April 1979 LIDS-P-912

Estimation of Roadway Traffic Density on Freeways Using Presence Detector Data+

by

Andrew Kurkjian*
Stanley B. Gershwin*
Paul K. Houpt**,*
Alan S. Willsky*
C. S. Greene*
E. Y. Chow*

---

* M.I.T. Laboratory for Information and Decision Systems, Rm. 35-308, Cambridge, MA 02139

** M.I.T. Department of Mechanical Engineering, Rm. 1-110, Cambridge, MA 02139

+ This research was supported by the U.S. Department of Transportation under Contract DOT-OS-60137

ABSTRACT

Existing methods of estimating section (link) density on freeways from data provided by electronic presence (loop) detectors typically require extensive knowledge of uncertainties and/or strong assumptions on prevailing flow conditions, such as homogeneity. Consequently, these methods are known to produce poor estimates in inhomogeneous conditions, or when a priori knowledge of traffic conditions is not available.

In this paper, a new data processing approach is presented which estimates density well over a wide range of traffic conditions. It does this by detecting spatially inhomogeneous traffic conditions and compensating the density estimation algorithm appropriately. The data processing algorithm is computationally simple, is not flow-level dependent, does not require any a priori knowledge of traffic conditions on the road and is insensitive to the types of uncertainty found in detector data. The algorithm uses both flow and occupancy data from adjacent (neighboring) detector stations to track the density on the link inbetween. A scalar Kalman filter formulation is used to provide the desired density estimate. The simplicity of the filter algorithm is achieved by using in tandem a scalar Generalized Likelihood Ratio (GLR) event detection algorithm to compensate the filter for spatially inhomogeneous conditions. Performance of the algorithm is demonstrated with a microscopic freeway simulation. Application of the same techniques as part of an overall incident defection scheme are also described.

1. INTRODUCTION

An electronic presence (loop) detector is a device buried in a road which provides a binary signal indicating the presence or absence of a vehicle in a well defined vicinity of the detector. Many hundreds of miles of freeway are equipped with these detectors. Typically, detectors are spaced at one half mile intervals along the road and, normally, each lane has a detector. Presence detectors were originally designed for the purpose of counting the number of vehicles to cross a point along the road over some time interval (i.e., the flow rate). However, it is currently being shown that traffic surveillance systems can be based solely on the data provided by these detectors. The extraction of information concerning traffic conditions on a freeway from the relatively restrictive information provided by presence detector signals is a difficult and challenging problem. One solution of this problem is the subject of this paper. The need for an automatic traffic surveillance system is clear in the context of urban traffic control and in detection of accidents or other abnormal events.

Ideally, one would like to continuously estimate the distribution and speed of vehicles on the road. This is impractical for several reasons. First, the computational burden of trying to track each vehicle's position and speed is overwhelming. Second, it cannot be done with the data provided by conventionally spaced detectors. Although the data is microscopic in nature (i.e., it contains information concerning individual vehicles), it is impossible to track individual vehicles as they travel from detector station to detector station down the road. This is because the presence pulses in the detector signal cannot be associated with the vehicles which

produced them. Moreover, estimation of a vehicle's speed from passage over a single detector is not a simple matter due to variations in vehicle lengths.

As an alternative, one can divide the freeway spatially into sections and estimate the number of vehicles (related to local density) and the average speed of the vehicles (space-mean speed) on each section, periodically in time. Such an approach is indeed feasible and, in fact, has received considerable attention in the literature. Existing methods for estimating density resort to some type of assumption concerning the homogeneity of the traffic flow. Because of this, the density estimates produced in inhomogeneous conditions often contain large errors.

The algorithm presented in this paper is unique in that it estimates the section (link) density accurately in all types of traffic conditions. It does this by detecting spatially inhomogeneous traffic conditions and compensating the density estimate appropriately for the adverse effects of the inhomogeneities. Furthermore, the data processing algorithm is computationally simple, is not flow-level dependent, does not require any a priori knowledge of traffic conditions on the road, and is insensitive to the types of imperfections found in the detector data. Because the system can detect spatially inhomogeneous conditions, it can be of some use for accident detection purposes.

Essentially, the method bases its estimate on two easily obtained measures from presence detector data: the flow rate and occupancy. The occupancy of a detector during a particular time interval is the percent of that interval during which the detector signals the presence of

vehicles. Both of these measurements we averaged over five second intervals. From measurements at neighboring detector stations, the number of vehicles, and therefore the density, on the link between stations is tracked. The estimate is calculated using a Kalman filter. The detection of spatially inhomogeneous conditions is done using a Generalized Likelihood Ratio (GLR) event detection scheme. Both the filter and the GLR algorithm are simple discrete-time scalar equations. All testing and experimentation was done using simulated traffic data. A microscopic traffic simulation program developed at M.I.T. was used for this purpose.

2. APPROACH

2.1 The State and Observation Equation

The method for estimating density relies on both flow and occupancy measurements from neighboring detector stations to arrive at the density estimate on the link inbetween. The concept underlying the method is to count the number of vehicles entering and leaving the link by counting the presence pulses observed at the detector stations at each end of the link. Thus, if the initial density on the link were known, and the detectors count passing vehicles without error, then the actual link density could be tracked perfectly. Presence detectors were originally designed for the purpose of counting vehicles. However, they do not count perfectly. Errors in vehicle counts imply that vehicles are entering and leaving a section unnoticed and this causes unpredictable errors in the estimation of the section density in the form of bias or drift. If uncompensated, these errors grow in time, leading to meaningless density estimates. Also, it is unreasonable to demand that perfect, or even good, knowledge of the initial link density be provided. Thus, some method of circumventing these difficulties must be devised. With these issues in mind, consider the following formulation of the problem.

Let $\rho(k)$ denote the actual density on a particular link at the k^{th} timestep (in veh/mile/lane). (That is, k is an index used for discrete time so that $k\Delta t = t$ where t is clock time and Δt is a specified time interval.) Then we can write the following exact conservation-of-vehicles

equation which is valid in all traffic conditions

$$\rho(k+1) = \rho(k) + u(k) + \nu(k) \tag{2.1}$$

Here, $u(k)$ represented the measured change in link density which occurs from timestep k to timestep k+1. The term $\nu(k)$ is a noise term used to model the difference between the actual and measured change in density. We will refer to $\rho(k)$ as the system state and to eq (2.1) as the state equation.

The measurement u(k) is obtained from vehicle count information, provided by the detector stations at each end of the link, and knowledge of the link length, Δx, and the number of lanes, L, as follows

$$u(k) = \frac{IN(k) - OUT(k)}{L(\Delta x)} \tag{2.2}$$

Here, IN(k) (OUT(k)) denotes the number of vehicles to enter (leave) the link as measured by the upstream (downstream) detector station between timesteps k and k+1. Knowledge of the statistics of the errors made in measuring IN(k) (or OUT(k)) will allow us to derive the statistics of $\nu(k)$. These will be developed in the next section. Note that eq (2.1) assumes that the link has no entrance or exit ramps and that the number of lanes does not change along the link. Our method can be directly adapted to the case of changing numbers of lanes by basing all of our calculations on cars/mile as opposed to cars/mile/lane. In addition, entrance and exit ramps can be taken into account as long as presence detectors are placed on the ramps. Since both of these possibilities can be taken care of without any major modification, we will not treat

them here in order to simplify our development.

Also note that knowledge of $u(k)$, $k=0,1,2,\ldots$ provides no information about the initial density, $\rho(0)$, and future values of $u(k)$ don't provide any information about past errors in the $u(k)$. These problems lead to difficulties in simply using the $u(k)$ to track $\rho(k)$. To overcome this difficulty, occupancy measurements are used to provide a rough measurement of density. This is a reasonable approach since occupancy appears, intuitively, to be related to density. The specific nature of this relationship must be explored. Specifically, in Section 4 we develop a model of the form

$$z(k) = \rho(k) + \eta(k) \tag{2.3}$$

where $z(k)$, the measurement, is

$$z(k) = \frac{\alpha}{2}\,[\mathrm{OCCUP}(k) + \mathrm{OCCDOWN}(k)] \tag{2.4}$$

Here $\mathrm{OCCUP}(k)$ is the upstream occupancy measured at the upstream detector station over the interval $[k, k+1]$. Similarly, $\mathrm{OCCDOWN}(k)$ is the downstream occupancy over $[k, k+1]$. The parameter, α, is a proportionality constant whose significance and value is developed in Section 4. The term, $\eta(k)$, is a noise term used to model the error in the conversion of occupancy into a measurement of density. Eq. (2.3) is referred to as the observation equation.

Equations (2.1) - (2.4) provide the basis for our density estimation method. The detector station data is used, via equations (2.2) and (2.4), to provide values for $u(k)$ and $z(k)$ respectively. Such measurements are

readily attainable from the data. The state and observation equations, (2.1) and (2.3), form the basis of a Kalman filter which is used to provide the desired estimate of the density $\rho(k)$. We refer to the estimate of $\rho(k)$ as $\hat{\rho}(k)$.

Before developing the filter algorithm, we must complete two tasks:

1) Determine the statistics of the state noise, $\nu(k)$;

2) Determine the value of α and the associated statistics of the observation noise, $\eta(k)$.

In addition, we must decide when equation (2.4) can be relied on. In Section 4 we find that this equation is only valid under homogeneous traffic conditions and breaks down in the presence of accidents. Our results in Section 6 provide an effective manner in which to compensate for this. Before continuing our development, let us briefly discuss several previously developed density estimation techniques.

2.2 Other Density Estimation Methods

Nahi [2], [3] proposes a density estimation scheme also based on conservation of vehicles. He uses equation (2.1) as a state equation, but his method treats the unobservability problem differently. He uses a homogeneity (i.e., smoothness of flow in space) assumption. This assumption is not valid during accident conditions, or when transient traffic inhomogeneities occur. Therefore this system can be expected to lead to significant estimation errors. Furthermore, his method also demands that a good initial guess of the link density be available to start the algorithm. Finally, the issue of imperfect vehicle counts is ignored.

Given a good initial density estimate, however, this method shows the ability to track the density closely in homogeneous conditions. No results are presented for inhomogeneous conditions. The performance with poor initial estimates is not discussed. The method proposed here is partially based on the same concept as that of Nahi's method, but manages to overcome the important issues of estimation in inhomogeneous conditions and *a priori* knowledge of the density.

Another, less direct, estimation procedure has been proposed by Gazis and Knapp [4] and Gazis and Szeto [5]. In these papers, a procedure for estimating density by first estimating travel time is introduced. The density estimate is then obtained from the travel time. The method is complicated because it requires the solution of a two point boundary value problem. Furthermore, extensive lane changing or accidents may cause significant errors in the travel time algorithm.

3. STATISTICAL MODEL FOR THE STATE NOISE

In this section we present a new approach to modeling the state noise process, $\gamma(k)$, which accounts for measurement error in the variables IN(k) and OUT(k). As our model is based on first principles of vehicle-detector interaction, we begin with a brief review of some of the research that has been done into the nature of presence detector data.

Mikhalkin [6] experimented with conventional 6'x6' inductive loop detectors centered in standard 12' lanes. He found the detector station to have the detector regions shown in Figure 1. If any part of a vehicle covers this region, the detector will activate and produce a presence

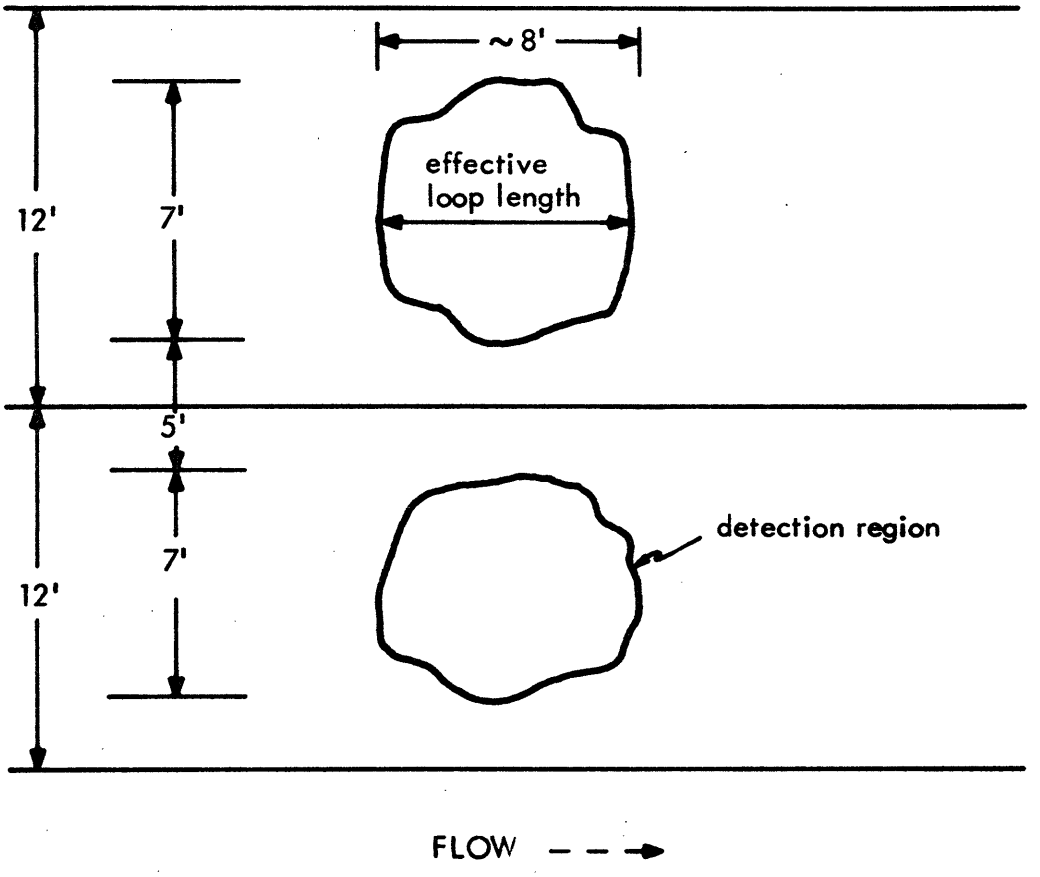


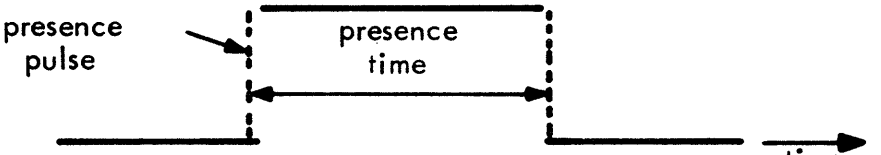
Figure 1. Detection Regions At A Detector Station

pulse (see Figure 2). No pulse is generated if no part of a vehicle is in this region.

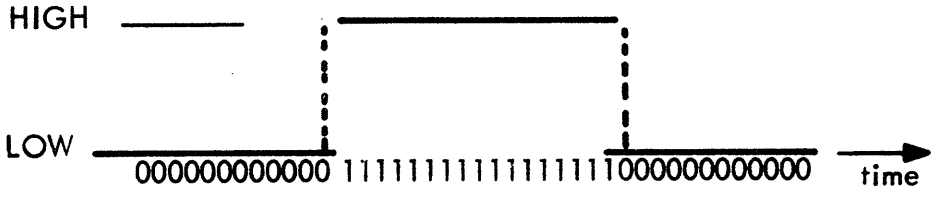
While the exact shape of the detection region can depend in a complex way on electronics sensitivity, loop installation (including surface and loop materials), the effective coverage can be approximated by nominal rectangular region shown in Figure 1. For the specific installation shown, Figure 1 shows that the only way for a vehicle to cross a detector station and not produce a presence pulse is for the vehicle to be less than five feet wide and traveling centered over a line separating lanes as it crosses the detector station. It is assumed that this has very small probability of occurrence and, therefore, all vehicles get counted at least once.

Figure 1 also shows that it is possible for a vehicle, changing lanes near a detector station, to activate presence detectors in both lanes and thus to produce two presence pulses. Figure 2 shows a top view of a vehicle of length ℓ [ft] and width w [ft] making a lane change. It is moving from center to center of adjacent 12' lanes. The lane changing operation is assumed to take place at a constant speed v [ft/sec] and requires t seconds to complete. Thus, z feet of road are needed for the change, where $z = vt$. Assuming that the detection regions of the loops in adjacent lanes are five feet apart, this vehicle will activate both detectors if and only if the detector station is located in the length X of road indicated in Figure 3. From the simple geometry of Figure 3, the following equation is obtained relating X to ℓ, w, v, and t

An analog signal resulting from the passage of a single vehicle



The analog signal is put through a threshold device to yield a binary signal in time



Signal is sampled (15 - 60 time sec.)

1: "vehicle present" bit

0: "vehicle absent" bit

Figure 2. Presence Detector Signal Associated With A Single Vehicle Passage

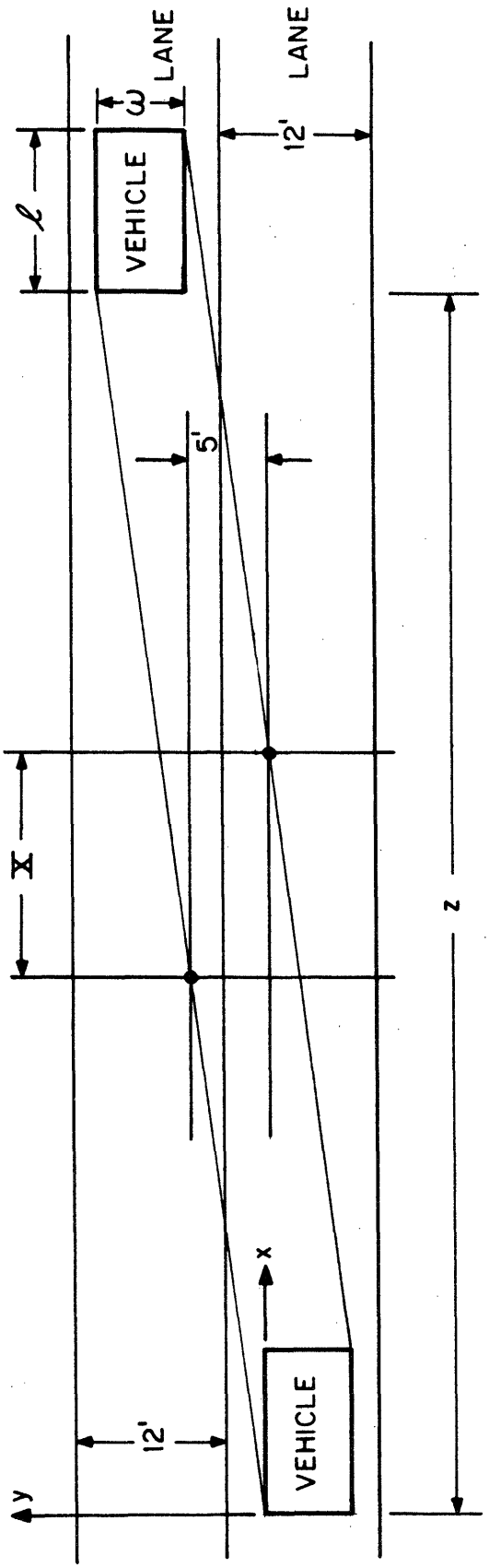


Figure 3. Top View Of Vehicle Changing Lanes

$$X = \frac{vt}{12} \quad (w-5) + \ell \qquad (3.1)$$

Suppose a vehicle 18' long and 6' wide makes a lane change at a constant speed of 88 ft/sec. and requires 4 seconds to complete the change. Using Eq. (3.1), this results in X = 47.3 feet. Assuming the lane change is equally likely to occur anywhere along the road, the probability of the lane change resulting in two presence pulses is

$$P = \frac{X}{2640 \text{ [ft./detector station]}} = \frac{47.3}{2640} = .0179$$

Thus, in this case it is rather unlikely that any given lane change will cause an extra count. Note that using this simple approach other detector geometries and/or vehicle types can easily be analyzed to find probabilities associated with spurious counts.

While this analysis is useful in order to analyze the probability of a single lane change causing an additional count, what is needed for the state model is a probabilistic description of the actual number of erroneous counts at a given detector station over a specified discretization interval. Such a statistical model clearly requires a model for driver lane changing behavior. Therefore, we have used the microscopic simulation detailed in [18], in order to obtain the desired statistics. This simulation includes a number of driver types, each of which has its own desired speeds (as a function of traffic conditions) and the simulation allows lane changes only if

1) The acceleration required is less than the maximum acceleration

capability of the vehicle at its current speed;

2) The speed computed with this acceleration is less than the desired speed of this vehicle

3) There is a gap of sufficient length available in the other lane. Only a vehicle restricted from driving its desired speed by other vehicles is eligible for a lane change.

When a changing of lanes takes place in the traffic simulation, the position of the vehicle as well as its length, width and speed are noted. This information is sufficient to compute the region X, using Eq. (3.1) and locate it on the road. If there is a presence detector station located within the region, then, as the vehicle crosses the detector station, a presence pulse is computed for each lane. The model used for computing the presence time is given by

$$\ell + d = vt + 1/2at^2 \qquad (3.2)$$

where ℓ = vehicle length [ft]

d = effective loop length [ft]

v = speed [ft/sec] of the vehicle when the front of the vehicle reaches the front of the detection region

a = acceleration of the vehicle [ft/sec^2] across the loop (assumed to be constant)

t = presence time (sec)

Equation (3.2) is a reasonable model for the presence time, based on the results of Mikhalkin [6]. Note that the effective loop length, d, used in Eq. (3.2) is not known exactly. Even though effective loop lengths

do depend on vehicle type, the range of variation is not large. Therefore, using Mikhalkin's results, it is assumed that d is a constant equal to 8 feet.

It is assumed that a vehicle cannot produce more than two pulses in crossing a detector station. Therefore, using lane changing near detector stations as the sole source of errors in vehicle counts and modeling this in the traffic simulation as described, the number of extra presence pulses generated was empirically examined. The results are shown in Table 1.

For reasons explained in Section 4, we use an interval duration, T, of five seconds throughout this study. That is, the time interval [t, t+T] is 5 seconds long. From Table 1, we can estimate the probability P_{ec} of one extra vehicle count occurring in a five second interval at a detector station. It is assumed that not more than one extra count occurs at a station in a five second interval. Such an assumption is reasonable because five seconds is a short time. That is, a lane change operation requires on the order of several seconds and involves two lanes. Therefore, in a five second interval, it is very unlikely that two lane changes can occur at the same point along the freeway.

Probability P_{ec} is approximately the probability that v(k) equals $\pm \frac{1}{L(\Delta x)}$. This is because, if we assume errors $\Delta IN(k)$ and $\Delta OUT(k)$ in the measurements of IN(k) and OUT(k), then equations (2.1) and (2.2) imply

$$\rho(k+1) - \rho(k) = u(k) + \nu(k) = \frac{IN(k) + \Delta IN(k) - OUT(k) - \Delta OUT(k)}{L\Delta x} \quad (3.3)$$

so that

$$\nu(k) = \frac{\Delta IN(k) - \Delta OUT(k)}{L\Delta x} \quad (3.4)$$

TABLE 1

VEHICLE COUNT ERROR STATISTICS ON A TWO LANE FREEWAY AT A DETECTOR STATION

AVERAGE FLOW RATE VEHICLES/HR. PER LANE	NUMBER OF MINUTES OF DETECTOR STATION DATA	AVERAGE NUMBER OF SECONDS PER EXTRA VEHICLE COUNT AT A STATION	STANDARD DEVIATION OF NUMBER OF SECONDS PER EXTRA VEHICLE COUNT AT A STATION	AVERAGE NUMBER OF EXTRA COUNTS PER HOUR AT A STATION
725	70	102	71	35
1000	112	97	93	37
1600	70	105	135	34

Since $\Delta IN(k) = 0$ or 1 and $\Delta OUT(k) = 0$ or 1, then

$$\nu(k) = \frac{\pm 1}{L\Delta x} \text{ or } 0 \tag{3.5}$$

The probability that $\nu(k)$ equals zero is, therefore, $1 - 2P_{ec}$. In this manner, the $\nu(k)$'s are modeled as discrete, independent, identically distributed, zero-mean random variables. Table 2 gives the variance of $\nu(k)$ for different flow levels. These results are needed in designing the Kalman filter.

4. A TRAFFIC MODEL AND ITS RELATION TO DETECTOR OBSERVATIONS

4.1 Definitions of Traffic Variables

In this section the relationship between occupancy and density is explored in detail. The objective is to determine the characteristics of the occupancy to density conversion given by equation (2.4) and to obtain a value for α. We begin by developing some notation and stating some key definitions. All "per lane" quantities (e.g., flow and density) are average values across all lanes.

The space-mean speed denoted by $\bar{v}_s(x, \Delta x, t)$, is the arithmetic average of the velocities, in miles/hr., of the vehicles in the section $[x, x + \Delta x]$ at time t.

The density in the section $[x, x + \Delta x]$ at time t is denoted $\rho(x, \Delta x, t)$ and is given by

$$\rho(x, \Delta x, t) = \frac{M(x, \Delta x, t)}{L\Delta x} \text{ [veh/mile per lane]} \tag{4.1}$$

where $M(x, \Delta x, t)$ is the number of vehicles in the section $[x, x + \Delta x]$ at

TABLE 2

OBSERVED STATISTICS OF ν AS A FUNCTION OF FLOW LEVEL IN MICROSCOPIC SIMULATION

AVERAGE FLOW RATE (VEH/HR PER LANE)	MEAN OF ν	SAMPLE VARIANCE OF ν
725	0	.097
1000	0	.103
1600	0	.094

time t, Δx is the length of the section in miles and L is the number of lanes. It is assumed that L is constant along the section.

The flow past a point x on the road during the time interval [t, t+T], denoted by $\phi(x, t, T)$, is given by

$$\phi(x, t, T) = \frac{N(x, t, T)}{TL} \quad \text{[veh/hr per lane]} \tag{4.2}$$

where $N(x, t, T)$ represents the number of vehicles to cross point x in the time interval [t, t+T]. Here, T is the duration of the interval in hours.

Each detector station has a presence detector in each lane, by assumption. The occupancy at a station is the arithmetic average of the coccupancies of the detectors in each lane. The occupancy of a presence detector in lane i, i = 1,2,...,L, located at point x in the time interval [t, t+T] is given by

$$occ_i(x,t,T) = \frac{100}{T}\left[t_{i,I} + \sum_{j=1}^{N_i(x,t,T)-1} t_{i,j} + t_{i,F}\right] \quad \text{[dimensionless]} \tag{4.3}$$

where $N_i(x,t,T)$ is the number of vehicles to cross point x in lane i in the interval [t, t+T] and where $t_{i,j}$, i = 1,2,...,L, j = 1,2,...,$N_i(x,t,T)$-1, is the presence time of the j^{th} vehicle to cross the detector in lane i. The effect of a vehicle already over the detector in lane i at time t is represented by $t_{i,I}$. Similarly, $t_{i,F}$ shows the effect of a vehicle over the detector in lane i at time t+T. (See Figure 4.) Define

$$\sum_{i=1}^{L} N_i(x,t,T) = N(x,t,T) \tag{4.4}$$

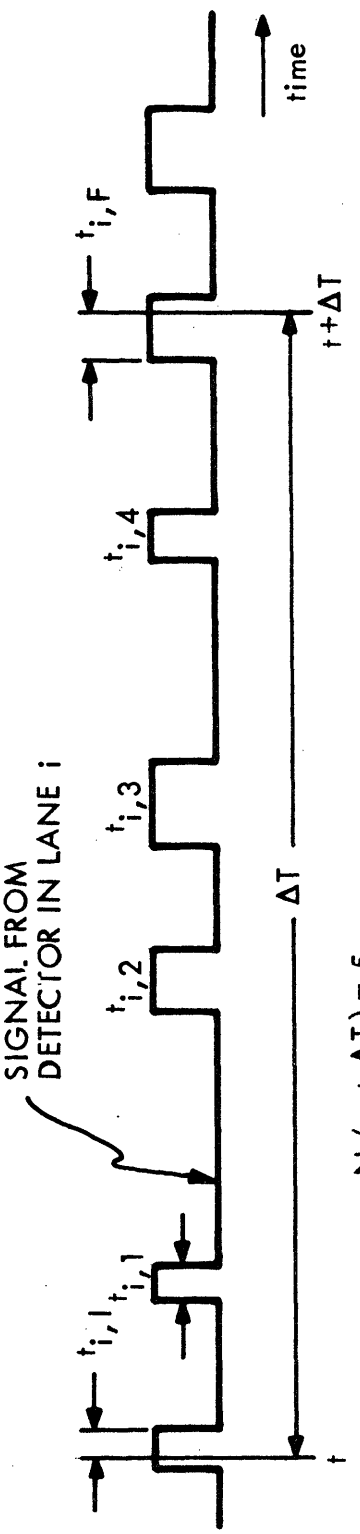


Figure 4. An Example of an Occupancy Computation

The occupancy of the detector station at fixed space point x over the interval [t, t+T] is given by

$$occ(x,t,T) = \frac{1}{L} \sum_{i=1}^{L} occ_i(x,t,T) \quad \text{[dimensionless]} \qquad (4.5)$$

In this paper, attention is restricted to the detector station occupancy of Eq. (4.5) as opposed to the specific detector occupancy of Eq. (4.3).

4.2 Traffic Model

Occupancy is a measurement obtained from data taken over time at a fixed point. Density, on the other hand, is a spatial quantity associated with a fixed time. In order to related fixed time spatial quantitites to fixed point temporal quantitites in traffic, a model describing the relationship between various instantaneous point variables in traffic is needed. Such models do exist (e.g., Phillips [7]). However, traffic flow is a complex process and these models typically take the form of nonlinear partial differential equations. Such models are not mathematically tractable but do reduce to simpler forms under some assumptions.

In our model, we assume that traffic conditions are homogeneous in time and space and show how spatial variables of interest can be related to available point observations in time. We subsequently show how the data processing algorithm developed can be used to compensate for inhomogeneities in actual traffic.

The traffic flow on a section [x, x+Δx] over an interval [t, t+T] is said to be space-time homogeneous if the space-mean speed and density on

any subsection of [x, x+Δx)] at any time within [t, t+T] is equal to the space-mean speed and density on any other subsection of [x, x+Δx] at any other time within [t, t+T]. (For a more rigorous definition, see Breiman [8].) Intuitively, the assumption of space-time homogeneous traffic flow means that the traffic conditions do not change either in time or in space. Thus, from observations at a point, spatial quantities can be inferred. Restricting our attention to this condition, the following simplification of notation is allowed

$$\left.\begin{aligned} \bar{v}_s(x,\Delta x,\ t) &= \bar{v}_s(x,\Delta x) \\ \rho(x,\Delta x,t) &= \rho(x,\Delta x) \end{aligned}\right\} \tag{4.6}$$

Breiman shows that for a point x_0 between x and x + Δx a relation exists between aggregate variables under space-time homogeneous conditions

$$\phi(x_0,t,T) = \rho(x,\Delta x)\bar{v}_s(x,\Delta x) \tag{4.7}$$

Thus the flow rate past any point in the section is the same. We can simplify the flow notation to

$$\phi(x,t,T) = \phi(t,T) \tag{4.8}$$

Under these same homogeneity assumptions, Wardrop [9] showed that the speeds of successive vehicles crossing a point should be <u>harmonically averaged</u> to yield the space-mean speed on the road. (See also Breiman [8], Gershwin [10], Kurkjian, *et al.*[17].) That is

$$\bar{v}_s(x,\Delta x) = \frac{N(t,T)}{\sum_{j=1}^{N(t,T)} \frac{1}{v_j}} \tag{4.9}$$

where, as before, v_j, $j = 1,2,\ldots,N(t,T)$ represents the sequence of successive vehicle speeds crossing a detector station located anywhere within the section $[x, x+\Delta x]$.

Substituting Eq. (4.9) and Eq. (4.2) into Eq. (4.7) results in

$$\rho(x,\Delta x) = \frac{1}{TL} \sum_{j=1}^{N(t,T)} \frac{1}{v_j} \left[\frac{\text{veh}}{\text{mile}} \text{ per lane}\right] \tag{4.10}$$

Substituting Eq. (3.2) without the acceleration term into Eq. (4.10) yields

$$\rho(x,\Delta x) = \frac{1}{TL} \sum_{j=1}^{N(t,T)} \frac{t_j}{\ell_j+d} \left[\frac{\text{veh}}{\text{ft}} \text{ per lane}\right] \tag{4.11}$$

The omission of acceleration results in little loss in accuracy. Only extremely slow speeds (i.e., under 5 miles/hr) or extremely rapid acceleration causes this term to become significant.

In Eq. (4.11) the presence times, t_j, and the (average) effective loop length, d, are known quantities but the vehicle lengths, ℓ_j, are unknown. In order to circumvent this problem, the ℓ_j are viewed as samples of a random variable, ℓ, with a <u>known</u> probability density function, $f_\ell(\ell)$, and Eq. (4.11) is replaced with its expected value over ℓ. This results in

$$\rho(x,\Delta x) = \frac{5280}{TL} E_\ell\left[\frac{1}{\ell+d}\right] \sum_{j=1}^{N(t,T)} t_j \left[\frac{\text{veh}}{\text{mile}} \text{ per lane}\right] \tag{4.12}$$

where $E_\ell[.]$ denotes expectation over $f_\ell(\ell)$. Note that the 5280 converts

the density value from vehicles/foot to vehicles/mile. Comparing Eq. (4.12) with the definition of occupancy, Eq. (4.3) and Eq. (4.5), (ignoring the end effects $t_{i,I}$ and $t_{i,F}$) an approximate relationship between occupancy and density is seen to exist.

$$\rho(x,\Delta x) = (52.8)\ \mathrm{occ}(t,T)\ E_{\ell}\left[\frac{1}{\ell+d}\right]\left[\frac{\text{veh}}{\text{mile}}\ \text{per lane}\right] \tag{4.13}$$

The density obtained using Eq. (4.13) is actually a time averaged density at a fixed space point and not the desired spatial average density at a fixed time. It is the space-time homogeneity assumption which allows time averages to be equated to spatial averages.

The assumptions and approximations made in deriving Eq. (4.13) are restated and discussed here.

1) The traffic is assumed to be space-time homogeneous. Such an assumption is restrictive, and will be compensated for in the sequel.

2) The harmonic average, Eq. (4.9), is actually an approximation of an expected value. (See Breiman [8]). The accuracy of such an approximation increases with N(t,T). This implies that large time intervals, T, are needed for a given level of accuracy when there are low flow rates.

3) It is assumed that $E_{\ell}\left[\frac{1}{\ell+d}\right]$ in Eq. (4.13) can be determined, given a value of d. A more accurate conversion than Eq. (4.13) could be obtained if d were known as a function of ℓ.

4) Sampling quantization errors $t_{i,I}$ and $t_{i,F}$, are ignored. These should only be significant at low densities or if one is using small averaging time intervals.

5) The vehicle accelerations are assumed to be zero while crossing the detector.

4.3 Evaluation of the Model

A study was made with the microscopic traffic simulation program to see how the density, computed using Eq. (4.13) compares with the actual traffic density. In particular, the study examines the accuracy of Eq. (4.13) as a function of, (1) the section size, Δx, (2) the duration of the time interval, T, and, (3) the type of traffic conditions on the road.

The experiment consisted of an examination of the statistics of the error between the actual spatial density and the density predicted by Eq. (4.13). The test used

1) Values of T ranging from 5 sec. to 1 minute
2) Values of Δx ranging from 100' to 1 mile
3) Traffic flow conditions ranging from low flow (~750 veh/hr per lane) to high flow (~1600 veh/hr per lane) and included homogeneous and inhomogeneous traffic.
4) A value of $E_{\ell}\left[\frac{1}{\ell+d}\right]$ equal to .034 feet. This was obtained from vehicle type distribution information (see [18]) assuming d = 8 feet.

Let $\tilde{\rho}(k)$ denote the density at time step k obtained using Eq. (4.13) and $\rho(k)$ denote the actual density at time step k obtained from the traffic simulation program, and the error between the two densities $e(k) = \tilde{\rho}(k) - \rho(k)$.

The results of the test were:

1) for 5 sec $\leq T \leq$ 10 sec and 100' $\leq \Delta x \leq$ 500'

The error process e(k) was observed to have the following characteristics under all traffic conditions

a) it has zero mean

b) it appears to be uncorrelated in time (i.e., a white process).

Depending on the traffic conditions the variance of the process ranged from 100 to 200 (veh./mile per lane)^2. The small value of T is the cause of the high variance. With small values of T, very few vehicles contribute to the occupancy used in Eq. (4.13). This gives rise to large statistical fluctuations and consequently a large variance. However, the small value of T is also the cause of the apparent whiteness of the process. The fact that this process has zero mean and is white under all traffic conditions will prove to be crucial to the density estimation scheme.

2) for T > 10 sec or Δx > 500',

As T or Δx are increased and traffic remains space-time homogeneous over [x, x+Δx] and [t, t+T], then Eq. (4.13) becomes more accurate. The larger value of T results in more vehicles contributing to the averaging approximation used in Eq. (4.13). Consequently, the variance of the error process drops and remains zero mean. The error process becomes more correlated in time (less nearly white) as T increases.

From these results we see the following:

1) In order to use Eq. (4.13), and to have uncorrelated errors, we must use timesteps on the order of 5 or 10 seconds, and

2) The value of α (in Eq. (2.4)) is 52.8 $E_{\ell}\left[\frac{1}{\ell+d}\right]$.

We know the variance of the error associated with the conversion is between 100 and 200 (veh/mil per lane)^2. Thus, averaging two such occupancies together, as in Eq. (2.4) results in an error, $\eta(K)$, which is a white, zero-mean process with a varaince which may range from 50 to 100 (veh/mile per lane)^2 in homogeneous conditions. This computation assumes that the five second occupancies at neighboring detector stations are independent random variables.

It is not sufficient to characterize $\eta(k)$ under homogeneous conditions only. Inhomogeneities do occur, as a result of accidents or sudden onsets of traffic when sports events or factory shifts end or other causes. Such inhomogeneities can result in biases developing in the occupancy-density relationship. For the approach to density estimation described here to be useful, it must be possible to identify such inhomogeneities and to prevent them from adversely affecting the estimate. A method for doing this is discussed in Section 6.

5. THE KALMAN FILTER

5.1 Filter Design

In this section a Kalman filter for the estimation of $\rho(k)$ is constructed which is based on equations (2.1) - (2.4), the statistical properties of $v(k)$ (Section 3) and $\eta(k)$ (Section 4) and the value of α (Section 4). The Kalman filter is simply a recursive system which produces the optimal estimate of $\rho(k)$ based on the measurements ($z(j)$ up to time $j=k$, the model (2.1) - (2.4) and the statistics of the noise. The interested reader is referred to [11] for the development of the Kalman

filter in general. Applying these results here, we obtain the following scalar filter equation:

$$\hat{\rho}(k+1) = [1 - H(k)]\,\hat{\rho}(k) + H(k)z(k) + u(k) \tag{5.1}$$

Here, $\hat{\rho}(k)$ is the estimate of $\rho(k)$, and $u(k)$ denotes the observed density change from detector counts and $Z(k)$ is an observation derived from (4.13) The time varying Kalman gain, $H(k)$, is given by the recursive relations

$$H(k) = \frac{\sigma^2(k)}{\sigma^2(k) + R} \tag{5.2}$$

$$\sigma^2(k) = \sigma^2(k-1) + Q - \frac{[\sigma^2(k-1)]^2}{R + \sigma^2(k-1)} \tag{5.3}$$

in which $\sigma^2(k)$ denotes the variance of the density estimation error at time k. Hence, $\sigma^2(0)$ is an indicator of the initial uncertainty about $\rho(0)$. Also, Q is the variance of $v(k)$ and R is the variance of $\eta(k)$, as discussed in Table 2 and Section 4.4, respectively.

Other parameters which are of interest are the steady-state Kalman gain, H, and $\Sigma(k)$, the variance of the filter residuals, $r(k)$, where $r(k) \triangleq z(k) - \hat{\rho}(k)$. These are given by

$$H = \lim_{k\to\infty} H(k) = \frac{Q + \sqrt{Q^2+4QR}}{Q + \sqrt{Q^2+4QR} + 2R} \tag{5.4}$$

and

$$\Sigma(k) = \frac{R}{1-H(k)} \tag{5.5}$$

5.2 Filter Characteristics

One can see from Eq. (5.1) that the Kalman gain H(k) determines the weighting between the old estimate, $\hat{\rho}(k)$ and the new observation which is to be used in arriving at the new density estimate, $\hat{\rho}(k+1)$. The value of H(k) depends upon Q,R, k and $\sigma^2(0)$. Thus, if the initial uncertainty about $\rho(0)$ is high, then H(k) is very close to unity for small values of k. This means that the filter algorithm (Eq. 5.1) weights the observations z(k), obtained from occupancy, heavily at first, until the system locks on to the actual density level. Because $\eta(k)$ has zero mean, we are assured that the filter will lock on to the true density.

As time increases (5.3) implies that $\sigma^2(k)$ decreases and H(k) will tend toward the steady state value given by Eq. (5.4). Using R = 100 and a = .01, we see that H = .031. Thus, in steady state, the filter almost ignores the observations (2.4) and relies almost entirely upon the vehicle count information given by u(k). Simulation studies have shown that, using five second time steps and large initial uncertainty in the initial density, the filter can lock on to the correct density within one minute (see Section 7 for a discussion of experimental results). This eliminates any need for *a priori* knowledge of the initial link density, $\rho(0)$. This is viewed as a major advantage of our system when compared to previous systems such as those of Nahi [2], [3].

A well known property of the Kalman filter algorithm when the observation noise sequence is white is that the innovations or residual error sequence

$$r(k) = z(k) - \hat{\rho}(k) \tag{5.6}$$

is also a zero-mean white uncorrelated random process. If it is known that $\eta(k)$ develops a bias during periods of inhomogeneous flow, $r(k)$ will differ from a white process in a predictable way. Note that this can result both from heavy recurrent traffic flow conditions or more severe incident conditions. In either case, we show in Section 6 how techniques of generalized likelihood ratio (GLR) can be exploited to (1) Detect the onset of a bias and (2) Compensate the Kalman filter algorithm, (5.1). In Section 8 we present some results that indicate that our technique (Kalman filter plus GLR system) is capable of tracking $\rho(k)$ under any traffic conditions.

6. COMPENSATION FOR INHOMOGENEOUS CONDITIONS VIA THE GLR METHOD

In this section, generalized likelihood ratio (GLR) failure detection methods developed in [12, 13] are adapted to traffic density estimation. As mentioned in Section 4, traffic inhomogeneities may cause the system modeled by (2.1) and (2.3) to develop an unmodeled bias in the observations $z(k)$. In this section we develop a technique for detecting such a bias and for compensating the Kalman filter based estimation system to account for the bias. In the next section we present the results of some of our studies into the nature and magnitude of biases that do develop in the occupancy-density relationship, while in Section 8 we present results for the Kalman filter-GLR system under homogeneous and inhomogeneous conditions.

The essence of the GLR method is as follows. Under no-bias conditions, the residuals (Eq. 5.6) of the Kalman filter (Eq. 5.1) are a zero-mean, white, Gaussian process [11]. If a bias suddenly occurs in Eq. (2.3), then

the residual process will change in character. The exact nature of this change is called a signature and can be computed off-line. The residual process is monitored and statistically tested for the presence of the bias signature. If a bias is detected, then the current estimate and future observations are corrected.

In the development of the GLR system to follow, it is assumed that the Kalman filter is in steady state. That is, the Kalman gain, $H(k)$, and the residual variance, $\Sigma(k)$, are both constants (see Equations (5.4) - (5.5). This steady state assumption greatly simplifies the analysis of the algorithm and is equivalent to assuming that the initial transient in the density estimate, due to uncertain initial density knowledge, has died out.

Adapting the development in [12] to our problem, we assume that it is possible that at some unknown time θ, the measurements develop a bias of an unknown magnitude b. That is

$$z(k) = z_1(k) \qquad k<\theta \tag{6.1}$$

$$z(k) = z_1(k) + b \qquad k\geq\theta$$

where the subscript "1" is used to denote the value of $z(k)$, according to (2.3), that would be observed if no bias were present. Because the Kalman filter is linear, the residuals and estimates can also be broken up into two parts

$$\hat{\rho}(k) = \hat{\rho}_1(k) + \hat{\rho}_2(k) \tag{6.2}$$

$$r(k) = r_1(k) + r_2(k)$$

where again $\hat{\rho}_1$ and r_1 are the estimate and residuals that would occur if no bias occurs, while $\hat{\rho}_2$ and r_2 are the effects of the bias, as defined in (6.1), on $\hat{\rho}$ and r, respectively,. Thus, if no bias is present in z, $\hat{\rho}=\hat{\rho}_1$, $r=r_1$, the estimate $\hat{\rho}$ is a good estimate, and the residuals r_i are zero mean and white. However, if a bias b at time θ develops, biases $\hat{\rho}_2$ and r_2 develop in both the estimate and residuals. Using the linearity of (5.1), these biases can be readily calculated as a function of b and θ from (6.1) and (5.1):

$$\hat{\rho}_2(k) = \sum_{j=\theta}^{k} (1-H)^j H \quad b \triangleq F(k-\theta)b, \quad k \geq \theta \tag{6.3}$$

$$r_2(k) = [1-F(k-\theta)]b \triangleq G(k-\theta)b, \quad k \geq \theta \tag{6.4}$$

Of course $\hat{\rho}_2(k)$ and $r_2(k)$ are both zero before the bias develops in $z(k<\theta)$.

The GLR algorithm is based on the solution of the problem of deciding between two hypotheses H_0, that no bias is present, or H_1, that a bias has occurred. Note that from what we have just seen, these hypotheses are equivalent to

$$H_0: \quad r(k) = r_1(k) \tag{6.5}$$

that is, the residuals are zero mean and white, and

$$H_1: \quad r(k) = G(k-\theta)b + r_1(k) \tag{6.6}$$

that is, the residuals contain a bias that depends upon θ and b. Following [12] and [14], the solution to this decision problem is the following: assuming that H_1 is true and assuming a bias initiation time θ, calculate the most likely bias size $\hat{b}(k,\theta)$, based on data up to time θ. Then,

assuming this value is valid, calculate the likelihood that such a bias actually did develop at the time θ. The required calculations are as follows: let

$$c(k-\theta) = \frac{1}{\Sigma}\left[G^2(0) + G^2(1) + \ldots + G^2(k-\theta)\right] \tag{6.7}$$

$$d(k;\theta) = \frac{1}{\Sigma}\left[G(0)r(\theta) + G(1)r(\theta+1) + \ldots + G(k-\theta)r(k)\right] \tag{6.8}$$

Here $d(k;\theta)$ is simply a correlation of the failure signature $G(j)$ with the residuals. Note that if a bias did develop at time θ, then $bG(0)$ would be exactly the bias in $r(\theta)$, $bG(1)$ the bias in $r(\theta+1)$, etc. Thus, $d(k,\theta)$ should be large in magnitude if a bias actually did develop. Also, since Σ is the variance of $r_1(k)$, $d(k;\theta)$ is scaled by the ambient level of background noise in $r_1(k)$. Also $c(k-\theta)$, which can be precalculated, essentially measures the ratio of energy in the bias part of the residual sequence $r(\theta), r(\theta+1), \ldots, r(k)$ (assuming a bias is present) to the energy in the background noise. Thus it is a measure of the amount of information available in this set of residuals concerning the occurrence of a bias at time θ).

The estimate $\hat{b}(k,\theta)$ is given by

$$\hat{b}(k,\theta) = \frac{d(k,\theta)}{c(k-\theta)} \tag{6.9}$$

and the measure of the actual likelihood that a bias did develop is the normalized simplified GLR statistic [14]

$$\ell_s(k,\theta) = \frac{d(k,\theta)}{\sqrt{c(k-\theta)}} \tag{6.10}$$

The validity of (6.9) follows from (6.8) and the observation that if a bias of size b did develop at time θ, then the expected value of $d(k;\theta)$ would be $bc(k-\theta)$. Equation (6.10) simply normalizes the correlation $d(k;\theta)$ by the deterministic signal to noise ratio $c(k-\theta)$. Intuitively if the signal to noise ratio is very large, then unless d is also very large, the likelihood of a bias having developed is quite small.

The preceding discussion presents an intuitive picture of what the GLR algorithm does in generating estimates (6.9) and likelihood measures (6.10). For a detailed technical derivation, see [12, 14]. We now turn to the issue of how to use these variables in detecting biases and in compensating for them in our density estimation system. First note that, assuming that $\eta(k)$ and $r(k)$ are Gaussian, then $d(k,\theta)$ and $\ell_s(k,\theta)$ are also Gaussian random variables. Using the fact that $r_1(k)$ is zero mean and white with variance Σ; we can calculate the mean and variance of d and ℓ under the hypothesis H_0 and H_1 for any assumed value of b and θ:

$$E[d_0(k,\theta)\,|\,H_0] = E[\ell_s(k,\theta)\,|\,H_0] = 0 \tag{6.11}$$

$$E[\ell_s(k,\theta)\,|\,H_1,b,\theta] = \frac{E[d_s(k,\theta)\,|\,H_1,b,\theta)]}{c(k-\theta)} = b\,c(k-\theta) \tag{6.12}$$

$$\text{var}[d(k,\theta)] = c(k-\theta) \tag{6.13}$$

$$\text{var}[\ell_s(k,\theta)] = 1 \tag{6.14}$$

where (6.13), (6.14) hold under either hypotheses H_0 or H_1. Therefore, we see that the effect of a bias on $\ell_s(k,\theta)$ is simply to shift its mean from

zero.

This suggests a decision rule

$$|\ell_s(k,\theta)| > \varepsilon \Rightarrow H_1 \tag{6.15a}$$

$$|\ell_s(k,\theta)| \leq \varepsilon \Rightarrow H_0 \tag{6.15b}$$

where the sign of ℓ_s is the same as that for d and hence the same as that for the estimate $\hat{b}(k,\theta)$ in (6.9).

The decision rule (6.15) is not quite the final one we want, as there are still several questions:

- In principle we must calculate $\ell_s(k,\theta)$ and $d(k,\theta)$ for every value of $\theta > k$. This not only means that we have a growing amount of computation but also that (6.15) must be modified since we have $\ell_s(k,\theta)$ for <u>many</u> values of θ.
- A criterion for choosing ε must be developed.

The choice of ε involves the tradeoff of probability of false alarm v. probability of missed detection. For example, if we assume for simplicity that we are looking for a positive bias, then

$$\beta = P[\text{declare } H_1 | H_0 \text{ is true}] = P[\ell_s(k,\theta) > \varepsilon | H_0] \tag{6.16}$$

$$\gamma = P[\text{declare } H_0 | H_1,\theta,b] = P[\ell_s(k,\theta) < \varepsilon | H_1,\theta,b] \tag{6.17}$$

These probabilities are illustrated in Figure 5. From this figure we see that the two quantities that control β and r are ε and the mean value,

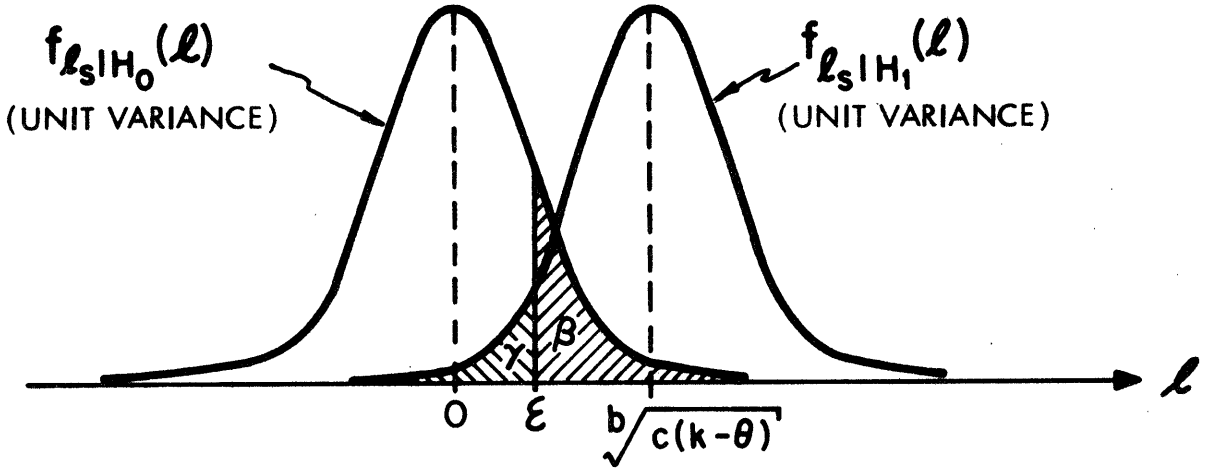


Figure 5. Probability Density Functions of the Likelihood Ratio Under Hypotheses H_0 and H_1

$b\sqrt{c(k-\theta)}$, of ℓ_s under the hypothesis H_1. As the bias is the unknown we wish to determine and ε is to be chosen, only $c(k-\theta)$ is fixed. Therefore, let us examine the behavior of $c(k-\theta)$.

From Eq. (6.7), it is evident that $c(k-\theta)$ is a monotonic increasing function of $k-\theta$. However, $c(k-\theta)$ does not diverge, but converges to a limiting value as $k-\theta$ increases. This is intuitively clear from the following argument.

A bias starting at time θ in the observations will cause the filter estimate to develop a bias gradually. That is, the filter estimate will eventually follow the observations, and consequently the effect of the bias, as measured by $G(k-\theta)$ will tend to zero. This can be seen from direct calculations of $G(k-\theta)$ from (6.3), (6.4)

$$g(j) = (1-H)^j \tag{6.18}$$

Since $0 < H < 1$, $G(j) \to 0$ as $j \to \infty$, and the following limit can be calculated from (5.4), (5.5), and (6.18):

$$C_\infty = \lim_{n\to\infty} C(n) = \lim_{n\to\infty} \frac{1}{\Sigma} \sum_{j=0}^{n} G^2(j) = \frac{1}{\Sigma[1-(1-H)^2]} = \frac{1-H}{RH(2-H)} \tag{6.19}$$

Further C_∞ can be expressed entirely in terms of Q and R using (5.4).

From Table 2 we see that a typical value of Q is .10. For a value of R taken to be 100, this results in H = .031 and C_∞ = .16. Figure 6 is a plot of C(n) versus n. Since the expected value (under hypothesis H_1) of $\ell_s(k,\theta)$ is proportional to $c(k-\theta)$, we see that up to a point, the longer we wait, the larger the mean value, and thus, from Figure 5, the

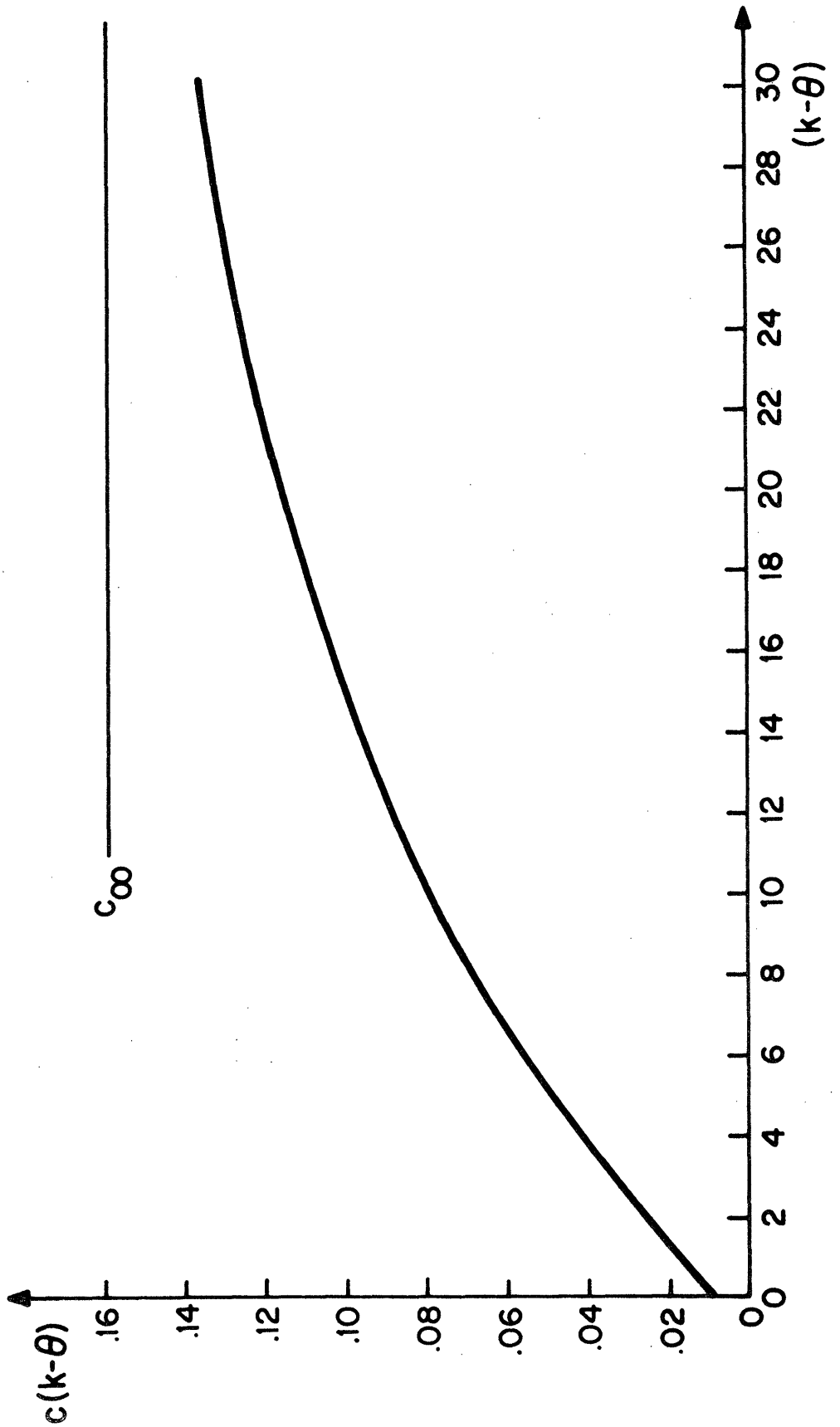


Figure 6. Plot of $c(k-\theta)$ Versus $k-\theta$ ($Q=.1$, $R=100$)

hypotheses H_0 and H_1 separate, reducing β for a fixed ε, or equivalently allowing us to increase ε, thereby reducing r while not increasing β. However, the rate of increase in C(n) decreases rather markedly as n increases, reflecting the fact that there is not much information in r(k) about a bias that develops at time θ for $k-\theta$ large. Therefore, there is some middle range of values of $k-\theta$ for which $c(k-\theta)$ is large but such that we have not yet reached the region of diminishing returns. Then at any time k we will only look for the onset of a bias at times θ so that $k-\theta$ is in this range. In our implementation of the GLR algorithm, bias detections are only made for

$$9 \leq k-\theta \leq 13 \tag{6.20}$$

Rewriting this in the form

$$k-13 \leq \theta \leq k-9 \tag{6.21}$$

we see that the GLR system has a "sliding window" of times θ at which it looks for a bias. The placement of the window can be interpreted as saying that we need at least 9 time steps but no more than 13 to be able to be certain of our decision (H_0 or H_1). This implies a mean time to detection of approximately 50 sec. Therefore the actual GLR system calculates at each time k $\ell_s(k,\theta)$ for θ in the range given in (6.21), thereby eliminating the problem of growing calculations. The largest of these is chosen, and the corresponding value of θ is denoted by $\hat{\theta}(k)$. Then the decision rule is

$$\begin{aligned} |\ell_s(k,\hat{\theta}(k))| > \varepsilon \quad &\Rightarrow H_1 \\ |\ell_s(k,\hat{\theta}(k))| < \varepsilon \quad &\Rightarrow H_0 \end{aligned} \tag{6.22}$$

We now turn to the issue of threshold determination, that is the evaluation of (6.16), (6.17) for θ in the range given by (6.21). Since $c(k-\theta) \simeq .08$ over this range, this constant value was used in our analysis. Fixing a value of ε directly fixed the value of β and specifies r as a function of b(given that $c(k-\theta) \simeq .08$). Figure 7 is a plot of β as a function of a, while Figure 8 is a plot of γ versus b for different values of ε. From the analysis described in the next section we found that the size of biases due to spatial inhomogeneities vary with flow level, and that thresholds ranging from 2.5 at low flows to 3.6 in heavy flow produced good β-r tradeoffs. Since flow can be measured quite accurately from car count data, flow-scheduled gains are quite feasible.

Once a bias has been detected at time $\hat{\theta}(k)$, we must adjust the estimate $\hat{\rho}(k)$ to remove the effect of the bias and must compensate by subtracting the bias from future incoming measurements. From (6.9) we have that our best estimate of b is

$$\hat{b}(k) = \hat{b}(k,\hat{\theta}(k)) \tag{6.23}$$

and thus, from (6.3), our best estimate of the induced bias in our estimate of $\hat{\rho}$ is

$$\hat{\rho}_2(k) = F(k-\hat{\theta}(k))\hat{b}(k) \tag{6.24}$$

Note that this is our <u>estimate</u> of the bias in the estimate, and therefore we should remove it. Thus, following detection of a bias we correct the estimate

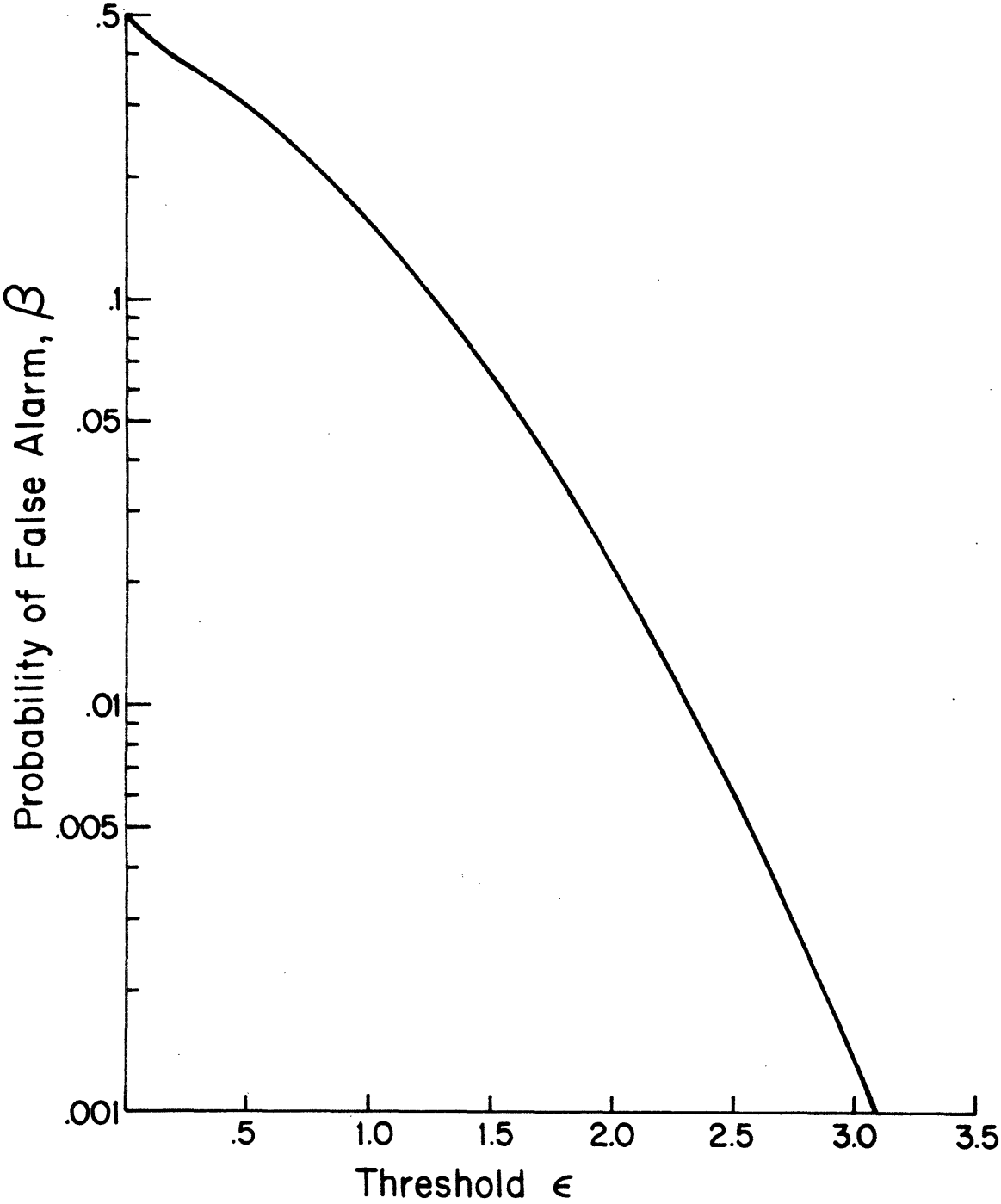


Figure 7. Probability of False Alarm Versus Threshold

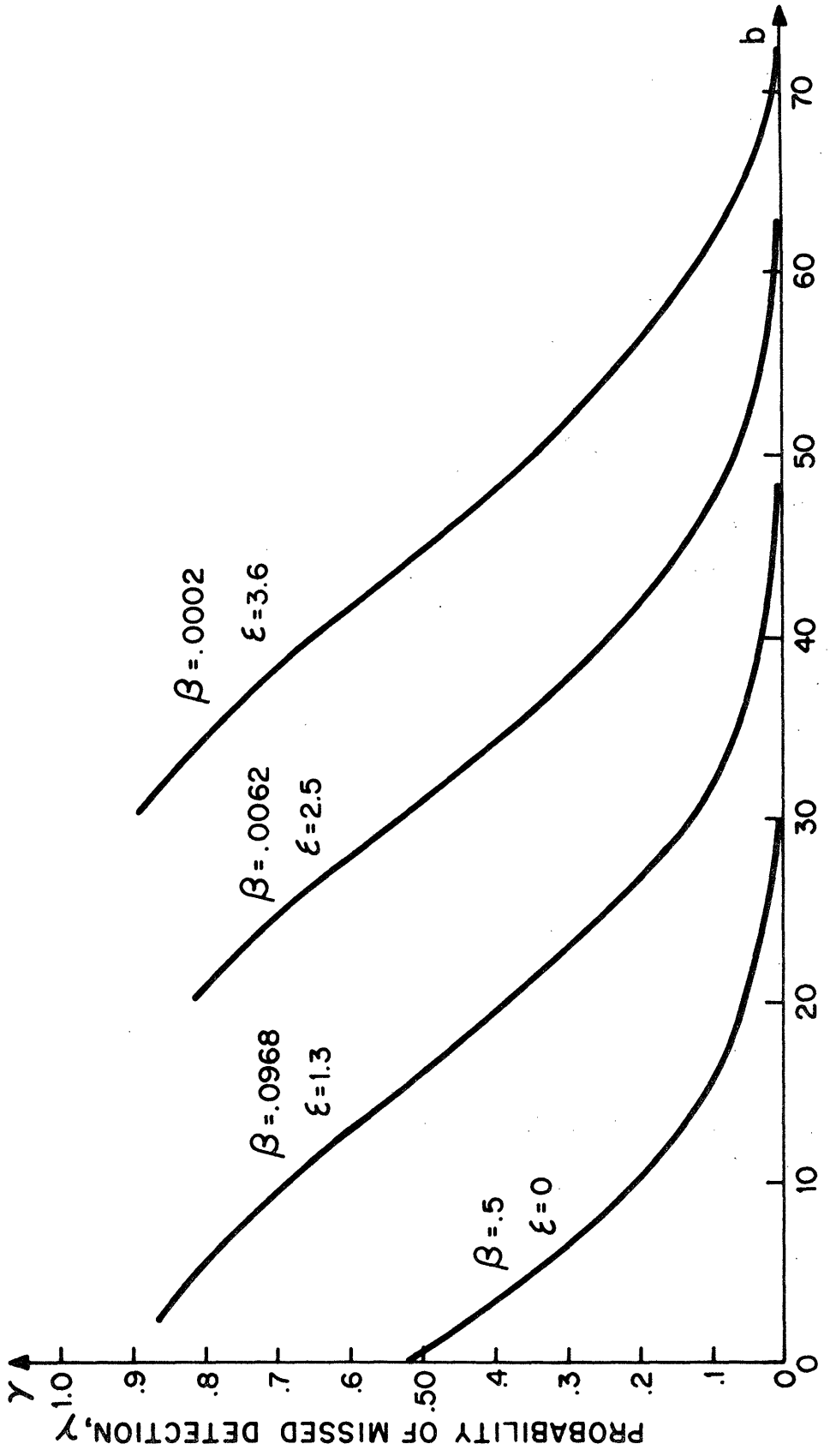


Figure 8. GLR Detection Performance (Time-To-Detect = 50 Sec.)

$$\hat{\rho}_{new}(k) = \hat{\rho}_{old}(k) - \hat{\rho}_2(k) \tag{6.25}$$

and subtract the bias from the failure observations

$$z(j) = \frac{\alpha}{2}[OCCUP(j) + OCCDOWN(j)] - \hat{b}(k), \quad j \geq k \tag{6.26}$$

The system can then continue to track density and to detect additional biases due to further spatial inhomogeneities, due to incorrect estimation using (6.23) of this first bias, or due to a return to homogeneous conditions. In this way, the system can adapt to continually changing conditions.

Computationally, this GLR algorithm requies that the likelihood ratio at time k, $\ell_s(k,\theta)$, be computed for $\theta = 0,1,2,\ldots,k$. This results in a continually increasing number of calculations per time step. However, it makes no sense to try to detect a bias when $k-\theta$ is, say, 120 (i.e., 10 minutes) since the $c(k-\theta)$ curve has essentially leveled off by this time. Therefore, waiting longer will not result in a detection. Similarly, one should not declare a bias to have occurred when $k-\theta$ is only 2 (i.e., 10 seconds) because it is not necessary to respond this quickly and a small amount of additional delay greatly reduces the false alarm probability. This imples that detections should be restricted to a sliding time window. In the version of this system simulated, detections are only made after 9 time steps but before 13. This corresponds to a time-to-detect between 45 and 65 seconds, and a fixed number of calculations at each time step.

In conclusion, we have developed an algorithm which monitors the residuals of the Kalman filter and looks for a signature characteristic of a bias

occurring in the observations. The simplified generalized likelihood ratio, ℓ_s, is the statistic which indicates the occurrence of the bias. If ℓ_s crosses a predetermined threshold the bias is detected and its magnitude and time of occurrence are estimted. The error in the current estimate due to the bias is corrected for, and the system continues to estimate section density. By using a small sliding window for GLR, the resulting algorithm is quite simple computationally. Figure 9 is a block diagram of the entire density estimation algorithm.

7. ANALYSIS OF THE SIZE OF BIASES THAT DEVELOP IN THE OCCUPANCY - DENSITY RELATIONSHIP

It was noted in Section 4 that the biases which occur in the observations are due to spatial inhomogeneities in the traffic. The GLR bias detection algorithm is able to detect the occurrence of a bias, or, equivalently, of a spatial inhomogeneity. Although biases can result from heavy recurrent congestion, most bias-producing inhomogeneities are associated with accidents or similar lane blocking incidents. Hence, the detection of a bias is, most often, the detection of an incident. An example of a non-incident traffic condition which causes a bias is a spatial inhomogeneity such as a curve or grade in the road. Just upstream of such a section, the traffic typically decelerates, giving rise to a higher density than that on the rest of the freeway. A detector station on this section would report a density consistently higher than the actual traffic density and thus a bias would result. Such constant topological sources of spatial inhomogeneity, once identified, can be directly accounted for in the system. For this reason, we are not concerned with them here.

During an incident at low or moderate flow levels, the associated region of congestion grows and reaches a steady-state length. As the region grows, the section density increases proportionally , while the observations z(k) from detector stations may not. This implies that the bias in the observations does not appear suddenly, but grows with time to its final value, b. Thus, modeling the bias as a sudden event is not strictly correct. However, the typical time required for the bias to reach its final value is only about fifteen seconds, so that this error in modeling is not serious. The bias magnitude during a simulated incident in low or medium flow typically ranges from 5 to 20 (veh/mile per lane) and is dependent upon the length of the region of congestion as well as its location relative to the detector stations.

When incidents occur in sufficiently heavy flow, the region of congestion grows without limit. As the region grows, the density increases until the section is totally congested on the upstream side of the accident. The section density then remains approximately constant. The modeling error here is more serious because more time is required for the bias to reach a steady-state value. In fact, it may take in excess of a minute. The bias magnitude during a simulated incident of this type is large and can be as high as 80.

The density estimation algorithm can determine, approximately, the magnitude of the bias that an incident would produce, if one were to occur, by identifying the level of traffic flow. That is, for the incidents simulated in this study, the bias is dependent mainly upon the flow level.

For example, if recent density estimates are, say, around 14 veh/mile/lane, then simulation results indicate that if an accident were to occur, a bias of around 8 would be expected. Knowledge of the expected bias magnitude greatly increases the GLR detection system performance since this information can be used to aid in selecting the threshold, ε. That is, if we expect biases of around 50, then the threshold can be set high so that very few false alarms result. Alternatively, if we expect a bias of only 5, we are forced to lower the threshold in order to detect it and thereby suffer a rise in the false alarm probability. In the simulations of this system, thresholds ranging from 2.5 at low flow levels to 3.6 in heavy flow, were found to produce good detection performance.

Recall that at least 45 seconds but less than 65 seconds are allowed to elapse before a bias is declared. In a heavy flow incident, when the bias requires more than a minute to grow to its final value, more than one bias detection may result. The first will occur before the bias reaches its final value. The estimated bias will be some intermediate value and the compensation will be only temporarily correct. The second detection will occur some 45-65 seconds later and another bias value will be estimated. This second bias estimate, when added to the first, will equal the final bias value, assuming it has been reached by this time. Similarly, when an accident in heavy flow clears, a series of negative bias detections will result if the congestion slowly disappears.

8. PERFORMANCE RESULTS

In this section, the simulation results of the density estimation algorithm shown in Figure 9 are presented. The scenarios selected span a wide variety of traffic conditions. Shown graphically in this section is the estimation performance in accident and non-accident conditions and over a wide range of flow levels. The detections made by the GLR algorithm are examined and shown graphically. It should be realized that the vehicle count data from presence detectors used by the density estimation system are corrupted in the manner discussed in Section 3.

The estimated and actual link density on Link 3 of Simulation 29 are shown in Figure 10. Although the initial estimated density is off by a factor of 4, the filter weighs the observations heavily at first and the estimate drops rapidly down to the actual density. The traffic on Link 3 is extremely light and homogeneous until t-115 sec at which time a large flow of traffic begins to enter the link. Because the vehicle count data is relatively good, the estimate is able to track the sudden rise in density accurately.

Figure 11 is associated with Link 5 of Simulation 28. Although the traffic is inhomogeneous, there is no incident and the GLR bias detection system did not detect a bias. Again, there is a large error in initial conditions. The density drops drastically at t=190 due to two slow upstream

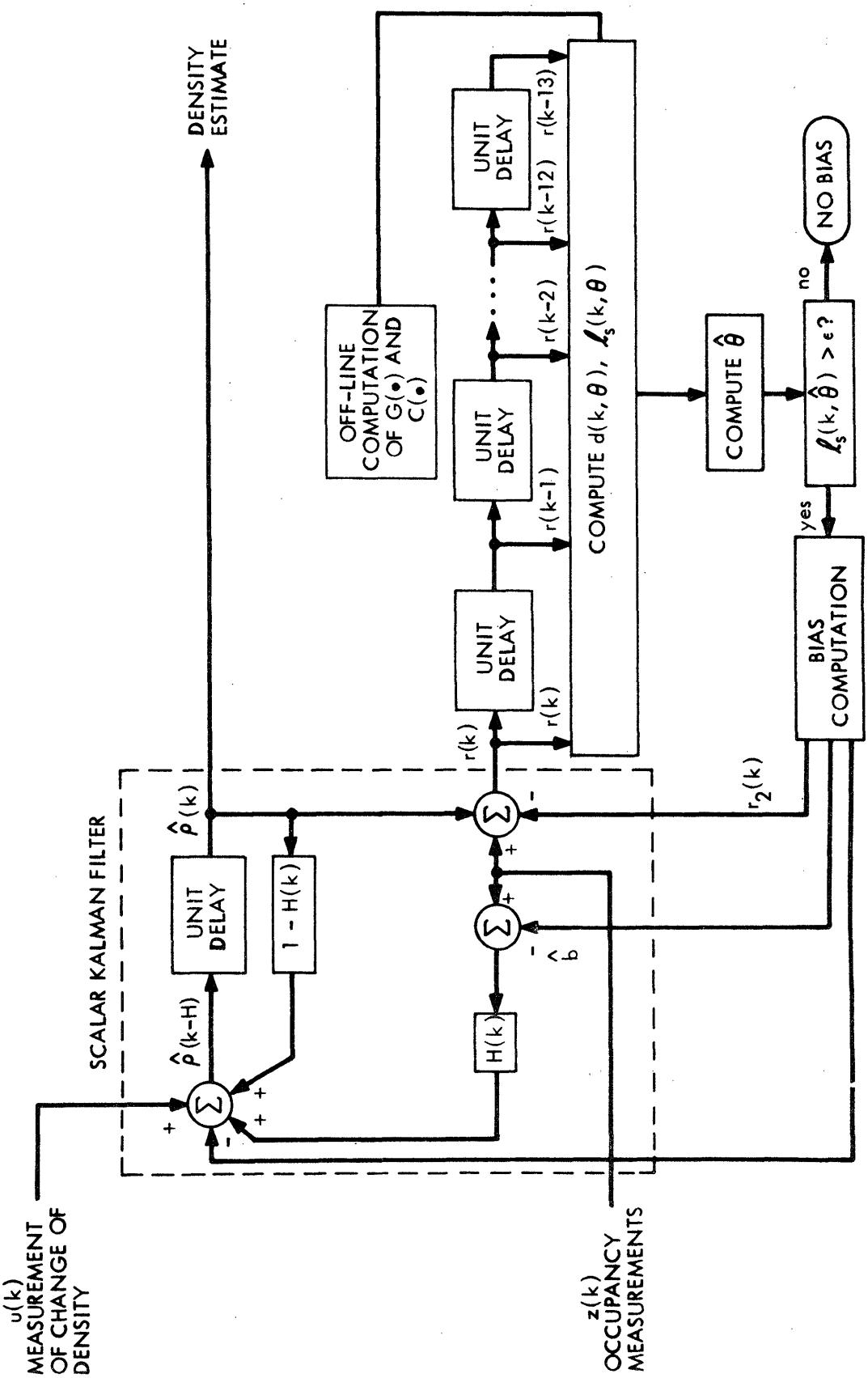


Figure 9. Block Diagram of Density Estimation Algorithm

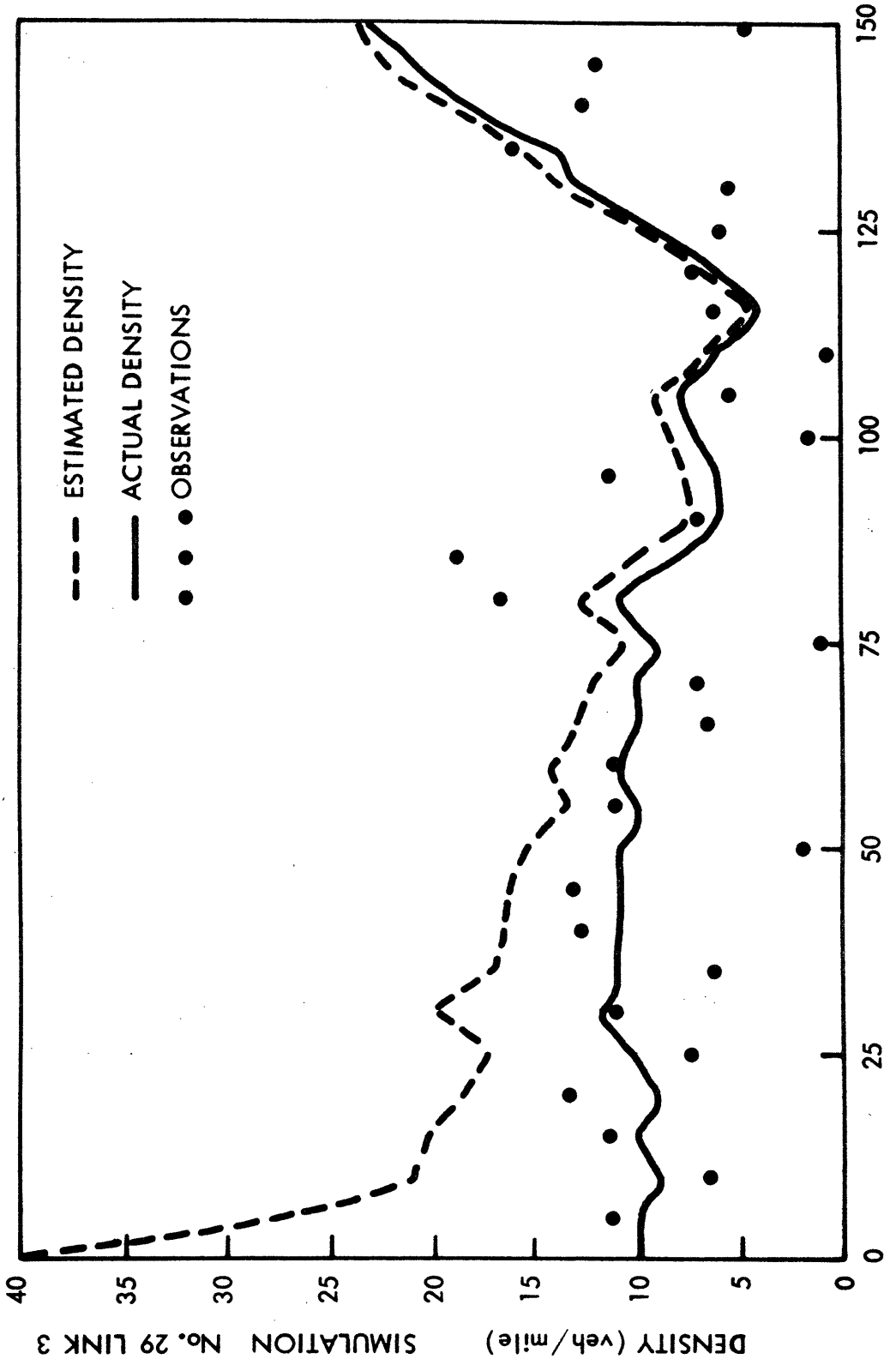


Figure 10. Estimation Performance: Link 3, Simulation 29

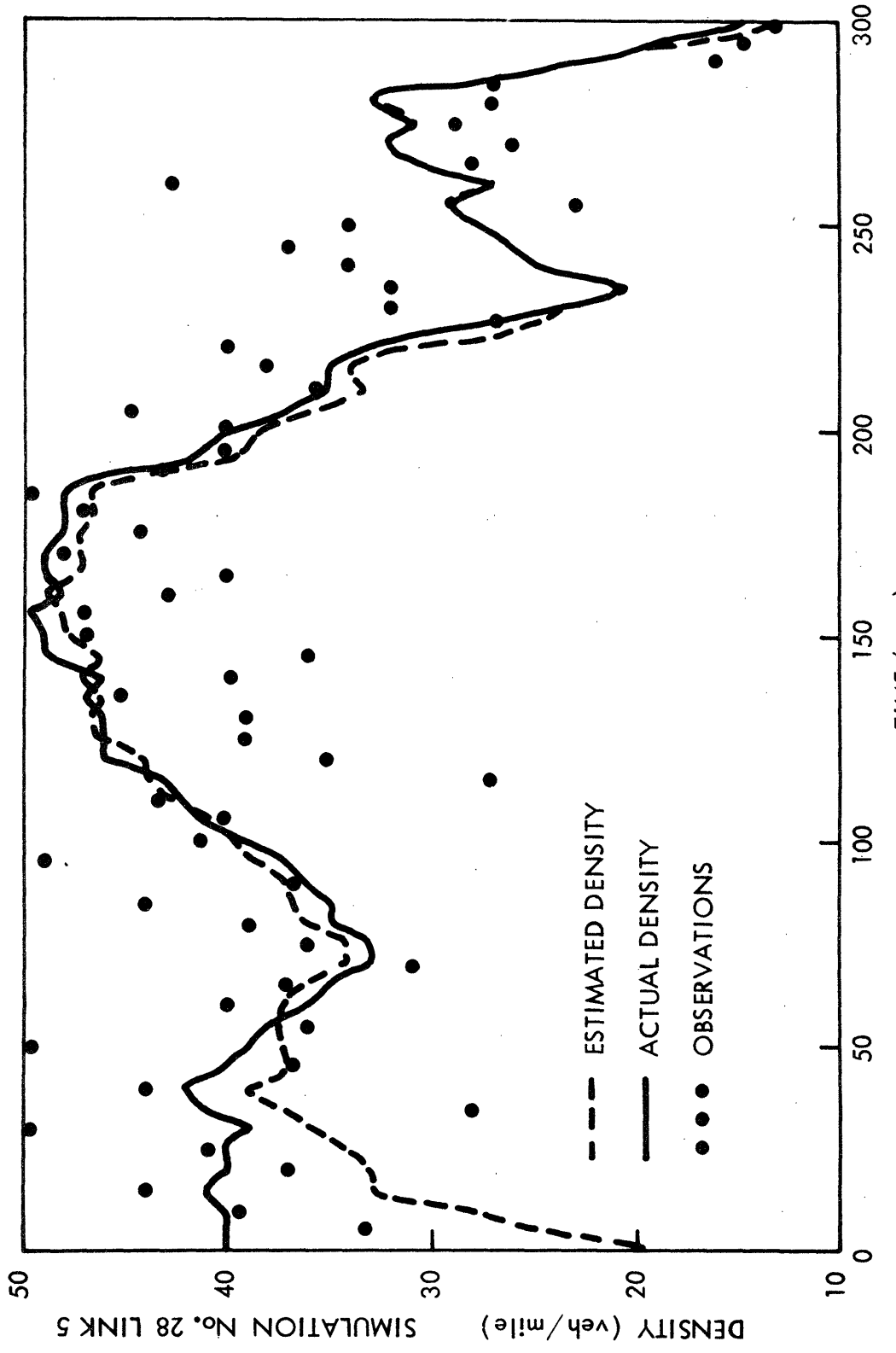


Figure 11. Estimation Performance: Link 5, Simulation 28

drivers clogging up traffic.

The estimated and actual link densities on Link 4 of Simulation 21 are shown in Figures 12 and 13. The traffic is initially very heavy. An incident occurs at t=180 (Figure 12) and the incident clears at t=540 (Figure 13).

It is interesting to note the behavior of the observations in this example. Before the incident, they are scattered above and below the link density, as they were in the non-incident examples of Figures 10 and 11. The occurrence of the incident immediately results in a drastic bias in the observations. This bias is detected at t=240 to be of magnitude 36. The incident occurrence time, θ, was estimated to be 190. The compensation to the estimate was 10.5 and is clearly evident in Figure 12. Because the congestion associated with the incident continued to grow with time, so did the bias and it as detected again at t=295 and again at t=345. The repeated detection and compensation was able to track the density as shown. The estimated bias is subtracted out of the observations at each detection (Section 6.5) which accounts for its step-like rise with time. After the incident is cleared, the observations became biased in the other direction and the detections and compensations made are shown in Figure 13. Thus, the end of the incident was signalled.

Figure 14 is associated with Link 4 of low flow incident Simulation 26. Note that the incident occurs at t=120 but does not really have much effect on the link density until t=250. However, a bias is seen to quickly develop in the observations and a detection and compensation is first made at t=185. Note also that another detection is made at t=310, but that the compensation

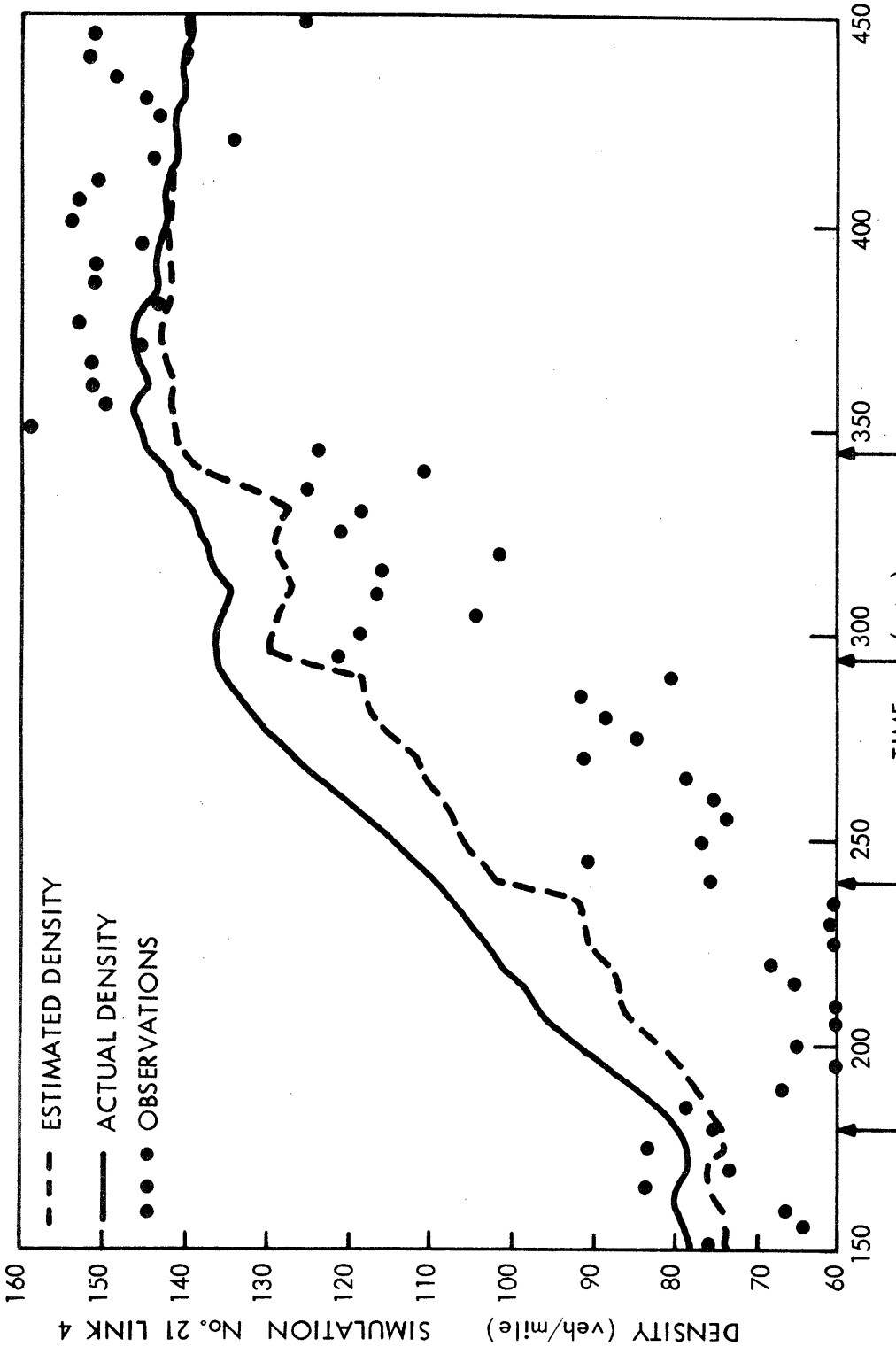


Figure 12. Estimation Performance: Link 4, Simulation 21

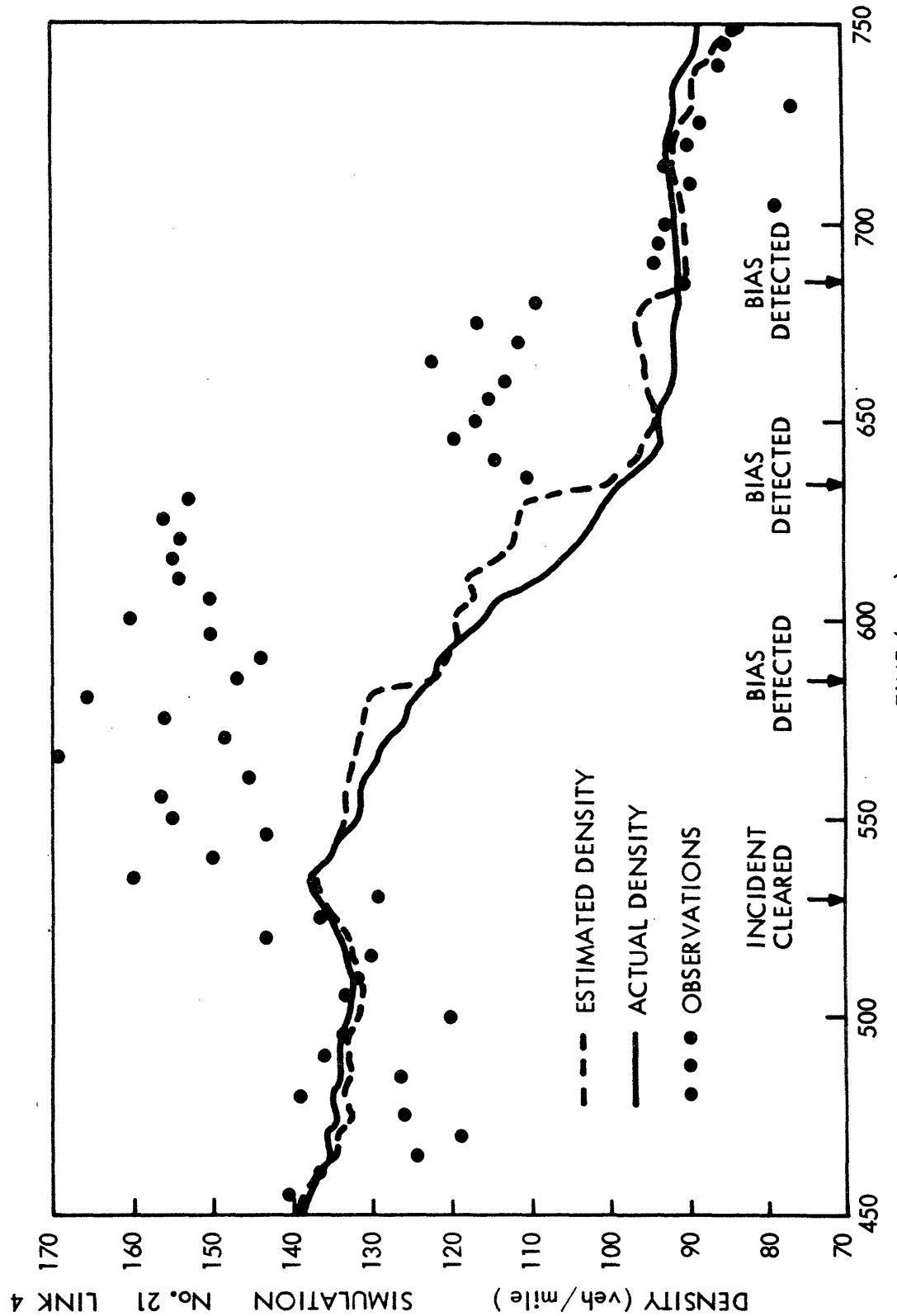


Figure 13. Estimation Performance: Link 4, Simulation 21 (continued)

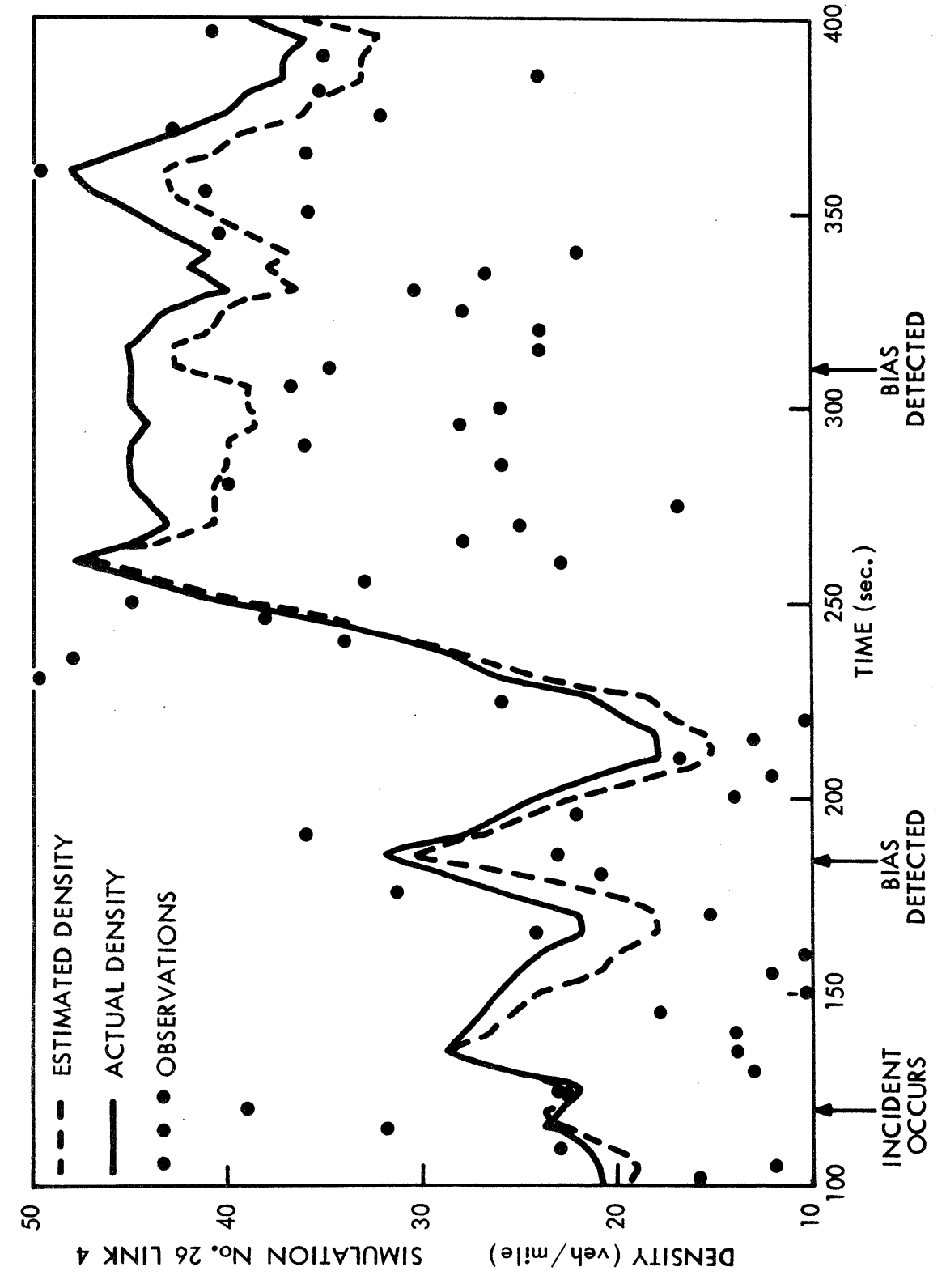


Figure 14. Estimation Performance: Link 4, Simulation 26

results in a bias in the estimated density. If the bias, b, is accurately estimated, then the mean value of the observations will be the actual density and the bias in the estimate will disappear with time.

Table 4 shows the error in the estimates for the simulations of Table 3. It is evident from Table 4 that this density estimation system provides very good estimates in a wide range of flow conditions and in homogeneous as well as inhomogeneous conditions.

Figure 8 apparently indicates that using a threshold between 2.5 and 3.6, as we did, there is a very high probabilitiy of missed detection (especially for small biases). However, no accidents were missed by the GLR bias detection system in the simulation studies. Thus, the simulated system performance seems to be much better than what was predicted analytically. The reason for this inconsistency is in the interpretation of Figure 8. Suppose that one has selected a threshold (i.e., a false alarm probability) and a bias of magnitude b suddenly appears. Figure 8 gives the probability that this bias will not be detected <u>exactly</u> 50 seconds later (assuming it has not already been detected). The actual missed detection probability of the system is the probability that the bias will occur, persist and disappear and not be detected. Using a time detection window from 45 to 65 seconds, the missed detection probability of the system γ_s, with fixed b and ε (or β) is actually:

$$\gamma_s = \text{Prob[missed at t=45]} \cdot \text{Prob[missed at t=50}\,|\,\text{missed at t=45]} \cdot \text{Prob[missed at t=55}\,|\,\text{missed at 45 and 50]} \cdot \text{Prob[missed at t=60}\,|\,\text{missed at t=45, 50 and 55]} \cdot \text{Prob[missed at t=65}\,|\,\text{missed at t=45, 50, 55 and 60]}.$$

TABLE 3

DESCRIPTION OF TRAFFIC SIMULATIONS

SIMULATION IDENTIFICATION NUMBER	INITIAL CONDITIONS ON THE FREEWAY		AVERAGE FLOW RATE (veh/hr per lane)	ACCIDENT INFORMATION			DESCRIPTION
	Density (veh/mile per lane)	Space-Mean Speed (mph)		Link	Lane	Time (sec)	
29	10	63	980	3	2	240	Initially, traffic is very light. At T=40 sec., the input flow becomes heavy. The heavy flow travels downstream and an accident occurs. The queue behind the incident grows at a rate of 8.5 ft/sec. The region of congestions remain entirely within the link.
28	40	40	1590	-	-	-	No accident occurs. From T=140 to T=400 sec., two vehicles travel the length of the freeway abreast of each other at a speed of 40 mph. Faster vehicles upstream are prevented from passing and a vacant gap as much as 1/2 mile long forms in front of the two drivers.

TABLE 3 DESCRIPTION OF TRAFFIC SIMULATIONS (cont.)

SIMULATION IDENTIFICATION NUMBER	INITIAL CONDITIONS ON THE FREEWAY		AVERAGE FLOW RATE (veh/hr per lane)	ACCIDENT INFORMATION			DESCRIPTION
	Density (veh/mile per lane)	Space-Mean Speed (mph)		Link	Lane	Time (sec)	
27	15	55	725	4	2	120	Traffic is light. The region of congestion reaches a steady state length of only 300 ft. The congestion is entirely within the link.
26	20	55	1000	4	2	120	Traffic is light. The region of congestion reaches a steady state length of 650 feet. The congestion is entirely within the link.
22	15	60	815	4	1	180	Traffic is light. The region of congestion reaches a steady state length of only 170 ft. The congestion is entirely within the link.
21	80	25	1625	4	2	180	Traffic is very heavy. The accident causes a region of congestion which grows endlessly. At T=600, the two links downstream of the accident are nearly vacant, while the two links upstream of the accident are filled with stopped vehicles.

TABLE 4

DENSITY ESTIMATION ERROR STATISTICS

SIMULATION IDENTIFICATION NUMBER	THRESHOLD ε	SAMPLE MEAN OF THE ESTIMATION ERROR ON LINK						SAMPLE VARIANCE OF THE ESTIMATION ERROR ON LINK					
		6	5	4	3	2	1	6	5	4	3	2	1
29	3.6	-.5	-2.1	-1.9	-.05	-.94	-3.2	21.	19.	17.	43.	23.	15.
28	3.6	-.4	1.6	3.1	0.34	-.22	-1.5	27.	7.9	3.3	1.4	2.0	6.3
27	3.0	.80	1.4	3.2	.66	-.5	.001	1.2	1.3	5.5	.26	2.0	2.7
26	2.5	-0.2	.73	2.3	2.6	-.39	-1.6	2.0	1.4	2.3	7.6	2.4	6.5
22	2.8	1.9	.13	3.3	1.2	.70	-.66	5.0	5.9	5.7	6.0	10.7	16.8
21	3.5	-.13	3.7	7.4	-1.7	5.6	7.9	8.1	26.8	17.4	19.4	10.2	6.9

Thus, γ_s is much less than the γ given in Figure 8. The calculation of γ_s is difficult due to the correlation between terms in Eq. (7.1). (See [13].)

In conclusion, the GLR bias detection system shows promise as an accident detection system.

9. CONCLUSIONS

A freeway traffic density estimation scheme has been presented in this paper. This scheme has four characteristics which promise to make it practical.

1. Accuracy. The performance results in Chapter 8 indicate that the method estimates density accurately and responds to changes in density quickly. These results, however, were obtained by using simulation. Firm conclusions cannot be drawn without the use of real roadway data.

2. Simplicity. Although the derivation of the method may appear to be complicated, the actual on-line computation is simple. In fact, only the following equations must be evaluated at each time step k for $k-13 \leq \theta \leq k-9$: (2.2),(2.4),(5.1),(6.8),(6.10). Thus at each step, only 69 multipliers must be performed, as well as some adds and compares. (One multiply in (2.2), one in (2.4), two in (5.1). The rest are needed for calculating $\ell_s(k,\theta)$ in (6.8) and (6.10) since for each k and θ, $k-\theta+2$ multiplications are required. Note that a few additional calculations are required when biases are detected.) This means that the method can be implemented in a decentralized fashion, with remote microprocessors communicating only with their neighbors (as well as transmitting results to where they are needed).

3. Robustness. The method does not require homogeneous conditions in which to work. In fact, it detects inhomogeneities and automatically adjusts to them. It does not require accurate initial conditions. This is because the system uses two independent measures of density. That is, car counts (equation (2.1)) and occupancy (equation (2.4)) together provide two different perspectives on the traffic process.

4. Accident detection. As a consequence of the method's ability to detect inhomogeneities, it can be used as an accident detection system. This is because an accident, which blocks lanes and influences driver behavior, causes a disruption in the orderly, homogeneous flow of traffic. However, to look for accidents with this method alone may be risky since there are other causes of inhomogeneity, such as slow drivers or the arrival of a sudden pulse of traffic such as after a popular sports event. It is clear that accidents have characteristics which distinguish them from other disruptions. The methods described in [14] and [15] exploit these characteristics. Thus an attractive accident detection system would be to use the method described here to trigger [14] or [15].

REFERENCES

[1] H.J. Payne, et al., "Evaluation of Existing Incident Detection Algorithms," Interim Report, Technology Services Corp., February 1975.

[2] N.E. Nahi, "Freeway Data Processing," Proc. IEEE, Vol. 61, May 1973, pp. 537-541.

[3] N.E. Nahi and A.N. Trivedi, "Recursive Estimation of Traffic Variables: Section Density and Average Speed," Univ. of Southern California, Los Angeles, California, p. 269-286.

[4] D.C. Gazis, and C.K. Knapp, "On-Line Estimation of Traffic Densities from Time-Series of Flow and Speed Data," Trans. Sci., Vol. 5, No. 3, pp. 283-301, August, 1971.

[5] D.C. Gazis and M.W. Szeto, "Design of Density Measuring Systems for Roadways," pp. 44-52, Highway Research Board Record, No. 388, 1972.

[6] B. Mikhalkin, "Estimation of Speed from Presence Detectors," Highway Research Board Record, No. 388, pp. 73-83, 1972.

[7] W. Phillips, "Kinetic Model for Traffic Flow," Utah State University, Final Report, Contract DOT-OS-40097, 1976.

[8] L. Breiman, "Space-Time Relationships in One-Way Traffic Flow," Transportation Research, Vol. 3, pp. 365-376, Permagon Press, 1969.

[9] J.G. Wardrop, "Some Theoretical Aspects of Road Traffic Research," Proc. Inst. Civil Engrs., Part II, 1, No. 2, pp. 325-362, 1952.

[10] S.B. Gershwin, "On the Relation between Vehicle Flow, Vehicle Density and Velocity Distribution," MIT Electronic Systems Laboratory, Cambridge, Technical Memorandum, ESL-TM-570, September 24, 1974.

[11] A. Gelb, Applied Optimal Estimation, MIT Press, 1974.

[12] A.S. Willsky and H.L. Jones, "A Generalized Likelihood Ratio Approach to State Estimation and Linear Systems Subject to Jumps," Proc. IEEE Conference on Decision and Control, Phoenix, 1974.

[13] E.Y. Chow, "Analytical Studies of the Generalized Likelihood Ratio Technique for Failure Detection," MIT Electronic Systems Laboratory, Report ESL-R-645, 1976.

[14] E.Y. Chow, A.S. Willsky, P.K. Houpt, S.B. Gershwin, "Dynamic Detection and Identification of Incidents on Freeways: Volume IV: Generalized Likelihood Ratio," MIT Electronic Systems Laboratory, Cambridge, 1977.

[15] C.S. Greene, P.K. Houpt, A.S. Willsky, S.B. Gershwin, "Dynamic Detection and Identification of Incidents on Freeways: Volume III: The Multiple Model Method," MIT Electronic Systems Laboratory, Cambridge, 1977.

[16] H.J. Payne, E.D. Helfenbein, H.C. Knobel, "Development and Testing of Incident Detection Algorithms," Final Report, Vol. 2, Federal Highway Report No. FHWA-RD-76, February 1976.

[17] A. Kurkjian, Stanley B. Gershwin, Paul K. Houpt and Alan S. Willsky, "Dynamic Detection and Identification of Incidents on Freeways: Volume II: Approaches to Incident Detection Using Presence Detectors," MIT Electronic Systems Laboratory (Now LIDS) report ESL-R-765, September 1977.

[18] W. Mitchell, "The Estimation and Simulation of Freeway Traffic Flow Using Car-Following and Fluid-Analog Models," MIT Electronic Systems Laboratory, M.S. Thesis, June 1977.