sec intervals, were aggregated into 5-min observations for purposes of analysis. Each 5-min observation was the sum of 10 measurements. Using the central limit theorem, it can be assumed that each of these 5-min observations may become distributed normally because the 5-min observation on one day should not be correlated with the same observation on another day. To verify this assumption, a 5-min estimate of flow for the 22 core days in the eastbound direction were estimated and stratified into bins. The observed probabilities were then tested using a chi-square goodness-of-fit test in order to establish whether the normal distribution assumption was valid, as illustrated in Figure 6. The chi-square type of analysis showed that the observed 5-min flows were not statistically different from the expected outcome of a normal distribution at the 95 percent confidence level. The test was repeated for higher average flows in the range of 1,800 vphpl, and similar results were found. Tests conducted for speed and occupancy 5-min observations had similar outcomes. Thus, it appears that the normal distribution assumption is valid.

The three average weekday surfaces were obtained by averaging each cell of the matrix over the nonincident weekdays using Equations 1 and 2. For each cell of these matrices, the standard deviation of the mean observation was estimated using Equation 8 and upper and lower bounds were estimated assuming a normal distribution using Equations 9 and 10, respectively. The proportion of similar observations was estimated as the ratio of observations within the upper and lower bounds to the total number of good observations using Equation 11. An average proportion of cells within the average weekday confidence limits subsequently was estimated for the weekdays using Equation 12. Using this proportion of successful observations, a lower confidence limit was estimated using Equation 13 (6):

$$\sigma_{i,j} = \sqrt{\frac{\sum_{n=1}^{nd} (\bar{x}_{i,j} - x_{i,j}^n)^2}{(n\mathrm{day} - 1)}} \qquad \forall \bar{x}_{i,j}, x_{i,j}^n \geq 0 \tag{8}$$

$$x_{i,j}^u = \bar{x}_{i,j} + 1.96 \times \sigma_{i,j} \qquad \forall \bar{x}_{i,j} \geq 0 \tag{9}$$

$$x_{i,j}^l = \bar{x}_{i,j} - 1.96 \times \sigma_{i,j} \qquad \forall \bar{x}_{i,j} \geq 0 \tag{10}$$

$$p^n = \frac{n_0^n}{n^n} \tag{11}$$

$$\bar{p} = \frac{\sum p^n}{nd} \tag{12}$$

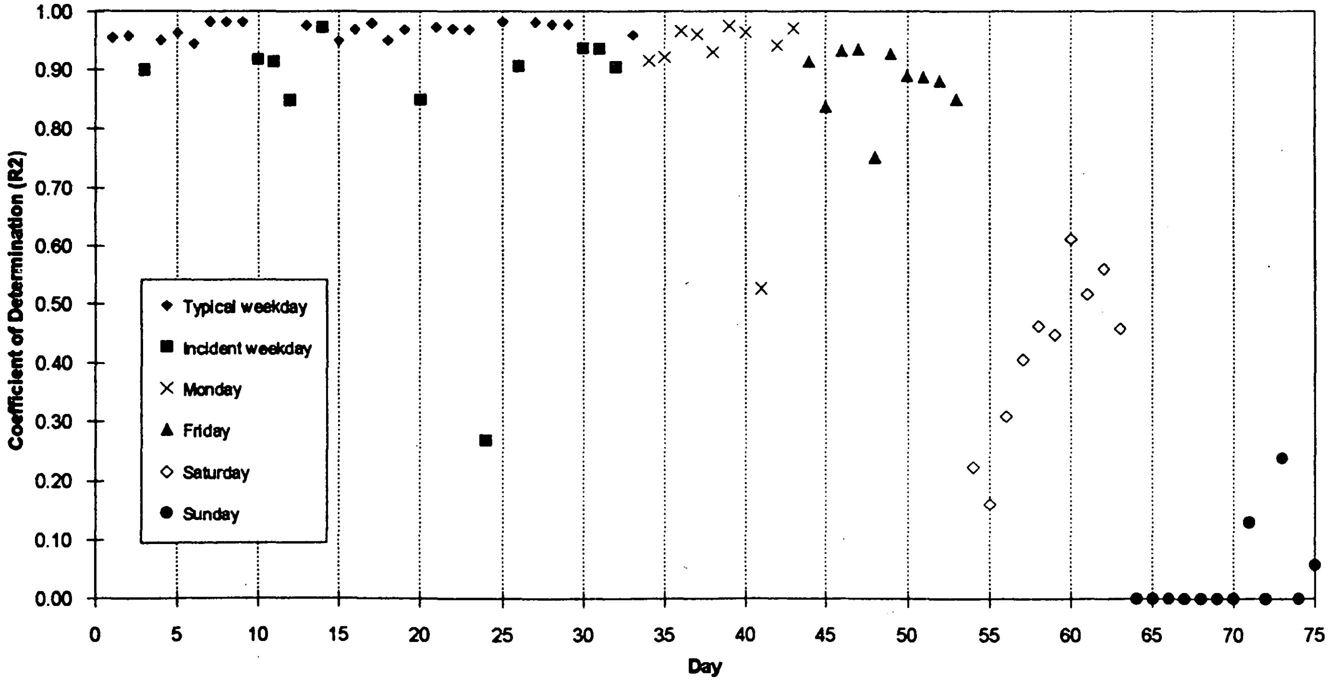


FIGURE 5 R^2 variation for eastbound flow.

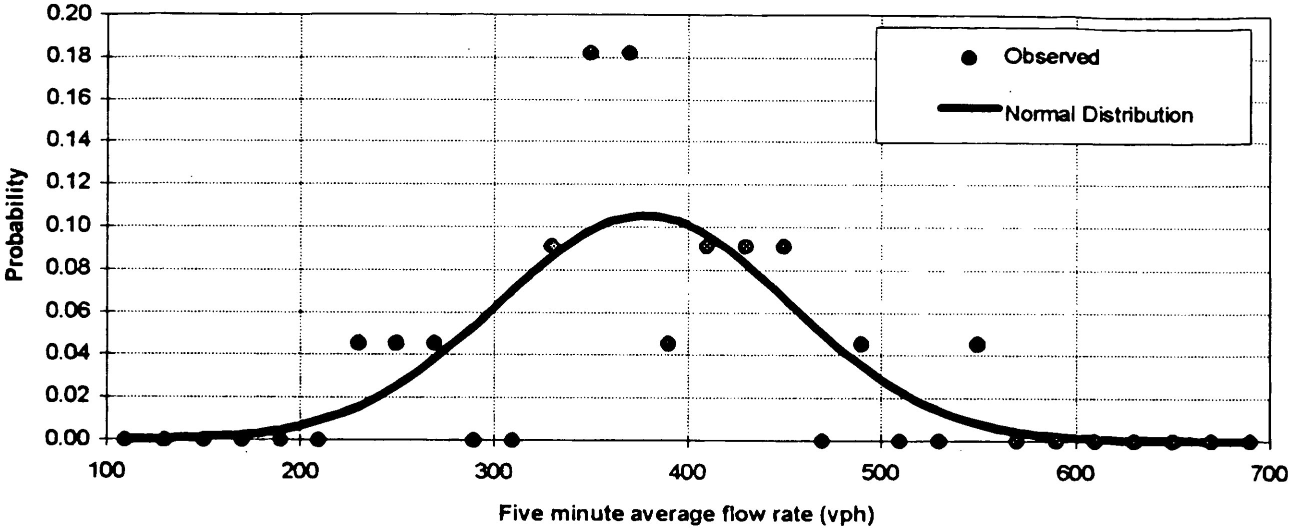


FIGURE 6 Observed and normal distribution 5-min flow rate estimates for 22 weekdays.

$$p^{l} = \bar{p} - 3\sqrt{\frac{\bar{p}(1-\bar{p})}{n\text{obs}}} \qquad (13)$$

where

- $n$obs = 24 × 288 = 6,912,
- $\sigma_{i,j}$ = standard deviation of 5-min observation distribution at station i at time interval j,
- $x^{u}_{i,j}$ = upper 95 percent confidence limit of 5-min observation at station i at time interval j,
- $x^{l}_{i,j}$ = lower 95 percent confidence limit of 5-min observation at station i at time interval j,
- n^{n}_{o} = number of observations for day n within confidence limits of average weekday surface,
- n^{n} = number of good observations for day n (observation ≥ 0),
- p^{n} = proportion of observations for day n within confidence limits of average weekday surface,
- $\bar{p}$ = average weekday proportion of observations within confidence limits, and
- p^{l} = lower bound of proportion of observations within confidence limits.

Figure 7 illustrates how flow p^{n} varied for the different days of the analysis period in the eastbound direction. It appears that most of the nonincident weekdays were within the confidence limit (16 of 22 observations). The high number of observations outside the range occurred because the number of good observations (nonnegative) for these extreme nonincident weekdays was less than $n$obs (used in estimating the confidence limits), and thus the lower confidence limit should have decreased to reflect the smaller number of observations. However, this was not done. The major incident (Day 24) did not have a major influence on p^{n}, which was 78 percent, indicating that traffic conditions were similar, based on this index, to typical core weekday conditions before and after the effects of the incident were removed. This high p^{n} resulted because this measure is not affected by how much the observation is outside the confidence limits, and thus the fact that the incident had an extreme effect on traffic flow was not reflected. It is important to note that except for a single incident day, all the incident days fell outside the preceding confidence range.

The Monday flows appeared to be near the borderline of the weekday flows (20 percent of the observations fell within the confidence range). Fridays differed from the weekday conditions, and so did Saturdays and Sundays (0 percent of the days fell within the confidence range). The westbound direction experienced a similar trend in variation of the flow p^{n}.

Summary

Two methods for distinguishing typical traffic conditions from atypical traffic conditions were investigated. The regression method, which uses the flow and occupancy surfaces, could distinguish typical from atypical weekday traffic conditions. However, the noise in the speed surface was too large to enable the identification of any systematic underlying variations. In the regression method it was not possible to determine any statistical confidence limits, which limits the practical usage of the method.

The success measure of the flow had the advantage of yielding confidence limits in order to distinguish statistically between significant and insignificant variations from the typical traffic conditions. This method could be developed further as an on-line incident detection routine by decreasing the averaging process from 5 to 2 min and estimating a p-value on-line for each station. A value outside the confidence limits would indicate a suspicious observation, and a second p-value outside the confidence range could set off an alarm. Such an approach to incident detection differs from techniques that detect incidents on the basis of the traffic state at upstream and downstream detector stations (*7*) rather than the deviation of the current observation from some bounds based on time of day.

CONCLUSIONS AND RECOMMENDATIONS

The premise of most equilibrium traffic assignments is that drivers base route selection on the assumption that in the absence of incidents, temporal traffic patterns are very similar from one day to the

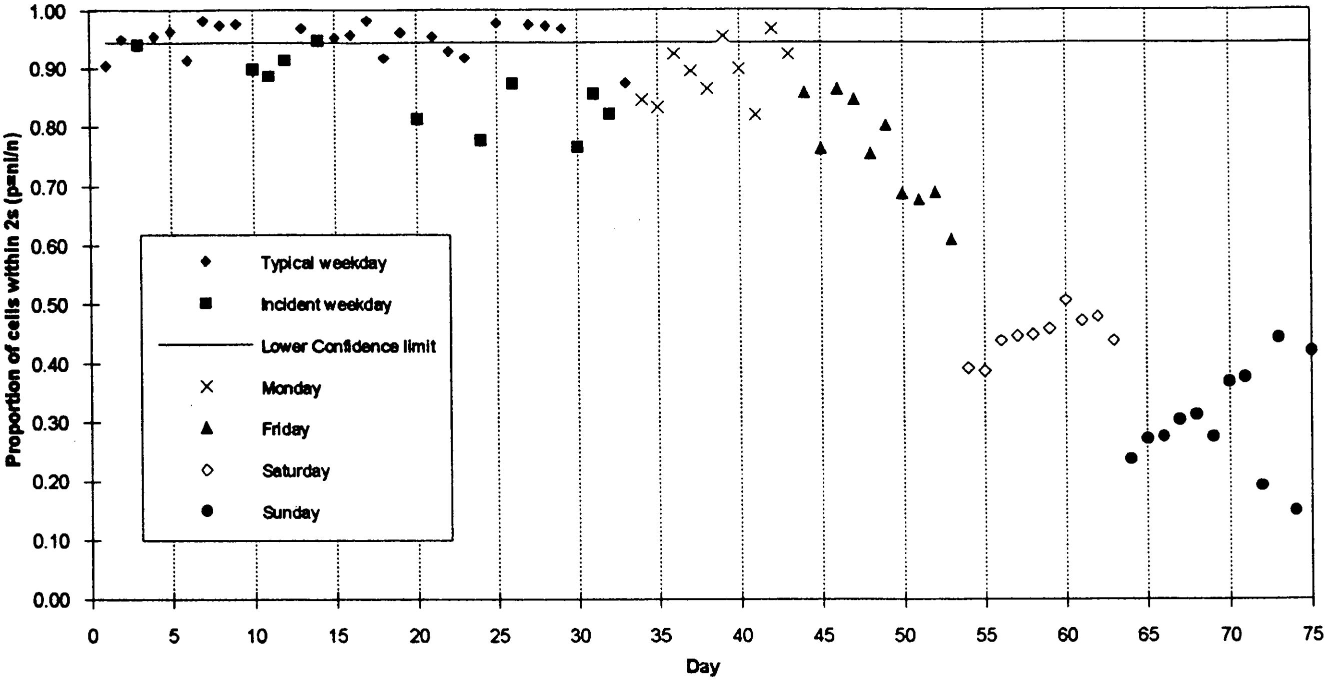


FIGURE 7 Variation of p for eastbound flow.

next. Many IVHS technologies attempt to explore the fact that even in the absence of incidents, traffic conditions on one day may be quite different from a similar previous day. This paper attempted to quantify these similarities and differences, both for incident and nonincident days.

It is recommended that the quantification of these similarities and differences be incorporated directly in any IVHS designs and benefit simulations. The present frequent use of hypothesized similarities or differences of day-to-day traffic may lead to designs or benefit estimates that are not consistent with the actual behavior of traffic. In this paper, such behavior has been quantified for at least one location, and a potential step toward a standardized procedure for analyzing others in a comparable fashion can be adopted.

ACKNOWLEDGMENTS

The authors would like to acknowledge the comments, critiques, and suggestions made by Hana Suleiman at Queen's University.

REFERENCES

1. Rakha, H., M. Van Aerde, E. R. Case, and A. Ugge. Evaluating the Benefits and Interactions of Route Guidance and Traffic Control Strategies Using Simulation. *Proc. IEEE Vehicle Navigation and Information Systems Conference,* Toronto, Ontario, Canada, Sept. 1989, pp. 296–303.
2. Rilett, L. R. *Modelling of TravTek's Dynamic Route Guidance Logic Using the INTEGRATION Model.* Ph.D. thesis. Queen's University, Kingston, Ontario, Canada, 1992.
3. Shbaklo, S., F. Koppelman, and C. Bhat. *Static Prediction Models of Flow and Occupancy.* ADVANCE Report TRF-TT-05, ADVANCE, Transportation Center, Northwestern University, Evanston, Ill., 1993.
4. Bhat, C., F. Koppelman, A. Sen, P. Thakuriah, P. Li, and N. Rouphail. *Short-Term Travel Time Prediction.* Report TRF-TT-02, ADVANCE. Transportation Center, Northwestern University, Evanston, Ill., 1992.
5. *SYSTAT for Windows: Statistics,* Version 5 Edition, SYSTAT, Evanston Ill., 1992.
6. Crow, E. L., F. A. Davis, and M. W. Maxfield. *Statistics Manual.* Dover, 1960.
7. Gall, A. I., and F. L. Hall. Distinguishing Between Incident Congestion and Recurrent Congestion: A Proposed Logic. In *Transportation Research Record 1232,* TRB, National Research Council, Washington, D.C., 1989.

Statistical Analysis and Validation of Multipopulation Traffic Simulation Experiments

SHIRISH S. JOSHI AND AJAY K. RATHI

Computer simulation has become a very powerful decision aid for varied facets of traffic engineering. Simulation experiments are often used to fit a metamodel of interest between the mean response and a selected set of input factors. This is done by carefully designing statistical experiments under alternative system designs, which are referred to as multipopulation simulation experiments. Validation and statistical analysis procedures are presented on linear metamodels from multipopulation traffic simulation networks under the common random number (CRN) strategy on three sample networks using the TRAF-NETSIM model. Under the CRN strategy, positive correlations are induced among the observations, and hence the usual statistical analysis cannot be applied to obtain point estimates and confidence intervals; therefore it must be modified. Before the statistical analysis is conducted, certain assumptions of the CRN strategy should be validated—those that, if violated, render the modified statistical analysis invalid.

Variance reduction techniques (VRTs) reduce the variance of the estimates of interest by replacing the original sampling procedure with a new procedure that yields the same expected value but with a smaller variance. Among the various correlation-induction techniques used as VRT, such as the common random numbers, antithetic variates, and Schruben-Margolin strategy, the common random number (CRN) strategy is perhaps one of the easiest to employ. Rathi (*1*) illustrated the effectiveness of the CRN strategy for the TRAF-NETSIM simulation model developed by FHWA.

TRAF-NETSIM is a microscopic, stochastic simulation model of traffic operations on urban street networks. This program has been applied extensively to a wide variety of problem areas by both practitioners and researchers and is the most widely used traffic simulation model (*2*). The availability of this model has enabled the development and testing of innovative traffic management concepts and designs (*3*). An important feature of this model is its amenability to control randomness from one simulation run to the next. This control can be used to induce desired correlations among the outputs and reduce the variance of estimates on the statistics of interest.

This paper presents validation and statistical analysis procedures on linear metamodels for multipopulation traffic simulation networks under the CRN strategy on three sample networks using the TRAF-NETSIM model. Under the CRN strategy, positive correlations are induced among the observations, and hence the usual statistical analysis cannot be applied to obtain point estimates and confidence intervals; therefore, it must be modified. Before the statistical analysis is conducted certain assumptions of the CRN strategy should be validated—assumptions that, if violated, render the modified statistical analysis invalid.

MULTIPOPULATION SIMULATION EXPERIMENTS

Often the purpose of a simulation experiment is to estimate a *metamodel* of a selected response, that is, a linear or nonlinear model of the mean response in terms of relevant decision variables for the simulated system. This fitted metamodel can be used in several ways. For example, it can be used to perform a sequential search in order to obtain better response values or make inferences on the behavior of the system. Consider a situation in which each simulation run yields a univariate response y. A particular run, j, called a *design point*, and denoted by x_{jl} ($l = 1,2, \dots ,k$), is identified by a setting of k factors or decision variables that are used as inputs to the simulation model. Suppose there are r replications of the simulation experiment across the m design points composing the experiment; then the relation of the response y_{ij} for the ith replication and the jth design point to the level of the k factors can be represented as a linear-metamodel having the form

$$y_{ij} = \beta_0 + \sum_{l=1}^{k} \beta_l x_{jl} + \epsilon_{ij} \qquad \text{for } i = 1, 2, \dots, r \text{ and } j = 1,2,\dots,m \tag{1}$$

where β_l ($l = 1,2, \dots ,k$) are the metamodel parameters and ϵ_{ij} is the experimental error at the ith replication at the jth design point. Across all m design points in the experiment, the metamodel in Equation 1 for the ith replicate can be written in matrix notation as

$$\mathbf{y}_i = X\beta + \epsilon_i \qquad \text{for } i = 1, 2, , \dots , r \tag{2}$$

where

$\mathbf{y}_i = (y_{i1}, y_{i2} , \dots , y_{im})'$,
$\beta = (\beta_0, \beta_1')' = (\beta_0, \beta_1, \dots , \beta_k)'$,
$\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \dots , \epsilon_{im})'$, and
$\mathbf{X} = (\mathbf{1_m T})$ is the $m \times (k + 1)$ design matrix with ones in the first column and x_{jl} in the jth row and $(l + 1)$st column.

In classical statistics, a design point is also referred to as a population. Since this experimental setup has more than one design point, it is called a *multipopulation simulation experiment.* Often, the variances of the estimates of the metamodel coefficients can be reduced by inducing correlations of a desired sign between observations obtained from different runs. The induced correlations are obtained

Center for Transportation Analysis, Oak Ridge National Laboratory, P.O. Box 2008, MS 6206, Oak Ridge, Tenn. 37831-6206.

by controlling the random number streams that drive the simulation model. Unfortunately there is no general guarantee that the correlation-induction strategies produce the desired variance reduction. Therefore, careful implementation of these techniques is needed. CRN is one such useful correlation-induction technique.

CRN STRATEGY

The idea of the CRN strategy is to compare alternative simulation models under similar experimental conditions in order to improve confidence that observed differences in performance are due to the differences in the model structure rather than to differences in the experiment itself (*4*, p. 61). Under the CRN strategy, the same set of random number streams, $\mathbf{R}_i = (r_{i1}, r_{i2} \ldots, r_{ig})$ is applied to all m design points in the ith replicate where g is the number of streams used to drive the simulation model. Also, independent random number streams are used across replicates of the experimental design. Replications reduce the variance of the outputs and present means of computing pure error. For the CRN strategy applied to simulation experiments, the following assumptions are made:

1. The response variance is constant across all design points, so that

$$\sigma_j^2 = \text{var}[y_{ij}(R_i)] = \sigma^2 \qquad \text{for } j = 1,2,\ldots,m \text{ and } i = 1,2,\ldots,r \tag{3}$$

2. There is a constant nonnegative correlation between all pairs of responses within a given replicate, y_{ij} and y_{ik} $(j{\neq}k)$. That is,

$$\begin{aligned} &\text{corr}(y_{ij}y_{ik}) = \rho_+ \text{ for } j \neq k \\ &1 < j,k < m \end{aligned} \tag{4}$$

where $0 < \rho_+ < 1$.

3. The vector of responses composing the ith replicate has a multivariate normal distribution. Under the first two assumptions, the covariance matrix between observations within a replicate is given by

$$\sum{}^{(CRN)} = \sigma^2 \begin{bmatrix} 1 & \rho_+ & \cdot & \cdot & \cdot & \rho_+ \\ \rho_+ & 1 & \cdot & \cdot & \cdot & \rho_+ \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \rho_+ & \rho_+ & \cdot & \cdot & \cdot & 1 \end{bmatrix} \tag{5}$$

VALIDATION AND STATISTICAL ANALYSIS PROCEDURES

The results for validation and statistical analysis procedures developed by Tew and Wilson (*5*) are presented. For a detailed theoretical framework, the reader is referred to Tew and Wilson (*5*). To perform statistical analysis and validation of the fitted metamodel under the CRN strategy, it is useful to transform the model to one with independent observations within each replicate. This is done by applying the $m \times m$ orthogonal transformation $\mathbf{\Gamma}^{(\text{CRN})}$ given by

$$\Gamma^{(CRN)} = \begin{bmatrix} m^{-1/2} & 1'_m \\ & C' \end{bmatrix} \tag{6}$$

where $\mathbf{C}$ is an $m \times (m - 1)$ matrix such that $m^{-1/2}\,\mathbf{1}_m\,\mathbf{C}$ is orthogonal. (Note that in this case, the term $\mathbf{1}_m\,\mathbf{C}$ does not indicate matrix multiplication; instead, it indicates the $m \times m$ matrix whose first column is given by $\mathbf{1}_m$, and whose remaining $m - 1$ columns comprise the matrix $\mathbf{C}$.)

Validation

The validation consists of a three-step procedure in which each step checks a key assumption across all design points. The test in each step depends on validation of hypothesized properties of the previous steps; hence, these diagnostic checks on the experimental design and analysis must be performed in order. At each step a highly significant test statistic generally will indicate the need for some corrective action by the analyst. The following three diagnostic tests must be performed.

1. Test for multivariate normality.

H_0: $\mathbf{y}_i \sim N_m\,(\boldsymbol{\mu}, \Sigma)$ where Σ is positive definite but otherwise μ and Σ are unspecified

versus

H_1: $\mathbf{y}_i$ has any nonnormal, nonsingular m-dimensional distribution (7)

2. Test for induced covariance structure.

H_0: $\text{cov}(\mathbf{y}_i) = \Sigma^{(\text{CRN})}$ with σ^2 and ρ_+ as in Equation 5 so that $\Sigma^{(\text{CRN})}$ is positive definite and $0 < \rho_+ < 1$; otherwise σ^2 and ρ_+ are unspecified

versus

H_1: $\text{cov}(\mathbf{y})_i$ is positive definite but different from $\Sigma^{(\text{CRN})}$ (8)

3. Test for lack of fit in the linear model.

H_0: $E(\mathbf{y}_i) = \mathbf{X}\,\beta$

versus

H_i: $E(\mathbf{y}_i) \neq \mathbf{X}\,\beta$ (9)

The Shapiro-Wilk test (*5*, Section 2.1) will be used to test the normality of responses $\mathbf{y}_i$.

To test the covariance structure, the conventional likelihood ratio test statistic for H_0 has the form

$$L = \left[\frac{\det(r^{-1}A)}{\hat{\lambda}_1^2(\hat{\lambda}_2^2)^{(m-1)}} \right]^{r/2} \tag{10}$$

where

$$A = \sum_{i=1}^{r} (y_i - \bar{y})(y_i - \bar{y})' \tag{11}$$

and $\hat{\lambda}_1^2$ and $\hat{\lambda}_2^2$ are explicitly defined elsewhere (*25, 27* respectively). Also,

$$\bar{y} = \frac{1}{r} \sum_{i=1}^{r} y_i \tag{12}$$

is the sample mean of the original m-dimensional response vectors. If the responses are multinormal with the prescribed covariance structure given by Tew and Wilson (*5*), then the test statistic $N \equiv -2\ \ln(L)$ asymptotically has a chi-squared distribution with 1/2 $m(m+1) - 2$ degrees of freedom as $r \longrightarrow \infty$ (*6*). However, the rate of convergence to this limiting distribution can be slow. To achieve adequate convergence to this limiting distribution of N with moderate values of r, Joshi and Tew (*6*) developed a modified likelihood ratio statistic, M, whose definition and use are described in the following.

Note that all the tests and analyses presented hereafter were derived using the transformed responses. They are, however, presented in terms of original responses to illustrate their application and ease of use to the simulation practitioner.

Reject the null hypothesis in Equation 8 if

$$M > \chi^2_{1-\alpha}\left[\frac{1}{2}m(m+1) - 2\right] \tag{13}$$

$$M = -2\psi_0 \ln(L) \tag{14}$$

with

$$\psi_0 = \frac{\dfrac{m(m+1)}{2} - 2}{\dfrac{m^2 - 3m + 2}{2\psi_1} + \dfrac{m-1}{\psi_2} + \dfrac{m-2}{\psi_3}} \tag{15}$$

and

$$\psi_1 = 1 - \frac{2m+3}{6r} \tag{16}$$

$$\psi_2 = 1 - \frac{3m^2 - 1}{6r(m-1)} \tag{17}$$

and

$$\psi_3 = 1 + \frac{m}{3r(m-1)} \tag{18}$$

The last stage of the validation procedure is to test for the lack of fit in the model. It uses a standard lack-of-fit test only applied to transformed responses. Define the error sum of squares, S_E, as

$$S_E = \sum_{i-1}^{r} \left\| y_i - X\hat{\beta}_1 \right\|^2 \tag{19}$$

where

$$\hat{\beta} = (X'X)^{-1} X'\bar{y} \tag{20}$$

is the ordinary least squares estimate of β. Also define S^*_E, the error sum of squares for the transformed responses, and S^*_{PE}, the pure error of the transformed responses respectively, as

$$S^*_E = S_E - m\sum^{r}(\bar{y}_{i.} - \bar{y}_{..})^2 \qquad \text{with } \nu^*_E = mr - k \tag{21}$$

$$S^{\cdot}_{PE} = r(m-1)\lambda^{\acute{}}_2 \qquad \text{with } \nu^{\cdot}_{PE} = m(r-1) \tag{22}$$

where $\bar{y}_{i.}$ is the average response vector at the *i*th replication taken over across all design points, and $\bar{y}_{..}$ is the average response taken over all design points and over all replications.

Reject H_0 in Equation 9 if

$$\frac{(S^*_E - S^*_{PE})/(\nu^*_E - \nu^*_{PE})}{S^*_{PE}/\nu^*_{PE}} > F^{\alpha}_{(\nu^*_E - \nu^*_{PE}, \nu^*_{PE})} \tag{23}$$

where F is the quantile of order $1-\delta$ for the F-distribution with $\nu^*_E - \nu^*_{PE}$ and ν^*_{PE} degrees of freedom. Results on statistical analysis are presented next.

Statistical Analysis

The statistical analysis involves estimation of β and construction of simultaneous confidence intervals for the elements of β. The uniformly minimum variance unbiased (optimal) estimator of β is $\hat{\beta}$ and is as given by Equation 20. The model independent estimator of σ^2 is given by

$$\hat{\sigma}^2 = [m(r-1)]^{-1} \sum_{i=1}^{r} \sum_{j=1}^{m} (y_{ij} - \bar{y}_{.j})^2 \tag{24}$$

where $\mathbf{y}_{.j}$ is the average response at the *j*th design point taken over all replicates. The variance on the estimate of β_0 is given by $\hat{\lambda}^2_1$ and can be viewed as the *between replicate variation;* it is

$$\hat{\lambda}^2_1 = m\sum_{i=1}^{r} \frac{(\bar{y}_{i.} - \bar{y}_{..})^2}{r} \tag{25}$$

We have

$$\frac{(mr)^{1/2}(\beta_0 - \hat{\beta}_0)}{\hat{\lambda}_1} \sim t_{r-1} \tag{26}$$

which can be used to construct $100(1 - \alpha)$ percent confidence interval for β_0.

Next, define $\hat{\lambda}^2_2$, the estimate of the pure error variance σ^2, as

$$\hat{\lambda}^2_2 = \frac{(mr - m)\hat{\sigma}^2 - m\sum_{i=1}^{r}(\bar{y}_{i.} - \bar{y}_{..})^2}{r(m-1)} \tag{27}$$

The joint $100(1 - \alpha)$ percent simultaneous confidence interval for $\mathbf{l'H}\beta_1$ for all $l \in \mathbf{R}^h$ under the prescribed covariance structure where H is a known $h \times k$ matrix of constants with rank $h \leq (k + 1)$ and is given by

$$\mathbf{l'H}\beta_1 \in \mathbf{l'H}\hat{\beta}_1 \pm \hat{\lambda}^2_2 \left[\frac{hF^{\alpha}_{h,(m-1)r-k-1} l'H(\mathbf{T'T})^{-1}\mathbf{H'l}}{r}\right]^{1/2} \tag{28}$$

ILLUSTRATIVE EXAMPLES

For the purpose of illustration, three sample TRAF-NETSIM networks were selected. The geometric conditions for Networks 1, 2, and 3 are depicted in Figure 1, 2, and 3, respectively. These data sets

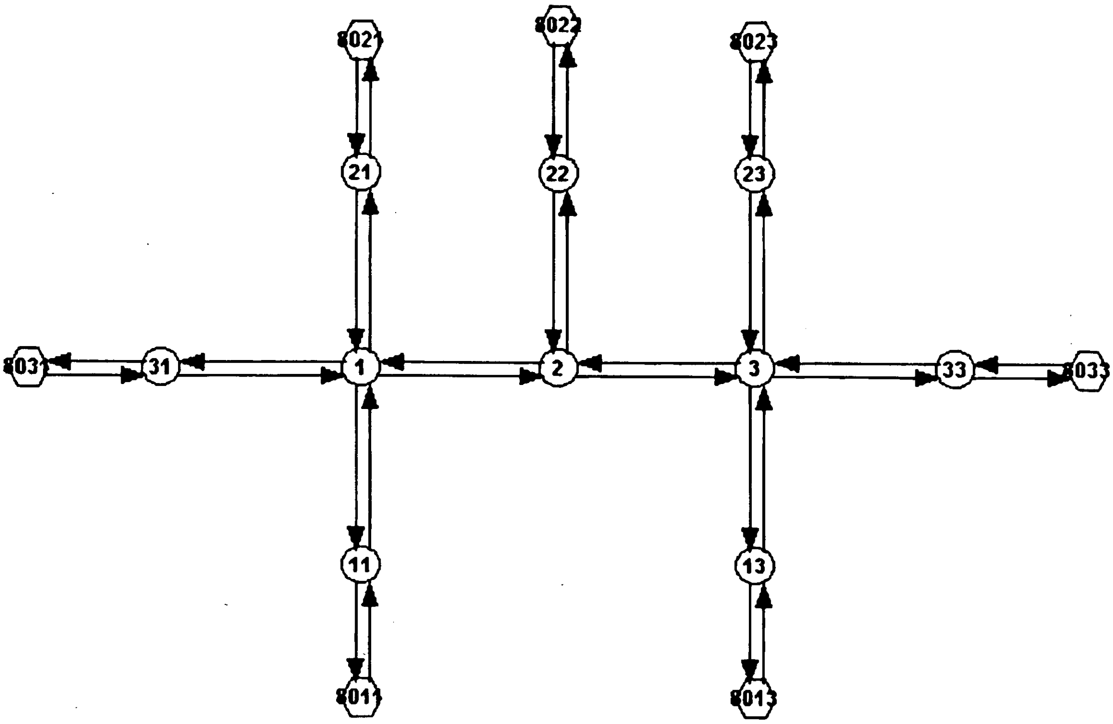


FIGURE 1 Graphical representation of Network 1.

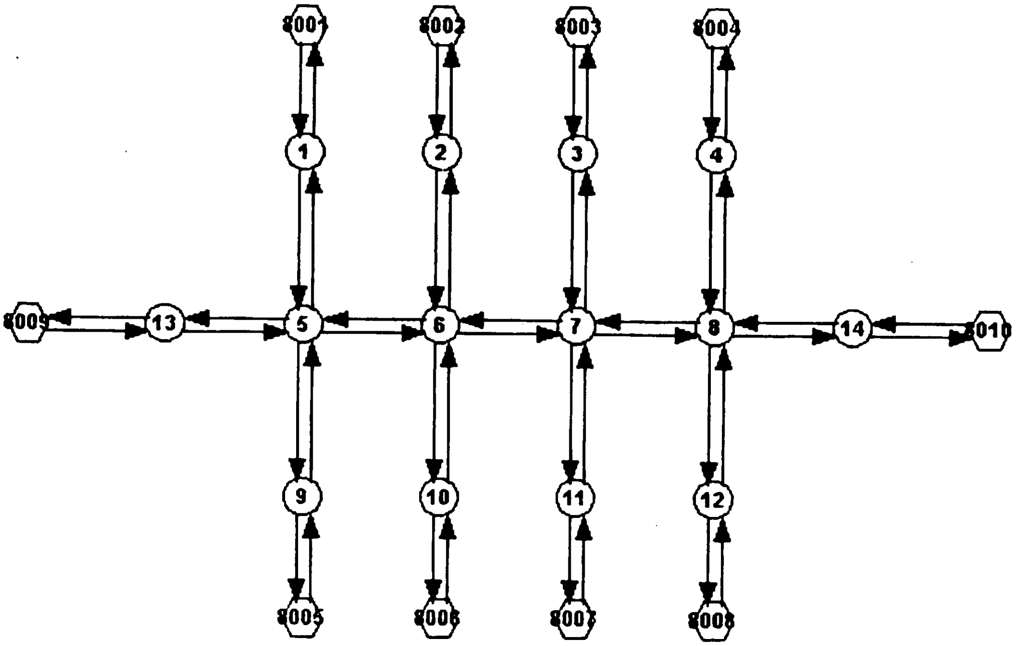


FIGURE 2 Graphical representation of Network 2.

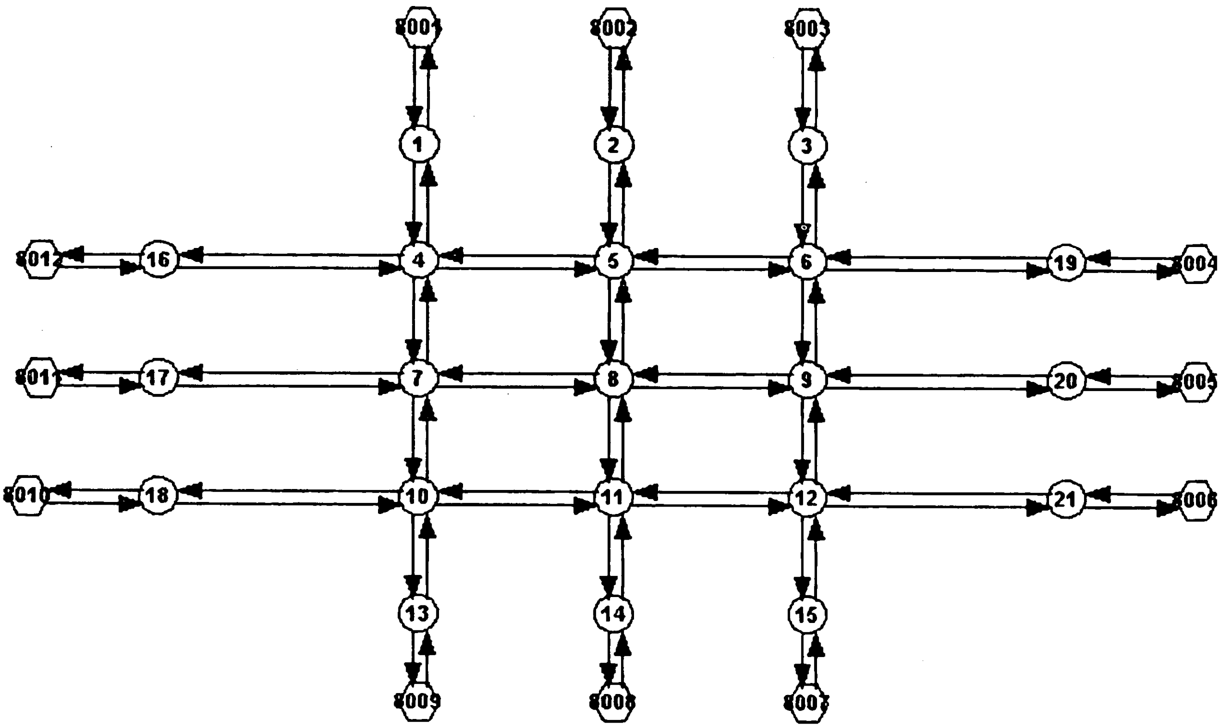


FIGURE 3 Graphical representation of Network 3.

represent traffic networks consisting of an isolated intersection (Figure 1), an arterial (Figure 2), and a grid (Figure 3). The input data for these networks regarding geometric length, signal control, number of lanes on each link, turning movements, and volume information are presented in Tables 1, 2, and 3 respectively. Because of the difference in characteristics of these three networks, different measures of effectiveness (MOEs) were chosen. The basic information pertaining to simulation experiments for the three networks is summarized in Table 4.

For all these networks, nodes numbered 8XXX, where X is an integer between 0 and 9, are called *entry/exit* nodes. That is, traffic enters and exits through these nodes *only.* The rest of the nodes are all *internal* nodes. The combination of selected traffic volume, network geometry, and control represents congested network. The purpose of the study was to estimate and validate the relationship between the MOE (y), and the decision variables (x's), for each network. Simulation experiments specified in Tables 1 through 4 were conducted, and MOEs, recorded.

For all these networks, the following first-order metamodel was used to describe the relationship between the response, y (MOE), and the decision variables, x_1 and x_2 ($i = 1, 2, \ldots, 10$ and $j = 1, 2, 3, 4$):

$$y_{ij} = \beta_0 + \beta_l x_{j1} + \beta_2 x_{j2} + \beta_3 x_{j1} x_{j2} + \epsilon_{ij} \qquad (29)$$

where β_0 is the average delay of interest across all design points and replicates.

EXPERIMENTAL SETUP

A sample experimental setup for Network 2 under the CRN strategy is described to help the practitioner appreciate the ease of its use. Since there are two decision variables, or factors (green splits at Nodes 5 and 8), denoted by x_1 and x_2, a 2^2 factorial experiment is conducted—that is, the two factors can be set at two levels, low and a high. The respective low and high levels for this example are 27 and 37 sec as described in Table 4. There can, then, be four combinations to set these factors: x_1 at its low level and x_2 at its low level, x_1 at its high level and x_2 at its low level, and so on. Each of these four combinations is called a design point, and there are four design points for this simulation experiment. Under the CRN strategy, each design point is driven by the same random number stream. Replications of such simulations at each design point, however, use independent random number streams. If five replications are performed, five random number streams are used. However, the same five random number streams are used across the four design points. In practice, applying the CRN strategy is easier than performing "normal simulation" (independent streams), since fewer random number streams are required under the CRN strategy. Note that in this case, the "normal simulation" would need 20 random number streams.

SIMULATION ANALYSIS AND RESULTS

The results of the validation for Networks 1, 2, and 3 are presented in Tables 5, 6, and 7 respectively, and those for statistical analysis procedures (if applicable) are presented in Tables 8, 9, and 10, respectively.

A visual inspection of the correlation matrix in Table 5 indicates consistently induced positive correlations among responses between all pairs of design points in Network 1. It therefore guarantees that variance reduction is achieved, and using the CRN strategy is better for this network than using independent streams. The test for multivariate normality of responses fails to be rejected, and

TABLE 1 Input data for Network 1

Link length:	All links	500 ft.
Signal Control		
Nodes 1 and 3		Signal control
Node 2		Two-phase actuated control
All other nodes		Perpetual Green
Number of Lanes		
Links with 3 lanes		8031-31, 3-33, 31-1, 1-2, 2-3
Links with 1 lane		8022-22, 2-22, 22-2
Links with 2 lanes		All other links.
Entry Volume (Follows a Uniform Distribution)		
node 8011		700 vph
node 8013		750 vph
node 8021		500 vph
node 8022		300 vph
node 8023		650 vph
node 8031		1350 vph
node 8033		650 vph
Turning Movement		
Link 31-1		30% left, 60% through, 10 % right
Link 2-1		30% left, 54% through, 16% right
Link 11-1		14% left, 36% through, 50 % right
Link 21-1		40% left, 40% through, 20% right
Link 3-2		71% through, 29% right
Link 2-3		82% through, 18% right
Link 13-3		19% left, 48% through, 33% right
Link 23-3		27% left, 50% through, 23% right
All other links		100% through

so does the test for the covariance structure from equation 8. The linear model representation in Equation 29 is found to be an adequate representation for Network 1. Since all three stages of the validation procedure failed to reject the null hypotheses in Equations 7, 8, and 9, the analyst proceeds with the statistical analysis. This analysis yields point estimates and confidence intervals for the unknown parameters of the model in Equation 29. From the preceding confidence intervals, it is observed that the main effect for x_1, the green split at Node 1, and the interaction effect do not appear to influence the delay in any significant manner. However, increasing x_2, the green split at Node 3 will decrease the vehicle delay, at least in the vicinity of the current setting of the decision variable.

Table 6 illustrates consistently induced positive correlations among responses between all pairs of design points in Network 2. The use of CRN strategy is therefore justified for this network. The test for multivariate normality of responses fails to be rejected, and so does the test for the covariance structure from Equation 8. The linear model representation in Equation 29 is found to be an adequate representation for Network 2. As for the previous network, all three stages of the validation procedure failed to reject their respective null hypotheses. Statistical analysis can therefore be conducted, as prescribed. From the previous confidence intervals, it is observed that the main effects, x_1, the green split at Node 5, and x_2, the green split at Node 8, are both significant. The interaction term does not appear to contribute significantly to the fitted metamodel. Therefore, decreasing x_1, and increasing x_2, will decrease the vehicle delay, at least in the vicinity of the current setting of the decision variables.

Finally for Network 3, Table 7 indicates consistently induced positive correlations among responses between all pairs of design points in Network 3. Employing the CRN strategy will therefore improve metamodel estimation, provided that the assumptions are validated. The test for multivariate normality of responses fails to be rejected, and so does the test for the covariance structure from Equation 8. The linear model representation in Equation 29 is found to be an adequate representation for Network 3. Since all three stages of the validation procedure failed to reject their null hypotheses, the statistical analysis could be conducted. It yields point estimates and confidence intervals for the unknown parameters of the model in Equation 29. From the confidence intervals, it is observed

TABLE 2 Input data for Network 2

Link length:	All links	500 ft.
Signal Control		
	Nodes 5,6, 7, and 8	Signal control
	All other nodes	Perpetual Green
Number of Lanes	All links	Two lanes
Entry Volume (Follows a Uniform Distribution)		
	All nodes	1600 vph
Turning Movement		
	All links at four-way intersections	25% left, 50% through, 25 % right
	All other links	100% through

TABLE 3 Input data for Network 3

Link length:	All links	500 ft.
Signal Control		
	Nodes 4 through 12	Signal control
	All other nodes	Perpetual Green
Number of Lanes		
	All links	Two lanes
Entry Volume (Follows a Uniform Distribution)		
	nodes 8001, 8005, 8009, and 8011	1000 vph
	All other nodes	1600 vph
Turning Movement		
	All links at four-way intersections	25% left, 50% through, 25 % right
	All other links	100% through

TABLE 4 Information for TRAF-NETSIM Experiments, Networks 1–3, 2^2 Factorial

	Network		
	1	2	3
No. of replications	10	5	5
Duration (sec)	1,800	1,800	1,800
Decision variables[a]			
X_1	Nodes 1 and 31	Nodes 5 and 13	Nodes 4 and 16
X_2	Nodes 3 and 33	Nodes 8 and 14	Nodes 10 and 18
Levels (sec)			
X_1 (low/high)	8/12	27/37	27/47
X_2 (low/high)	13/17	27/37	27/47
MOE: y (sec)	Avg. delay for vehicles entering at Node 31 and exiting at Node 33	Avg. delay in network	Avg. delay in network

[a] Decision variables denote green split and approach nodes.

TABLE 5 Validation results for Network 1

Correlation matrix of responses	
$corr(y) =$	$\begin{bmatrix} 1.0000 & 0.5924 & 0.7259 & 0.7354 \\ 0.5924 & 1.0000 & 0.7339 & 0.7630 \\ 0.7295 & 0.7339 & 1.0000 & 0.7211 \\ 0.7354 & 0.763\ 0 & 0.7211 & 1.0000 \end{bmatrix}$
Test for multivariate normality	
W^*	0.67
$w^*_{0.05}(4,10)$	0.598
Test for the correlation matrix	
M	7.86
$\chi^2(8)$	20.09
Test for lack-of-fit of postulated model	
Test statistic	0.54
$F^{1-0.05}_{(1,26)}$	4.23
Validation complete. Fail to reject all null hypotheses. Therefore proceed with statistical analysis.	

that the main effect for x_1, the green split at Node 16 does not appear to influence the delay in any significant manner. However, increasing x_2, the green split at Node 10 will decrease the vehicle delay, at least near the current setting of the decision variable. The interaction term, however, can play a role in this situation, and hence the analyst should proceed with caution. Conducting a pilot study to explore the effects of increasing the green split at Node 10 may be a suitable alternative before reaching to any meaningful conclusions about this network.

The goal of this paper is to demonstrate the application of the validation and statistical analysis procedures of linear metamodels from multipopulation simulation experiments under the CRN strategy, so the authors do not conduct further analysis on the effect of the decision variables on the response but instead point out the ease of application of such procedures to the practitioner. The three networks selected for this study exhibit three different characteristics in their statistical analyses. Network 1 has only one factor significant, which is the main effect, x_2. The other main effect and the interaction terms

TABLE 6 Validation results for Network 2

Correlation matrix of responses	
$corr(y) =$	$\begin{bmatrix} 1.0000 & 0.4728 & 0.5992 & 0.9048 \\ 0.4782 & 1.0000 & 0.0785 & 0.3184 \\ 0.5992 & 0.0785 & 1.0000 & 0.6362 \\ 0.9048 & 0.3184 & 0.6362 & 1.0000 \end{bmatrix}$
Test for multivariate normality	
W^*	0.72
$w^*_{0.05}(4,10)$	0.5
Test for the correlation matrix	
M	1.6873
$\chi^2(8)$	20.09
Test for lack-of-fit of postulated model	
Test statistic	0.16
$F^{1-0.05}_{(1,26)}$	4.23
Validation complete. Fail to reject all null hypotheses. Therefore proceed with statistical analysis.	

TABLE 7 Validation results for Network 3

Correlation matrix of responses

$$corr(y) = \begin{bmatrix} 1.0000 & 0.8373 & 0.7235 & 0.3418 \\ 0.8373 & 1.0000 & 0.9531 & 0.7992 \\ 0.7235 & 0.9531 & 1.0000 & 0.8538 \\ 0.3418 & 0.7992 & 0.8538 & 1.0000 \end{bmatrix}.$$

Test for multivariate normality	
W^*	0.64
$w^*_{0.05}(4,10)$	0.5
Test for the correlation matrix	
M	7.9986
$\chi^2(8)$	20.09
Test for lack-of-fit of postulated model	
Test statistic	0.16
$F^{1-0.05}_{(1,11)}$	4.84

Validation complete. Fail to reject all null hypotheses. Therefore proceed with statistical analysis.

are not significant. Network 2 exhibits the significance of both major factors on the metamodel, but the interaction term is not significant. Network 3 has one main effect, x_2, and the interaction term to be significant in the fitted metamodel, but the other main effect appears insignificant. These networks are therefore interesting for further exploration in their own way. For example, exploring Network 3 along increasing values of the variable x_2 can be self-defeating if the interaction term increases the delay with an increase in x_2.

In applying the CRN strategy for metamodel estimation and analysis to any simulation experiment, a note of caution is warranted. There is no guarantee that the CRN strategy will produce the desired variance reduction. For this reason, the analyst should conduct a pilot study of the system before conducting an exhaustive simulation analysis. This can be done by computing the correlations obtained across design points for a smaller study and validating the assumptions. If negative correlations are observed across design points, then it is an indication that the CRN strategy may not be amenable for this particular problem.

CONCLUSIONS AND FUTURE RESEARCH

This paper demonstrates the statistical analysis and validation techniques for linear metamodels in multipopulation traffic simulation networks under the CRN correlation-induction strategy. This illustration comprises a three-stage validation procedure and comprehensive postvalidation statistical analysis. The examples show the ease of applying the CRN strategy for traffic simulation experiments and of performing estimation and analysis on a fitted metamodel. The TRAF-NETSIM model can be used to perform efficient simulation experiments that could help a traffic simulation analyst gain more confidence in the results.

TABLE 8 Statistical analysis results for Network 1

Optimal estimator of β

$$\hat{\beta} = \begin{bmatrix} 27.1339 \\ -0.0933 \\ -3.6866 \\ -0.9689 \end{bmatrix}.$$

95% Confidence Interval for β_0:	$25.82 \le \beta_0 \le 28.44$

Simultaneous 95% confidence interval for elements of β

$-1.94 \le \beta_1 \le 1.76$, $-5.54 \le \beta_2 \le -1.84$, $-2.82 \le \beta_3 \le 0.88$

TABLE 9 Statistical analysis results for Network 2

Optimal estimator of β

$$\hat{\beta} = \begin{bmatrix} 533.1096 \\ 18.3969 \\ -4.9419 \\ 3.7534 \end{bmatrix}.$$

95% Confidence Interval for β_0: $524.47 \leq \beta_0 \leq 541.74$

Simultaneous 95% confidence interval for elements of β

$$14.53 \leq \beta_1 \leq 22.24, \quad -8.79 \leq \beta_2 \leq -1.09, \quad -0.1 \leq \beta_3 \leq 7.6$$

TABLE 10 Statistical analysis results for Network 3

Optimal estimator of β

$$\hat{\beta} = \begin{bmatrix} 526.6619 \\ 1.0044 \\ 14.3064 \\ -11.0461 \end{bmatrix}.$$

95% Confidence Interval for β_0: $517.8 \leq \beta_0 \leq 535.51$

Simultaneous 95% confidence interval for elements of β

$$-2.46 \leq \beta_1 \leq 4.46, \quad 10.84 \leq \beta_2 \leq 17.76, \quad -14.5 \leq \beta_3 \leq -7.58$$

Future development in the TRAF-NETSIM model could include the statistical analysis procedures incorporated within the model for different types of simulation experimentation. Another avenue for research would be to develop statistical analysis and validation techniques for multiple responses instead of just a single response.

REFERENCES

1. Rathi, A. K. The Use of Common Random Numbers to Reduce the Variance in Network Simulation of Traffic. *Transportation Research,* Vol. 26B, 1992, pp. 357–363.
2. Traffic Network Analysis with NETSIM—A User Guide. Report FHWA-IP-80-3. FHWA, U.S. Department of Transportation, 1980.
3. Rathi, A. K., and E. B. Lieberman. Effectiveness of Traffic Restraint for a Congested Urban Network: A Simulation Study. In *Transportation Research Record 1232,* TRB, National Research Council, Washington, D.C., 1989, pp. 95–102.
4. Law, A. M., and W. D. Kelton. *Simulation Modeling and Analysis,* 2nd ed. McGraw Hill, New York, 1991.
5. Tew, J. D., and J. R. Wilson. Validation of Simulation Analysis Methods for the Schruben-Margolin Correlation-Induction Strategy. *Operations Research,* Vol. 40, 1992, pp. 87–103.
6. Joshi, S. S., and J. D. Tew. Validation and Statistical Analysis Procedures Under the Common Random Number Correlation-Induction Strategy for Multipopulation Simulation Experiments. *European Journal of Operational Research* (in preparation).

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 (*1*). 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-

Systems and Industrial Engineering Department, Engineering Building 20, University of Arizona, Tucson, Ariz. 85721.

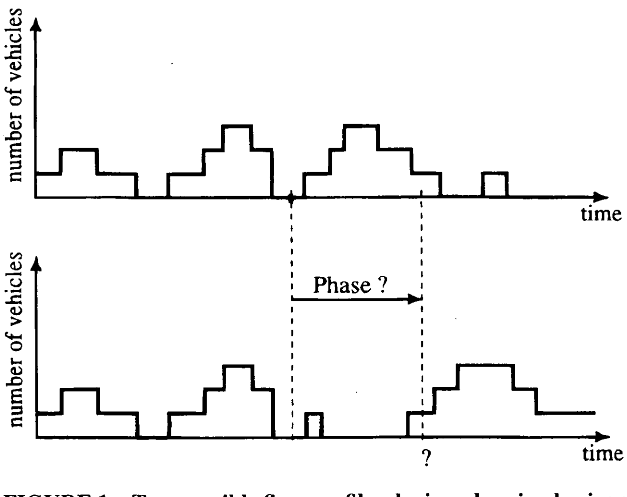


FIGURE 1 Two possible flow profiles during planning horizon.

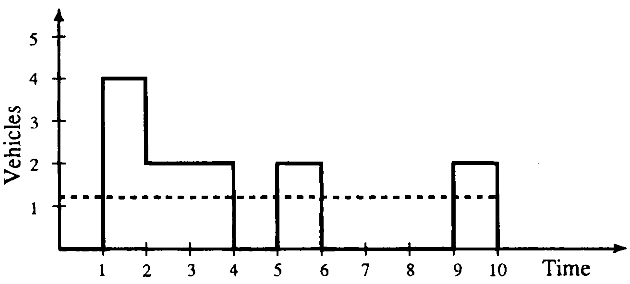


FIGURE 2 Depiction of difference between prediction frequencies.

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-by-cycle 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 1 Comparison of Existing Traffic Demand Prediction Algorithms (*2*)

	Characteristic			
Algorithm	Horizon	Frequency[a]	Sources	Performance
UTCS-2	5-15 min	1	Single	Best for 5 min
UTCS-3	5-15 min, cycle	1, 1	Single	Poor Overall
Historical Average	5-15 min	1	Single	Best for cycle
Current Measurement	5-15 min, cycle	1, 1	Single	Poor due to time delay
PID	5-15 min, cycle	1, 1	Single	Sensitive to Parameters

[a]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,

$$n(t) = n_p(t) + n_u(t) \tag{1}$$

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:

$$SNR = \frac{\int_{t}^{t+T} n_p(t)dt}{\int_{t}^{t+T} n_p(t) + n_u(t)dt} \tag{2}$$

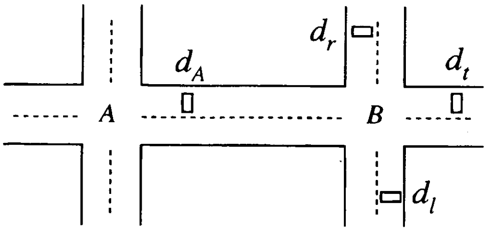


FIGURE 3 Geometric layout of prediction scenario.

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 $SNR \approx 0.5$, 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_l, d_t, and d_r, 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_i where $i \in \{l,t,r\}$, at time t_{d_i}. Let this event be denoted $e_i(t_{d_i}.)$ Several factors affect when and if the vehicle will arrive at d_A, including

- Travel time from d_i 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 that the vehicle will travel along a route that includes location d_A.

Figure 4*a–d* depict the delay associated with the first four factors. In Figure 4*a* the vehicle arrives at detector d_i and passes freely to detector d_A. The arrival time, denoted t_a, at d_A can be estimated as

$$t_a = t_{d_i} + T_{d_i,S_B} + T_{S_B,d_A} \tag{3}$$

where $T_{d_i.S_B}$ is the travel time from d_i to the stop bar at intersection B and T_{S_B,d_A} is the travel time from the stop bar at intersection B to the detector at d_A. Each of these travel times can be estimated on the basis of the approach speed and link flow speed, respectively.

The approach speed can be estimated, for conventional detectors, from the state of the signal and the occupancy of the detector. The link flow speed can be estimated from the occupancy at d_A. Both travel time estimates can be greatly improved using probe vehicle data and advanced surveillance technologies such as video surveillance. However, it cannot be assumed that these information sources will be available, so the prediction model must perform well enough using only conventional detector information.

In Figure 4*b* the vehicle arrives at detector d_i and is delayed by the signal at intersection B. Hence the travel time from d_i to d_A must account for the travel time from d_i to the stop bar, the delay due to the signal, and the travel time from the stop bar to d_A. The arrival time at d_A can then be estimated as

$$t_a = t_{d_i} + \max\{T_{d_i,S_B}, T_{u_B}\} + T_{S_B,d_A} \tag{4}$$

where T_{u_B} is the delay until the signal timing plan advances to a phase that will serve the desired movement. The travel time from d_i to the stop bar can be estimated on the basis of approach speed, assuming that the vehicle will have to stop. The signal timing delay can be determined using the signal timing plan and the travel time from the stop bar to d_A can be estimated from the link flow speed, assuming that the vehicle starts from a stop at B.

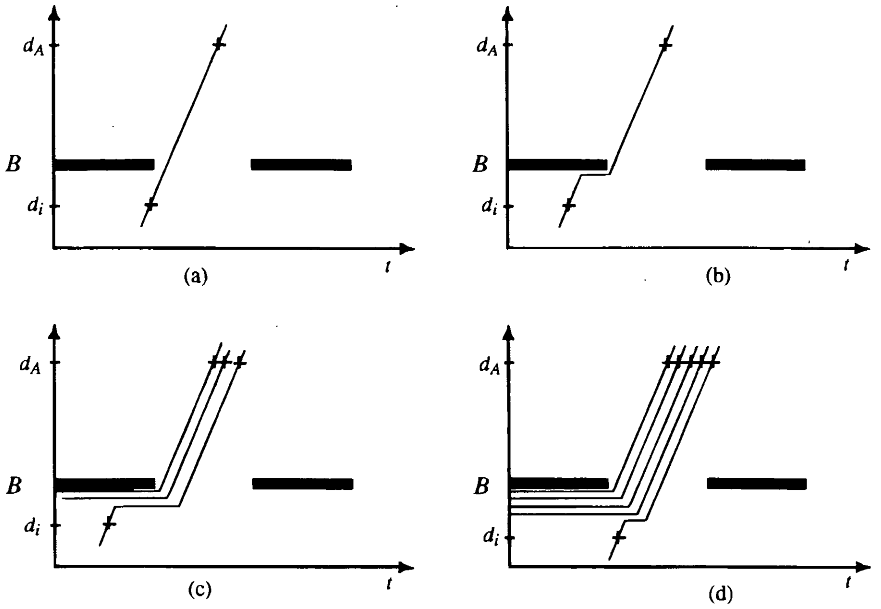


FIGURE 4 Delay associated with predicted travel time: *a*, detected vehicle passes freely through intersection; *b*, detected vehicle arrives during red signal—signal delay; *c*, detected vehicle arrives during red signal and a queue exists—signal and queue delay; and *d*, detected vehicle arrives during the green signal and a queue exists—queue delay.

In Figure 4*c* 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_A. The travel time is defined as

$$t_a = t_{d_i} + \max\left\{T_{d_i,S_B}, T_{u_B} + T_{q_i}\right\} + T_{S_B,d_A} \tag{5}$$

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

$$T_{q_i} = a_0 + a_1 N_{q_i} \tag{6}$$

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 4*d* 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

$$t_a = t_{d_i} + \max\left\{T_{d_i,S_B}, T_{q_i}\right\} + T_{S_B,d_A} \tag{7}$$

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, $e_i(t_{d_i})$, 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_i^{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 $n_i(t_{d_i})$ vehicles cross detector d_i in time interval t_{d_i}, then using Equations 3–7, the expected number of arrivals at d_A at time t_{d_A} can be predicted to be

$$\hat{n}_A(t_{d_A}) = \sum_{i\in\{l,t,r\}} \sum_{\forall t_p = t_{d_A}} p_i^{BA} n_i(t_p) \tag{8}$$

The inner summation estimates that the expected number of arrivals at d_A predict that these arrivals will occur at future time $t_p = t_{d_A}$ 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.

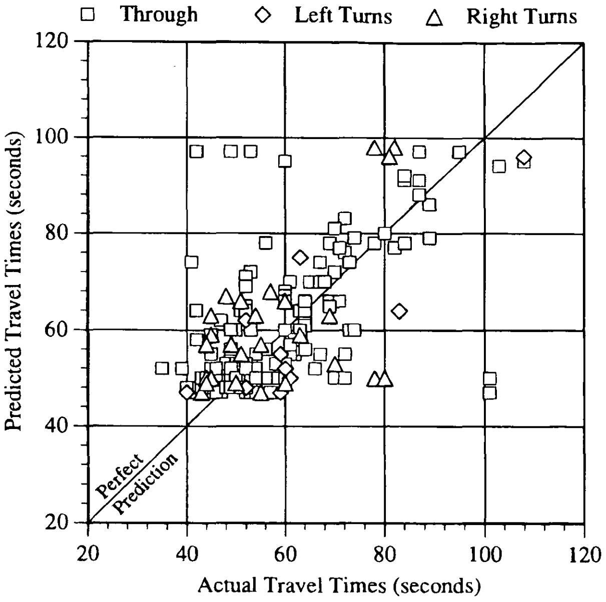


FIGURE 5 Actual versus predicted travel times.

Figures 6*a–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 6*a*, 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 6*b*. 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 6*c* 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, $\hat{n}_A(t_{d_A})$, 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

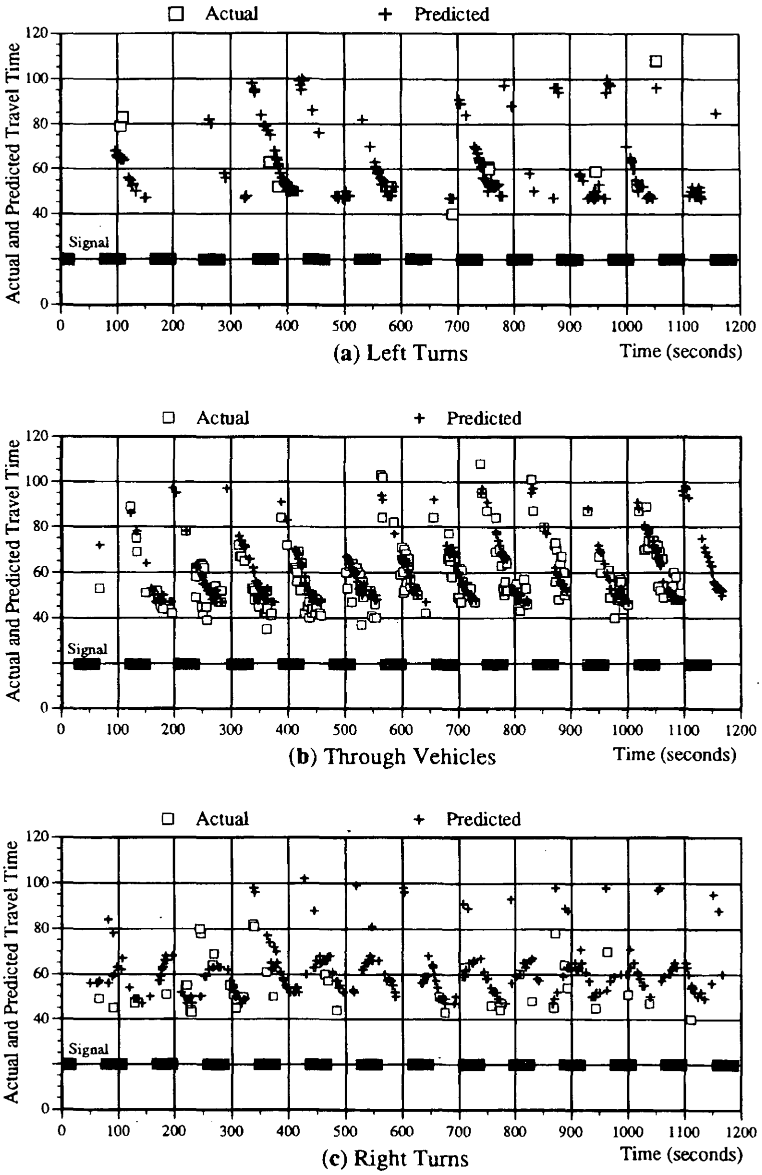


FIGURE 6 Predicted and actual travel times as function of prediction time.

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.

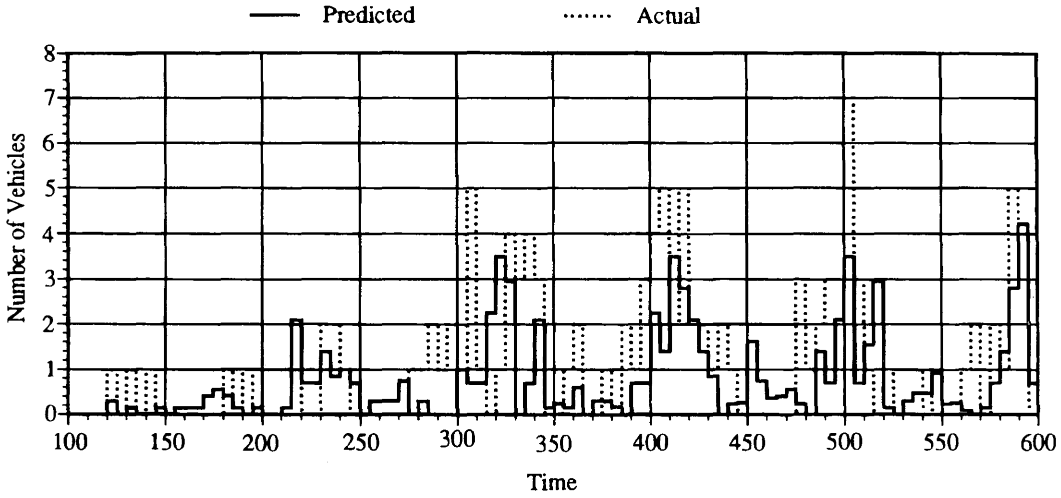


FIGURE 7 Flow profile showing actual and predicted number of arrivals during time intervals of 5 sec.

Figure 9 shows the ability of the prediction model and the traffic-adaptive signal control logic to work together to reduce the delay at the intersection.

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

Measure	Value
ME	0.09
SSE	358.13
MSE	0.31
MAE	0.29
MARE	0.27
D-W	1.94

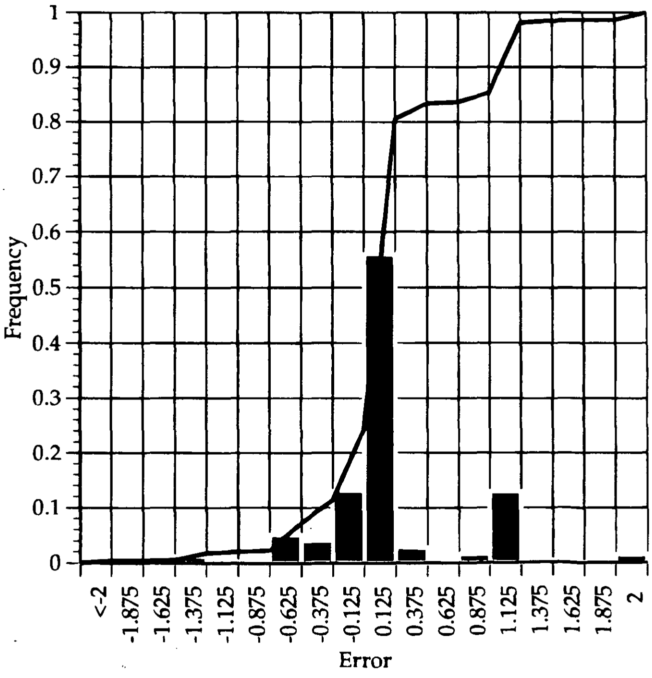


FIGURE 8 Histogram of prediction errors.

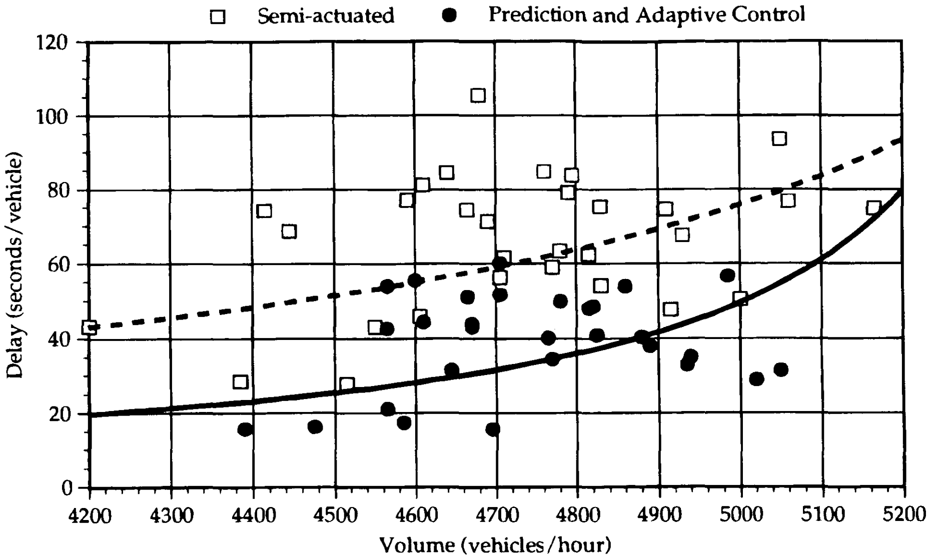


FIGURE 9 Comparison of prediction model with a real-time traffic adaptive signal timing optimization logic and semiactuated control.

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

1. Tarnoff, P. J. The Results of FHWA Urban Traffic Control Research: An Interim Report. *Traffic Engineering,* Vol. 45, 1975, pp. 27–35.
2. Stephanedes, Y. J., P. G. Michalopoulos, and R. A. Plum. Improved Estimation of Traffic Flow for Real-Time Control. In *Transportation Research Record 795,* TRB, National Research Council, Washington, D.C., 1981, pp. 28–38.
3. Okatani, I., and Y. J. Stephanedes. Dynamic Prediction of Traffic Volume Through Kalman Filtering Theory. *Transportation Research Part B: Methodological,* Vol. 18B, No. 1, 1984, pp. 1–11.
4. Gartner, N. H. Discussion of Improved Estimation of Traffic Flow for Real-Time Control. In *Transportation Research Record 95,* TRB, National Research Council, Washington, D.C., 1981, pp. 38–39.
5. Hunt, P. B., D. I. Robertson, R. D. Bretherton, and R. I. Winton. *SCOOT—A Traffic Responsive Method of Coordinating Signals.* Report LR 1014. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1981.
6. Gartner, N. H., P. J. Tarnoff, and C. M. Andrews. Evaluation of Optimized Policies for Adaptive Control Strategy. In *Transportation Research Record 1324,* TRB, National Research Council, Washington, D.C., 1991.
7. Mauro, V., and D. Di Taranto. UTOPIA. *Proc., 6th IFAC/IFIP/IFORS Symposium on Control, Computers and Communication in Transportation,* Paris, France, 1990.
8. *UTOPIA Control Strategies Description.* Draft Report. Mizar Automazione S.p.A., Turin, Italy, April 1992.
9. Sims, A. G. The Sydney Coordinated Adaptive Traffic System. *Urban Proc., Engineering Foundation Conference on Research Priorities in Computer Control of Urban Traffic Systems, Urban Transport Division, ASCE,* 1979, pp. 12–27.
10. McShane, W. R., and R. Roess. *Traffic Engineering.* Prentice-Hall, New York, 1990.
11. Baras, J. S., W. S. Levin, and T. L. Lin. Discrete-Time Point Processes in Urban Traffic Queue Estimation. *IEEE Transactions on Automatic Control,* AC-24, No. 1, 1979, pp. 12–27.
12. Head, K. L., P. B. Mirchandani, and D. Sheppard. A Hierarchical Framework for Real-Time Traffic Control. In *Transportation Research Record 1360,* TRB, National Research Council, Washington, D.C., 1992, pp. 82–88.
13. *TRAF-NETSIM User's Manual.* FHWA, U.S. Department of Transportation, 1992.
14. Makridakis, S., S. C. Wheelwright, and V. E. McGee. *Forecasting: Methods and Applications.* John Wiley and Sons, New York, pp. 43–54.
15. Sen, S., and K. L. Head. Controlled Optimization of Phases (COP) at an Intersection. SIE Department Working Paper, University of Arizona, 1994. *Transportation Science,* (in preparation).
16. Farradyne Systems, Inc. *Functional Specifications: Final Task A Interim Report.* FHWA, U.S. Department of Transportation, April 1993.

Estimating Intersection Turning Movement Proportions from Less-Than-Complete Sets of Traffic Counts

GARY A. DAVIS AND CHANG-JEN LAN

Estimated turning movement proportions are used in a number of traffic simulation and traffic control procedures to predict the turning movement flows at intersections. Historically, these proportions have been estimated by manual counting, but the ongoing deployment of real-time adaptive traffic control strategies indicates that the ability to automatically estimate these proportions from traffic detector data is becoming increasingly important. When it is possible to count the vehicles both entering and exiting at each of an intersection's approaches, methods based on ordinary least squares can produce usable estimates of the turning movement proportions, but when the number or placement of the detectors does not support complete counting, these methods fail. The feasibility of estimating turning movement proportions from less-than-complete sets of traffic counts is assessed, and the statistical properties of less-than-complete count estimates are compared with estimates generated from complete counts. It turns out that estimation from less-than-complete counts can be done as long as the detector configuration satisfies an identifiability condition. A numerical test is presented to assess whether or not this condition is satisfied, along with some simple rules for designing detector configurations that are likely to satisfy this condition. A Monte Carlo experiment suggests that estimates generated from less-than-complete counts can be more variable than those generated from complete counts.

A commonly used representation of the demand for travel on a bounded network of urban streets requires specifying (*a*) the arrival flows at each input point on the boundary of the network, and (*b*) the turning movement proportions at each of the network's intersections. Both arrival flows and turning movement proportions may vary in time. When coupled with a method for estimating the travel times on street segments, knowledge of the arrival flows and turning movement proportions allows a traffic engineer to predict the turning movement flows at each intersection in the network, and these in turn are needed to evaluate the effectiveness of all but the most simple intersection signal control plans. Not surprisingly, this representation of demand has a long history of practical application, including use by classical methods for computing pretimed controls for isolated intersections (e.g., Webster's method), the *Highway Capacity Manual*'s method for evaluating level of service at intersections and along arterials (*1*), and computer models used for off-line optimization and evaluation of timing plans for networks of intersections (e.g., TRANSYT, NETSIM). More recently, on-line adaptive control schemes (e.g., SCAT, CARS) have also used this representation.

In the past, a major limitation on the timely updating of signal control plans was that the only reliable method for estimating the turning movement flows was time-consuming and costly manual counting. This limitation became even more burdensome when one wished to adapt a control plan in real time, and often it led to reliance on a stored library of "typical" turning movement patterns, which were determined by off-line counting. It is no surprise, then, that over the past 15 years, a number of researchers have investigated methods for estimating turning movement proportions automatically from the traffic count data collected by real-time traffic control systems, which typically are gathered using detectors embedded in the pavement. Almost without exception, however, this work has assumed that it is possible to count the total number of vehicles entering the intersection from each of its approaches as well as the total number of vehicles exiting from each exit leg. For example, the intersection of two two-lane, two-way streets would require a minimum of eight detectors. It is now well-established that when time series of an intersection's input and output counts are available, estimation methods based on ordinary least squares will produce usable estimates of the turning movement proportions, both off-line and in real time (*2–6*). However, such a rich density of detectors tends to be the exception rather than the norm, at least in the United States, and the slow application of automatic turning movement estimation in the United States can in part be blamed on the added expense imposed by the additional detectors. The functional specifications for real-time traffic adaptive control systems (RT-TRACS), recently prepared for FHWA, explicitly recognizes this limitation by calling for a maximum of 20,000 detectors for a total of 5,000 intersections.

Before proceeding, it is useful to specify more completely the relation between this paper and past work. For the case in which counters are placed at each entry and exit point of an intersection, it has been recognized that the problem of estimating turning movement flows or turning movement proportions from the counts is a special case of the more general problem of estimating an origin-destination (OD) matrix from traffic counts, and reviews of this problem can be found elsewhere (*7–9*). As noted by Davis (*10*), OD estimation methods can be classified as either over- or underdetermined, depending on whether the traffic count data at hand are sufficient to produce a unique estimate of the OD elements. For underdetermined approaches, an infinite number of OD estimates consistent with the count data will exist, and one of these is selected by specifying a prior estimate of the OD matrix and then selecting as the new estimate the OD matrix that is consistent with the count data and "closest" to the prior estimate (*7*).

Three general approaches to underdetermined OD estimation have appeared to date, defined primarily by how they define "closeness" to the prior estimate: the information minimizing (IM)

Department of Civil Engineering, University of Minnesota, 500 Pillsbury Drive, S.E., Minneapolis, Minn. 55455.

approach developed by Van Zuylen and Willumsen (*11*) and Bell (*12*), the weighted least squares (WLS) approach initiated by Maher (*13*) and Cascetta (*14*), and a maximum likelihood (ML) approach described by Speiss (*15*). Speiss also assumes that the prior estimate comes from a survey with known sampling properties. All of these approaches are subject to the criticism that despite more than 15 years of research, none has been shown to yield estimators that are consistent, in the statistical sense of becoming increasingly accurate as the amount of traffic count data become arbitrarily large. In fact, Davis and Nihan (*16*) have shown that an underdetermined least squares OD estimator remains underdetermined, and hence not consistent, even with an infinite time series of traffic count data. The IM, WLS, and ML approaches all have specializations to the problem of estimating an intersection's turning movement flows from traffic counts, and Maher (*17*) has provided a concise summary of these methods, where he found that for a particular computational example, these three approaches tended to produce similar estimates. As with general OD estimators, the underdetermined methods for estimating intersection turning flows will fail if a good prior estimate is not available, so they are unable to "bootstrap" good estimates from traffic count data alone. This dependence on prior information makes them particularly ill-suited for real-time implementation.

The limitations of underdetermined approaches were described by Cremer and Keller (*2*), who also described the first overdetermined method for estimating intersection turning movement proportions. Here it was assumed that time-series data of the intersection's entering and exiting counts were available, and an estimate of the turning movement proportions was coupled with the entering counts to produce predictions of the exiting counts. Those values of the turning movement proportions that minimized a measure of error between the predicted and observed exit counts were then selected as the best estimates. Subsequent papers (*3–5*) located this work within the framework of the systems identification paradigm (*18–20*), and general results on systems identification have been used to show not only that ordinary least-squares estimates of turning movement proportions are consistent (*5*), but also that consistent estimates of more general OD matrices can be computed from time series of traffic counts (*10*). A particular advantage of the systems identification approach is that real-time implementation of the estimation algorithms is often straightforward.

When considering the problem of estimating turning movement proportions for a network of intersections and complete entry and exit counts are not available, the estimation problem is no longer a special case of OD estimation, and to date no underdetermined methods have been proposed for this problem. When time series of traffic counts are available, however, the overdetermined estimation problem again falls within the systems identification paradigm, for which a reasonably general statistical theory (*20*) and real-time implementations (*18*) have been described. This paper considers the problem of estimating intersection turning movement proportions in networks where time series of traffic counts are available from automatic traffic detectors but the number or placement of the detectors may not be sufficient for the standard least-squares estimation methods. Although it is recognized that method of moments, least squares, and ML approaches are applicable to this problem, the focus will be on a nonlinear least squares (NLS) approach because *(a)* it leads to a straightforward generalization of the methods that use complete sets of counts, and *(b)* the basic ideas behind this approach can be developed with the least amount of statistical jargon. Thus the authors believe that the NLS approach is more likely to be accessible to interested practitioners. The primary focus in this paper is on determining feasibility, so the authors concentrate on off-line computation of the turning proportion estimates and simply note that on-line versions of NLS estimation, using state-space models, have been described in the literature *(18,21)*. This restriction to off-line methods is justified by the fact that an approach that performs poorly off-line will also perform poorly on-line, and the pathologies of an approach are usually easier to diagnose off-line.

TRAFFIC FLOW MODEL

To date, all methods for automatic estimation of turning movement proportions have used prediction error minimization methods, in which one first specifies a model for predicting the intersection's exit counts using the intersection's input counts and a trial set of turning movement proportions. One then selects as the estimated proportions those values that minimize some measure of the difference between the predicted and the actual exit counts. The prediction model thus is essential for estimating, or identifying the turning proportions. The first requirement then is a prediction model that is capable of handling several intersections simultaneously and that allows for a variety of detector configurations.

Consider a set of street intersections surrounded by a cordon boundary. Traffic counters are located at each point where traffic can enter the cordon area; they count the number of vehicles crossing into the cordon area at that point. Suppose there are m of these input counters, and let $q_i(t)$ equal the traffic count at input counter i during time interval t, $i = 1, \ldots, m$.

Next, suppose the streets within the cordon have been divided into s sections, or compartments, according to the following rules:

1. Traffic flow within a compartment is unidirectional,
2. The stop lines at intersections always mark the downstream boundaries of a compartment, and
3. The exit line on an intersection leg always marks the upstream boundary of a compartment.

A segment of a two-way street connecting two intersections must be divided into at least two compartments, one for each direction, with the compartment boundaries being the intersection stop and exit lines. These two compartments may be divided further. At a total of n compartment boundary points are placed additional detectors that count the number of vehicles crossing that boundary point. Call these the output detectors, and let $y_j(t)$ equal the number of vehicles crossing output detector j during time interval t, $j = 1, \ldots, n$.

Next, let

$x_k(t)$ = number of vehicles in compartment k at beginning of time interval t;
$\mathbf{q}(t), \boldsymbol{x}(t), \mathbf{y}(t)$ = m-, s-, and n-dimensional vectors, respectively, containing individual elements $q_i(t), x_k(t)$, and $y_j(t)$;
b_{lk} = proportion of vehicles currently in compartment l that desire entry into compartment k, if compartment l is adjacent to compartment k, or 0, if compartment k is not adjacent to compartment l;
$\boldsymbol{b}$ = d-dimensional vector containing turning movement proportions;

$p_k(\mathbf{x})$ = proportion of vehicles that can physically exit compartment k during time interval t, as a function of the current distribution of vehicles in the system;

g_{ki} = 1, if input counter i is at the upstream boundary of compartment k, and 0, otherwise:

The distribution of vehicles over the compartments then evolves in time according to the mass balance equations

$$x_k(t+1) = \{1 - p_k[\mathbf{x}(t)]\}x_k(t) + \sum_l x_l(t)p_l[\mathbf{x}(t)]b_{lk} + \sum_i g_{ki}q_i(t) \quad k = 1,\ldots,s \qquad (1)$$

Thus the quantity $x_l(t)\, p_l\, [\mathbf{x}(t)]$ gives the number of vehicles actually exiting compartment l during time interval t, and these are then distributed to the neighboring compartments in proportion to the b_{lk}, with $\Sigma_k b_{lk} = 1.0$. At this point no assumptions are made concerning specific functional forms for the exit probabilities $p_k(\mathbf{x})$, but note that plausible forms can be derived from traffic flow models, so that the quantity $x_k p_k$ behaves like a traffic flow—that is, as the product of space-mean speed and traffic density *(22–24)*. Additional generality can be achieved by letting these exit functions depend explicitly on time or on the destination compartment as well as the origin compartment, or on other dynamic variables, such as compartment mean speeds, making this class of models roughly coextensive with macroscopic traffic models based on continuum theory. Such enhancements do not affect the main conclusions of this paper, but they tend to obscure the drift of the argument with notational complexities and so will not be dealt with here. It is noted, though, that actual application requires specification of the exit functions.

Finally, for a given sequence of input counts $\mathbf{q}(1), \mathbf{q}(2), \ldots, \mathbf{q}(N)$ and a given vector of turning movement proportions $\mathbf{b}$, predicted output counts can be generated by solving the mass balance equations recursively while computing the predicted output counts via

$$\hat{y}_j(t,\mathbf{b}) = \begin{cases} x_k(t)p_k[\mathbf{x}(t)] & \text{if detector } j \text{ counts exits from compartment } k \\ \sum_l x_l(t)p_l[\mathbf{x}(t)]b_{lk} & \text{if detector } j \text{ counts entries into compartment } k \end{cases} \qquad (2)$$

Equations 1 and 2 define a nonlinear state-space model: the first describes the state dynamics and the second gives predictions of the observations.

The simplest example of such a model would be a network consisting of a single intersection and its adjacent compartments, with the input counters located at the upstream boundaries of the intersection's approaches, the output counters located at the intersection's exit points, and $p_k(\mathbf{x}) = 1.0$ for all k and $\mathbf{x}$. Since each proportion b_{kl} corresponds to exactly one input/output pair, these can be reindexed as b_{ij}, and they give the intersection's turning movement proportions as defined elsewhere *(2–6)*. In this case, given the input counts, the prediction of an output count is given by the simple linear relationship

$$\hat{y}_j(t,\mathbf{b}) = \sum_i b_{ij}q_i(t-1) \qquad (3)$$

and constrained ordinary least squares (CLS) estimates of the turning movement proportions can be computed by minimizing the sum of squares function

$$S_1(\mathbf{b}) = \sum_t \sum_j [y_j(t) - \sum_i b_{ij}q_i(t-1)]^2 \qquad (4)$$

subject to the constraints

$$0 \le b_{ij} \le 1.0 \qquad (5a)$$

$$\sum_j b_{ij} = 1.0 \qquad i = 1,\ldots,m \qquad (5b)$$

This problem is well-defined as long as the matrix

$$\mathbf{Q} = \sum_t \mathbf{q}(t)\mathbf{q}(t)^T \qquad (6)$$

is nonsingular. This is the basic model used by Cremer and Keller *(2,3)* and Nihan and Davis *(4,5)* in developing their numerous variants of least-squares estimators of turning movement proportions, whereas letting $p_k \le 1$ produces the platoon dispersion model proposed by Bell *(6)* to account for travel time lags between the input and output counters.

IDENTIFIABILITY OF TURNING MOVEMENT PROPORTIONS

Returning now to the nonlinear prediction model defined in Equations 1 and 2, for a given sequence of input counts and an estimate of the turning movement proportions $\mathbf{b}$, this model can be used to generate a sequence of predicted output counts, which in turn can be used to compute the sum-of-squares function

$$S_2(\mathbf{b}) = \sum_t [\mathbf{y}(t) - \hat{y}(t,\mathbf{b})]^T[\mathbf{y}(t) - \hat{y}(t,\mathbf{b})] \qquad (7)$$

where $\hat{y}(t, \mathbf{b})$ denotes the vector of predicted outputs produced by Equation 2. The dependence of the predicted outputs on the unobserved state vector $\mathbf{x}(t)$ makes $\hat{y}(t, \mathbf{b})$ a nonlinear function of the turning movement proportions, so that attempting to minimize S_2 with respect to $\mathbf{b}$ leads to an NLS problem. This can be solved using any of a number of standard routines as long as the problem is well-defined, in the sense that at least a locally unique minimizing value of $\mathbf{b}$ exists. It may be, though, that the number or placement of the output detectors is not sufficient to produce a well-defined problem, leading to a situation analogous to the underdetermined OD estimation problem.

The problem of determining in advance whether a data collection experiment will support estimation of a model's parameters is an example of the system identifiability problem, to which a substantial research effort has been devoted *(24,25)*. It is straightforward to verify that when the output count predictions are differentiable functions of the turning movement proportions (which is true for prediction model used here), and when there exists a vector $\mathbf{b}_0$ that produces "good" predictions (in the sense that the prediction errors are uncorrelated with the input counts), then the problem will be well-defined as long as the matrix $\mathbf{J}(\mathbf{b})^T\mathbf{J}(\mathbf{b})$ is nonsingular, where

$$\mathbf{J}(\mathbf{b}) = \begin{bmatrix} \dfrac{\partial \hat{y}_1(1,\mathbf{b})}{\partial b_1} & \cdots & \dfrac{\partial \hat{y}_1(1,\mathbf{b})}{\partial b_d} \\ \cdots & \cdots & \cdots \\ \dfrac{\partial \hat{y}_n(N,\mathbf{b})}{\partial b_1} & \cdots & \dfrac{\partial \hat{y}_n(N,\mathbf{b})}{\partial b_d} \end{bmatrix} \qquad (8)$$

is the Jacobian matrix giving the derivatives of the predicted output counts with respect to the turning movement proportions. In practice, one would test whether or not a particular configuration of output counters will support identification of the turning movement proportions by computing the determinant of $\mathbf{J}(\mathbf{b})^T\mathbf{J}(\mathbf{b})$ at a sample of values for **b**, using a typical sequence of input counts. Analytic expression for the partial derivatives appearing in $\mathbf{J}(\mathbf{b})$ is not needed, as these can be evaluated numerically; as long as one can generate an *a priori* reasonable set of input counts, no actual data are needed to perform these tests. This makes this test suitable for use in designing detector configurations. A justification for testing only a few sample values for **b** is given by a result attributable to Eisenfeld (*26*): suppose $\mathbf{J}(\mathbf{b})$ is a polynomial function of **b** (as is the case for the prediction model described by Equations 1 and 2). Then if there exists one value **b** such that the determinant of $\mathbf{J}(\mathbf{b})^T J(\mathbf{b})$ does not equal 0, the determinant of $\mathbf{J}(\mathbf{b})^T\mathbf{J}(\mathbf{b})$ does not equal 0 for almost all values of **b.**

Experience from the identification of compartment models in biology and medicine indicates that this property, known as local identifiability, is useful for determining which data collection configurations can support parameter estimation (*24,25*).

DESIGN OF IDENTIFIABLE DETECTOR CONFIGURATIONS

The Jacobian test provides a method for assessing the ability of a given detector configuration to provide enough information for estimating turning movement proportions, but it provides no guidance as to how one might arrive at plausible configurations in the first place, nor does the test indicate how to correct an unidentifiable configuration. Ideally one would like to have identifiability conditions that are both necessary and sufficient, where the necessary conditions give guidance on how to design the detector configuration while the sufficient conditions verify that the design is in fact adequate. In the current state of the art, useful necessary and sufficient conditions have yet to be found, even for linear, time-invariant models. For linear models, however, there do exist necessary conditions that indicate how to avoid certain common reasons for nonidentifiability, and although the traffic model described previously is nonlinear, because of the dependence of the exit flows on the current traffic distribution $\mathbf{x}(t)$, it shares many of the structural features of linear models, becoming a time-invariant linear model when the exit probabilities are constant. Thus it can be recommended that following the conditions for linear systems should provide good starting points for designing identifiable detector configurations for the nonlinear model.

- A configuration of detector placements will be said to produce an input-reachable model if there is a route to each compartment from at least one input detector. Similarly, the configuration is output-reachable if there exists a route from each compartment to at least one output detector.
- A pair of turning movement proportions will be called inseparable if every route connecting an input detector to an output detector that involves one of these turning movements also involves the other.

For linear models, it has been established that models that are not input- and output-reachable are unidentifiable whereas two inseparable parameter values will be identifiable only in special cases (*25*). Thus input and output reachability and separability can be regarded as highly desirable properties for a detector configuration, and for very simple networks it is usually possible to verify input and output reachability and separability by inspecting a graphical representation of the network (*25*). For larger networks, input and output reachability can be verified by computing reachability matrices for the network (*27*), but separability is more difficult to check. The task becomes much simpler if the network shows the graph theoretic properties of strong connectedness and degree-2 vulnerability. [By strongly connected, the authors mean that it is possible to travel from any internal compartment to any other internal compartment; by degree-2 vulnerability, they mean that the network remains strongly connected even if any one of its turning movements is forbidden. Roberts (*27*) gives a more detailed discussion of these properties.]

- *Proposition.* Suppose a network of intersections is bounded by a cordon line, with no internal origins or destinations. Suppose the network is strongly connected and degree-2-vulnerable and that detectors are placed so that a complete cordon count of both entering and exiting vehicles is achieved. Then this detector placement is both input- and output-reachable and separable.
- *Proof.* Since the vehicles entering from the cordon line must enter an internal compartment, and since the vehicles exiting at the cordon line must exit from an internal compartment, strong connectivity implies input and output reachability. Now let $(k, l_1, l_2 \ldots, l_r l)$ denote a sequence of compartments that when traversed, form a route from input point k to output point l. Let $[(k, l_1)(l_1, l_2), \ldots ,(l_r, l)]$ denote the sequence of turning movements used in traversing this route, and select any two turning movements from this sequence, denoting them by (l_a, l_b) and (l_α, l_β). Since the network is degree-2-vulnerable, it is possible to forbid movement (l_α, l_β) and still construct one route from an input from an input point to compartment l_a and another route from compartment l_b to an output point. Joining these routes with the movement (l_a, l_b) creates a route from an input to an output that uses (l_a, l_b) but not (l_α, l_β), so the configuration is separable.

Although it is easy to construct networks that are not strongly connected (the network shown in Figure 1 is an example), the authors believe that most well-designed street systems should have this property. For if a network is not strongly connected, it will be possible to divide it into two or more components, some of which are inaccessible from others (*27*). That is, a vehicle that is one part of the network can find it impossible to travel to other parts. Degree-2 vulnerability also appears plausible but less general, so that some networks will have this property and some will not. One exception would arise from a T-intersection formed by two one-way streets, where, for instance, vehicles turn left from the cross of the T into the stem of the T. Forbidding this left turn would make it impossible to enter the stem of the T (and hence destroy the network's strong connectivity), and this also makes it impossible to construct a route using a movement exiting the stem of the T without using this left turn. The solution for this problem would be to place an additional output detector to count vehicles entering the stem of the T, so that routes terminating at this detector would separate the left turn into the stem of the T from the movements exiting the stem. Finally, for networks with internal origins or destinations, placing detectors to count the vehicles exiting or entering these points will convert them to "internal" cordon points, and the preceding results will still hold.

To summarize, a detector configuration that is input- and output-reachable and separable is not guaranteed to be identifiable, but it

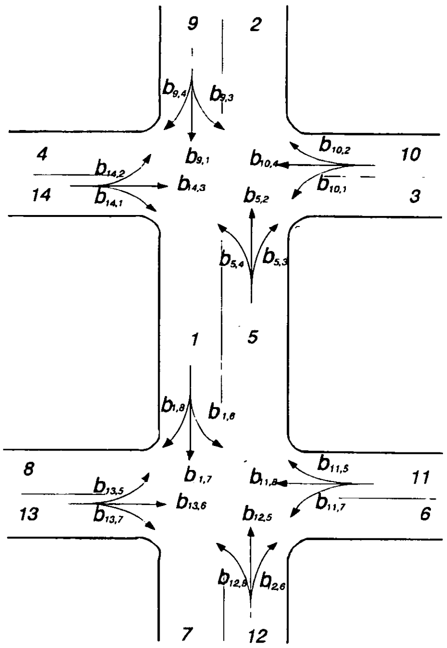


FIGURE 1 Simple signalized network.

will avoid two common causes of nonidentifiability. If a network is strongly connected and this connectivity is relatively invulnerable to disruption, then a complete cordon count will give an input- and output-reachable and separable configuration. Finally, internal compartments that can be entered from only one other internal compartment are likely to cause separability problems unless additional detectors are used.

MONTE CARLO EXPERIMENT

A system that is identifiable in the preceding sense is one for which the data collection configuration will not, by itself, prevent estimation of the turning movement parameters. However, the quality of the resulting estimates will depend at least in part on factors such as quality and quantity of the available data, the algorithm used to solve the NLS problem, and the choice of NLS as opposed to some other estimation approach, such as method of moments or ML. A comprehensive answer to the questions raised here is not available, but to illustrate these issues, consider the simple network depicted in Figure 1, showing two intersections of two-way streets. The various compartments are numbered from 1 to 14, and the figure also shows the 24 separate turning movement proportions, indexed according to their exit and entry compartments. Since for any given approach the proportions for left turns, right turns, and through movements must add up to 1.0, there are in fact only 16 linearly independent turning movement parameters in this network, and the vector **b** containing these independent parameters will have dimension $d = 16$. Figure 2 shows two different configurations of detector placements for this network. Placement Scenario 1 corresponds to the complete detectorization assumed by the linear model for estimating turning movement proportions, and Scenario 2 corresponds to a cordon count placement. It is straightforward to verify that under Scenario 2, the detector configuration is both input- and output-reachable and separable.

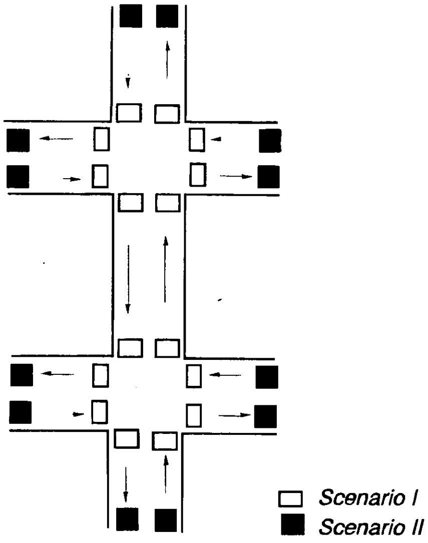


FIGURE 2 Configurations of detector placements.

The primary objective of this paper was to generate a sample of turning proportion estimates computed by minimizing the nonlinear sum of squares function S_2 and then to compare it with a sample of estimates generated by minimizing the linear least-squares function S_1. To this end, simulated traffic counts for both the Scenario 1 and the Scenario 2 detectors were generated using a stochastic version of the prediction model described by Equations 1 and 2. Simulated input counts at each of the six input points for time interval t were generated as Poisson outcomes with time-varying means $\bar{q}_i(t)$, and the number of vehicles exiting compartment k during interval t was generated as binomial random variable with parameters $x_k(t)$, $p_k[\mathbf{x}(t)]$. The exiting vehicles were then allocated to adjacent compartments as multinomial random outcomes with classification probabilities b_{kl}. The exit probability functions were of the same form as those presented and tested elsewhere (*23,24*) to describe freeway traffic flow, but with free-flow speeds, capacities, and jam densities selected to make them more representative of arterial travel. The traffic signal at each intersection was given a standard two-phase timing plan, with a 60-sec cycle length and 30 sec of green allocated to each phase (i.e., no yellow intervals were used). The effect of red time on a movement was simulated by setting the exit probability to 0.0 during the red interval. Fifty simulated data sets were generated, each consisting of 180 1-min traffic counts for each of the detectors depicted in Figure 2. Under Scenario 1, it was assumed that data from the white detectors were available, and estimates of the turning movement proportions were computed using the equality-constrained least-squares algorithm (*28*). Under Scenario 2, it was assumed that data from the black detectors were available, and predicted values for the cordon output detectors were computed recursively using the prediction model described in Equations 1 and 2, with the 1-min input detector counts as inputs. This recursion was implemented as a subroutine called by the NAG optimization routine E04JBF (*29*), which computed those estimates of the turning movement proportions that minimized the nonlinear sum-of-squares function S_2. For the nonlinear estimation, only the left and right turning proportions at each approach were treated as independent parameters, with the through proportion then being computed as $b_{\text{through}} = 1 - b_{\text{left}} - b_{\text{right}}$.

As noted earlier, even when a detector configuration supports identification of the turning movement proportions, the statistical properties of these estimates remain to be assessed. The least-squares estimates generated by an identifiable configuration may still show enough bias or variability to limit their practical usefulness. Estimated turning proportions were computed for each of the 50 simulated data sets, giving a pseudorandom sample of the estimates under each scenario. Table 1 presents the results of this experiment.

The mean columns in Table 1 give the average, across the 50 data sets, of the estimates for that parameter, whereas the "std" column gives the standard deviation of the estimates. The "*t*" columns give the *t*-statistic testing the hypothesis that the sample average for that parameter is equal to its true value (*i.e.*, a test for whether that estimate is biased). For each approach, the "true" parameter values used in generating the simulated data were $b_{\text{through}} = 0.6$, $b_{\text{left turn}}=0.3$, and $b_{\text{right turn}} = 0.1$. For the NLS estimates, the *t*-statistics for the through movements are omitted since they are actually deterministic functions of the estimates for the right and left turn proportions. The results for the CLS estimates are consistent with those reported by Nihan and Davis (*5*), being unbiased with moderately low standard deviations. As would be expected, the NLS estimates show an increase in variability, because the NLS estimator is working with less information than the CLS estimator. The first set of NLS estimates also shows a substantial number of instances of bias, but this appears to be due in large part to numerical difficulties experience by E04JBF. In 21 of 50 instances, E04JBF terminated with a message indicating that it was unable to satisfy all convergence criteria; in the remaining 29 cases, satisfactory convergence was achieved. Computed means, standard deviations, and *t*-statistics for only those cases showing satisfactory convergence are displayed in the three rightmost columns of Table 1, and these show removal of a number of instances of bias. This result suggests that careful attention to the numerical properties of one's optimization algorithm may result in improved estimator performance.

From a practical standpoint, probably the most interesting result is the increased variability shown by NLS estimates when compared with CLS estimates. To interpret this, the results in Table 1 suggest that with 180 1-min traffic counts, roughly 95 percent of the time one could expect to have an estimate of $b_{5,4}$ that would fall in the interval [0.25, 0.35], whereas with NLS one would need the interval [0.18, 0.42] for the same degree of confidence. A similar result is shown for each of the turning movement proportions. Thus, shifting to fewer detectors does not guarantee something for nothing. The cost savings can be offset by a loss of precision.

CONCLUSION

The first objective of this paper was to assess the feasibility of estimating intersection turning movement proportions from automatic traffic counts, when the number or placement of the detectors cannot provide complete counts for each intersection. It was determined that such estimation was possible for detector configurations

TABLE 1 Results of Monte Carlo Experiments

Para-	CLS			NLS (1)			NLS (2)		
meters	mean	std	t	mean	std	t	mean	std	t
$b_{5,4}$	0.3015	0.0274	0.38	0.2978	0.0584	0.27	0.3011	0.0613	0.10
$b_{5,2}$	0.5926	0.0312	1.67	0.5769	0.0535	-	0.5854	0.0526	-
$b_{5,3}$	0.1059	0.0297	1.40	0.1254	0.0426	4.21*	0.1136	0.0402	1.82
$b_{9,3}$	0.2994	0.0190	0.21	0.2913	0.0259	2.38*	0.2981	0.0264	0.40
$b_{9,1}$	0.5953	0.0308	1.08	0.6045	0.0495	-	0.5985	0.0398	-
$b_{9,4}$	0.1052	0.0232	1.60	0.1041	0.0401	0.73	0.1035	0.0306	0.61
$b_{10,1}$	0.3096	0.0453	1.50	0.3080	0.1120	0.50	0.3139	0.1029	0.73
$b_{10,4}$	0.5888	0.0404	1.97	0.5825	0.1095	-	0.5832	0.1094	-
$b_{10,2}$	0.1016	0.0338	0.34	0.1095	0.0710	0.94	0.1029	0.0657	0.23
$b_{14,2}$	0.3067	0.0331	1.44	0.2819	0.0577	2.21*	0.2812	0.0583	1.74
$b_{14,3}$	0.5970	0.0296	0.72	0.5883	0.0525	-	0.5908	0.0564	-
$b_{14,1}$	0.0963	0.0323	0.81	0.1298	0.0612	3.44*	0.1280	0.0681	2.22*
$b_{12,8}$	0.3032	0.0286	0.80	0.3106	0.0415	1.80	0.3160	0.0429	2.01
$b_{12,5}$	0.5933	0.0264	1.80	0.5751	0.0544	-	0.5711	0.0611	-
$b_{12,6}$	0.1035	0.0287	0.85	0.1144	0.0442	2.30*	0.1129	0.0498	1.39
$b_{1,6}$	0.3010	0.0343	0.21	0.2838	0.0416	2.76*	0.2842	0.0432	1.97
$b_{1,7}$	0.5973	0.0424	0.45	0.5859	0.0477	-	0.5852	0.0515	-
$b_{1,8}$	0.1017	0.0315	0.38	0.1303	0.0422	5.08*	0.1306	0.0501	3.29*
$b_{11,7}$	0.3020	0.0214	0.65	0.2993	0.0280	0.18	0.3011	0.0292	0.21
$b_{11,8}$	0.5975	0.0243	0.72	0.5468	0.0490	-	0.5494	0.0362	-
$b_{11,5}$	0.1005	0.0191	0.19	0.1474	0.0377	8.90*	0.1495	0.0399	6.68*
$b_{13,5}$	0.3038	0.0295	0.91	0.3162	0.0643	1.78	0.3131	0.0518	1.36
$b_{13,6}$	0.6002	0.0384	0.04	0.6036	0.0549	-	0.6046	0.0389	-
$b_{13,7}$	0.0960	0.0414	0.69	0.0802	0.0553	2.54*	0.0823	0.0536	1.78

providing a requisite minimum amount of information. The authors described a numerical test of whether a given pattern of detector placements could provide this information and recommended a minimal placement pattern that is likely (but not guaranteed) to produce adequate information. Overall, it appears plausible that there is more information about turning movement proportions in limited detector configurations than is being used.

The second objective was to obtain some idea of the effects on the statistical properties of turning movement estimates that result from a reduced detector configuration. A Monte Carlo study using a simple two-intersection network showed a noticeable increase in a tendency toward bias and in estimate variability when one shifted from a complete set of counts to cordon counts. This suggests that minimal identifiable detector configurations might not provide the precision needed for real-time tracking of turning movement proportions. If full detectorization is not possible, one could begin with a minimal configuration, such as cordon counters, and add as many detectors as is economically possible. One compromise might be to divide a large network into a number of smaller cordoned areas, allowing some detectors to do double duty on the boundary between two areas. Doing so would also facilitate direct verification of input and output reachability and separability.

For practitioners, the fact that a residual amount of uncertainty remains in the estimates of the turning movement proportions, even after processing 3 hr of data, should cause them to question the standard practice of "certainty equivalent" control, in which estimated quantities are used as if they were known with certainty. For a given identifiable detector configuration, some of this uncertainty might be eliminated by switching to a more efficient estimation approach, such as ML; the feasibility of such a switch is currently under investigation. It does not appear likely, however, that all uncertainty can be eliminated, and genuinely optimal control of traffic signal systems may need to take uncertainty into explicit account.

ACKNOWLEDGMENT

This research was supported by the ITS Institute at the Center for Transportation Studies, University of Minnesota.

REFERENCES

1. *Special Report 209: Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 1985.
2. Cremer, M., and H. Keller. Dynamic Identification of Flows from Traffic Counts at Complex Intersections. *Proc., 8th International Symposium on Transportation and Traffic Theory* (V.F. Hurdle, ed.), University of Toronto Press, Ontario, Canada, 1983, pp. 121–142.
3. Cremer, M., and H. Keller. A New Class of Dynamic Methods for Identification of Origin-Destination Flows. *Transportation Research,* Vol. 21B, 1987, pp. 117–132.
4. Nihan, N., and G. Davis. Recursive Estimation of Origin-Destination Matrices from Input/Output Counts. *Transportation Research,* Vol. 21B, 1987, pp. 149–163.
5. Nihan, N., and G. Davis. Application of Prediction-Error Minimization and Maximum Likelihood to Estimate Intersection O-D Matrices from Traffic Counts. *Transportation Science,* Vol. 23, 1989, pp. 77–90.
6. Bell, M. The Real Time Estimation of Origin-Destination Flows in the Presence of Platoon Dispersion. *Transportation Research,* Vol. 25B, 1991, pp. 115–125.
7. Nguyen, S. Estimating Origin-Destination Matrices from Observed Flows. In *Transportation Planning Models* (M. Florian, ed.), Elsevier Science, Amsterdam, the Netherlands, 1984, pp. 363–380.
8. Cascetta, E., and S. Nguyen. A Unified Framework for Estimating or Updating Origin/Destination Matrices from Traffic Counts. *Transportation Research,* Vol. 22B, 1988, pp. 437–455.
9. Davis, G. *A Dynamic, Stochastic Model of Traffic Assignment and its Application to the Maximum Likelihood Estimation of Origin-Destination Parameters.* Ph.D. dissertation. Department of Civil Engineering, University of Washington, 1989.
10. Davis, G. A Statistical Theory for Estimation of Origin-Destination Parameters from Time-Series of Traffic Counts. In *Transportation and Traffic Theory* (C. Daganzo, ed.), Elsevier, Amsterdam, The Netherlands, 1993, pp. 441–464.
11. Van Zuylen, H., and L. Willumsen. The Most Likely Trip Matrix Estimated from Traffic Counts. *Transportation Research,* Vol. 14B, 1980, pp. 281–293.
12. Bell, M. The Estimation of an Origin-Destination Matrix from Traffic Counts. *Transportation Science,* Vol. 17, 1983, pp. 198–217.
13. Maher, M. Inferences on Trip Matrices from Observations on Link Volumes: A Bayesian Statistical Approach. *Transportation Research,* Vol. 17B, 1983, pp. 435–447.
14. Cascetta, E. Estimation of Trip Matrices from Traffic Counts and Survey Data: A Generalized Least-Squares Estimator. *Transportation Research,* Vol. 18B, 1984, pp. 289–299.
15. Speiss, H. A Maximum Likelihood Model for Estimating Origin-Destination Matrices. *Transportation Research,* Vol. 21B, 1987, pp. 395–412.
16. Davis, G., and N. Nihan. A Stochastic Process Approach to Estimating OD Parameters from Time-Series of Traffic Counts. In *Transportation Research Record 1328,* TRB, National Research Council, Washington, D.C., 1991, pp. 36–42.
17. Maher, M. Estimating the Turning Flows at a Junction: A Comparison of Three Models. *Traffic Engineering and Control,* Jan. 1984, pp. 19–22.
18. Ljung, L., and T. Soderstrom. *Theory and Practice of Recursive Identification.* MIT Press, Cambridge, Mass., 1983.
19. Caines, P. *Linear Stochastic Systems.* Wiley and Sons, New York, 1989.
20. Gallant, A., and H. White. *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models.* Basil Blackwell, Ltd., Oxford, England, 1988.
21. Chen, Y. Convergence Study of Two Real-Time Parameter Estimation Schemes for Nonlinear Systems. In *Nonlinear Stochastic Problems* (R. Bucy and M. Moura, eds.), D. Reidel Publishing, 1983.
22. Davis, G., and G. Kang. Filtering and Prediction of Freeways Using Markov Models. *Proc., 3rd International Conference on Applications of Advanced Technologies in Transportation Engineering,* (C. Hendrickson and K. Sinha, eds.), ASCE, New York, 1993, pp. 43–49.
23. Davis, G., and J. G. Kang. Estimating Destination-Specific Traffic Densities on Urban Freeways for Advanced Traffic Management. In *Transportation Research Record 1457,* TRB, National Research Council, Washington, D.C., 1994.
24. Jacquez, J., and P. Greif. Numerical Parameter Identifiability and Estimability: Integrating Identifiability, Estimability and Optimal Sampling Design. *Mathematical Biosciences,* Vol. 77, 1985, pp. 210–227.
25. Eisenfeld, J. Remarks on Bellman's Structural Identifiability. *Mathematical Biosciences,* Vol. 77, 1985, pp. 229–243.
26. Eisenfeld, J. A Simple Solution to the Compartmental Structural-Identifiability Problem. *Mathematical Biosciences,* Vol. 79, 1986, pp. 209–220.
27. Roberts, F. *Discrete Mathematical Models.* Prentice-Hall, Englewood Cliffs, N.J., 1976.
28. Lawson, G., and R. Hanson. *Solving Least Squares Problems.* Prentice-Hall, Englewood Cliffs, N.J., 1974.
29. *NAG Workstation Library, Version 1.* Numerical Algorithms Group, Oxford, England, 1986.

All facts, conclusions, and opinions expressed here are strictly the responsibility of the authors.

Arterial Incident Detection Integrating Data from Multiple Sources

NIKHIL BHANDARI, FRANK S. KOPPELMAN, JOSEPH L. SCHOFER, VANEET SETHI, AND JOHN N. IVAN

An integrated incident detection system for an arterial street network being implemented for the *ADVANCE* project, an advanced traveler information system demonstration in the northwest suburbs of Chicago, Illinois, is described. Incidents will be detected using three distinct data sources: loop detectors, probe vehicles, and anecdotal sources. Specialized incident detection algorithms will process each of these data types separately. The outputs from the fixed detector, probe vehicle, and anecdotal source algorithms will be integrated by a data fusion process to determine the overall likelihood that an incident has occurred at any particular location. The incident detection system will also estimate the expected duration of the incidents and their effects on link travel times as a function of the type of incident.

Incidents are unexpected events that disrupt the flow of traffic on a segment of a roadway link and have significant effects on link travel times; examples are stalled vehicles, collisions, and materials spills. The effect of an incident is to reduce the capacity of the segment; if demand volume is high enough, this can result in queues, delays, and increased travel time on the link. Early detection of incidents can help traffic management agencies respond quickly, dispatch emergency vehicles to the incident site, and perhaps divert traffic to reduce delay. Detection of incidents also helps agencies warn the oncoming traffic and thereby reduce the danger of secondary incidents (*1*).

Recently, there has been much interest in increasing the efficiency of existing roadways by developing intelligent transportation systems (ITS) and particularly advanced traveler information systems (ATIS). One aim of these systems is to provide road users with real-time information on travel times and roadway status to reduce their individual travel times. An important component of these systems will be the ability to detect traffic flow disruptions on the road network and alert and divert potential users of the affected links.

The ITS field tests in the United States and Europe [e.g., *ADVANCE* (*2*), Pathfinder (*3*), TravTek (*4*), ALI-SCOUT (*5*), EURO-SCOUT (*6*)] used or will use multiple data sources such as traffic sensors, probe vehicles, video cameras, and anecdotal sources to collect real-time traffic information. The largest of these demonstrations is the *ADVANCE* project in suburban Chicago. *ADVANCE* will provide approximately 3,500 participants with real-time route planning information based on up-to-date travel times and incident information in the test area. *ADVANCE* drivers will be local residents, and information about recurrent congestion or navigational guidance will be of limited value as they will have considerable experience with the network. Incidents—unexpected nonrecurrent events on the network—can have significant effects on link travel times. Real-time information about an occurrence and its impact on travel time will be valuable even to travelers who are familiar with the network structure under normal conditions. Evidence from recruitment studies conducted for *ADVANCE* suggests that incident detection will be important for attracting participants to the project and sustaining their interest over the period of the demonstration (*7*). In general, knowledge about incident locations and their travel time effects will enable the *ADVANCE* project to give drivers more accurate estimates of link travel times for route planning and also to provide information on the reasons for increases in travel time.

ADVANCE will integrate information from three distinct data sources to detect incidents:

- *Fixed detectors,* which provide occupancy and volume data averaged over a fixed time interval (e,.g., 5 min) for a specific section of selected network links;
- *Probe vehicles* participating in the demonstration project, which travel freely on the network and automatically report link travel times by radio; and
- *Anecdotal sources,* reports of particular events affecting traffic flow provided by people traveling on or monitoring the road network, including emergency services workers.

On the *ADVANCE* network, made up primarily of suburban arterial streets, fixed detectors providing volume and occupancy data are located approximately 350 ft upstream from selected signalized intersections on major arterials. A fraction of these will be connected by telephone line to the *ADVANCE* traffic information center (TIC) to support incident detection. Probe vehicles and anecdotal sources will provide data intermittently at locations determined by the location of the probe-equipped vehicle and the reporting source, respectively. Specifically, probe vehicles, driven by drivers volunteering to participate in the operational test for up to 2 years, will automatically report travel times by radio to the TIC each time they complete link traversals. During any time interval, there may be no probe report for many links in the network because of the relatively small number of probe vehicles. Similarly, anecdotal data will be available only when emergency personnel or motorists report a traffic incident on the link. These reports will be collected from a centralized emergency services dispatch center responsible for police, fire, and ambulance services in the test area; from the cellular telephone emergency reporting center; and from the state department of transportation's emergency patrol fleet.

To obtain the best determination of incident versus nonincident conditions under different conditions of data availability, the

N. Bhandari, F.S. Koppelman, J.L. Schofer, and V. Sethi, Transportation Center, Northwestern University, Evanston, Ill. 60208. J. N. Ivan, Transportation Institute, Department of Civil Engineering, University of Connecticut, Storrs, Conn. 06269.

authors adopt a hierarchical structure for the incident detection system in which data are first processed by specialized algorithms for each data source (fixed detector, probe vehicle, and anecdotal algorithms), then data fusion processes integrate all the available data to determine the overall likelihood that an incident has occurred at any particular location. This approach provides the flexibility to use the identification (ID) system when information on *ADVANCE* links is available from only one data source by employing the relevant algorithm in isolation; for a majority of incidents data from only one (or two) sources will be available because only a small fraction of links in the test area have fixed detectors and the small probe vehicle fleet can provide information from only a limited number of links during any short time interval. When more than one data source is available, this system will extract the most useful information from all available sources to make a more accurate determination of incident presence rather than select a single best source for each incident detection (*3*).

This paper presents a fully integrated incident detection system that is being implemented for the *ADVANCE* project and describes the individual components of the ID system, the input and output requirements of each component, and the relationships among the components. The development, refinement, and calibration of the individual algorithms used by the system are documented elsewhere (*8–12*).

OVERVIEW OF *ADVANCE* INCIDENT DETECTION SYSTEM

Figure 1 shows the relationships among the different components of the *ADVANCE* incident detection system; the components are as follows:

- *Fixed detector algorithm* uses the real-time and historical data provided by fixed detectors located on major arterials to classify conditions on the detectorized streets as incidents or nonincidents.
- *Probe vehicle algorithm* uses travel time reports by probe vehicle and historical travel times on these links to interpret traffic conditions as incident or nonincident.
- *Anecdotal algorithm* uses information provided by emergency personnel and other motorists on the network to detect incidents in real time.
- *Data fusion algorithms* combine the output from the fixed detector, probe vehicle, anecdotal algorithms, and the operator interface.

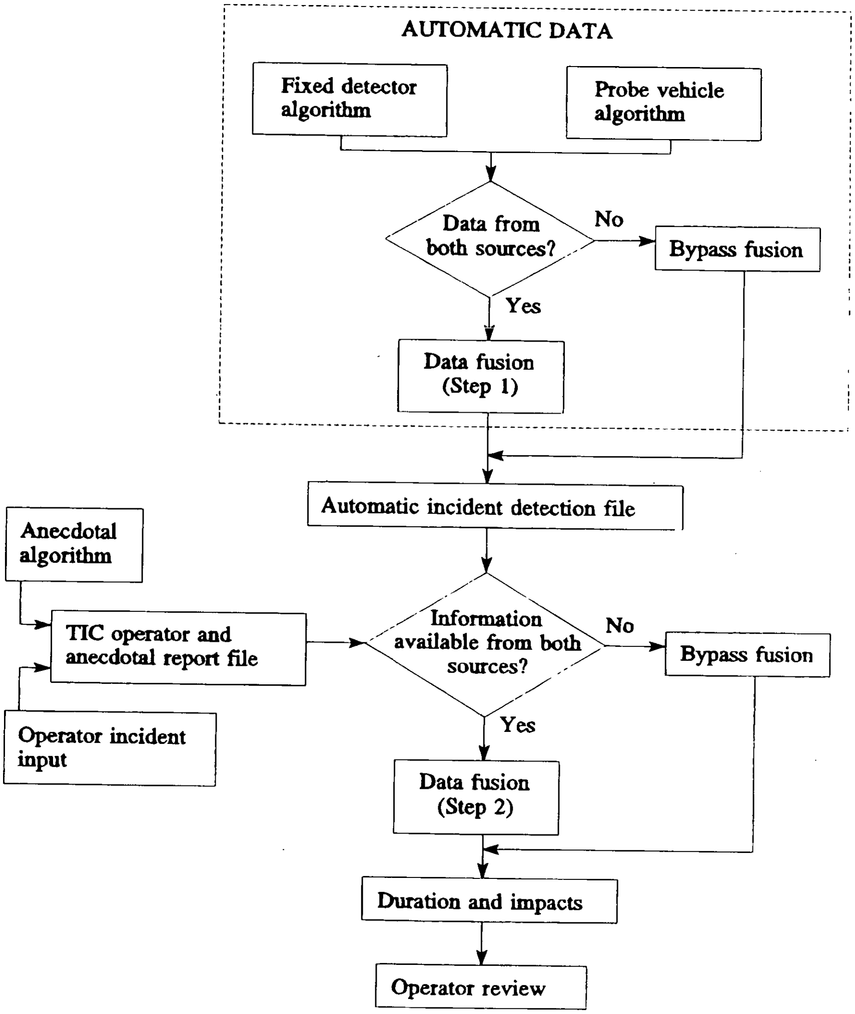


FIGURE 1 *ADVANCE* **incident detection system.**

- *Duration and impacts module* determines the expected duration of the incident and the impacts on the incident link travel times.
- *Operator interface* allows the TIC personnel to view the output from the data algorithms and to key in incident reports from other sources. The output from the duration and impacts module will be available to the operator for review.

The automatic data algorithms (the fixed detector and probe vehicle algorithms) classify conditions on links for which current data are available at the end of a prespecified period (*e.g.,* 5 min); "current data" refers to probe reports or detector output received by the TIC during the current period. If both probe vehicle and fixed detector data are available for a link, the output from both algorithms is combined in the data fusion (Step 1) module; the fusion process is bypassed if data are available from only one source, and the output from the corresponding algorithm is used alone. When all the links with current data have been processed, the classification results are saved in the automatic detected incident file containing an unique tag for each link and a flag indicating incident presence or absence.

A new algorithm for incident detection using fixed detector data was developed instead of adapting existing pattern recognition (*13–15*) or time series methods (*1,16*). The primary reasons for this were (*a*) the existing algorithms were developed primarily for freeway environments whereas the *ADVANCE* network consists mostly of arterial streets (it is important to recognize that traffic flow characteristics on arterials are very different from freeways because of the presence of traffic signals, parking, and such, which results in greater variability in traffic flow measures such as occupancy and travel time), and thus the freeway algorithms are not readily transferable to arterials; (*b*) they require loop detector data for short intervals of time (approximately 30 sec to 2 min), and it will not be possible to get these data for less than 5- or 15-min intervals for *ADVANCE;* and (*c*) many of the existing algorithms use data from adjacent pairs of detectors, which generally are not available on *ADVANCE* arterials in the test area. Further, no well-established methods are available for using probe vehicle and anecdotal data and data fusion processes for incident detection.

The fixed detector, probe vehicle, and data fusion algorithms were estimated using simulated data because no actual field data and corresponding incident confirmations were yet available. When it is deployed, the *ADVANCE* operational test will generate field data that will be used to recalibrate the algorithms. Each of these models was estimated using discriminant analysis (*17*), which produced a function of the traffic flow parameters, the value of which is used to identify incident or nonincident conditions. Discriminant analysis uses prespecified values of prior probabilities of incidents (priors) to control the classification output and develop more realistic models. Incident conditions will exist during only a small fraction of time periods on any given link; this low probability of incidents in the real world is taken into account by adopting incident priors of 0.0001 (*i.e.,* in the absence of any other information, any particular report has a probability of 0.0001 of being an incident report). This ensures that the number of false alarms (incident reports generated in nonincident conditions) will be small.

The anecdotal algorithm uses data from anecdotal sources (*e.g.,* computerized emergency dispatch systems) and produces a link-specific output that is expected to be more detailed than the automatic classification output. The operator will be able to key in incident reports from other sources using a menu-based interface. The link-specific anecdotal algorithm output and operator inputs will be stored in the TIC operator/anecdotal report file containing a link identification tag, incident indicator, incident type, and a vector of variables representing the incident intensity.

At the end of every period, the incident information from TIC operator/anecdotal report file and the automatic detected incident files will be matched by link; the data fusion (Step 2) will be performed for links that have classification information available from both the files. Fusion will be bypassed for links having incident information from only one of the files.

The duration and impacts module uses the output from the data fusion (Step 2) process to estimate the duration and impacts of the incidents based on the incident type and intensity. For links with missing values for the incident type and intensity information, default estimates of duration and impacts will be used. For many links and periods, no data will be available; for these links the default conditions will be assumed to be nonincident. Some incidents will last for more than one period; when such incidents are detected in consecutive periods, the incident duration will be updated in the duration and impacts module for each period that the incident is detected.

The TIC operator will be able to review the final output before it is passed to other *ADVANCE* processes. The operator will be given a limited time window for review to avoid a backlog of detected incidents not reported to participating drivers. Eventually, reports of incidents and their effects on travel time will be transmitted by radio to route planning computers in participating vehicles.

ALGORITHM COMPONENTS

Fixed Detector Algorithm

The fixed detector algorithm compares current and historical volume and occupancy data from fixed detectors at the end of every period (*8*); Figure 2 shows the flow diagram for the algorithm. The historical (nonincident) volume and occupancy data are aggregated over a fixed time interval for each detector location by day type (weekday, weekend, *etc.*) and time of day.

The algorithm uses current and corresponding historical volume and occupancy data to compute two variables:

$$\text{Occupancy deviation} = \text{occupancy}_{\text{observed}} - \text{occupancy}_{\text{historical}}$$

$$\text{Volume/occupancy deviation} = (\text{volume/occupancy})_{\text{observed}} - (\text{volume/occupancy})_{\text{historical}}$$

A discriminant score is then computed using Equation 1, which determines incident presence in the proximity of each detector:

$$\text{Discriminant score} = -14.880 + 0.0192 * \text{occupancy deviation} - 4.088 * \text{volume/occupancy deviation} \quad (1)$$

If the discriminant score is greater than 0, an incident is flagged for the link associated with the detector for the corresponding time period; if the discriminant score is less than 0, normal conditions are assumed. Incident classification, discriminant scores, and values of the deviation variablès are provided to the data fusion module for every link that is processed by the algorithm.

Both the fixed detector and probe vehicle algorithms identify incidents that occur either on the link from which the detector output or probe report is obtained or on the upstream portion of the adjacent downstream link. However, because of the typical traffic flow impacts of arterial street incidents, the incidents detected are

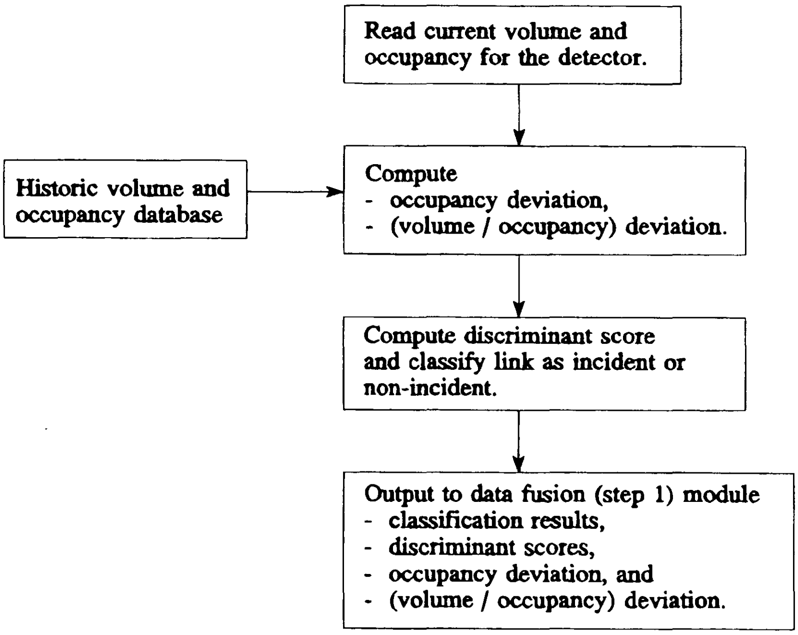

Read current volume and occupancy for the detector.
Historic volume and occupancy database
Compute
- occupancy deviation,
- (volume / occupancy) deviation.
Compute discriminant score and classify link as incident or non-incident.
Output to data fusion (step 1) module
- classification results,
- discriminant scores,
- occupancy deviation, and
- (volume / occupancy) deviation.

FIGURE 2 Fixed detector algorithm.

most likely to be in the downstream section of the report link; in a few cases incidents will be detected that are on the upstream or mid-block section of the link that is downstream of the reporting link. The current versions of both fixed detector and probe vehicle algorithms do not differentiate between these cases; they will attribute all such incidents to the reporting link. This will be the correct assignment in most cases, and it will tend to have the correct impact on drivers in other cases, diverting them from the incident links.

Probe Vehicle Algorithm

The probe vehicle algorithm requires current and historical (nonincident) probe travel time data; it operates in two stages as shown in Figures 3 and 4 (*9*). In the first stage (Figure 3), average link travel times are computed by aggregating individual probe reports at the end of the period. Aggregation of probe reports results in a more accurate representation of the traffic conditions by averaging out aberrant nonincident probe reports. The aggregation procedure will depend on the number of reports received in the current time period. If only one report is available in the current period, probe reports from the previous period are averaged with the current reports. If more than one report is available, the average travel time is computed using all the probe reports received during the current period. If no report is received in the current period, or if in two consecutive periods only one report is received from a link, that link is not processed because of the unreliability of individual reports.

The second stage of the probe algorithm is the application of the different models to classify conditions on the links as incident or normal; Figure 4 shows this procedure. Travel time ratio and speed ratio are computed using the observed travel time on the links and the corresponding historical travel times stored in a data base. Incident presence is determined by computing the discriminant score; the effective cut-off travel time ratios for declaring an incident is dependent on the number of reports received during the detection interval to recognize the increased reliability of the average link travel time with increasing number of probe reports (*9*). The cut-off points are given in Table 1. The output from the probe vehicle algorithm consists of the classification results, discriminant scores, travel time ratios, and speed ratios for every link processed by the algorithm.

Data Fusion—Step 1

The data fusion algorithm reviews each link in the network once in every period and determines incident presence or absence for each link for which reports from both probe vehicle and fixed detector algorithms are available (*11*). The classification result for those links will be stored in the automatic detected incident file. If a report is available from only one of the algorithms, it will be used directly as the output of the automatic data fusion. Once all the links are evaluated for that time period, the automatic detected incident file will be combined with the TIC operator/anecdotal report file.

Two approaches, discriminant analysis and artificial neural networks, were tested for this fusion task. Comparison of incident detection results with discriminant analysis and neural network data fusion models showed that the model estimated using discriminant analysis had better detection rates for less extreme priors, whereas the neural network model performed better for more extreme priors (*11*). It is useful to evaluate both of these models with real data before selecting one over the other. For the current version of the *ADVANCE* incident detection system, the discriminant analysis model will be adopted because of its simpler structure. However, all the data required by the neural network model will be saved for offline comparative testing of the two models.

The best data fusion model using the discriminant analysis approach uses occupancy deviation and volume/occupancy deviation from fixed detector data, and travel time ratio and speed ratio from probe vehicle data (Figure 5) for computation of the discriminant scores as follows:

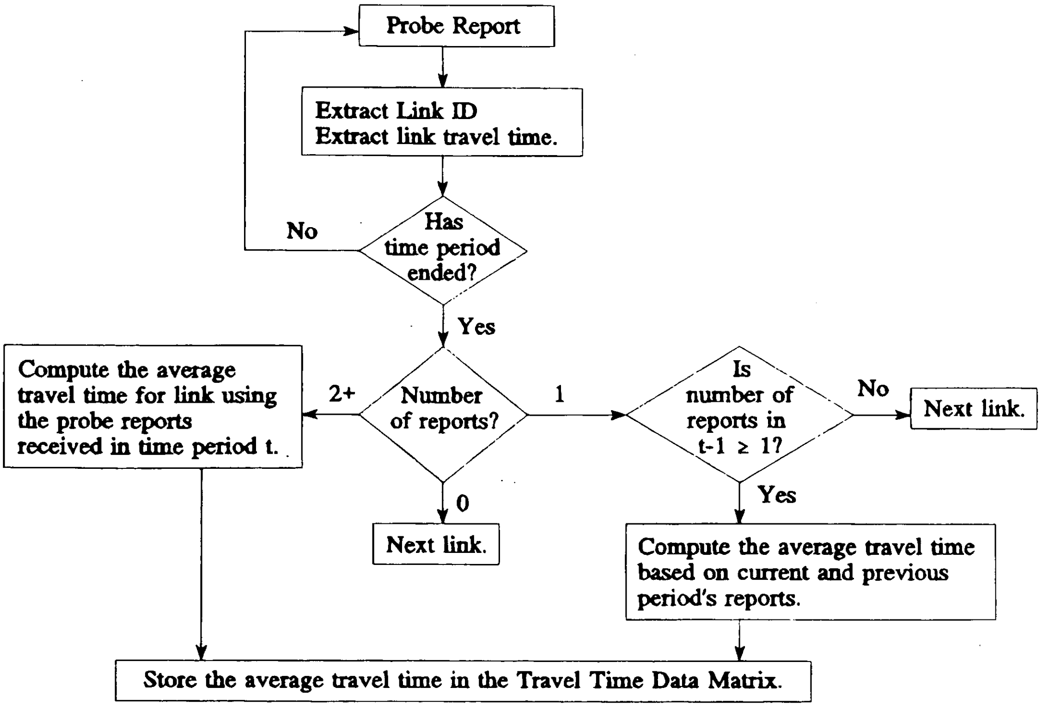

Probe Report
Extract Link ID
Extract link travel time.
Has time period ended?
No
Yes
Number of reports?
2+
1
0
Compute the average travel time for link using the probe reports received in time period t.
Next link.
Is number of reports in t-1 ≥ 1?
No
Next link.
Yes
Compute the average travel time based on current and previous period's reports.
Store the average travel time in the Travel Time Data Matrix.

FIGURE 3 Computation of average travel time using probe reports.

$$\text{Discriminant score} = 3.005 - 0.255 * \text{occupancy deviation} - 4.523 * (\text{volume/occupancy}) \text{ deviation} - 24.573 * \text{speed ratio} + 1.834 * \text{travel time ratio} \quad (2)$$

An incident is flagged if the discriminant score is greater than 0. The performance of this model was substantially better than either the fixed detector or probe vehicle algorithm when applied to the same data set (*11*), showing that overall detection ability can be improved by using detector and probe vehicle data together (when available).

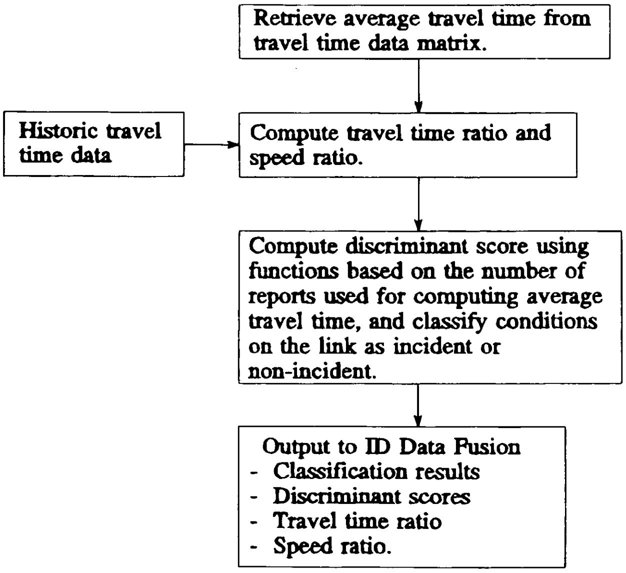

Retrieve average travel time from travel time data matrix.
Historic travel time data
Compute travel time ratio and speed ratio.
Compute discriminant score using functions based on the number of reports used for computing average travel time, and classify conditions on the link as incident or non-incident.
Output to ID Data Fusion
- Classification results
- Discriminant scores
- Travel time ratio
- Speed ratio.

FIGURE 4 Application of probe vehicle algorithm.

Anecdotal Information Algorithm

The anecdotal incident detection algorithm uses a qualitative description of incidents reported by field observers, both trained and untrained, to detect incidents in real time (*10*). The two primary sources of anecdotal data will be

1. The Northwest Central Dispatch System (NWCD), a computer-aided emergency service dispatch agency serving six communities in the center of the *ADVANCE* test area; and
2. The *999 center, which receives toll-free calls from cellular telephone users voluntarily reporting roadway incidents and other problems.

Reports from the *999 center, operated for the Illinois State Toll Highway Authority, originate primarily from lay citizens; they will be in the form of qualitative descriptions of events and nominal location references. A simple manual data connection to the *999 center is planned for late in the *ADVANCE* operational test. The rest of this section focuses on the use of NWCD anecdotal inputs and inputs from other sources through the TIC operator.

TABLE 1 Cut-Off Point for Declaring Incidents Used by Probe Vehicle Algorithm

Number of Reports	Travel Time Ratio
2	3.45
3, 4	2.80
5, 6, 7	2.60
8, ..., 15	2.40
15, 16, ...	1.45

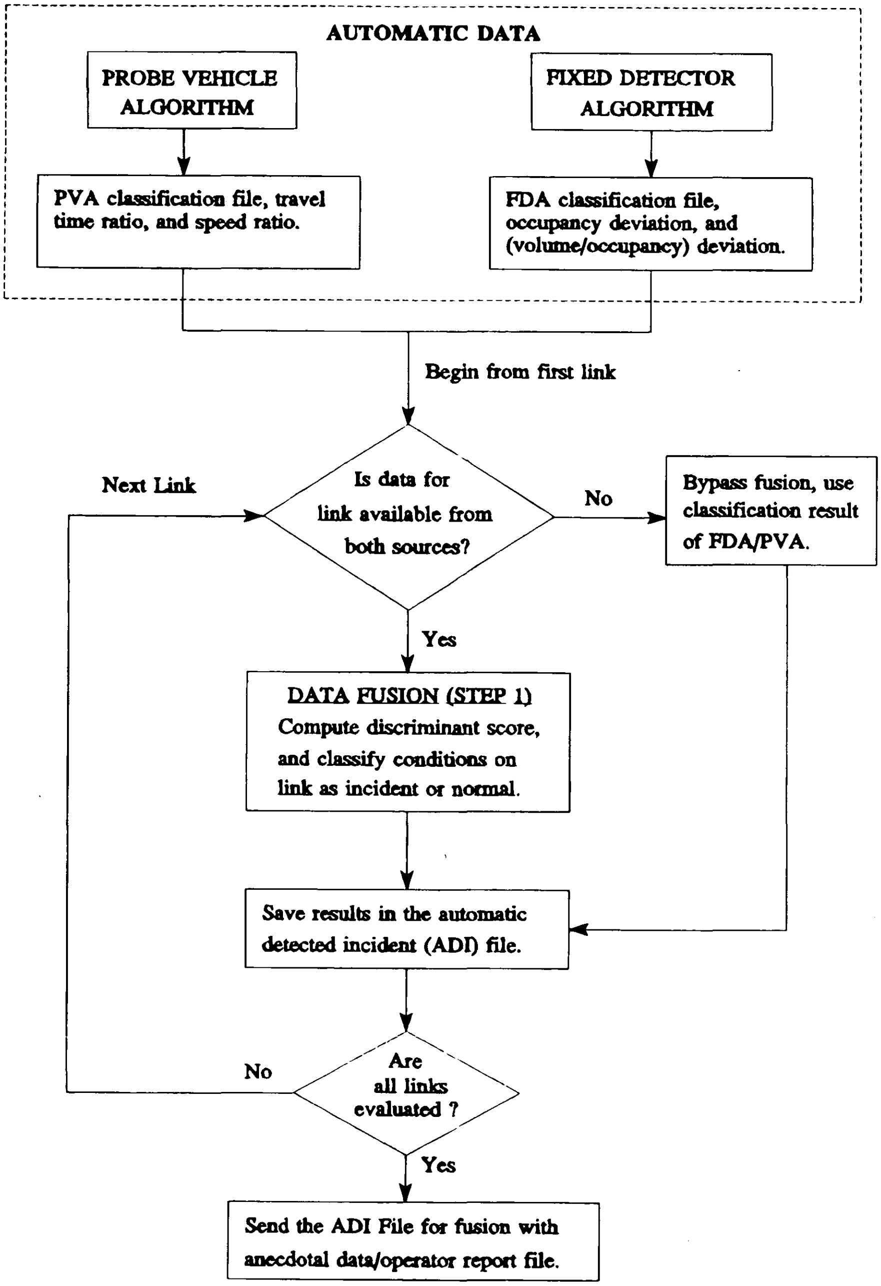

AUTOMATIC DATA
PROBE VEHICLE ALGORITHM
FIXED DETECTOR ALGORITHM
PVA classification file, travel time ratio, and speed ratio.
FDA classification file, occupancy deviation, and (volume/occupancy) deviation.
Begin from first link
Next Link
Is data for link available from both sources?
No
Bypass fusion, use classification result of FDA/PVA.
Yes
DATA FUSION (STEP 1)
Compute discriminant score, and classify conditions on link as incident or normal.
Save results in the automatic detected incident (ADI) file.
No
Are all links evaluated ?
Yes
Send the ADI File for fusion with anecdotal data/operator report file.

FIGURE 5 Data fusion algorithm using discriminant analysis model.

Reports from NWCD will be captured from the computer system in the dispatch center, where all incoming calls, emergency services dispatches, and other communications are entered into a data base. The *ADVANCE* anecdotal algorithm will use descriptions of emergency vehicle dispatches to roadway incidents, explicit or implicit incident confirmation and clearance reports provided by on-scene emergency service personnel, incident type descriptions by standard codes (accident with property damage, accident with personal injuries, hazardous material spills, motorist assist, *etc.*), and (in some cases) incident intensity as reflected by number of service units on the scene or other qualitative descriptions. NWCD receives incident location information from callers in various forms, including street addresses, intersections, and landmark names; this information is geocoded to street addresses within the NWCD computer system.

Figure 6 shows the main components of the anecdotal ID algorithm. Initially, the only source of anecdotal data will be NWCD. Data from NWCD will be *preprocessed* at NWCD before transmission to the *ADVANCE* TIC to extract only roadway incidents, and only those descriptor variables of interest to *ADVANCE*. The NWCD preprocessor will

- Distinguish new incidents from update reports on incidents already identified,

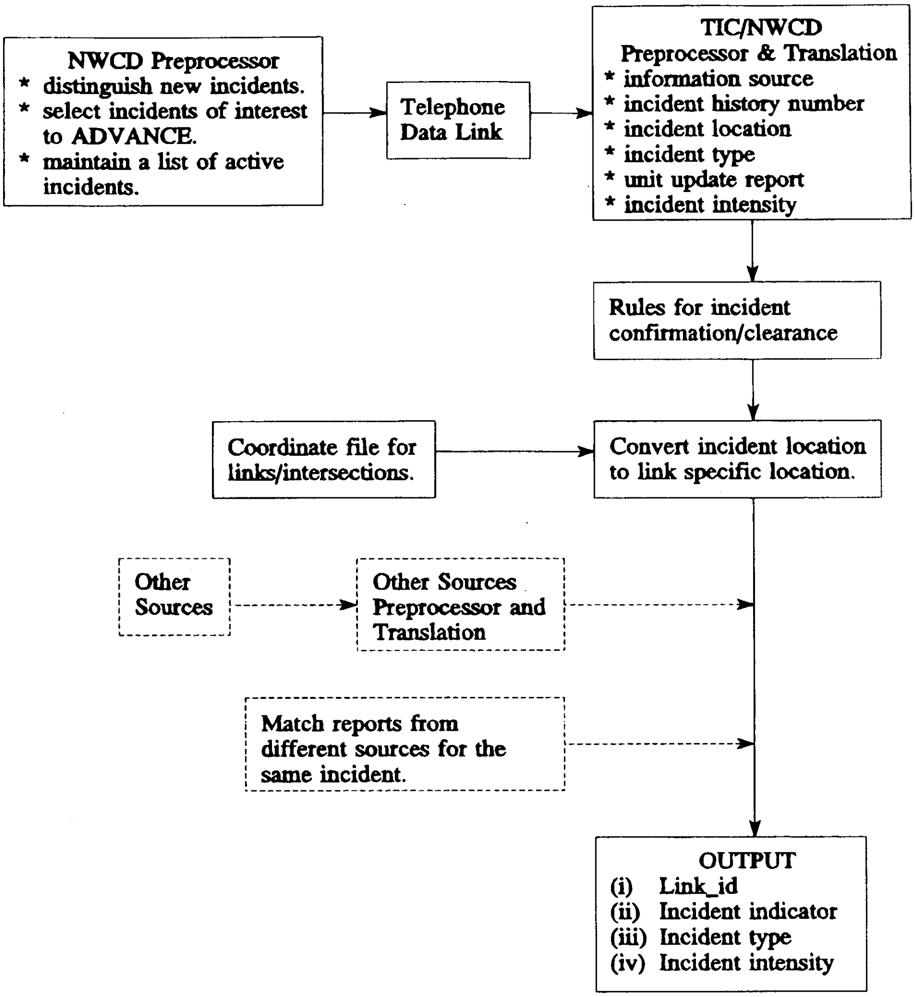

NWCD Preprocessor
* distinguish new incidents.
* select incidents of interest to ADVANCE.
* maintain a list of active incidents.
Telephone Data Link
TIC/NWCD Preprocessor & Translation
* information source
* incident history number
* incident location
* incident type
* unit update report
* incident intensity
Rules for incident confirmation/clearance
Coordinate file for links/intersections.
Convert incident location to link specific location.
Other Sources
Other Sources Preprocessor and Translation
Match reports from different sources for the same incident.
OUTPUT
(i) Link_id
(ii) Incident indicator
(iii) Incident type
(iv) Incident intensity

FIGURE 6 Different components of anecdotal incident detection algorithm.

- Separate incidents of interest to *ADVANCE* (*i.e.*, roadway blocking incidents) from others, and
- Maintain a list of active incidents; format messages to be sent by telephone to the TIC.

Data from other anecdotal sources will have a separate preprocessor and translation module that will be similar to the TIC/NWCD preprocessor and translation module. Further, a procedure for matching reports from different sources for the same incident will be required. These will be incorporated in the anecdotal ID system when data from other sources become available.

The anecdotal reports will be received by the TIC/NWCD preprocessor and translation module that assigns the information from the report into several fields (*i.e.*, information source, incident history number, incident location, incident type, unit update report, and incident intensity). The incident history number is defined by NWCD and used as a basis for maintaining unduplicated files of data for each active incident. Location will be in the form of street addresses. Incident type codes will be a short list of types of interest to *ADVANCE*. Mobile unit update reports will indicate emergency service unit radio call numbers and a status code (enroute, on-scene, clear). This information will be used to confirm incidents and determine clearance:

- Incidents will be confirmed 3 min after the first emergency responder arrives on the scene unless that unit reports clear,
- Incidents will be reported clear 5 min after the last emergency unit leaves the scene.

These criteria were established because responding units usually do not formally confirm incident presence or signal final clearance.

Next, each incident will be assigned to a link by converting the address to a segment ID using a file with coordinates of all links and intersections in the test area and finding the street name. Once the street name is found, the address will be used to identify the segment. The link-specific anecdotal reports (link identification, incident indicator, incident type, and incident intensity) on confirmed and cleared incidents will then be saved in the TIC operator and anecdotal report file.

Intensity data will be available for use in enhancing estimates of incident duration and traffic impacts; initial intensity data will be a simple count of emergency units on the scene. It may prove useful and feasible to parse free-form text reports from NWCD, and particularly from *999 (which will be rich in qualitative data from untrained observers) for use in more complex impact estimation efforts. Development of these capabilities must await the availability of field data from the *ADVANCE* implementation.

The TIC operator will also be able to enter anecdotal reports of incidents from telephone calls and other sources into the TIC computer system for use in the incident detection system. Inputs will match the data types used by the anecdotal algorithm: information source, location by address or intersection, incident type, and status (confirmed or clear). When the operator receives a report, he or she will open an incident report window on the TIC terminal. This window will display an input form, allowing data items to be keyed in or selected from short menus with a pointing device.

The operator inputs would bypass the anecdotal algorithm and go directly to the TIC operator/anecdotal report file. In this way, the operator will be able to clear an incident already detected or define a new incident not previously recognized by the ID system. The operator interface will also allow the operator to display a list of all active incidents by location, type, and start time.

Data Fusion—Step 2

The data fusion (Step 2) process combines the output from the automatic data fusion algorithm and TIC operator/anecdotal algorithm; Figure 7 shows the process. This data fusion is accomplished using a rule-based approach. For the current version of the *ADVANCE* ID system, the output from the anecdotal algorithm and TIC operator will override the output from the automatic algorithms (*i.e.,* fixed detector and probe vehicle algorithms). Initial anecdotal algorithm output will come only from NWCD, a highly reliable source originating with emergency services professionals. When other anecdotal sources are brought on-line, a fusion procedure that blends algorithm inputs along the lines of Step 1 fusion approach described earlier will be developed. The output from the data fusion (Step 2) process will be saved in the final classification file that will be passed on to the duration and impacts module.

Duration and Impacts

The *impact* of an incident is its effect on the travel time. The *duration* of an incident is the length of time during which that travel time impact occurs. The *clearance time* of an incident is the time from detection until the blocking event is removed from the roadway. The duration may last beyond the clearance time if a major queue must be dissipated. In other cases, the vehicle(s) involved in the incident can be moved off the road and the traffic conditions returned to normal, but the anecdotal algorithm will not declare the incident as cleared until the last emergency unit has left the site; in such cases the duration is less than the clearance time. The duration is more important to *ADVANCE* than the clearance time, but only the latter is typically available from anecdotal sources and may be used to estimate the former. Since current ID algorithms cannot distinguish between the two, the authors adopt a conservative perspective and set the incident duration to be 10 percent more than the average clearance time.

The duration and impacts module will receive the final classification file from the data fusion (Step 2) process. The incident type and intensity variables (number of police and fire units on scene), when available, will be used to compute the expected duration and impact of the incident. Figure 8 shows the procedure for determining incident duration and impact for different categories of incidents (*12*); some related incident types have been grouped together because they have similar characteristics. These incident types are a subset of incident categories used by NWCD; when data for other types of incidents become available, they will be incorporated in the algorithm.

The duration of incidents can be expressed in terms of minutes or algorithm cycle (5-min periods). Figure 8 reports duration in minutes; this can be converted to algorithm cycle units by dividing duration by cycle length (*i.e.,* 5 min) and increasing fractional time period values to the next highest integer. Since the algorithms oper-

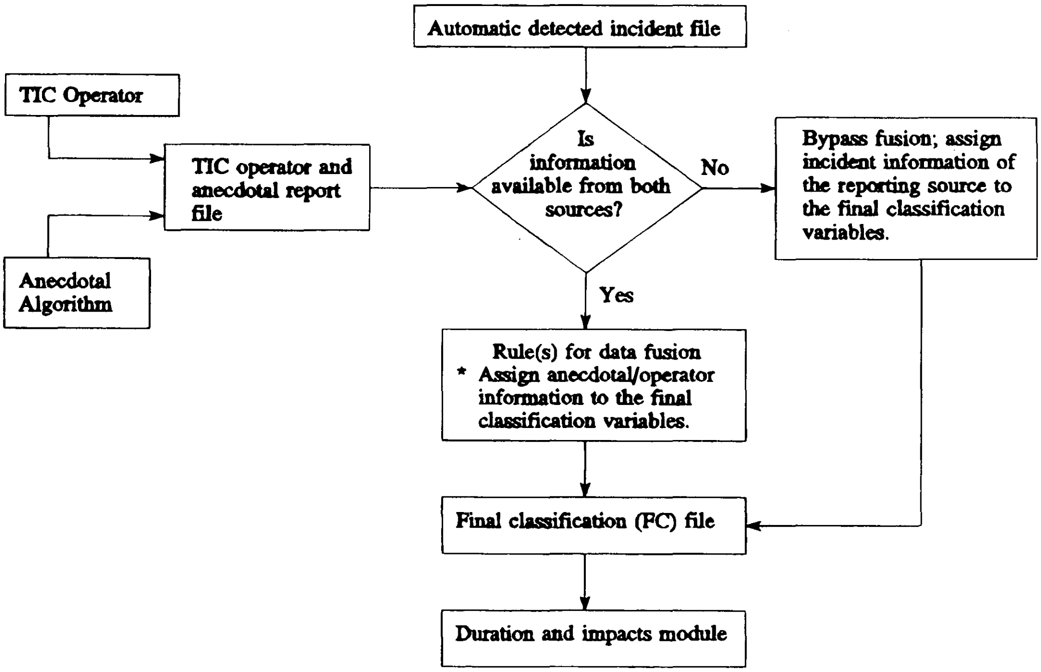
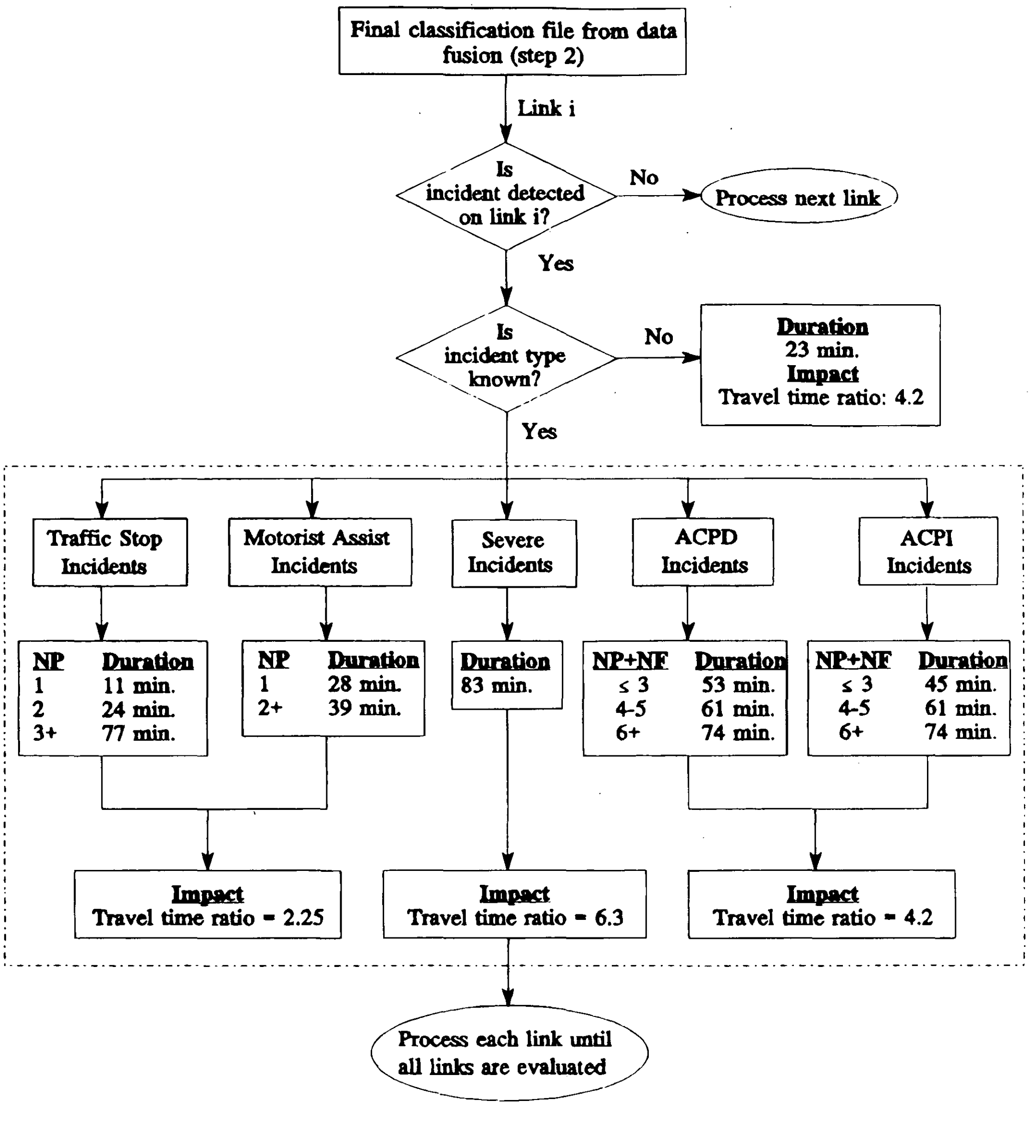

Automatic detected incident file
TIC Operator
TIC operator and anecdotal report file
Anecdotal Algorithm
Is information available from both sources?
No
Bypass fusion; assign incident information of the reporting source to the final classification variables.
Yes
Rule(s) for data fusion
* Assign anecdotal/operator information to the final classification variables.
Final classification (FC) file
Duration and impacts module

FIGURE 7 Data fusion (Step 2).



Notes:

- NP (NF) denotes the number of police (fire) units on scene.
- ACPD denotes accident with property damage, and ACPI denotes accident with personal injury.
- Severe incidents include accidents with entrapments, accidents involving hazardous materials, and fire related incidents.

FIGURE 8 Procedure for determining incident duration and impacts.

ate at the end of each period, it will be simpler to use periods as the unit of duration, especially for the duration updating process. Finally, when a specific clearance message is received from NWCD or other valid sources, the incident will be terminated.

Initially the incident impact, formulated as a travel time ratio, will be based on default values derived from the simulation studies (*12*). These values are based on the distribution of travel time ratios for identified incidents. The median of the travel time distribution is used for motorist assist and traffic stops, and the 75th percentile of the travel time distribution, for all other incidents except severe incidents and in cases when the incident type is unknown. For severe incidents (such as accidents with entrapments, accidents involving hazardous materials, and fire-related incidents), 1.5 times the values used for all other incidents (excluding motorist assist and traffic stops) are used. Default values of duration will be used for incident links for which the incident type is not known, and the links with no data from any of the sources will be assumed to be nonincident.

For incidents that last longer than one period, the duration will be updated every period by the time elapsed since the incident was first detected. If the incident lasts longer than the expected duration, then for every additional period that it is detected, the duration will be set to one more period until the incident is cleared or a nonincident message is generated by the algorithms that identified the incident. The procedural details for duration updating are provided elsewhere (*18*).

Operator Review

The operator review module will allow the operator to review the output of the ID system before it is passed to other *ADVANCE* processes. Through this option the operator can override the algorithm recommendations on the presence of incidents on the links. Initially, all the ID output will be confirmed manually by the operator. If the operator does not confirm the output or does not take any action, the links will be assigned nonincident status. In all cases (*i.e.,* when operator takes an action or does not respond to ID output), the ID output will be saved with the corresponding operator response; the operator responses will be evaluated to make this feature more effective.

CONCLUSIONS

This paper describes the incident detection system being implemented for the *ADVANCE* project. The ID system will use information from three distinct data sources: fixed detectors, probe vehicles, and anecdotal sources processed through specialized algorithms. The output from these algorithms will be integrated using a two-stage data fusion process to determine the overall likelihood that an incident has occurred at any particular location. On the basis of type of incident, the expected duration and travel time impacts of the incidents will be determined.

In contrast with other incident detection methods, this approach is designed to integrate information from multiple data sources. It is based on the concept that effective integration will result in an enhanced detection capability, making use of the special characteristics of each data source. Evaluation with field data during the *ADVANCE* operational test will provide an opportunity to verify this design concept.

ACKNOWLEDGMENTS

The information in this paper was partially funded by and developed for the *ADVANCE* project, a joint ITS undertaking by FHWA, the Illinois Department of Transportation, the University of Illinois at Chicago, and Northwestern University operating under the auspices of the Illinois Universities Transportation Research Consortium, Motorola, Inc., and other participating organizations.

REFERENCES

1. Thancanamootoo, S., and M. G. H. Bell. *Automatic Detection of Traffic Incidents on a Signal-Controlled Road Network.* Report 76. Transport Operations Research Group, University of Newcastle upon Tyne, England, June 1988.
2. Boyce, D. E., A. Kirson, and J. L. Schofer. Design and Implementation of ADVANCE: The Illinois Dynamic Navigation and Route Guidance Demonstration Program. *Proc., Vehicle Proc., Navigation and Information Systems Conference,* SAE, Warrendale, Pa., 1991.
3. Sumner, R. Data Fusion in Pathfinder and TravTek. *Proc., Vehicle Navigation and Information Systems Conference,* SAE, Warrendale, Pa., 1991.
4. Rillings, J. H., and J. W. Lewis. TravTek. *Proc., Vehicle Navigation and Information Systems Conference,* SAE, Warrendale, Pa., 1991.
5. Von Tomkewitsch, R. Dynamic Route Guidance and Interactive Transport Management with ALI-SCOUT. *IEEE Transactions on Vehicular Technology,* Vol. 40, No. 1, Feb. 1991.
6. Sodeikat, H. Cooperative Transport Management with EURO-SCOUT. In *Advanced Technology for Road Transport: IVHS and ATT* (I. Catling, ed.), Artech House, Boston, Mass., 1994.
7. Schnaiberg, A., and J. L. Schofer. *Driver Recruitment Focus Group.* ADVANCE Project Technical Report NU-le.2-2. Transportation Center, Northwestern University, Evanston, Ill., Nov. 1991.
8. Sethi, V. *Arterial Incident Detection Using Fixed Detector Data.* M.S.C.E. thesis. Department of Civil Engineering, Northwestern University, Evanston, Ill., May 1994.
9. Bhandari, N. *Detecting Arterial Incidents Using Probe Vehicles.* M.S.C.E. thesis. Department of Civil Engineering, Northwestern University, Evanston, Ill. May 1994.
10. Liu, P.-C., J. L. Schofer, and J. N. Ivan. *Anecdotal Incident Detection Algorithm: Northwest Central Dispatch Preprocessor.* ADVANCE Project Interim Report. Transportation Center, Northwestern University, Evanston, Ill., May 1993.
11. Koppelman, F. S., V. Sethi, and J. N. Ivan. *Calibration of Data Fusion Algorithm Parameters with Simulated Data.* ADVANCE Project Technical Report TRF-ID-152. Transportation Center, Northwestern University, Evanston, Ill.; June 1994 (revised).
12. Sethi, V., F. S. Koppelman, C. P. Flannery, N. Bhandari, and J. L. Schofer. *Duration and Travel Time Impacts of Incidents.* ADVANCE Project Technical Report TRF-ID-202. Transportation Center, Northwestern University, Evanston, Ill; Nov. 1994.
13. Payne, H. J., and S. C. Tignor. Freeway Incident Detection Algorithms Based on Decision Trees with States. In *Transportation Research Record 682,* TRB, National Research Council, Washington, D.C., 1978.
14. Levin, M., and G. M. Krause. Incident Detection—A Bayesian Approach. In *Transportation Research Record 682,* TRB, National Research Council, Washington, D.C., 1978.
15. Persaud, B. N., and F. L. Hall. Catastrophe Theory and Patterns in 30-second Traffic Data—Implications for Incident Detection. *Transportation Research,* Vol. 23A, No. 2 , 1989.
16. Dudek, C. L., G. M. Messer, and N. B. Nuckles. Incident Detection on Urban Freeways. In *Transportation Research Record 495,* TRB, National Research Council, Washington, D.C., 1974.
17. Klecka, W. R. *Discriminant Analysis.* Sage University Paper Series. Quantitative Applications in the Social Sciences, 07–019. Sage Publications, Beverly Hills, Calif., 1980.
18. Bhandari, N., F. S. Koppelman, V. Sethi, and J. L. Schofer. *Revised Documentation for ID Algorithms for Release 1.5.* ADVANCE Project Technical Report TRF-ID-155. Transportation Center, Northwestern University, Evanston, Ill; June 1994 (revised).

The contents of this paper reflect the views of the authors, who are responsible for the information and accuracy of the data presented. It does not constitute a standard, specification, or regulation. It does not necessarily reflect the official policy of any of the ADVANCE *parties, nor does any mention of manufacturers or products constitute an endorsement of the* ADVANCE *parties.*

Driver Deceleration Behavior on a Freeway in New Zealand

CHRISTOPHER R. BENNETT AND ROGER C. M. DUNN

The results of a study that monitored driver deceleration behavior on a freeway in New Zealand are presented. A series of axle detectors were placed over a 500-m interval and the speeds were recorded using a data logger. The speeds of the same vehicle at different stations along the road were established for more than 1,200 vehicles. The speed profiles showed that vehicles decelerated over the same distance irrespective of the initial speed. As a result, the deceleration rate was proportional to the initial speed. A relationship was developed to predict the speed at any time as a function of the approach speed.

To model traffic flow, most simulation programs resort to models of driver acceleration and deceleration behavior. These dictate the speeds adopted during the simulation and, thus, significantly influence the results.

This paper presents the results of a study of vehicle deceleration behavior on a freeway in New Zealand. It begins with an overview of the various techniques used to model acceleration and deceleration behavior, which is followed by the results of a specific study of decelerations on a freeway.

RESEARCH ON MODELING ACCELERATION AND DECELERATION

Given the importance of modeling driver deceleration and acceleration behavior, there are surprisingly few studies reported in the literature on this topic. The research that has been done essentially can be divided into four distinct areas: constant, linearly decreasing, polynomial, and driving power–based models.

Constant Acceleration Models

The simplest form of model is the constant acceleration model. [The generic term acceleration will be used to describe either acceleration (positive) or deceleration (negative) except when presenting specific equations or study results.] It assumes that the average acceleration is maintained throughout the acceleration maneuver. Table 1 presents some typical values reported in the literature for average acceleration rates (*1–7*).

Linearly Decreasing Acceleration Models

Constant acceleration models are not appropriate for developing detailed speed profiles. Accordingly, for these purposes researchers have tended to adopt a speed-dependent acceleration model. For example, Sullivan (*8*) presents curves showing the discretionary and maximum comfortable deceleration rates as a function of speed. These rates decrease linearly with increasing speed. This is an example of one of the most common forms of acceleration models: the linear-decreasing model.

Linear-decreasing models generally assume that the maximum acceleration occurs at the beginning of the maneuver, linearly decreasing to 0, or a constant value, at the final speed. Equation 1 is an example of such a model (*9*):

$$a = a_0 - a_1 v - Mg \frac{GR}{(M + M')} \qquad (1)$$

where

a = acceleration (m/sec^2),
a_0, a_1 = model coefficients,
v = vehicle speed (m/sec),
M = vehicle mass (kg),
M' = effective vehicle mass (i.e., the mass considering inertial effects) (kg),
g = acceleration due to gravity (m/sec^2), and
GR = gradient (%).

Many researchers have used linear-decreasing models (*1,6,7,9,10*). At higher speeds the model can become asymptotic, taking a long time to reach the final speed. This is illustrated in Figure 1, which shows the time-versus-speed profile for accelerating from 0 to 100 km/hr for four vehicle classes using Equation 1 with the parameter values from N-ITRR (*11*).

The medium and heavy commercial vehicles do not reach the 100-km/hr final speed within a reasonable time, because the acceleration decreases to a very small value, on the order of 0.02 m/sec^2, for heavy commercial vehicles as time increases. This rate compares with 0.57 m/sec^2 for the same vehicles when they begin to accelerate at the onset of the acceleration maneuver. It is therefore prudent to assume a minimum acceleration rate to eliminate this problem. Doing so, however, creates a second problem in that when the vehicles reach terminal speed, there will be an instantaneous change in the acceleration rate. In reality, drivers slowly reduce their rates so as to experience zero "jerk" at the end of the acceleration.

Polynomial Acceleration Models

Because of the problems outlined previously, other researchers have preferred polynomial model forms. Samuels (*12*) investigated the acceleration and deceleration of vehicles at an intersection. The data

C. R. Bennett, N.D. Lea International, Ltd., 1455 West Georgia Street, Vancouver, British Columbia V6G 2T3 Canada. R. C. M. Dunn, Department of Civil Engineering, University of Auckland, Private Bag 92019, Auckland, New Zealand.

TABLE 1 Values Used in Constant Acceleration Model

Source	Country	Acceleration or Deceleration Rate in m/s²	
		Acceleration.	Deceleration.
Lay (1)	Australia	1.00 to 4.00	
McLean (2)	Australia	0.34 to 1.18	-0.50 to -1.47
Watanatada, et al. (3)	Brazil		-0.40 to -0.60
Lee, et al. (4)	New Zealand	0.28 to 0.95	-0.28 to -0.96
Brodin and Carlsson (5)	Sweden		-0.50
Lay (1)	U.K.	0.50	
Bester (6)	U.S.A.		-0.60 to -1.90
St. John and Kobett (7)	U.S.A.		-1.07

indicated that a nonlinear speed-time relationship was applicable, and an equation of the following form was fitted to the data:

$$v = a_0 + a_1 t + a_2 t^2 \qquad (2)$$

where t is time in seconds and a_0, a_1, and a_2 are regression coefficients.

Samuels and Jarvis (*13*) investigated the maximum rates of deceleration and acceleration for a sample of 17 passenger cars. The models developed were of the following form:

- *Accelaration:*

$$v^2 = a_0 + a_1 t \qquad (3)$$

- *Deceleration:*

$$v = a_2 - a_3 t \qquad (4)$$

Jarvis (*14*) examined the acceleration behavior of drivers departing from a rural intersection. A regression was performed using Equation 3 along with a second-order model. The second-order term markedly improved the fit of the model, and parameters were presented for five classes of vehicles, from passenger cars to heavy trucks. These results were later modified (*15*) to consider speed as a function of distance.

In New Zealand a study on acceleration behavior was conducted in the small rural city Palmerston North at four roundabouts, five signalized, and four priority intersections using arrays of pneumatic tubes connected to a data logger (*16*). The analysis consisted of the fitting of a fourth-degree polynomial equation to the speed/distance profiles. This equation was of the form

$$S = a_0 + a_1 \,\text{DISPL} + a_2 \,\text{DISPL}^2 + a_3 \,\text{DISPL}^3 + a_4 \,\text{DISPL}^4 \qquad (5)$$

where DISPL is the cumulative distance traveled in meters, and a_0 through a_4 are regression constants.

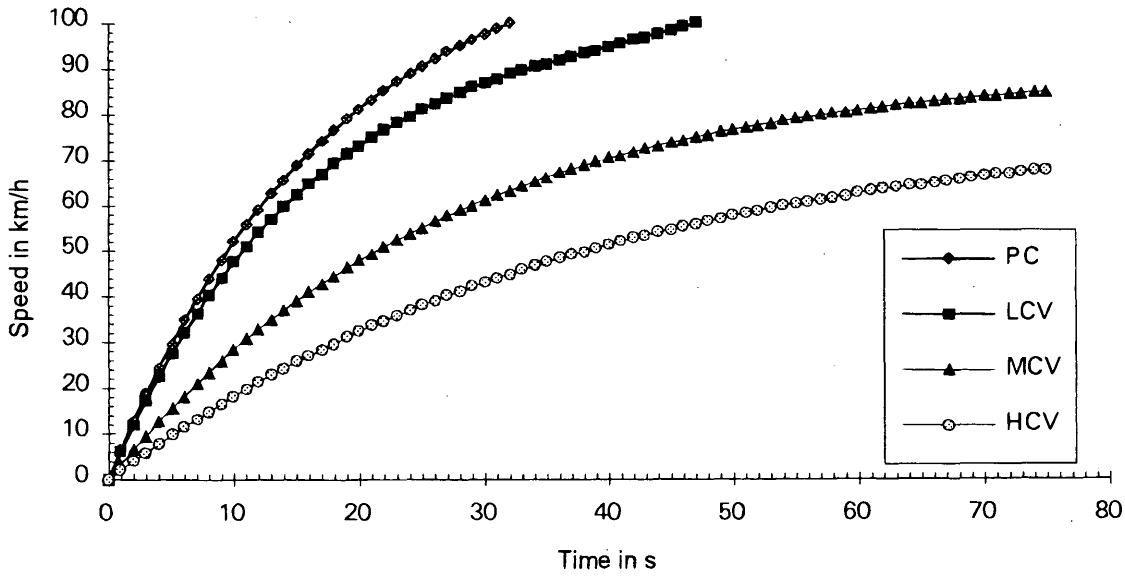


FIGURE 1 Speeds predicted by South African acceleration model.

Only vehicles with headways above 5.0 were included in the analysis, and equations were developed for each individual site and three composite equations for the different intersection types. Unfortunately, the model formulation does not lend itself to extrapolating for different approach speeds. A better method would have been to use the approach speed as an independent variable and to dispense with the constant a_0. Although limits for the equations are not given, it appears that the maximum approach speed in the study was on the order of 70 km/hr, so these equations are not appropriate beyond this speed.

Akçelik et al. (*10*) presented three models for passenger car acceleration profiles, a two-term sinusoidal, a three-term sinusoidal, and a polynomial model. These models were compared with constant and linear-decreasing acceleration models using data collected during fuel consumption testing. It was found that the polynomial model gave the best overall predictions and the linear-decreasing model the worst. This led to the development of the Australian Road Research Board (ARRB) polynomial model (*17*).

The ARRB polynomial model uses the time to accelerate or decelerate and the average, initial, and final speeds to predict a model parameter δ. This is a shape parameter that indicates whether the maximum acceleration occurs early or late in the profile. If δ is known (or assumed), only the times to accelerate and decelerate are required. A series of other model parameters are derived that result in a speed-time equation.

A series of equations were developed to predict the time to accelerate/decelerate and the acceleration/deceleration distances from field data collected in Australia (*17*). However, Bennett (*18*) indicates that there were problems with the ARRB equations in that their predictions were inconsistent for some speed combinations. Because of this, a linear model was adopted for acceleration in New Zealand (*18*). This linear acceleration model was less than ideal in that it predicted the same acceleration rate irrespective of speed (i.e., 0 to 20 km/hr would take as long as 80 to 100 km/hr).

Vehicle Power–Based Acceleration Models

The maximum acceleration of a vehicle is governed by the available acceleration reserve. Several researchers who have developed speed simulation models have used the acceleration reserve as the basis for predicting acceleration. (*3,5,8*). The underlying philosophy in this approach is that drivers use all the available power to accelerate their vehicle. (Since the acceleration reserve applies to positive power only this method is not used for deceleration.) Since the acceleration reserve decreases nonlinearly with increasing speed, this approach gives a nonlinear decreasing speed model. However, these sources (*3,5,8*) do not state explicitly whether an upper limit was used with the acceleration reserve to reflect the fact that drivers may use different power levels under acceleration than under steady-state driving.

GEIPOT (*19*) employed a variation of this approach. It adopted a nonlinear acceleration-speed relationship that gave the acceleration or deceleration as a function of gradient, roughness, and surface type. A single function was used that gave both acceleration and deceleration.

GRAFTON MOTORWAY DECELERATION STUDY

Introduction

A study was conducted at the Grafton Motorway exit ramp in Auckland, New Zealand, to monitor vehicle deceleration behavior. Seven pairs (stations) of axle detectors were installed on the ramp over a distance of 500 m upstream from the traffic signal at the end of the ramp. The first station was positioned to record approach speeds, and the last station was 10 m before a traffic signal.

The ramp was straight and had very high sight distances (>750 m) and a slight downgrade (<3 percent) over the initial 300 m. It had a single lane except for 75 m upstream from its end, where there were two lanes. The detectors at Stations 1 to 3 were spaced at 100-m intervals and thereafter at 50-m intervals. The experiment was conducted over a 24-hr period, but only data from daytime were used in the analysis.

Data were recorded at each station using a VDDAS data logger (*20*). The time of each axle crossing a detector was recorded to the nearest millisecond. The speeds were then calculated on the basis of these times and the distances between the detectors. VDDAS allowed for continuous sampling, which eliminated sampling biases in data collection that may arise with manual methods such as radar. More important, it allowed the speeds of the same vehicle to be tracked as it crossed successive detectors, thereby giving speed profiles for individual vehicles. The vehicles were classified into one of 44 classes based on the number of axles and their spacing. Special software was written both for the data reduction and establishing the speed profiles as described by Bennett (*22.*)

A total of 1,200 valid speed profiles were obtained in the study; they were stored in a FoxPro data base. A valid profile was considered to be one in which the same vehicle was identified and had its speed monitored at four or more stations.

Data Reduction

The speed profile data base contained the speed of the vehicle at each station along with the time between stations. It was necessary to manipulate these data into a format suitable for statistical analysis. It was postulated that the deceleration behavior would vary by vehicle type, so it was also necessary to disaggregate the data by vehicle type.

The data base was filtered so that only vehicles with a minimum headway of greater than 4.5 sec at all stations were included in the analysis. An upper limit of 15.0 sec was placed on the data to eliminate any unusually slow vehicles. This upper limit affected less than 0.1 percent of all available data. The profile data were filtered and converted into a sequential data base. Because of the limited amount of data available, the analysis could be conducted only for three vehicle types: (*a*) passenger cars and small light commercial, (*b*) medium commercial, and (*c*) heavy commercial vehicles. The total number of speed-time observations available by vehicle class were as follows:

Class	*No. of Speed-Time Observations*
Passenger cars and small LCV	1,604
Medium commercial vehicles	255
Heavy commercial vehicles	131

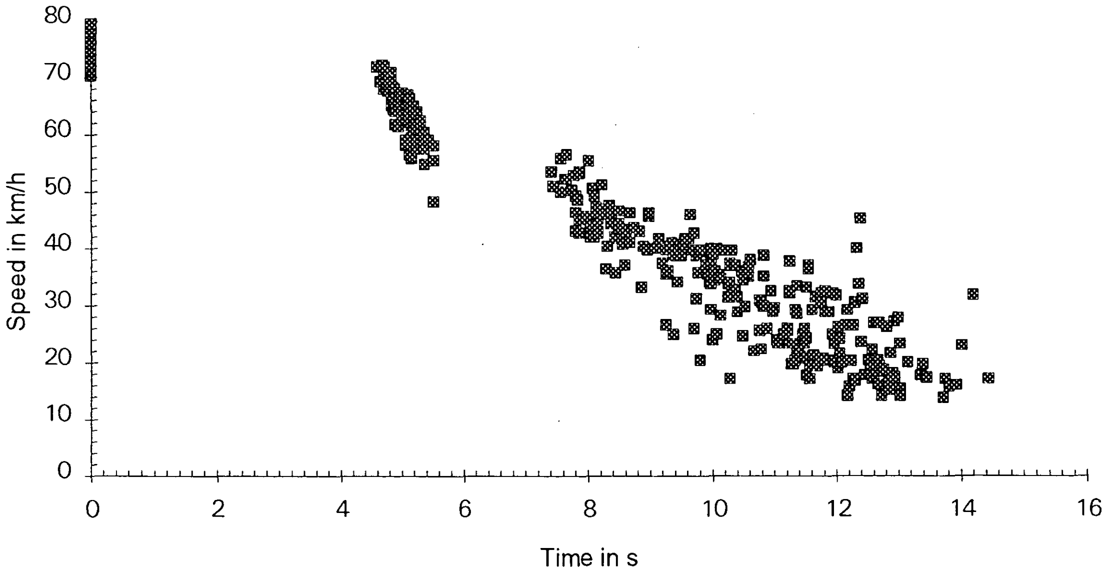


FIGURE 2 Elapsed time versus speed for vehicles approaching at 70 to 80 km/hr.

Results of Analysis

The literature review presented earlier indicated that deceleration behavior was probably a function of speed, so the analysis first concentrated on investigating such a relationship for passenger cars since these vehicles had the most data available. Figure 2 is an example of speed versus the elapsed time from the first detection for passenger cars traveling at an initial speed of between 70 and 80 km/hr at Station 3.

When the data were plotted for different approach speeds, it became apparent that deceleration behavior varied as a function of approach speed. Furthermore, there was little deceleration at the first two stations, so the initial station for deceleration purposes was treated as Station 3. The data were segmented into files from 50 to > 100 km/hr in 10-km/hr increments. For each file a regression analysis was conducted that investigated the effect of time on speed. A variety of linear and nonlinear models were tested with the equations in Table 2 being selected as the most appropriate for modeling deceleration behavior.

The deceleration equations used are all of the form $S = a_0 - a_1 t - a_2 t^2$. In comparing the models it can be observed that the coefficients a_1 and a_2 increase with increasing approach speed. This indicates that faster vehicles decelerate at a higher rate. However, there is a problem with these models in that they provide inconsistent predictions at lower speeds. The predictions cross because the faster drivers do not begin decelerating until late in the maneuver and thus have a different time base than the lower speeds.

Although the preliminary models developed were inadequate for general use, they did indicate that the higher the approach speed, the higher the rate of deceleration. This characteristic was further investigated by stratifying the data into various speed intervals and determining the average deceleration over these intervals for each approach speed group. It was not possible to use identical intervals with each approach speed group since there were often marked variations in the deceleration rate with time. Table 3 presents the average deceleration rates as a function of approach speed and deceleration speed for speeds below 100 km/hr.

Table 3 verifies that there is a marked difference in deceleration behavior by approach speed. Vehicles traveling at low speeds experienced a low deceleration rate, whereas those at high speeds had much higher rates. This suggests that rather than taking a much longer distance, or time, to decelerate, high-speed drivers prefer to decelerate more rapidly.

For comparative purposes, the New Zealand deceleration rates were assessed against those used in the ARFCOM model from Australia (*21*). For speeds below 80 km/hr, the observed New Zealand deceleration rates were similar to those in ARFCOM. However, in the 100 to 80 km/hr area, the New Zealand rates were approximately

TABLE 2 Preliminary Regression Models by Approach Speed

Approach Speed	Speed Model	R_a^2
60 - 70 km/h	$S = 66.66 - 0.96\ t - 0.18\ t^2$	0.96
70 - 80 km/h	$S = 75.68 - 1.64\ t - 0.22\ t^2$	0.96
80 - 90 km/h	$S = 84.46 - 2.59\ t - 0.25\ t^2$	0.96
90 -100 km/h	$S = 94.36 - 3.65\ t - 0.30\ t^2$	0.98
> 100 km/h	$S = 105.69 - 4.95\ t - 0.41\ t^2$	0.97

Note: R_a^2 = the adjusted coefficient of determination

TABLE 3 Mean Deceleration by Approach Speed and Speed During Deceleration

Mean Deceleration Rate by Approach Speed and Speed During Deceleration							
60 - 70 km/h		70 - 80 km/h		80 - 90 km/h		90 -100 km/h	
Decel. Speed (km/h)	Mean Decel. (m/s^2)	Decel. Speed (km/h)	Mean Decel. (m/s^2)	Decel. Speed (km/h)	Mean Decel. (m/s^2)	Decel. Speed (km/h)	Mean Decel. (m/s^2)
65 - 55	0.46	75 - 62	0.78	85 - 68	1.23	95 - 75	1.39
55 - 45	0.93	62 - 50	1.11	68 - 58	1.39	75 - 58	1.89
45 - 20	1.39	50 - 19	1.78	58 - 18	2.22	58 - 22	2.34

90 percent higher. This difference could reflect the fact that the ARFCOM data are primarily urban-based whereas the New Zealand data pertain to open road speeds.

It is interesting that the maximum deceleration observed in another New Zealand study conducted in the rural city of Palmerston North (*16*) was −1.72 m/sec^2. This result is similar to the *average* deceleration for the approach speed of 80 to 90 km/hr, verifying that drivers on open roads use a higher deceleration rate than do drivers in urban areas.

The Grafton Motorway data indicate that vehicles generally start decelerating at the same point on the road irrespective of the approach speed. Faster drivers then accept a higher deceleration rate than the slower drivers. This has the effect of producing deceleration times and distances that are of the same magnitude irrespective of the initial and final speeds.

A number of models were investigated for predicting the speed profile. These included sigmoidal models, the polynomial model from ARRB, as well as various polynomial equations. One of the main problems in developing a suitable model was the need to consider the variation in the deceleration rate as a function of approach speed and the predictions as vehicles approached stopping. It was found that the following formulation gave the most suitable overall predictions:

$$S = S_a + a_0 S_a t^2 \qquad (6)$$

where S is the speed of the vehicle at time t in kilometers per hour, and S_a is the approach speed of the vehicle in kilometers per hour. Taking the derivative of this equation with respect to time gives the following model for predicting acceleration:

$$a = a_1 S_a t \qquad (7)$$

Table 4 presents the coefficients and regression statistics for the previous two models by vehicle class. The values for coefficient a_0 indicate that light vehicles decelerate 22 percent faster than heavy vehicles. The differences between passenger cars and medium trucks is so small that it is negligible. Figure 3 illustrates the predicted speed profiles of passenger cars from different approach speeds using Equation 6.

Equation 7 predicts that the higher the approach speed, the greater the deceleration rate. This was observed from the raw data. It also indicates that the maximum deceleration will occur at the very end of the speed profile. This is a deficiency in the model since at the end of the profile the drivers will actually experience zero jerk.

CONCLUSIONS

This analysis has developed equations for predicting deceleration behavior of vehicles as a function of approach speed and the cumulative time. Although the equations pertain to a specific situation—vehicles decelerating from the open road speed toward a stop—the analysis has provided useful insight into driver deceleration behavior.

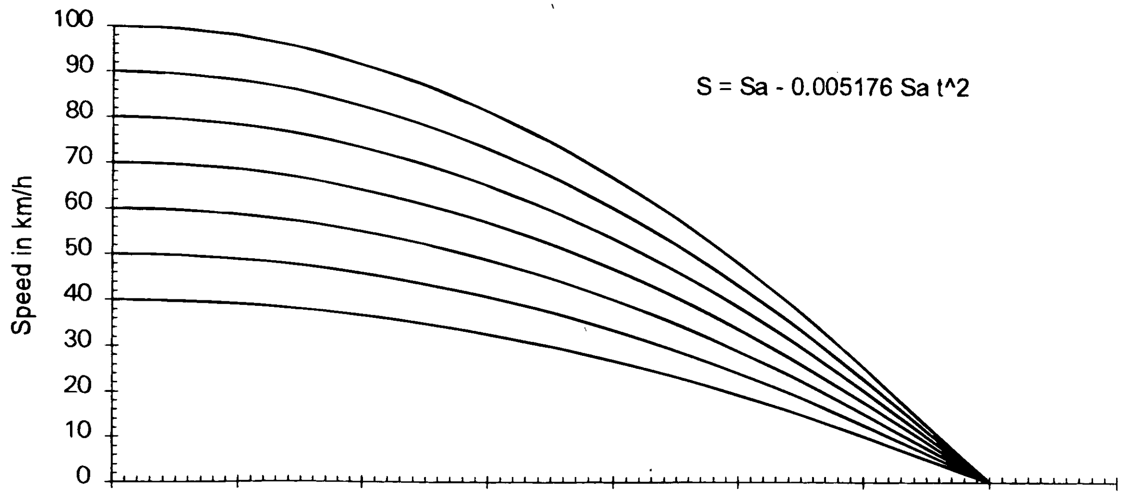


FIGURE 3 Predicted deceleration profiles for passenger cars and small commercial vehicles.

TABLE 4 Final Deceleration Model Regression Coefficients

Vehicle Class	Regression Model Coefficients		R_a^2
	a_0	a_1	
Passenger Cars and Small LCV	-0.005176	-0.002876	0.83
Medium Commercial Vehicles	-0.005129	-0.002849	0.86
Heavy Commercial Vehicles	-0.004244	-0.002358	0.83

Notes: a_0 and a_1 = model coefficients
R_a^2 = the adjusted coefficient of determination

It was found that higher-speed drivers decelerate over a short period of time, thereby experiencing high deceleration rates instead of gradually decelerating over a long period. This is different than what is predicted by the equations for applying the ARRB polynomial model (*10*), which imply that drivers decelerate over a longer distance with higher speeds.

The average deceleration rate for drivers with an approach speed of 80 to 90 km/hr was similar to the maximum deceleration rate observed in an urban study in New Zealand. This indicates that open road drivers have much higher deceleration rates than urban drivers employ.

ACKNOWLEDGMENTS

The work presented in this paper was undertaken at the University of Auckland as part of a research project to develop a speed prediction model for rural highways (*22*). The funding and support for the research were provided by Transit New Zealand.

REFERENCES

1. Lay, M. G. *Acceleration-Time Relationships.* Australian Road Research Board Internal Report AIR-454-2. Australian Road Research Board, Nunawading, 1987.
2. McLean, J. R. *Adapting the HDM-III Vehicle Speed Prediction Models for Australian Rural Highways.* Working Document TE 91/014. Australian Road Research Board, Nunawading, 1991.
3. Watanatada, T., A. Dhareshwar, and P. R. S. Rezende-Lima. *Vehicle Speeds and Operating Costs: Models for Road Planning and Management.* Johns Hopkins Press, Baltimore, Md., 1987.
4. Lee, K. C., G. W. Blanchard, and M. S. Rosser. *Driving Patterns of Private Vehicles in New Zealand.* Applied Research Office Report ARO/1529. University of Auckland, New Zealand, 1983.
5. Brodin, A., and A. Carlsson. *The VTI Traffic Simulation Model.* VTI Meddelande 321A. Swedish Road and Traffic Research Institute, Linköping, 1986.
6. Bester, C. J. *Fuel Consumption of Highway Traffic.* Ph.D. thesis. University of Pretoria, South Africa, 1981.
7. St. John, A. D., and D. R. Kobett. *NCHRP Report 185: Grade Effects on Traffic Flow Stability and Capacity.* TRB, National Research Council, Washington, D.C., 1978.
8. Sullivan, E. C. *Vehicle Operating Cost Model—User's Guide.* Research Report UCB-ITS-77-3. Institute of Transportation Studies, University of California, Berkeley, 1977.
9. Slavik, M. M., et al. *Development of Optimum Geometric Standards for South African Roads: Status of the Project a Year Before Target Date.* Technical Report RT/4/79. National Institute for Transport and Road Research, Pretoria, South Africa, 1979.
10. Akçelik, R., D. C. Biggs, and M. G. Lay. *Modelling Acceleration Profiles.* ARRB Internal Report AIR 390-3. Australian Road Research Board, Nunawading, 1983.
11. NITRR *Evaluation of Road User Costs: Technical Manual for RODES 2.* NITRR Manual P11. Council for Scientific and Industrial Research. Pretoria, South Africa, 1983.
12. Samuels, S. E. Acceleration and Deceleration of Modern Vehicles. *Australian Road Research,* Vol. 6, No. 2,1976, pp. 23–29.
13. Samuels, S. E., and J. Jarvis. *Acceleration and Deceleration of Modern Vehicles.* Australian Road Research Report 86. Australian Road Research Board, Nunawading, 1978.
14. Jarvis, J. R. In-Service Vehicle Performance. *Proc., 2nd Conference on Traffic Energy and Emissions,* Society of Automotive Engineers-Australasia/Australian Road Research Board, Melbourne, Australia, 1982.
15. Jarvis, J. R. *Acceleration Lane Design.* ARRB Internal Report AIR 281-1, Australian Road Research Board, Nunawading, 1987.
16. ATS. *Acceleration/Deceleration Profiles at Urban Intersections.* Transit New Zealand Australasian Traffic Surveys, Victoria, 1990.
17. Akçelik, R., and D. C. Biggs. Acceleration Profile Models for Vehicles in Road Traffic. *Transportation Science,* Vol. 21, No. 1, 1987.
18. Bennett, C. R. *The New Zealand Vehicle Operating Costs Model.* RRU Bulletin 82. Transit New Zealand, Wellington, 1989.
19. GEIPOT. *Research on the Interrelationships Between Costs of Highway Construction Maintenance and Utilisation: Final Report on Brazil-UNDP Highway Research Project* (12 vol.). Brasilia, Brazil, 1982.
20. Hoban, C. J., P. J. Fraser, and J. J. Brown. The ARRB Vehicle Detector Data Acquisition System (VDDAS). *Australian Road Research,* Vol. 17, No. 1, 1987, pp. 52–54.
21. Biggs, D. C. *ARFCOM—Models for Estimating Light to Heavy Vehicle Fuel Consumption.* ARRB Research Report ARR152. Australian Road Research Board, Nunawading, 1988.
22. Bennett, C. R. *A Speed Prediction Model for Rural Two-Lane Highways in New Zealand.* Ph.D. thesis. University of Auckland, New Zealand, 1994.