THE ESTIMATION OF TRAFFIC VARIABLES AND DETECTION OF INCIDENTS USING PRESENCE DETECTOR DATA

by

Andrew Loris Kurkjian

B.S. Catholic University of America
(1975)

SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE
DEGREES OF

MASTER OF SCIENCE
and
ELECTRICAL ENGINEER
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
January, 1978

Signature of Author .....
Department of Electrical Engineering and Computer Science, January 13, 1978

Certified by .....
Alan S. Willsky, Assoc. Prof. Stanley B. Gershwin, Lecturer

Accepted by .....
Chairman, Departmental Committee on Graduate Students

ACKNOWLEDGEMENT

This research was conducted at the MIT Electronic Systems Laboratory with support provided by the U.S. Department of Transportation under the OST Program of University Research, Contract No. DOT-OS-60137. We thank Dr. Diarmuid O'Mathuna of TSC and Dr. Michael Rabins of the DOT Office of University Research for their criticisms, remarks and support.

This thesis is based on the report entitled "Dynamic Detection and Identification of Incidents on Freeways, Volume II: Approaches to Incident Detection Using Presence Detectors", MIT Electronic Systems Laboratory, Report ESL-R-765, September 1977. I wish to thank Dr. Stanley B. Gershwin for his constant support, enthusiasm and suggestions. I also wish to thank Professor Willsky for his encouragement and patience.

ABSTRACT

The evolution of incident detection systems has resulted in California Algorithm #7. California Algorithm #7 is typical of existing systems in that it does not exploit a dynamic model of traffic behavior. In addition, raw presence detector data, in the form of occupancy, is used directly. In this report, California Algorithm #7 is shown to exhibit poor performance (e.g., failure to detect some incidents) in low flow conditions.

A study of the occupancy-density relationship reveals that occupancy can be converted into an estimate of traffic density on a section extremely local to the detector under all traffic conditions, and that occupancy may not relate to the traffic conditions as little as five hundred feet away. With detectors conventionally 1/2 mile apart, incidents may not affect the occupancy seen at a detector.

Density is a spatial quantity which is affected more consistently and predictably by an incident than is occupancy. A system is presented which uses vehicle count information as well as occupancy to estimate section density. The system is computationally simple, estimates well at all flow levels and in incident conditions, does not require apriori knowledge of the traffic conditions on the road and is insensitive to the errors in vehicle counts obtained from detectors. The statistics of these errors have been determined. These density estimates can then be used to detect incidents.

Remarks are also included on the estimation of space-mean speed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 PROBLEM STATEMENT

This thesis concerns itself with the general problem of quickly and automatically detecting automobile accidents of freeways. It is not the purpose of this thesis to solve this problem in its entirety. However, the research reported in this study comprises onepart of a larger research effort which, in fact, attacks an even more general problem [1-4]. The goal of the overall research effort is the detection and identification of incidents on freeways. Here, an incident is defined to be any event on the freeway whose occurrence significantly disturbs the flow of vehicles. Hence, accidents and similar capacity reducing events are simply one type of incident. Other types of incidents include debris on the road, disabled vehicles, the sudden appearance of large input flows onto the freeway, etc.

The need for the development of a reliable incident detection and identification for freeways is obvious. For example, such a system's ability to distinguish accidents from non-capacity reducing spatial inhomogeneities (e.g. pulses of flow) allows its incorportation with an emergency vehicle dispatch service. If the system detects accidents more quickly than current methods then injured people can be reached sooner, the road can be cleared sooner and delays to other motorists will be decreased. Also, a freeway traffic control system, whose purpose it is to control the traffic conditions on a freeway network, would require that any incidents be detected quickly, identified accurately and that very few incidents go undetected.

The detection and identification of incidents on freeways requires some type of sensing device along the road so that traffic conditions can be monitored. Aerial photographs, television cameras, radar and ultrasonic devices are types of sensors which are available for this use. Because these devices are expensive and typically require high-bandwidth communications, human operator intervention and complicated processing, they have been eliminated from consideration in this study. Instead, electronic presence detectors are assumed to provide all the observations of the traffic conditions. Presence detectors do not have any of these undesirable qualities and are presently installed on hundreds of miles of freeway in the United States [5].

The general question discussed in this thesis is what to do with the data

provided by presence detectors. In this thesis, presence detectors are modelled and the type of data they provide is examined. Existing incident detection systems, based on presence detector data, are studied to see how they incorporate the data and what the limitations of these systems are. Past efforts to estimate traffic variables from presence detector data are examined and considered for incident detection purposes. Conclusions are then drawn as to what should be done with the available data.

The result is that certain key traffic variables (i.e. spatial aggregate variables) are needed for incident detection purposes and should be estimated from the data. In this thesis methods for estimating two such variables, density and space-mean speed, are developed. It should be noted that these estimation methods stand on their own as important contributions to the literature, outside of the incident detection context. When used for detecting incidents, these variables can be used directly in a simple incident detection scheme. Alternatively, they could also be used as observations in a complex dynamic traffic behavior model. Such a model describes the temporal and spatial evolution of these variables in a modern system-theoretic context. This allows the application of a wide variety of theoretical tools to aid in the detection and identification of incidents. For example, in this research effort, the Multiple Model Identification Method (MM) and the Generalized Likelihood Ratio Method (GLR) have both been sucessfully adapted to this problem. These two systems are not presented in this thesis but can be found in reports by Greene [3] and Chow [4].

The specific topics discussed in this thesis are as follows.

(1) The limitations of existing incident detection techniques are identified.

(2) A study is made of the relationship between variables readily obtainable from presence detector data (i.e. flow and oocupancy) and the spatial aggregate variables of concern here (i.e. traffic density and space-mean speed).

(3) A system is developed for estimating density and space-mean speed that works under far wider conditions than previous systems. This can then be used as a preprocessor for a traffic control system or for a a more sophisticated incident detection system based on macroscopic models.

(See Greene [3] and Chow [4].) It also directly incorporates an "inhomogeneous conditions detector" that can be used as an initial warning of possible incidents.

The remainder of this introductory chapter is outlined as follows. Section 1.2 provides the reader with background information pertinent to the overall project. Presence detectors are described, approaches to incident detection are discussed and traffic flow modelling is summarized. Section 1.3 identifies the problems specifically examined in this thesis. Then, Section 1.4 describes the experimental facilities available for testing purposes.

1.2 BACKGROUND

1.2.1 Presence Detectors

A presence detector is a device buried in the pavement which provides a binary signal indicating the presence or absence of a vehicle in a well defined area on the road. These sensors are typically found at one half mile intervals along the road and normally each lane has a detector [5].

Presence detectors were originally designed for the purpose of counting the number of vehicles which cross them in some time interval. In fact, presence detectors do count vehicles, and consequently, measure the flow rate (vehicles/hour), quite accurately (see Section 2.4).

Occupancy is a measurement also easily obtainable from a presence detector signal. The occupancy, of a particular time interval, at a given detector, is the percent of the time interval that the detector signalled that a vehicle was present. Occupancy qualitatively behaves like density (i.e. concentration) of vehicles on the road in the vicinity of the detector.

Other measurements, such as vehicle speeds, lengths, etc., can be derived from presence detector signals but with much greater difficulty and with less accuracy.

1.2.2 Approaches to Incident Detection Using Presence Detectors

1.2.2.1 Two General Approaches

All incident detection algorithms share the following basic structure.

(1) One or more measurements are periodically derived from presence detector signals.

(2) Models are developed which describe the behavior of these meas-

urements under the hypothesized traffic conditions;

H_0: No incident has occurred

H_1: Incident has occurred.

(3) The actual measurements are then examined to see if they more closely exhibit the behavior described by the incident model or the non-incident model. In this manner, a decision between H_0 and H_1 is made.

This structure naturally implies two possible methods for designing an incident detection system.

Method #1

Use a previously developed traffic behavior model and extract from the presence detector signal the variables required by the model, or,

Method #2

Use any measurement from the presence detector signal and develop a model describing its behavior under incident and non-incident traffic conditions.

1.2.2.2 Our New Approach

Method #1 is actually a new approach to designing an incident detection system. It offers the advantage of using a sophisticated fully-developed model , of which there are several. (For a comprehensive survey of traffic models, see Edie [6].) The disadvantage of this approach is that the estimation of the variables required by the model may be quite difficult and probably requires a model of its own. Assuming that the estimation can be achieved at a reasonable expense, this method appears promising and, in fact, was selected for development in this research effort [3,4].

1.2.2.3 Past Approaches

Incident detection systems using presence detectors have been in existence for over ten years. A survey by Payne in 1975 [5] presented twenty-four detection algorithms, all of which are based upon flow and occupancy measurements from presence detectors. Simple models were developed for each of these algorithms. Thus, existing incident detection algorithms were designed using the philosophy of method #2.

1.2.2.4 California Algorithm #7

The California Algorithm #7 (referred to here as CA-7) is perhaps the most widely accepted incident system in use today. It is based solely

upon occupancy measurements and uses a simple model. Based on the analysis of Payne[5] et.al., the performance (defined in Section 1.2.3) of CA-7 is regarded as the effective upper limit for algorithms designed using this approach.

1.2.3 Performance Criteria for Incident Detection Systems

The false alarm rate, missed detection rate and mean time-to-detect comprise the three major indicators of an incident detection system's performance capabilities. These terms are defined here.

A false alarm occurs when the system signals an incident and there is none. The false alarm rate is the ratio of the number of false alarms to the number of decisions (either incident or non-incident) made by the system in non-incident conditions.

A missed detection occurs when an incident occurs, continues and ends without the system detecting it. The missed detection rate is the ratio of the number of missed detections to the number of decisions made by the system in incident conditions.

The mean time-to-detect is the average amount of time required by the system to make a detection, given that there is a valid detection.

In the overall research effort, the MM and GLR incident detection systems were compared to the California Algorithm #7 by subjecting each system to the same test. This test consists of running each algorithm in a series of incident and non-incident traffic scenarios covering a wide range of traffic flow conditions. The performance results allow a valid comparison among the systems [1].

1.2.4 Traffic Flow Models for Incident Detection

1.2.4.1 Dynamic Versus Static Models

Over the past 30 years a wide variety of traffic flow models have appeared in the literature. (See Edie [6]; see also [7-12].) The more precise models are based upon nonlinear differential equations which describe the spatial and temporal behavior of key traffic variables. Such models are termed dynamic. A static model does not describe spatial or temporal behavior of traffic variables.

Dynamic models, as opposed to static models, allow detailed flow des-

cription over a wide range of traffic conditions. Static models are, in general, computationally simpler than dynamic models. Most existing incident detection systems, including CA-7, are static in nature [5].

1.2.4.2 Microscopic vs. Macroscopic Models and Aggregate Variables

Any traffic model is classified as either a microscopic or a macroscopic model. A microscopic model takes individual vehicle behavior into account. An example of one such model is known as the car-following model [13]. In this model, the position, speed and acceleration of a vehicle is determined as a function of its type and its driver's characteristics as well as the position, speed and acceleration of the vehicle directly ahead. Because presence detector data is microscopic in nature (i.e., the detector signal contains information about individual vehicles), it would seem that a microscopic model is well suited for incorporation with the data. This is not the case since the pulses resulting from the passage of vehicles over the detectors cannot be identified with the vehicles which produced them. Even if each vehicle could be tracked, the computational burden of such an approach would be overwhelming and not necessary for the purposes of incident detection.

Instead of tracking each vehicle, consider dividing the freeway spatially into sections and periodically monitoring the behavior of the groups of vehicles within each section. This group behavior is commonly described by aggregate variables. Aggregate variables approximate the large number of individual vehicle characteristics by significantly fewer quantities which represent the average characteristics. The average speed of the vehicles within each section (space-mean speed) and the number of vehicles in each section (which is proportional to section density) are examples of aggregate variables

Macroscopic models describe the behavior of these aggragate variables and are a more natural choice for incident detection purposes.

1.2.4.3 The Dynamic Macroscopic Payne-Isaksen Model

The Payne-Isaksen Model [14] has become an extremely popular macroscopic model in recent years. It is dynamic and has been adopted for use in the MM and GLR incident detection systems. It requires measurements of section density and space-mean speed. The dynamic equations upon which

it is based are analogous to those which describe one-dimensional fluid flow.

1.3 CONTRIBUTIONS OF THIS THESIS

The objective of the overall research effort is the development of a dynamic model-based incident detection and identification system which exhibits significant performance improvements over California Algorithm #7. The contributions to this goal made by this thesis are as follows.

1.3.1 The Occupancy-Density Relationship

If there exists a section of road over which traffic flows smoothly for an interval of time, then an occupancy measurement is proportional to the density of traffic on that section of road. This type of flow condition is termed space-time homogeneity and is formally defined, along with the occupancy-density relationship, in Chapter 3.

An incident, by definition, disrupts the flow of traffic and any measurement of occupancy in inhomogeneous conditions will, in general, not be proportional to density. However, if the interval of time over which occupancy is computed is small enough (on the order of 5 or 10 seconds) then the occupancy converts, under all conditions, via the same proportionality constant, to the density of traffic on an extremely small section of road adjacent to the detector. Thus, the density of traffic in the immediate vicinity of each detector can be monitored using occupancy measurements taken over short time intervals. This result will be used in the density estimation system developed in Chapter 5.

Occupancy measurements taken on a one minute interval are used in CA-7, and, in general, such long time averages do not relate to section density. As we have mentioned previously, in general, occupancies cannot be used to monitor the traffic density a distance five or six hundred feet away from the detector. These results are used in the study of CA-7 presented in Chapter 4.

1.3.2 California Algorithm #7

The evolution of incident detection systems has resulted in California Algorithm #7. Chapter 4 reports on tests of CA-7's behavior under a wide range if incident and non-incident conditions. The algorithm misses many incidents in low flow conditions due mostly to its sole dependence upon

occupancy measurements. False alarms, although rare, are the result of the use of a simple static traffic model. The algorithm performs well in heavy flow conditions.

1.3.3 Density Estimation

The presentation of research into new incident detection methods is initiated in Chapter 5 with the development of a new density estimation scheme. This scheme is later incorporated into two dynamic model based incident detection and identification systems. The incident detection systems themselves are discussed by Greene [3] and Chow [4].

The estimation of section density from presence detector signals has received considerable attention in the literature. The existing density estimation systems are described in papers by Nahi [15], Nahi and Trivedi [16], Gazis and Knapp [17], and Gazis and Szeto [18]. The models upon which these systems are based are not valid under incident or other spatially inhomogeneous conditions. Consequently, such density estimates may not be adequate for incident detection purposes.

A new density estimation system is presented which is computationally simple, estimates well at all flow levels and in incident conditions, does not require any apriori knowledge of traffic conditions and is insensitive to the types of imperfections found in presence detectors. The method uses only vehicle count information and occupancy measurements from detectors, both of which are easily obtained.

1.3.4 Incident Detection Using Density Estimates Only

A new concept in incident detection has developed out of this research. Consider using occupancy to monitor the density on the road in the vicinity of each detector, as described in Section 1.3.1. Now compare this with the section density estimate, using the system described in Section 1.3.3. In steady flow conditions, the two estimates are nearly the same, but under incident conditions, large differences are observed. By monitoring these quantities, a class of incidents can be detected. This method is pursued in Chapter 5.

1.3.5 Vehicle Count Error Statistics

The statistics of the errors made by presence detectors in counting vehicles are determined uning a presence detector model. This model has been

developed from the experimental results of Mikhalkin [19] and is presented in Chapter 2. These new results are used in the estimation system described is Section 1.3.3.

1.3.6 Space-Mean Speed Estimation

The Payne-Isaksen model also requires space-mean speed estimates. Considerable effort has been expended in the past toward such estimation systems [15], [16], [19], [20]. Unfortunately, these systems assume that the speeds of vehicles crossing detectors can be accurately estimated in real time. This estimation of speeds is not a simple matter, as Mikhalkin has shown [19]. Furthermore, in order to use these speeds, a model is required to convert them into space-mean speed estimates. Typically this model resorts to some homogeneity assumptions which are not accurate in incident conditions.

A different method is proposed whereby flow information and the section density estimate are used to estimate space-mean speed.

1.4 FACILITIES FOR EXPERIMENTATION: MICROSCOPIC TRAFFIC SIMULATION PROGRAM

All testing and experimentation in this thesis has been accomplished using a traffic microscopic simulation computer program. This program, developed at MIT [21], is based on the St.John car-following equations [22] and can be used to simulate traffic under almost any conditions. The program operates in discrete time and offers the following features:

(a) a variety of vehicle and driver types can be modelled

(b) presence detectors can be placed as desired

(c) on-ramp and input flow rates can be specified

(d) accidents can be simulated by stopping a vehicle at any desired time and location.

Because the simulation is based on a microscopic model, the position, speed, acceleration, driver type and vehicle type for each vehicle on the road are known and available.

Appendix A describes the distribution of vehicle and driver types assumed in the simulation. Chapter 2 describes the simulation of presence detector signals from data provided by the traffic simulation.

2. SIMULATION OF PRESENCE DETECTOR SIGNALS

2.1 INTRODUCTION

The testing of incident detection systems and aggregate variable estimation systems in this thesis necessitates the simulation of presence detector signals from data provided by the traffic simulation computer program. In this chapter, a deeper look is taken into presence detectors and a model is developed. Section 2.2 surveys the relevant literature and Section 2.3 presents the simulation model. The statistics of vehicle counting errors are described in Section 2.4. These new results are used in the density estimation system of Chapter 5. Section 2.5 concludes Chapter 2.

2.2 LITERATURE SURVEY

There are two types of presence detectors in current use: magnetic detectors and inductive loop detectors [23]. The loop detector is by far the most widely used of the two [5, p. 8] and this study will restrict its attention to the simulation of loop detectors only.

A typical realization is a wire loop set in a square, 6' x 6', centered in a standard 12' lane with associated detector circuitry positioned off to the side of the road. It will be assumed in this study that detectors are configured this way and that detector stations are spaced at 1/2 mile intervals on the road with detectors in each lane (see Figure 2.1).

Mikhalkin [19] experimented with such a configuration and found that a loop detector has the detection region shown in Figure 2.2. If any part of the vehicle covers the region, the detector will activate and produce a presence pulse (Figure 2.3). Outside of this region no pulse is generated.

Mikhalkin stated the following characteristics of the detection region.

(1) For a particular vehicle travelling at speeds between 10 and 60 mph, the size and shape of the region do not change. No testing was performed outside of this range of speeds.

(2) The use of 3 different sets of detector electronics for the same loop resulted in a 4 percent change in the effective loop length. (See Figure 2.2.)

(3) Provided that a particular vehicle remains completely within the 12' lane, the effective loop length varies by less than 2 percent.

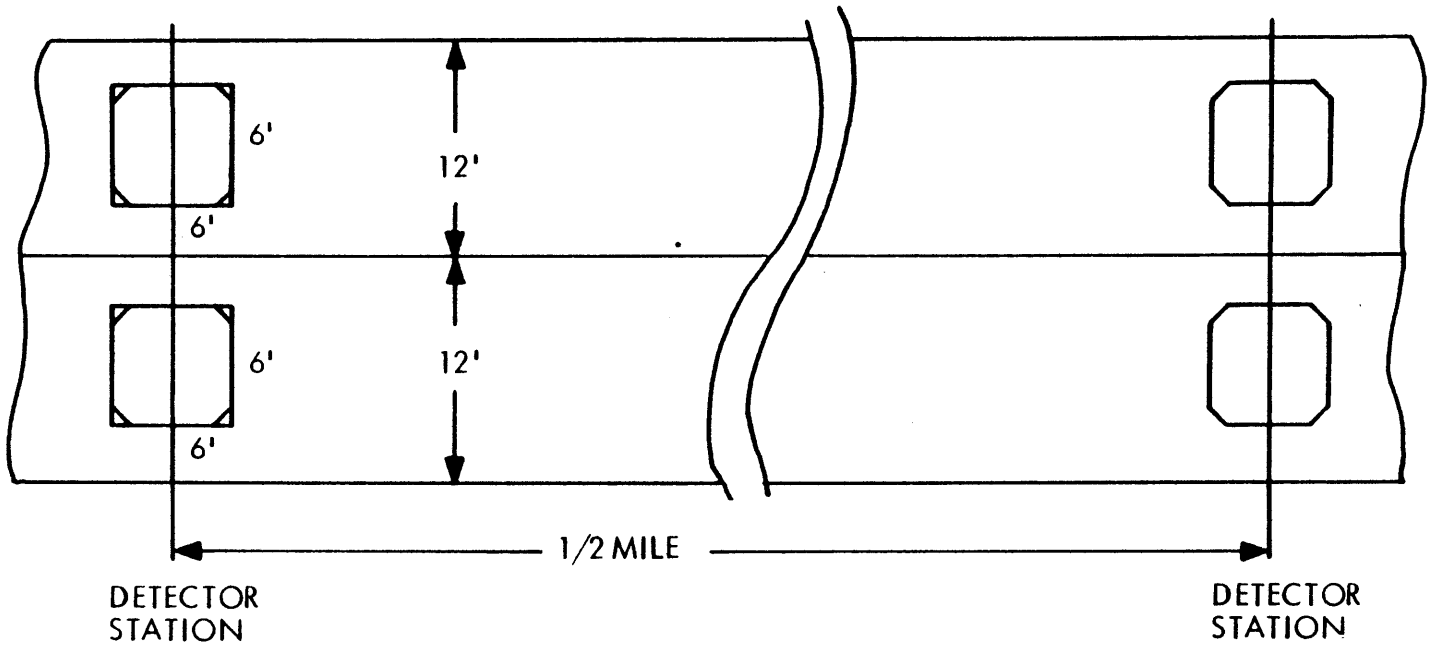


Figure 2.1 PRESENCE DETECTOR CONFIGURATION ON THE FREEWAY

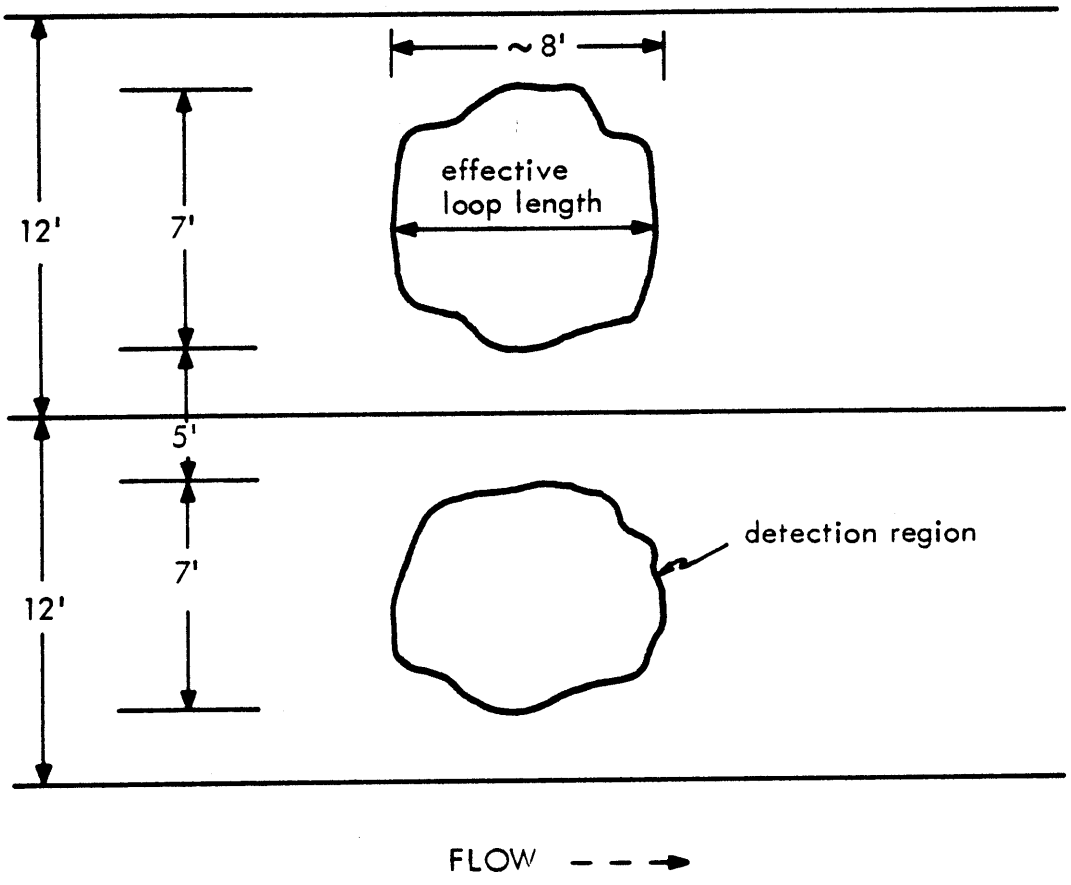
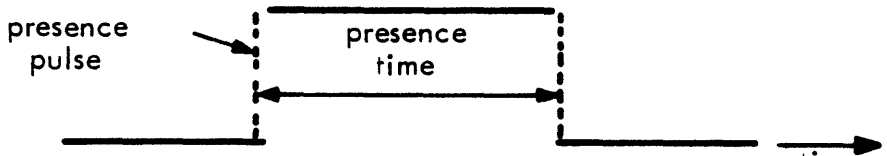
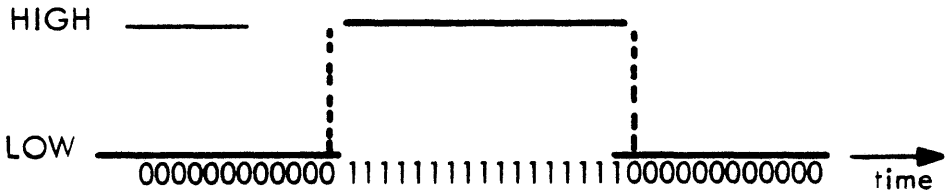


Figure 2.2 DETECTION REGIONS AT A DETECTOR STATION

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 2.3 PRESENCE DETECTOR SIGNAL ASSOCIATED WITH A SINGLE VEHICLE PASSAGE

(4) Comparisons of 4 different vehicle types resulted in a range of effective loop lengths of 1.8 feet. The average effective loop length was 8.0 feet.

It has been observed that certain large trucks are capable of producing two presence pulses in crossing a single detector [24]. The analog signal associated with the passage of such a truck (see Figure 2.3) peaks when the front and rear ends of the truck cross the loop and dips very low in between. This behavior is due to the large change in the proximity of iron in the truck body to the pavement which occurs when the front axles cross the loop, followed by the main body and then the rear axles. The dip, if low enough, can deactivate the detector only to be reactivated as the rear end crosses the loop.

Using this experimental information, a model can be developed which simulates presence detector signals.

2.3 LOOP DETECTOR MODEL

Based upon Mikhalkin's results, it will be assumed that

(1) All detectors have identical detection regions

(2) The size and shape of this region is independent of vehicle speed

(3) The effective loop length is only a function of vehicle type.

Therefore, a vehicle's presence time depends only upon speed, acceleration, and effective loop length, assuming that the vehicle is entirely within the lane. That is, the presence time, t, is found by solving Eq. (2.1)

$$\ell + d = vt + \frac{1}{2} at^2 \qquad (2.1)$$

where

ℓ = vehicle length [feet]

d = effective loop length for this vehicle [feet]

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

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

t = presence time [sec]

In the traffic microsimulation, the locations of presence detectors can be specified. The times that vehicles arrive at the detectors are available as well as the associated values of ℓ, v and a. The effective loop length, d, however, is not known exactly. Even though actual effective loop lengths do

depend on vehicle type, the range of variation is not large. Therefore, it will be assumed that d is a constant equal to 8 feet. Knowledge of the presence times and the arrival times of vehicles at the loops is sufficient to construct the continuous time binary presence detector signal. In reality, the detector signal is sampled between 15 and 60 times per second (see Figure 2.3); however, in this study, the signals are assumed to be continuous.

It will also be assumed in this study that the trucks which produce two presence pulses in crossing a loop are sufficiently scarce that they can be ignored.

2.4 ERRORS IN COUNTING VEHICLES

There is no literature available on the statistics of the errors that are made in counting vehicles as they cross detector stations. Vehicle counts are simply the result of counting the number of presence pulses in a specified time interval. In this section, the possibility of a vehicle crossing a detector station undetected as well as the possibility of a vehicle being detected more than once are examined. The statistics of these errors are used in the density estimation system of Chapter 5.

Figure 2.2 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 travelling centered over a line separating lanes as it crosses the detector station. It is assumed here that this event never occurs and, therefore, all vehicles get counted at least once.

Figure 2.2 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 <u>two</u> presence pulses. Figure 2.4 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 2.4. From the simple geometry of Figure 2.4, the following equation is obtained relating X to ℓ, w, v, and t

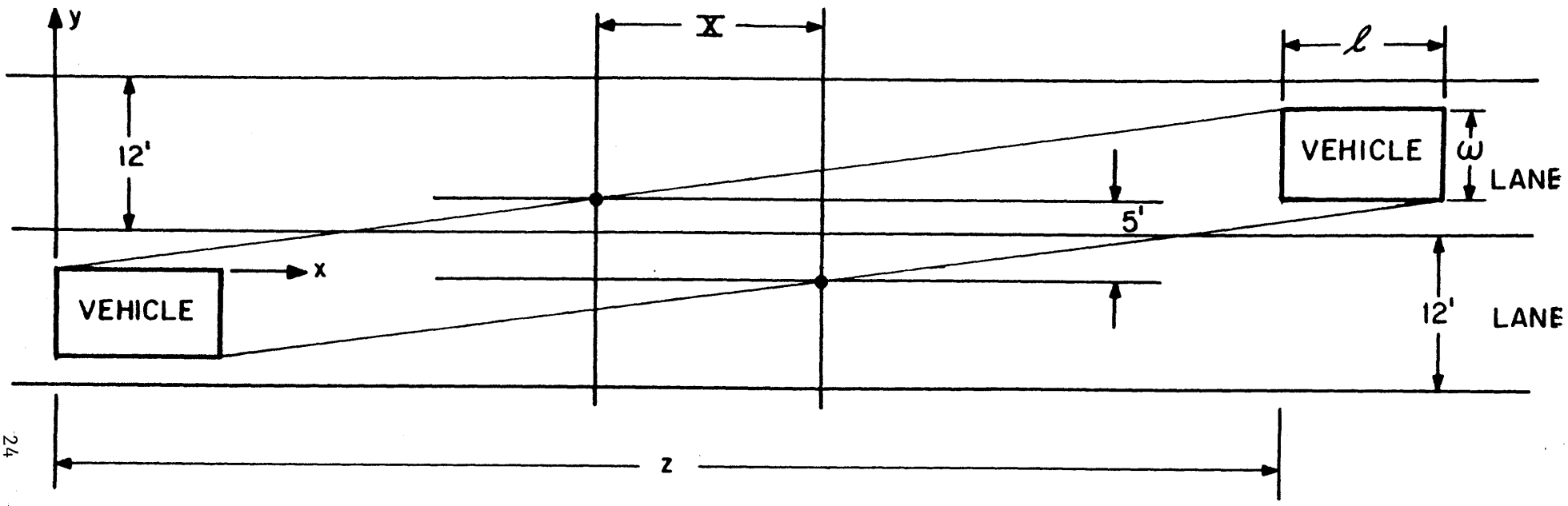


Figure 2.4 TOP VIEW OF VEHICLE CHANGING LANES

$$X = \left(\frac{vt}{12}\right)(w-5) + \ell \qquad (2.2)$$

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. (2.2) 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 it is rather unlikely that any given lane change will cause an extra count.

The traffic microscopic simulation contains a model for lane changing. A lane change can occur only if

(1) The acceleration required is less than the maximum acceleration capability at the current speed .

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

(3) There is a gap 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. (2.2), 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 in the manner described in Section 2.3 for each lane.

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 modelling this in the traffic simulation as described, the number of extra presence pulses generated was empirically examined as a function of the average flow rate over the detector. The results are shown in Table 2.1.

TABLE 2.1

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

UNITS OF 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

2.5 CONCLUSIONS

It was the attempt of this section to simulate detector signals from data provided by the traffic microscimulation in a manner as realistic as possible. The literature was found to contain little information on many modelling issues. The resulting model, although simple, is sufficient for the purposes of computing occupancies and estimating aggregate variables, as we will discuss in Chapter 5.

3. THE OCCUPANCY-DENSITY RELATIONSHIP

3.1 INTRODUCTION

The occupancy of a presence detector in a time interval is the percent of the interval that vehicles cover some part of the detection region. California Algorithm #7 is based entirely upon this measurement. Intuitively, occupancy indicates the density of vehicles on the roadway. In this section, the relationship between occupancy and density is explored. The results are used in Chapter 4 to lend insight into the behavior of CA-7, and in Chapter 5 in a density estimation system.

Section 3.2 defines section density, flow, space-mean speed and occupancy. The notation to be used throughout this report is also developed.

When traffic flow is very smooth, measurements over time from a fixed point (e.g., occupancy) are directly related to spatial quantities at a fixed time (e.g., density). In Section 3.3, this flow condition, called space-time homogeneity, is defined and the occupancy-density relationship is derived.

In Section 3.4 the relationship is examined in general traffic flow conditions. Section 3.5 summarizes the results obtained.

3.2 DEFINITIONS AND NOTATION

In this section the definitions of key traffic variables are stated. Any 'per lane' quantities (e.g., flow of 900 veh./hr. per lane) are average values across all lanes. No lane-specific variables are used.

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

$$\rho(x, \Delta x, t) = \frac{M(x, \Delta x, t)}{L\Delta x} \text{ [veh/mile per lane]} \qquad (3.1)$$

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

$$\phi(x, t, T) = \frac{N(x, t, T)}{TL} \text{ [veh/hr per lane]} \tag{3.2}$$

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 occupancies of the detectors in each lane. The occupancy of a presence detector in lane i, $i = 1,2,\ldots,L$, located at point x in the time interval [t, t+T] is given by

$$occ_i(x,t,T) = \frac{100}{T}\left[t_{i,I} + \sum_{j=1}^{N_i(x,t,T)-1} t_{i,j} + t_{i,F}\right] \text{ [dimensionless]} \tag{3.3}$$

where $N_i(x,t,T)$ is the number of vehicles to cross point x in lane i in the interval [t, t+T].

$$\sum_{i=1}^{L} N_i(x,t,T) = N(x,t,T) \tag{3.4}$$

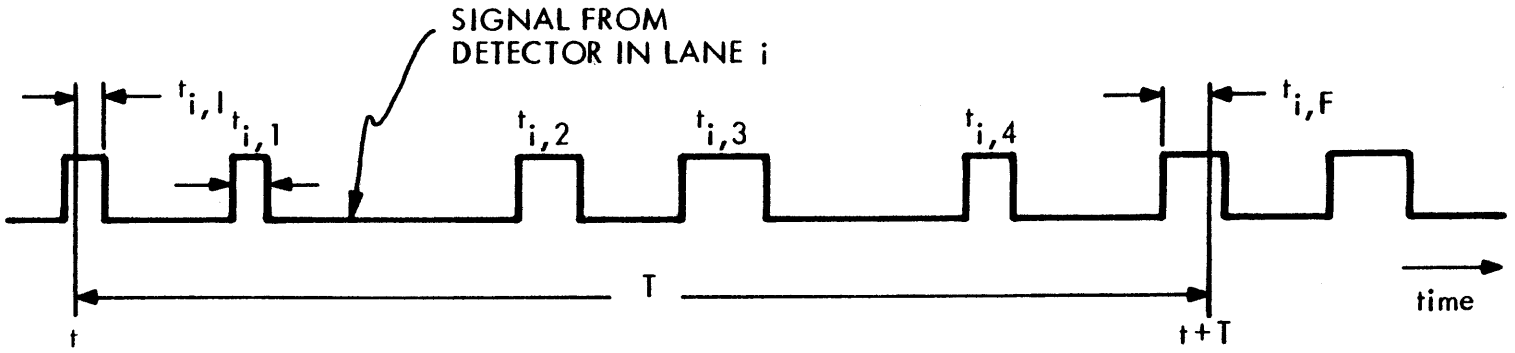
In Eq. (3.3) the $t_{i,j}$, $i = 1,2,\ldots,L$, $j = 1,2,\ldots,N_i(x,t,T)-1$, is the presence time of the j^{th} vehicle to cross the detector in lane i. The effect of a vehicle already over the detector in lane i at time t is represented by $t_{i,I}$. Similarly, $t_{i,F}$ shows the effect of a vehicle over the detector in lane i at time t+T. (See Figure 3.1.) The occupancy of the detector station at fixed space point x over the interval [t, t+T] is given by

$$occ(x,t,T) = \frac{1}{L}\sum_{i=1}^{L} occ_i(x,t,T) \quad \text{[dimensionless]} \tag{3.5}$$

In this report, interest is restricted to the detector station occupancy of Eq. (3.5) as opposed to the specific detector occupancy of Eq. (3.3).

3.3 SPACE-TIME HOMOGENEITY AND THE OCCUPANCY-DENSITY RELATIONSHIP

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 relate fixed time spatial quantities to fixed point



$$N_i(x,t,T) = 5$$

$$OCC_i(x,t,T) = 100/T\ (t_{i,I} + t_{i,1} + t_{i,2} + t_{i,3} + t_{i,4} + t_{i,F})$$

Figure 3.1 AN EXAMPLE OF AN OCCUPANCY COMPUTATION

temporal quantities in traffic, a model describing the relationship between various instantaneous point variables in traffic is needed. Such models do exist (e.g., Phillips [25]). 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.

The approach taken here is to restrict the traffic on a section of road $[x, x+\Delta x]$ to be very regular over an entire time interval $[t, t+T]$. Under this condition, called space-time homogeneity, a direct relationship between occupancy and density is shown to exist.

The traffic flow on a section $[x, x+\Delta x]$ over an interval $[t, t+T]$ is said to be space-time homogeneous if the space-mean speed and density on any subsection of $[x, x+\Delta x]$ at any time within $[t, t+T]$ is equal to the space-mean speed and density on any other subsection of $[x, x+\Delta x]$ at any other time within $[t, t+T]$. (For a more rigorous definition see Breiman [26].) 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

$$\left.\begin{aligned} \bar{v}_s(x,\Delta x,t) &= \bar{v}_s(x,\Delta x) \\ \rho(x,\Delta x,t) &= \rho(x,\Delta x) \end{aligned}\right\} \tag{3.6}$$

Breiman [24] showed the following relation to exist between aggregate variables under space time homogeneous conditions

$$\phi(x_0,t,T) = \rho(x,\Delta x)\bar{v}_s(x,\Delta x) \qquad \forall\, x_0 \in [x, x+\Delta x] \tag{3.7}$$

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

$$\phi(x,t,T) = \phi(t,T) \tag{3.8}$$

Under these same homogeneity assumptions, Wardrop [25] was the first to show that the speeds of successive vehicles crossing a point should be harmonically averaged to yield the space-mean speed on the road. (See also Breiman [26], Gershwin [28].) That is

$$\bar{v}_s(x,\Delta x) = \frac{N(t,T)}{\sum_{j=1}^{N(t,T)} \frac{1}{v_j}} \tag{3.9}$$

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 $[x, x+\Delta x]$. The harmonic average, Eq. (3.9), should not be confused with an arithmetic average. Appendix B gives insight into why a harmonic average of successive velocities of vehicles crossing a detector station results in the space-mean speed.

Substituting Eq. (3.9) and Eq. (3.2) into Eq. (3.7) results in

$$\rho(x, \Delta x) = \frac{1}{TL} \sum_{j=1}^{N(t,T)} \frac{1}{v_j} \left[\frac{\text{veh}}{\text{mile}} \text{ per lane}\right] \tag{3.10}$$

Substituting Eq. (2.1) without the acceleration term into Eq. (3.10) yields

$$\rho(x,\Delta x) = \frac{1}{TL} \sum_{j=1}^{N(t,T)} \frac{t_j}{\ell_j+d} \left[\frac{\text{veh}}{\text{ft}} \text{ per lane}\right] \tag{3.11}$$

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

In Eq. (3.11) 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, imagine that the ℓ_j are samples of a random variable, ℓ, with a <u>known</u> probability density function, $f_\ell(\ell)$, and replace Eq. (3.11) with its expected value over ℓ. This results in

$$\rho(x,\Delta x) = \frac{5280}{TL} E_\ell\left[\frac{1}{\ell+d}\right] \sum_{j=1}^{N(t,T)} t_j \left[\frac{\text{veh}}{\text{mile}} \text{ per lane}\right] \tag{3.12}$$

where $E_\ell[\cdot]$ denote expectation over $f_\ell(\ell)$. Note that the 5280 converts the density value from vehicles/foot to vehicles/mile. Comparing Eq. (3.12) with the definition of occupancy, Eq. (3.3) and Eq. (3.5), (ignoring the end effects

$t_{i,I}$ and $t_{i,F}$) an approximate relationship between occupancy and density is seen to exist.

$$\rho(x,\Delta x) = \left(\frac{5280}{100}\right) \text{occ}(t,T)\ E_{\ell}\left(\frac{1}{\ell+d}\right)\left[\frac{\text{veh}}{\text{mile}} \text{ per lane}\right] \tag{3.13}$$

It is important that the reader have an intuitive understanding of Eq. (3.13). The density obtained using Eq. (3.13) 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. (3.13) should be understood. They are restated and discussed here.

(1) The traffic is assumed to be space-time homogeneous. Such an assumption is restrictive.

(2) The harmonic average, Eq. (3.9), is actually an approximation of an expected value. (See Breiman [26]). 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. (3.13) can be determined, given a value of d. A more accurate conversion than Eq. (3.13) could be obtained if d was known as a function of ℓ.

(4) The end effects, $t_{i,I}$ 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.

Because of the restrictive assumptions and approximations used in deriving Eq. (3.13), the practical value of Eq. (3.13) may seem dubious. In Section 3.4 the accuracy of Eq. (3.13) is determined.

3.4 A STUDY OF THE OCCUPANCY-DENSITY RELATIONSHIP IN GENERAL TRAFFIC CONDITIONS

A study has been made with the microscopic traffic simulation program (Section 1.4) to see how the density, computed using Eq. (3.13), compares with the actual traffic density. In particular, the study examines the accuracy of Eq. (3.13) 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.

A density map provides insight into these issues. A density map is a three dimensional plot of point density versus a point on the road, x, and a point in time, t. Point density, denoted here by P(x,t) is related to section density via the following relations

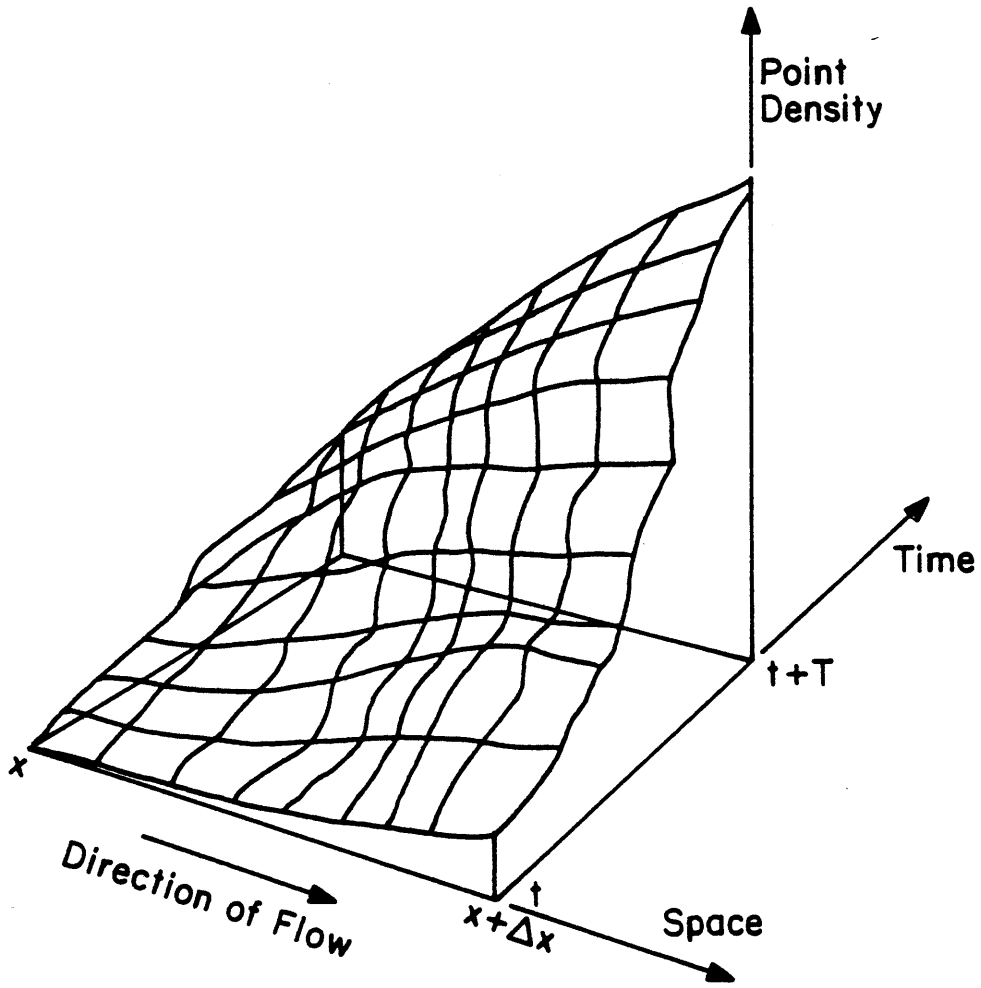
$$\rho(x,\Delta x,t) = \frac{1}{\Delta x}\int_{x}^{x+\Delta x} P(x_0,t)\,dx_0 \tag{3.14}$$

$$P(x,t) = \lim_{\Delta x \to 0} \rho(x,\Delta x,t) \tag{3.15}$$

Figure 3.2 is an example of a density map which describes an increase in traffic density with time along the entire section. This behavior could possibly be the result of an accident downstream. Because the density surface is not nearly level (flat), the traffic conditions on [x,x+Δx] during [t,t+T] are inhomogeneous. However, the traffic is approximately homogeneous on the smaller space-time intervals which are reasonably flat. From this observation, it appears that an inhomogeneous space-time interval can be subdivided into smaller intervals which are approximately homogeneous. The size of these smaller space time intervals depends upon the time and space constants which characterize temporal and spatial changes in traffic conditions. Breiman [29] examined the time scales involved in traffic flow and found that the flow rate past a point does not change appreciably over 5 or 10 seconds. It has been observed [30] that spatially, traffic conditions do not change significantly over several hundred feet, in the absence of an obstruction. The effect of an obstruction is a discontinuity in traffic conditions at the site of the obstruction. Thus, for T and Δx sufficiently small, it would appear that the space-time homogeneity assumption used in deriving Eq. (3.13) is always valid. For larger values of T and Δx, homogeneity cannot be assumed.

This conjecture concerning space-time homogeneity has, in fact, been verified using the traffic simulation program. The effects of the other approximations used in deriving Eq. (3.13) (see Section 3.3) have also been determined.

The testing of Eq. (3.13) consisted of an examination of the error between the actual spatial density and the density predicted by Eq. (3.13). The test used



Sigure 3.2 A DENSITY MAP

(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 Appendix A) assuming d = 8 feet.

Before discussing the results, let us first develop some notation. Let $\hat{\rho}(k)$ denote the density at time step k obtained using Eq. (3.13) and $\rho(k)$ denote the actual density at time step k obtained from the traffic simulation program. The error between the two densities, e(k), can be thought of as a random process, $e(k) = \hat{\rho}(k) - \rho(k)$. This study is actually a study of the error process e(k), k = 1,2,... .

The results of the test were

(1) $5 \text{ sec} \leq T \leq 10 \text{ sec}$ and $100' \leq \Delta x \leq 500'$.

The error process was observed to have the following characteristics under all traffic conditions

(a) it is zero mean

(b) when plotted, it visually appears to be uncorrelated in time (i.e., a white process). No statistical whiteness tests have been performed.

Depending on the traffic conditions the variance of the process ranged from 100 to 200 (veh./mile per lane)^2. The short value of T is the cause of the high variance. With short values of T, very few vehicles contribute to the occupancy used in Eq. (3.13). This gives rise to large statistical fluctuations and consequently a large variance. However, the short value of T is also the cause of the apparent whiteness of the process. The fact that this process is zero mean and white under all traffic conditions will prove to be crucial to the density estimation scheme presented in Section 5.

(2) $T > 10 \text{ sec}$ or $\Delta x > 500'$

As T or Δx are increased and traffic remains space-time homogeneous over $[x, x + \Delta x]$ and $[t, t+T]$, then Eq. (3.13) becomes more accurate. In the microscopic traffic simulation, space-time homogeneous conditions are identified

visually from a velocity map. The velocity map shows graphically the speeds and positions of the vehicles on the road at 5 second intervals. The larger value of T results in more vehicles contributing to the averaging approximation used in Eq. (3.9). 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.

Of course, the problem with increasing T or Δx is that the guarantee of space-time homogeneity is lost. Traffic conditions can and do change drastically at a point over a one minute interval and along a 1/2 mile section at a fixed time. Because the density obtained from Eq. (3.13) is actually a time averaged density localized at a point, this density can be highly dependent upon where in the section the station is located. Furthermore, the section density can change considerably over the time interval. Thus, the selection of a section and a time to which the density of Eq. (3.13) relates is not straightforward and sometimes results in large unpredictable errors.

3.5 SUMMARY

The conclusions of this section are:

(1) The temporal density variation on a section or road very local to the detector station can be obtained from occupancy measurements taken over short time intervals. Eq. (3.13) is the conversion from occupancy to density and is valid at all flow levels and in inhomogeneous conditions. The noise associated with the conversion (i.e., the difference between the true density and that predicted by Eq. (3.13)) is a zero mean white process with a large variance.

(2) Occupancy measurements taken over larger, e.g. one minute, intervals, do not, in general, convert to a section density using (3.13). This is due to the inhomogeneities (i.e., irregularities) that do occur in traffic over one minute intervals.

(3) Section densities on, say, a 1/2 mile section cannot be, in general, obtained from a static conversion of occupancy measurements at a detector station. This is due to the inhomogeneities that do occur in traffic along a 1/2 mile section.

Each of these conclusions will be referred to in later sections.

4. CALIFORNIA ALGORITHM #7 AND OTHER EXISTING INCIDENT DETECTION SYSTEMS

4.1 INTRODUCTION TO AND HISTORY AND DESCRIPTION OF EXISTING SYSTEMS

Incident detection systems which use presence detectors have been in existence for over 10 years. Most major cities have many miles of highway with 6' x 6' inductive loop detectors placed in all lanes every 1/2 mile. These loops are equipped to provide occupancy and flow measurements.

In a severe incident in heavy traffic, the capacity of the freeway at the incident site is reduced below the level of the approaching traffic volume. The result is that traffic quickly backs up upstream of the incident while a region of light traffic develops downstream. The congestion upstream of the incident will continue to grow until the incident condition disappears or until the oncoming flow decreases to less than the capacity of the incident site.

This developing traffic pattern is reflected in the occupancy and flow data from vehicle presence detectors. This is because the traffic conditions at the presence detector station immediately upstream of the incident are markedly different from those at the next downstream station. One might, therefore, consider an incident detection algorithm which is based on the following test

$$\begin{pmatrix}\text{occupancy}\\ \text{upstream}\end{pmatrix} - \begin{pmatrix}\text{occupancy}\\ \text{downstream}\end{pmatrix} \begin{matrix}\text{incident}\\ \gtrless\\ \text{no incident}\end{matrix} \text{threshold} \qquad (4.1)$$

The completion of the algorithm would require selecting the value of the threshold. An algorithm this simple would enjoy some success. However it will create false alarms [31]. Therefore, a more sophisticated algorithm seems to be needed where, given a desired false alarm rate, the missed detection rate and mean time-to-detect are acceptable.

Some 24 existing algorithms, all based on simple functions of flow and occupancy crossing calibrated thresholds, were studied by Payne in 1975 [5]. None of these algorithms employed a traffic model based on differential equations describing the temporal and spatial behavior of traffic variables (e.g., Payne-Isaksen [14], Phillips [25]). None of the algorithms made any attempt to derive spatial aggregate variables. All were intuitively appealing and computationally simple.

The conclusions of the study were [5]

(1) By exponentially smoothing the data, the need for very precise threshold calibration was reduced.

(2) When station-specific thresholds are used, the best performance was achieved. The "California Algorithm" appeared to perform as well as any existing system with such thresholds.

(3) The algorithms which perform best were based on occupancy and simple functions of occupancy (i.e., no flow information is used).

Since this study, extensive development has gone into the California Algorithm [31] and California Algorithm #7 (referred to here as CA-7) has emerged as the most widely accepted incident detection system in use today. Its performance can reasonably be regarded as the effective limit of algorithms which are based on simple functions of occupancy and flow crossing predetermined fixed thresholds, or which are based on the smoothing of occupancy and flow data.

4.2 CALIFORNIA ALGORITHM #7

California Algorithm #7 is based solely on occupancy measurements and is specifically designed for freeways with detector stations every 1/2 mile (see Figure 4.1). The occupancy variables are sixty second averages and are averaged across all lanes so that each detector station produces one measurement each minute. (It should be noted that there is a version of CA-7 designed to work with occupancies averaged over twenty seconds. We are restricting our attention here to the sixty second version.)

During non-incident conditions there should be no notable difference in occupancy values from station to station at any one time. However, during an incident on a link, the station downstream experiences a low occupancy and the upstream station sees a rise in occupancy. If this condition persists for at least two minutes, then an incident is signalled. Thus, in CA-7, tests are performed each minute on each link to see if various functions of occupancy cross predetermined, station-specific thresholds. The functions, or features, as they are sometimes called [5] are

$$\left.\begin{aligned} \text{occdf}\ (i,t) &= \text{occ}(i+1,\ t,\ 60) - \text{occ}(i,\ t,\ 60) \\ \text{occrdf}\ (i,t) &= \frac{\text{occdf}(i,t)}{\text{occ}(i+1,t,60)} \\ \text{docc} &= \text{occ}(i,\ t,\ 60) \end{aligned}\right\} \qquad (4.2)$$

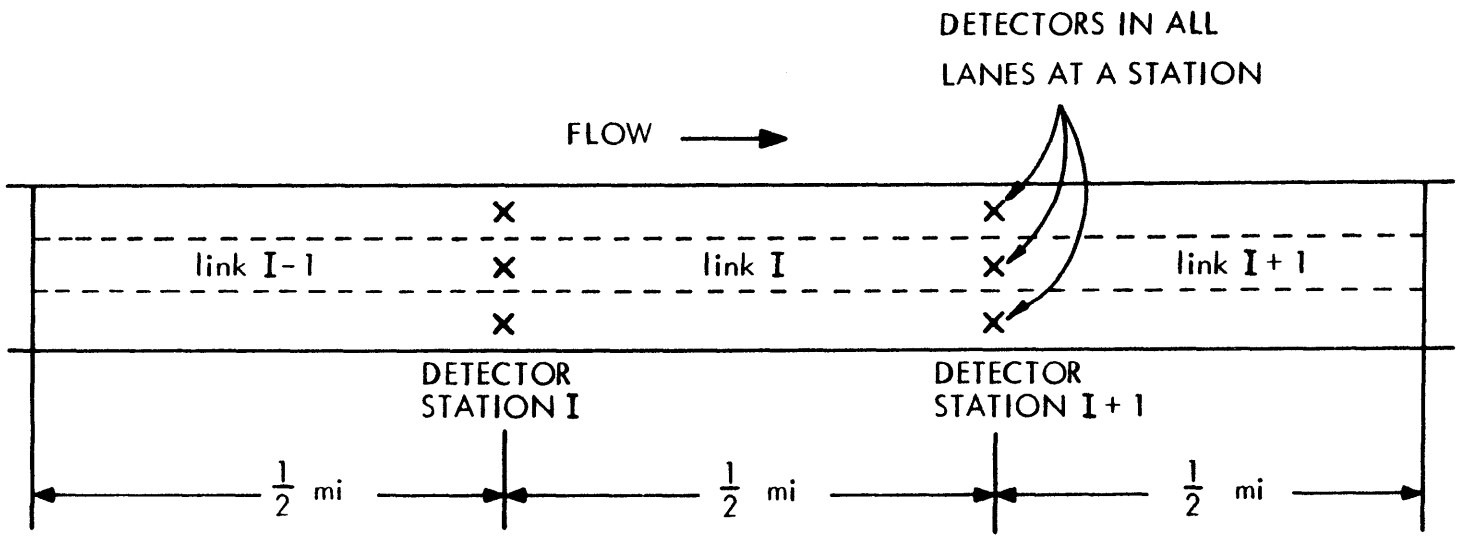


Figure 4.1 DETECTOR STATIONS AND LINKS

Figure 4.2 shows a block diagram of the CA-7 algorithm. The algorithm works as follows. Suppose, initially, CA-7 is in an incident-free state (i.e., state = 0). After one minute of detector data, the features occdf, occrdf and docc are computed. If occdf and occrdf exceed their respective thresholds, T_1 and T_2, and docc is below threshold T_3, then the algorithm switches to state = 1 (tentative incident). Otherwise, the system remains in state = 0 and the same test is performed the next minute. Now, suppose the system is in state = 1. The variable occrdf determines whether the next state is state = 2 (incident occurred) or state = 0, depending on whether it exceeds its threshold, T_2, or not. If the state = 2 is reached, state = 3 (incident continuing) will result if occrdf exceeds its threshold. Otherwise the system returns to the zero state.

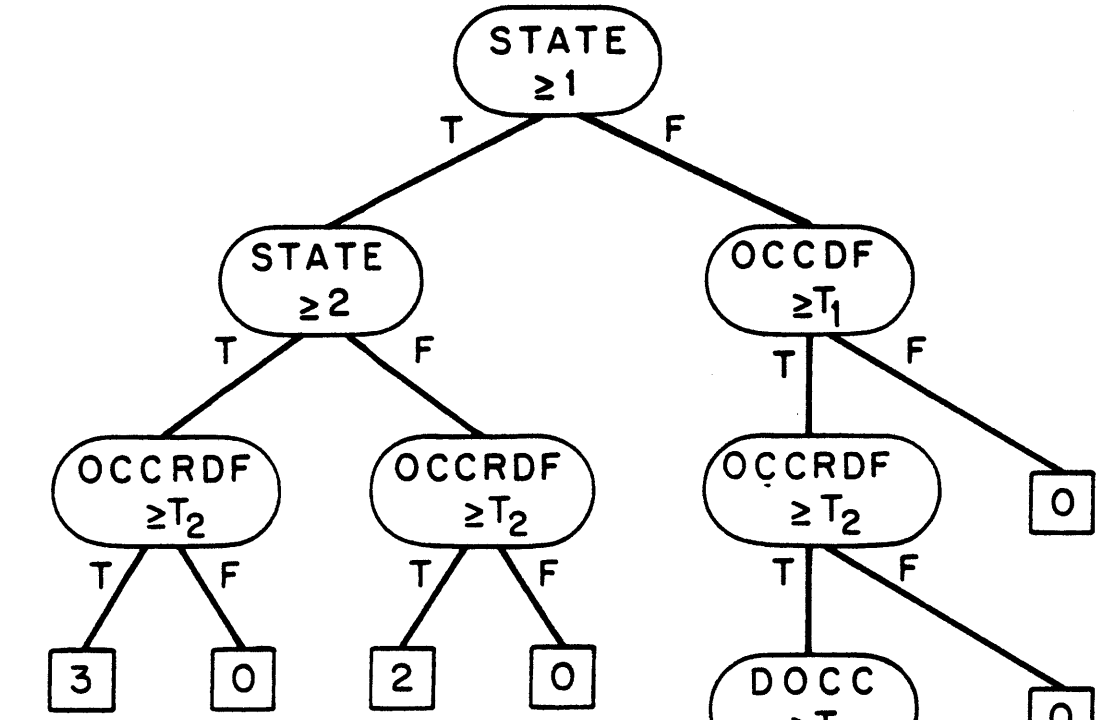
With a large data base of detector data from Los Angeles highways, the thresholds in Figure 4.2 were empirically calibrated, for any desired false alarm rate, to yield the best possible detection rate without regard for the mean time-to-detect. The incident data base included such incidents as traffic collisions, disabled vehicles, gawking and spilled loads [32]. These incidents occurred in a wide range of traffic flow levels.

Seven sets of thresholds, T_1, T_2, and T_3 are presented in Table 4.1. Each threshold set is associated with a false alarm rate. The detection rate and mean time-to-detect which results from the use of each threshold set is also tabulated in Table 4.1.

The performance results presented by Payne for CA-7 are shown in Figure 4.3. The curve shows, for each threshold set, the percent of incidents detected versus time-to-detect. Note that several threshold sets appear to detect incidents before they occur. This is because the precise occurrence times of the incidents on the data base are not known. Note that using the thresholds associated with the highest false alarm rate, CA-7 can only be expected to detect 60% of the incidents. No information is provided as to which types of incidents were not detected or which flow levels caused detection problems. At lower false alarm rates, CA-7 may only detect 30 to 50% of the incidents.

4.3 COMPUTER SIMULATION OF CALIFORNIA ALGORITHM #7

In this section, the traffic conditions under which California Algorithm #7 works well, yields false alarms and misses detections are determined. The



State	Designates
0	Incident-free
1	Tentative incident
2	Incident occurred
3	Incident continuing

Figure 4.2 BLOCK DIAGRAM OF CALIFORNIA ALGORITHM #7 (TAKEN FROM PAYNE [31])

TABLE 4.1

CALIFORNIA ALGORITHM 7 (CA-7) THRESHOLDS
(Taken from Payne, et al., [31])

Threshold Set	Detection rate (%)	False alarm rate (%)	Mean-time-to-detect (minutes)	Thresholds		
				T_1	T_2	T_3
1	59	.134	3.25	8.1	.313	16.8
2	51	.050	4.31	12.9	.360	16.6
3	49	.043	4.94	13.1	.358	15.8
4	41	.029	4.85	9.6	.359	12.3
5	37	.017	6.17	13.1	.393	12.5
6	31	.006	5.84	21.6	.301	13.9
7	20	.004	7.73	26.6	.322	13.4

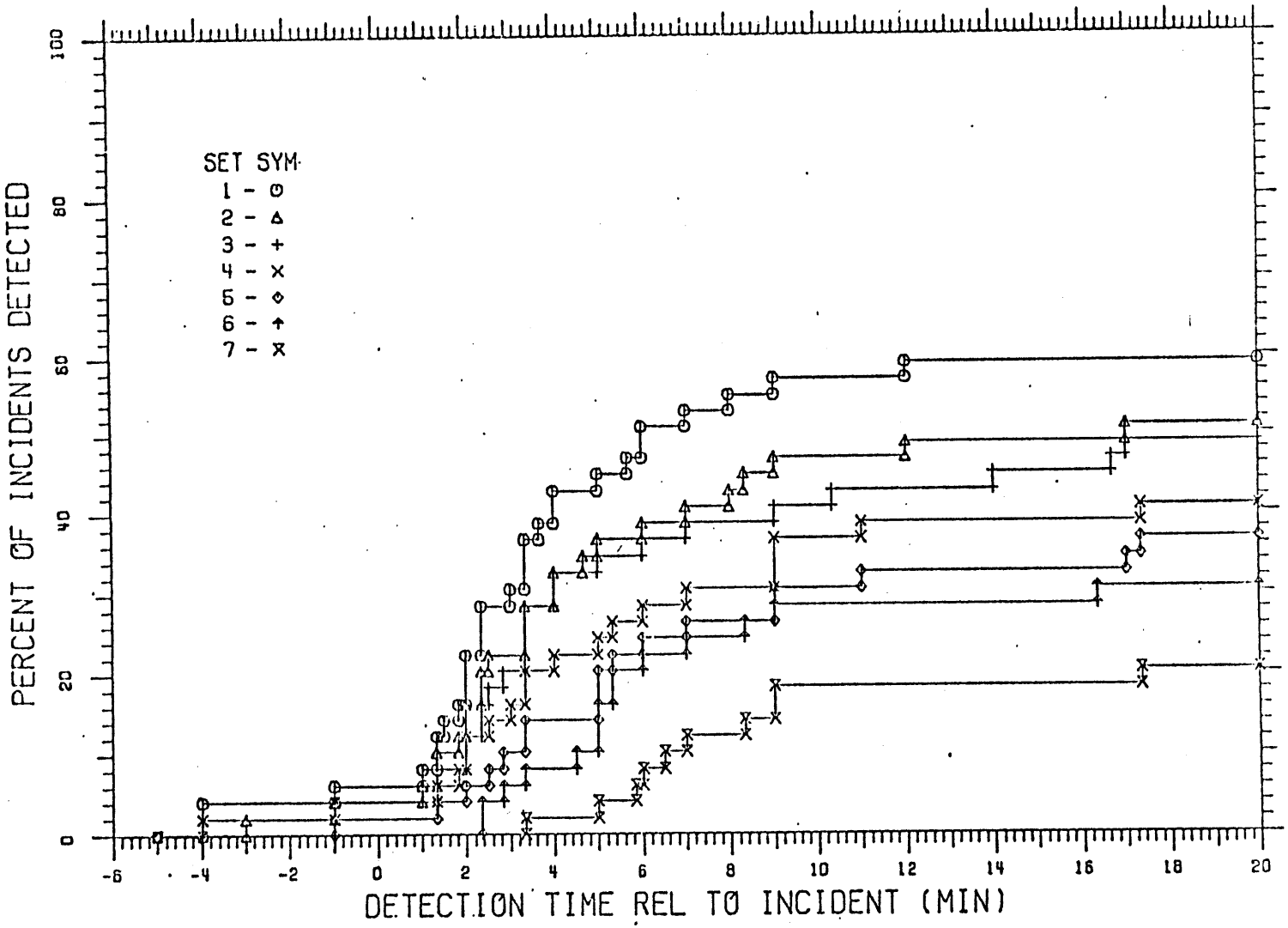


FIGURE 4.3 PERFORMANCE OF CALIFORNIA ALGORITHM #7
(TAKEN FROM PAYNE [31, p. 237])

microscopic computer simulation provides the sixty second occupancy input to run CA-7. The occupancy measurements are not corrupted by lane changing near detector stations because this corruption only degrades performance and may disguise the issues at hand. For any particular simulation, CA-7 is run seven times, each time using a different set of thresholds. Each threshold set is associated with a different false alarm rate. The thresholds used are exactly those shown in Table 4.1. All the incident simulations were modelled by stopping a vehicle and blocking a lane, thus leaving only one lane open for vehicles to pass the incident. This could represent a disabled vehicle or a traffic collision. All simulations have detector stations in both lanes and stations every 1/2 mile over a three mile stretch (see Figure 4.4). All totalled, CA-7 has been simulated on over forty traffic scenarios. Rather than present all forty, a selection of six representative scenarios have been chosen to include in this report. Table 4.2 describes these simulations. Table 4.3 describes the response of CA-7 to each of the simulations in Table 4.2.

4.4 CONCLUSIONS ON CA-7 PERFORMANCE

From the results shown in Table 4.3, the following conclusions can immediately be drawn.

(1) CA-7 detects incidents well in heavy flow but misses a large number of incidents at lower flow levels.

(2) CA-7 produces very few false alarms.

It is important that the reasons for this behavior be brought out. CA-7 uses a set of fixed thresholds chosen to keep the false alarm rate at an acceptable level. This use of fixed thresholds restricts CA-7 in the range of traffic conditions in which it is effective. For example, the occupancy difference between stations (occdf) during an incident is much larger in heavy flow conditions than it is in low flow conditions. By selecting the occdf threshold, T_1, to be too large, the false alarm rate drops but incidents in low flow are missed. Similarly, by lowering T_1, the false alarm rate rises and more low flow incidents are detected. This observation implies that flow dependent thresholds would increase the range of traffic conditions in which CA-7 is effective.

During an incident, there is a certain reduced capacity at the incident site. If the oncoming flow is greater than this capacity, then the backup from the incident will, in theory, continue to grow endlessly. In this case, the upstream detector station will be, eventually, entirely covered by a queue of

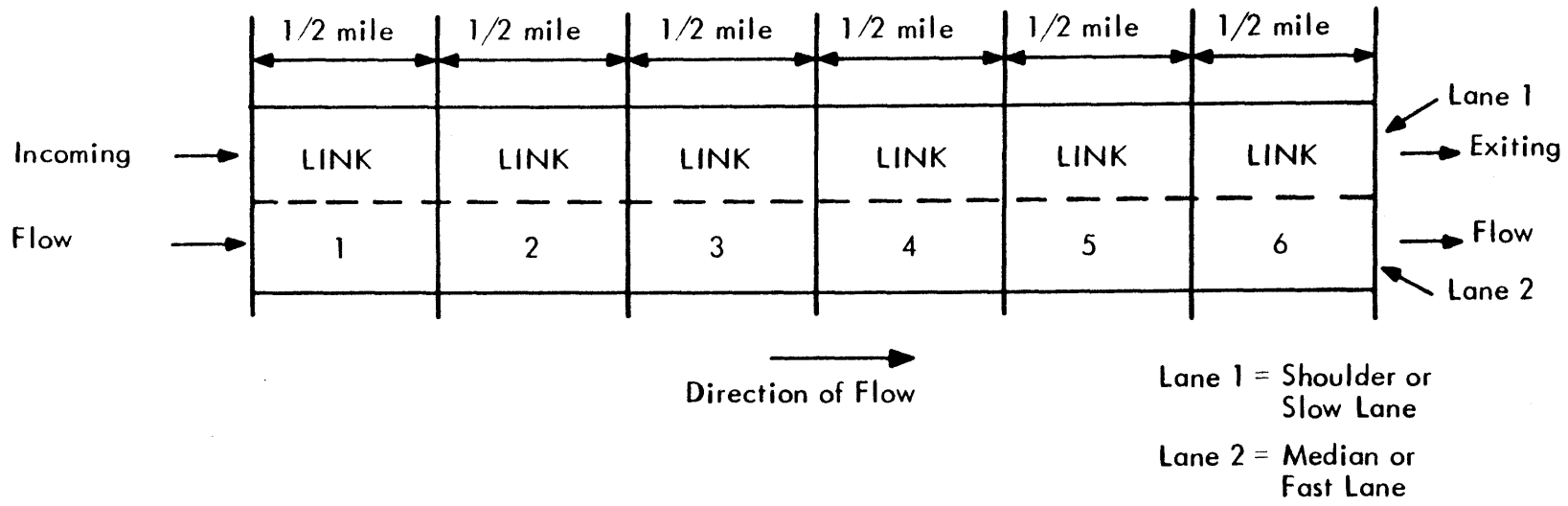


Figure 4.4 SIMULATED FREEWAY

TABLE 4.2

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 band of congestion* remains 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.
27	15	55	725	4	2	120	Traffic is light. The band of congestion reaches a steady state length of only 300 ft. The congestion is entirely eithin the link.
26	20	55	1000	4	2	120	Traffic is light. The band of congestion reaches a steady state length of 650 feet. The congestion is entirely within the link.

*The term band of congestion is defined to be the region on the roadway outside of which traffic is unaffected by an incident.

TABLE 4.2 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)	
22	15	60	815	4	1	180	Traffic is light. The band 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 band 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 4.3

SIMULATION RESULTS OF CALIFORNIA ALGORITHM #7

Simulation Identification Number	Response of California Algorithm #7
29	The sudden appearance of the heavy input flow caused a state=1 at T=120 by two of the 7 threshold sets on link 1. The accident was not detected by T=360 which is when the simulation ended. A state=1 was reached on link 3 at time 360 by one of the 7 threshold sets.
28	The vacancy in front of the vehicles holding up traffic caused a state=1 by five of the seven threshold sets on link 5 at T=360 and on link 6 at T=420.
27	No detection. State=0 for all threshold sets on all links for the entire 10 minute simulation.
26	No detection. State=1 reached by two of the seven threshold sets on link 1 as a result of inhomogeneities in the input flow.
22	No detection. State=0 for all threshold sets on all links for the entire 10 minute simulation.
21	Incident detected (state=2) on correct link (link 4) by 3 threshold sets at T=300 and the remaining 4 threshold sets at T=360. Six of the seven threshold sets also detected an incident on link 5 at T=660. Several tentative incident signals were received from link 6.

vehicles while downstream traffic will be light. CA-7 has no trouble detecting such behavior. However, suppose the oncoming flow is less than the capacity at the incident site. In this case, the queue behind the incident will not grow endlessly but, instead, will reach a steady-state length. The result is a region on the roadway, called a band of congestion, outside of which traffic is not affected by the incident. (See Figure 4.5.) (The term band of congestion is new and was defined for this report to aid in discussions.) Certainly if the length of the band is greater than 1/2 mile, then having detector stations 1/2 mile apart would imply that one station is inside the heavy density region and CA-7 could be expected to trigger. However, our simulations have shown that an input flow (oncoming flow) of 1300 vehicles/hr per lane causes a steady state queue length of about 1/4 mile and any higher flow levels cause the queue to grow endlessly with time. (Thus the capacity at the incident site in our simulated incidents is about 1300 veh/hr per lane.) With detector stations 1/2 mile apart, the approximate probability of this incident being detected, assuming the incident is equally likely to occur anywhere, is $^{1/4}/_{1/2} = .5$. It should be noted that 1300 veh/hr per lane is not a light traffic condition. In fact, the density is high enough that vehicles are restricted from travelling at their desired speed and are forced to drive at 48 miles/hr on the average. Table 4.2 includes a description of the congestion behavior of the simulations. A density map for a low flow incident which would not be detected by CA-7 is shown in Figure 4.6. The congestion associated with the incident does not affect the occupancy measured at each detector station.

California Algorithm #7 was designed to maintain an acceptable false alarm rate. This involved choosing the thresholds such that only incidents could cause a detection to occur. Unfortunately the tradeoff was to raise the missed detection rate. In light traffic conditions, one can expect almost no false alarms, because, as mentioned earlier, the occupancy difference threshold (occdf) is so high, that, in order to cross the threshold, vehicles have to actually be stopped over the detector station. Only an incident could cause this behavior.

Simulations in heavier traffic have also produced very few false alarms. Payne [31] recognized the sources of false alarms to be (a) bands of congestion which move downstream slowly (i.e., 15 miles/hr or less) and (b) bands of congestion which move upstream slowly (called compression waves). The two

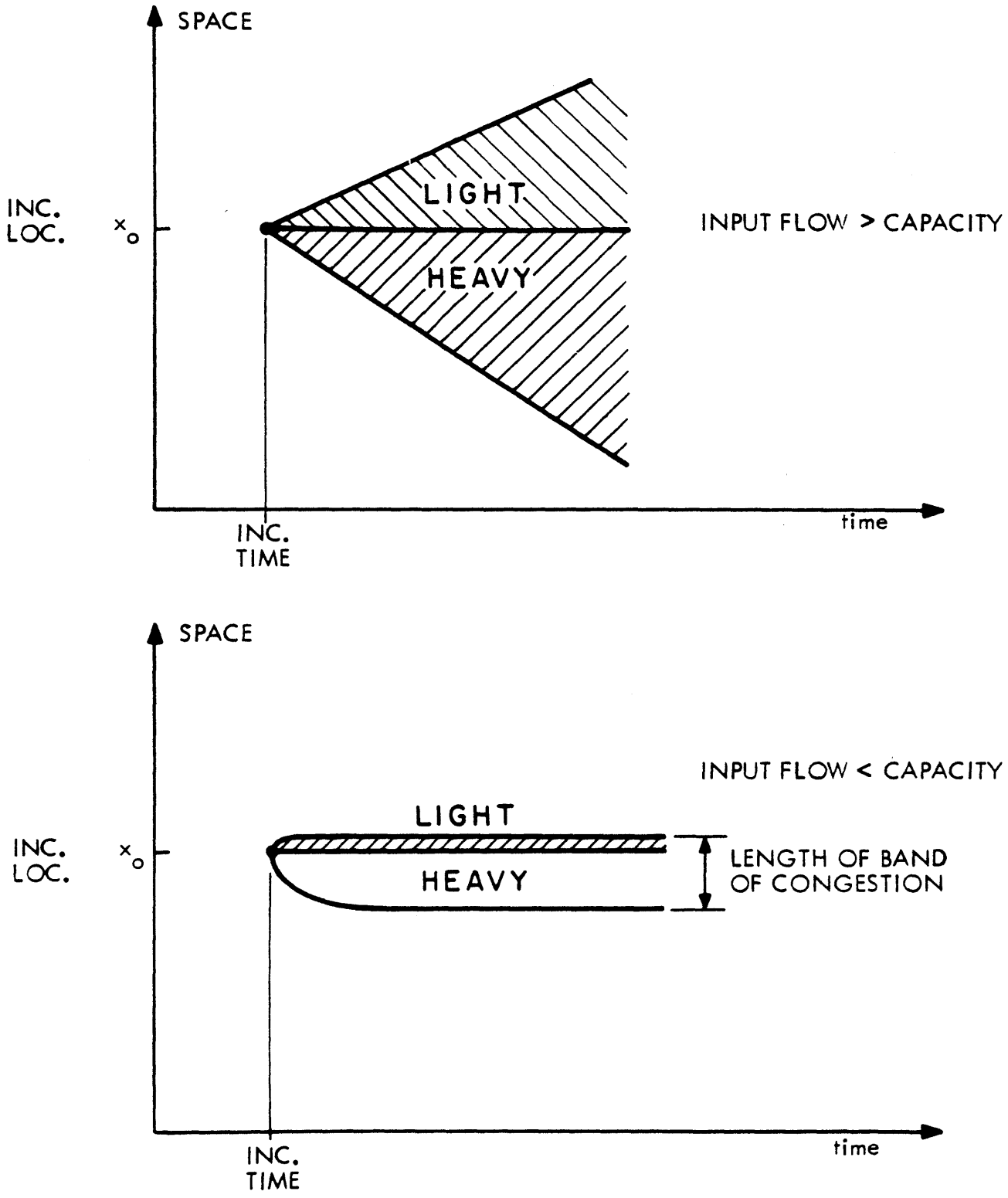


Figure 4.5 ACCIDENT CONGESTION BEHAVIOR WHEN (a) FLOW > CAPACITY AND (b) FLOW < CAPACITY

conditions differ from an incident condition only by the fact that they move. They occur in heavy traffic only when the vehicles are travelling close together. If, at the downstream and upstream ends of the congestion there is a sharp density gradient, then a false alarm (state=2 in Figure 4.2) will result.

Although Payne [31] acknowledged the fact that compression waves may produce false alarms, the microscopic computer simulation is incapable of generating such a condition. The best that can be done is a fast (speed greater than 18 miles/hr) upstream moving band of congestion with gradual density gradients at each end of the band. The resulting behavior from CA-7 is that the state=1 level is reached but not state=2.

Bands of congestion which move downstream and cause false alarms also cannot be simulated. If a sudden surge of traffic enters a freeway which has a light traffic condition, a state=1 condition will result. If the heavy traffic were to travel downstream more slowly, a false alarm would occur. Also, two slow drivers, travelling abreast of each other will cause congestion behind them and a vacancy in front of them. This, again, is similar to the sudden onflow of heavy traffic in that it yields a state=1 but moves too quickly to reach state=2.

In conclusion, the performance of CA-7 is relatively poor in light and moderate conditions and adequate in heavy conditions. CA-7 has difficulties distinguishing between incidents and normal fluctuations in density. Because traffic dynamics are not modelled into the algorithm. A dynamical traffic behavior model, such as the Payne-Isaksen model, is needed to predict the occurrence of moving waves of congestion so as to differentiate them from incidents. As mentioned earlier, two systems using the Payne-Isaksen model have been developed for the purpose of avoiding such false alarms and for detecting and identifying incidents in a much wider dynamic range of flow conditions [3], [4].

Occupancy, as we have seen, is a local measurement which may have no relevance to traffic conditions as little as five hundred feet away. In light and moderate flow, when the impact of an incident is restricted to a small section of roadway, many missed detections result because the congestion most likely will not be located at a detector station. Therefore, for incident detection with a CA-7-type algorithm, some quantity other than occupancy should be used. Such a quantity must reflect an incident condition at all flow levels and on

all parts of the highway. Simulation results have shown that link density, a spatial quantity, is more consistently affected by an incident than is occupancy, a local quantity. Figure 4.7 shows the behavior of link density as compared to the density estimate obtained from occupany, using Eq. (3.13), for the low flow incident of Figure 4.6. Figures 4.6 and 4.7 were taken from link #4 data of simulation #26 (see Table 4.2). In Chapter 5 a method for converting presence detector data into link density estimates is presented. These density estimates are then used for incident detection purposes.

Of course, there are extremely minor incidents, such as a stopped vehicle in the shoulder lane, whose occurrence is not sufficiently observable from presence detector data alone. Because the incident occurrence information is not contained in the detector signal, this type of incident cannot be detected using any quantities (e.g., aggregate variables) which are derived from detector signals.

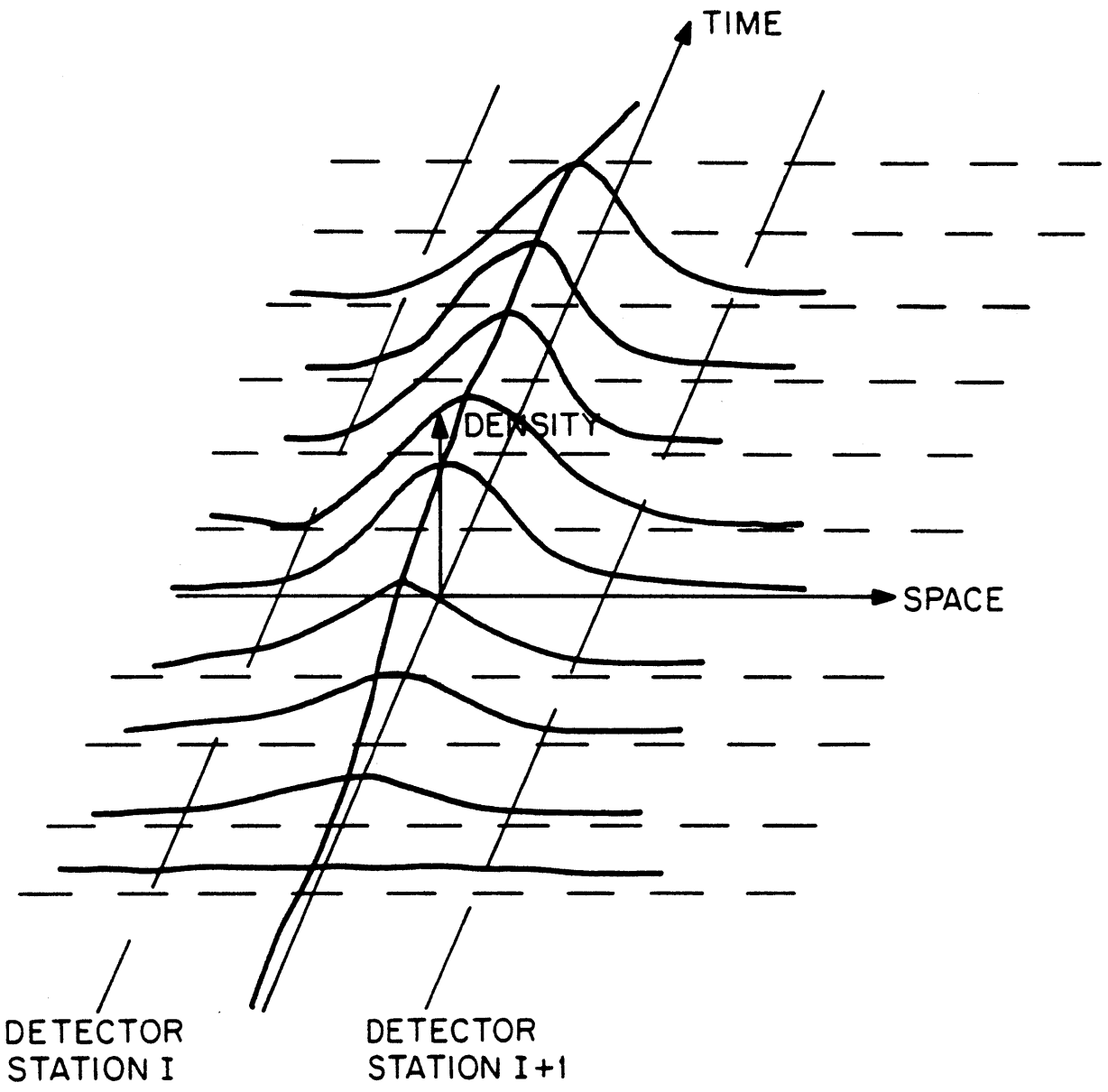


Figure 4.6 LOW FLOW INCIDENT DENSITY MAP

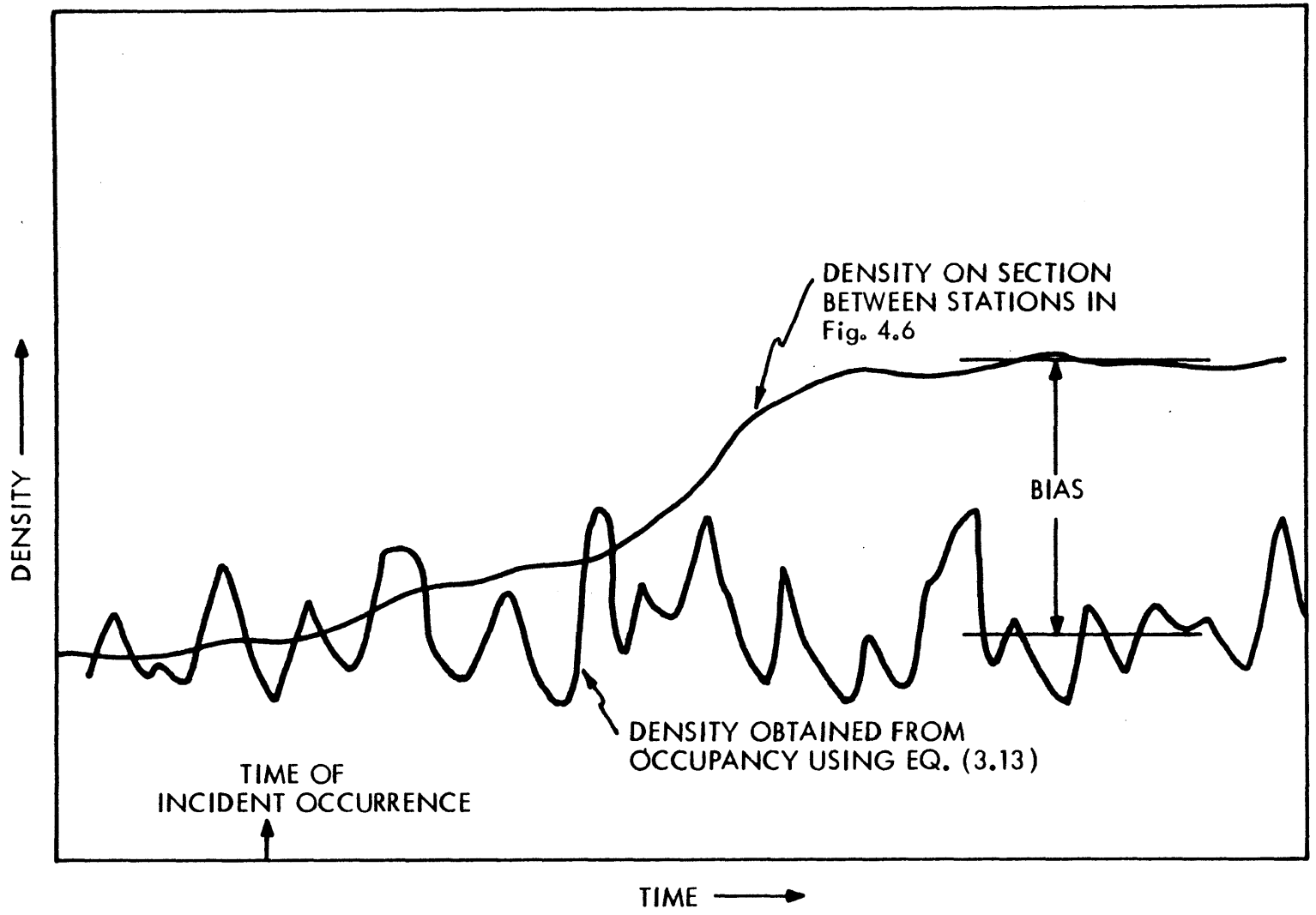


Figure 4.7 BEHAVIOR OF OCCUPANCY AND DENSITY FOR LOW FLOW INCIDENT OF FIGURE 4.6

5. THE ESTIMATION OF TRAFFIC VARIABLES AND INCIDENT DETECTION

5.1 LINK DENSITY ESTIMATION AND INCIDENT DETECTION

5.1.1 Introduction

The dynamic model-based incident detection systems described in reports by E. Chow [4] and C. Greene [3] require estimates of section density as observations. In Section 5.1, a system is presented which converts presence detector data into these estimates. Several approaches to this problem have been proposed but the approach presented here is the first with all of the following properties:

(1) It is computationally simple.
(2) It estimates density well at all flow levels and under inhomogeneous conditions.
(3) An accurate initial section density estimate is not required.
(4) The estimates are insensitive to the imperfections found in presence detectors.

There are other reasons for estimating density besides providing model-based techniques with observations. Recall that California Algorithm #7, which is based solely on the use of occupancy data, performs poorly (i.e., does not detect incidents) under low flow conditions. One major reason for this is that an occupancy measurement may only have relevance to the traffic conditions on the road as little as five hundred feet away. (See Section 3.4.) Section density, on the other hand, is a spatial quantity which is more consistently affected by an incident that is occupancy. Hence, a measurement of density would offer the possibility of detecting accidents in low flow conditions.

A new idea for incident detection is also introduced in this section. While occupancy and density are closely related in homogeneous conditions, this relationship fails when inhomogeneities (such as those resulting from an incident) occur. By comparing the estimated density with the density predicted by occupancy, inhomogeneous conditions can be detected.

In Section 5.1.2 a survey of past efforts at density estimation is presented. In Sections 5.1.3 and 5.1.4 the density estimation system is developed. Section 5.1.5 discusses the possibilities of detecting incidents using density estimates alone. The performance of the density estimation system is presented

in Section 5.1.6. Section 5.1.7 concludes Section 5.1.

5.1.2 Past Efforts

The existing density estimation systems are described in papers by Nahi [15], Nahi and Trivedi [16], Gazis and Knapp [17] and Gazis and Szeto [18]. In the two Gazis papers, a procedure for estimating section 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 and accidents may cause significant errors in the travel time algorithm.

The papers by Nahi estimate the density on a link (defined here as a section bounded at each end with presence detectors in all lanes). The recursive estimation system is based upon counting vehicles as they enter and exit the link. Given a good initial density estimate, Nahi's method showed the ability to track the density very closely in homogeneous conditions. No results were presented for inhomogeneous conditions. The issue of imperfect vehicle count information was not considered. The performance with poor initial estimates was also not discussed. Furthermore, an explicit homogeneity assumption was made in the development of the system. This type of assumption is clearly not valid for incident conditions and can be expected to lead to large estimation errors.

5.1.3 The Link Density Estimation System

Motivated by the papers by Nahi [15], [16], a new link density estimation system was developed and is presented in the next two sections. The system uses vehicle count information as well as occupancy measurements from neighboring detector stations to arrive at a density estimate on the link between stations. Detector stations are assumed to have detectors in each lane.

Let $\rho(k)$ denote the actual density (veh/mile per lane) of vehicles on the link at time k. An interval duration of five seconds will be used throughout this study. Vehicle count information can be used to write a conservation-of-vehicles equation:

$$\rho(k+1) = \rho(k) + u(k) + \nu(k) \qquad (5.1)$$

where u(k) represents the measured change in link density between time steps k and k+1. This measurement is obtained from the vehicle count information provided by the detector stations at each end of the link and knowledge of the link length,

Δx, and number of lanes, L, as follows

$$u(k) = \frac{IN(k) - OUT(k)}{L\Delta x} \tag{5.2}$$

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. The noise, $\nu(k)$, models the discrepancies between the actual and measured change in density. In Chapter 2, the source of these discrepancies was found to be from lane changing in the vicinity of the detector station. Using a lane changing behavior model, the statistics of vehicle count errors at a station were found. (See Table 2.1.) From Table 2.1, we obtain the approximate probability of one extra vehicle count occurring in a five second interval at a detector station. Call this probability P_{ec}. It is assumed that not more than one extra count will occur at a station in a five second interval. This probability is approximately the probability that $\nu(k)$, $k = 0,1,2,\ldots$, equals $\pm\frac{1}{XL}$ (see Eq. (5.2)). The probability that $\nu(k)$, $k = 0,1,2,\ldots$, equals zero is, therefore, $1 - 2P_{ec}$. In this manner, the $\nu(k)$'s are modelled as discrete, independent, identically distributed, zero-mean random variables. The variance as a function of flow level is given in Table 5.1.

Note that Eq. (5.1) assumes that the link has no entrance or exit ramps and that the number of lanes does not change along the link. Note also in Eq. (5.1) that knowledge of $u(k)$, $k = 0,1,2,\ldots$ provides no information on the initial density $\rho(0)$. That is, the system is unobservable. To overcome this problem, occupancy measurements are used to provide a rough measurement of ρ. Specifically, the model is

$$z(k) = \rho(k) + \eta(k) \tag{5.3}$$

where the measurement is

$$z(k) = \alpha\left[\frac{occup(k) + occdown(k)}{2}\right] \tag{5.4}$$

Here, occup(k) is the occupancy measured at the upstream detector station over the interval [k, k+1]. Similarly, occdown(k) is the downstream occupancy over [k, k+1].

The parameter α required to convert occupancy to density is taken from Eq. (3.13). Recall from the discussion of Eq. (3.13) that, using 5 second timesteps, (3.13) converts occupancy into a noisy measurement of the density in the

TABLE 5.1

STATISTICS OF ν AS A FUNCTION OF FLOW LEVEL

Average Flow Rate (veh/hr per lane)	Mean of ν	Sample Variance of ν
725	0	.097
1000	0	.103
1600	0	.094

immediate vicinity of the detector station. The noise is zero-mean in homogeneous as well as inhomogeneous conditions. Recall also that the variance of the error in using Eq. (3.13) was found to range from 100 to 200. Thus, averaging two such occupancies together to yield $z(k)$, as in Eq. (5.4), results in an error, $\eta(k)$, which is a white, zero-mean process with a variance which may range from 50 to 100 in homogeneous conditions. This computation assumes that the five second occupancies at neighboring detector stations are independent random variables.

In incident conditions, $\eta(k)$ was observed to have a non-zero mean. This behavior is easily explained. Equation (5.4), in essence, equates the link density with the average of the densities at each link end point. In non-incident conditions, any heavily congested or sparse areas on the road do not remain stationary. That is, these areas move, which prevents the density at the link end points from being consistently higher or lower than the link density. Consequently, the fact that $\eta(k)$ was found to be zero-mean in these conditions is not surprising.

Under incident conditions, a heavily congested area and a sparse area form which do not move. In this case, Eq. (5.4) yields different results depending on the location of the detector stations relative to the incident location. For example, consider the incident depicted in Figure 4.6. The measurements, $z(k)$, obtained using Eq. (5.4) are consistently below the link density because the congestion does not move in time. (See Fig. 4.7.) Thus under incident conditions $\eta(k)$ may have a non-zero mean. This has been observed in many incident simulations.

Eq. (5.1) and Eq. (5.3) are combined to construct a Kalman filter for the estimation of ρ. In Eq. (5.3), $\eta(k)$ is assumed to be zero-mean. The scalar filter equation is

$$\hat{\rho}(k+1) = [1 - H(k)]\hat{\rho}(k) + H(k)z(k) + u(k) \tag{5.5}$$

where $\hat{\rho}(k)$ is the estimate of $\rho(k)$ and $H(k)$, the time-varying Kalman gain, is given by the relations

$$\left.\begin{aligned} H(k) &= \frac{\sigma^2(k)}{\sigma^2(k) + R} \\ \sigma^2(k) &= \sigma^2(k-1) + Q - \frac{(\sigma^2(k-1))^2}{R + \sigma^2(k-1)} \end{aligned}\right\} \tag{5.6}$$

In Eq. (5.6), $\sigma^2(k)$ denotes the variance of the density estimation error at time k. Hence, $\sigma^2(0)$ is an indicator of one's initial uncertainty about $\rho(0)$. Also in Eq. (5.6), Q is the variance of $\nu(k)$ and R is the variance of $\eta(k)$. Other relationships which will be called upon shortly 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

$$H = \lim_{k\to\infty} H(k) = \frac{Q+\sqrt{Q^2+4QR}}{Q+\sqrt{Q^2+4QR}+2R} \tag{5.7}$$

and

$$\Sigma(k) = \frac{R}{1-H(k)} \tag{5.8}$$

Equations (5.5)-(5.8) are well known results (e.g., see Gelb [30]).

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 a minute. This eliminates any need for apriori 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 [15], [16].

Using five second time steps and large initial uncertainty in the initial density, the filter can lock on to the correct density within a minute. This eliminates any need for apriori knowledge of the initial link density, $\rho(0)$.

By monitoring the residual errors

$$r(k) = z(k) - \hat{\rho}(k) \tag{5.9}$$

The advent of a bias in $\eta(k)$ can be detected. Any stationary inhomogeneity could cause such a bias but incidents are the most common cause. A generalized likelihood ratio (GLR) [33], [34] system has been developed to detect the occurrence of such a bias. This GLR system also removes biases in z under such conditions, thus making the ρ estimation system operable under <u>all</u> situations. The GLR bias detection and compensation system is described in the following section.

5.1.4 <u>GLR Bias Detection and Compensation</u>

5.1.4.1 Introduction

In this section, generalized likelihood ratio (GLR) failure detection methods developed by Willsky [33], Chow [34] and others are adapted to the following problem. The system, defined by Equations (5.1) and (5.3), may suddenly develop an unmodelled bias in the observations, z(k). In order to estimate density accurately, this bias must be <u>detected</u> quickly <u>and</u> future estimates must be appropriately <u>compensated.</u>

The essence of the GLR method is as follows. Under no-bias conditions, the residuals (Eq. (5.9)) of the Kalman filter (Eq. (5.5) are a zero-mean, white, Gaussian process [32]. If a bias suddenly occurs in Eq. (5.3), 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 calculations are made to see how far off this unmodelled bias has caused the current density estimate to be. The current estimate and future observations are corrected and the estimation continues.

The detection of and compensation for a bias allows accurate estimation of density in all traffic conditions. It should be emphasized at this point that, besides estimation, the bias detection system also has another important purpose: incident detection. That is, the detection of a bias may very well be the detection of an incident, because a stationary spatial inhomogeneity is the cause of a bias (see Section 5.1.3). Although this issue is dealt with in detail in Section 5.1.5, it should be kept in mind while reading this section.

Certain equations will be referred to repeatedly by name in this section. They are

Eq. (5.1) State Equation

Eq. (5.3) Observation Equation

Eq. (5.1) and (5.3) System Equations

Eq. (5.4) Measurement Equation

Eq. (5.5) Kalman Filter

Eq. (5.9) Residual Equation

In Section 5.1.4.2, the computation of the signature is explained. In Section 5.1.4.3 the statistic upon which a bias or no-bias decision is made is defined and its statistics are determined. In Sections 5.1.4.4 and 5.1.4.5, the selection of a decision threshold is discussed and the time-of-occurrence and magnitude of the bias are estimated. Compensation and computational issues are discussed in Sections 5.1.4.6 and 5.1.4.7 respectively. Finally, Section 5.1.4.8 concludes Section 5.1.4.

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 (5.6) - (5.8)). This steady state assumption greatly eases the analysis of the system and is equivalent to assuming that the initial transient in the density estimate, due to uncertain initial density knowledge, has died out.

5.1.4.2 Computation of the Signature

Because the system equations are linear, the residuals, estimates and observations can be decomposed into two parts.

$$\left.\begin{aligned} r(k) &= r_1(k) + r_2(k) \\ \hat{\rho}(k) &= \hat{\rho}_1(k) + \hat{\rho}_2(k) \\ z(k) &= z_1(k) + z_2(k) \end{aligned}\right\} \tag{5.10}$$

The variables subscripted by 1 denote the value of the residual, estimate or observation under no-bias conditions. The variables subscripted by 2 denote the effect of a non-zero bias on these variables. Thus, prior to the occurrence of a bias, all the variables subscripted with a 2 are equal to zero. During a bias situation $z_2(k)$ is equal to b, the magnitude of the bias, and $r_2(k)$ is the deterministic signature which results from this bias. The variable θ is used to denote the time of occurrence of the bias. The following hypotheses are defined

H_0: no bias exists

H_1: bias exists.

It is then evident that

$$\left.\begin{aligned} E[r(k)\,|\,H_0] &= 0 \\ E[r(k)\,|\,H_1;\, b,\theta] &= r_2(k) \end{aligned}\right\} \tag{5.11}$$

In Eq. (5.10), $\hat{\rho}_2(k)$ denotes the error in the estimate induced by the unmodelled bias, b, which occurred at time θ.

The signature, $r_2(k)$ depends upon the magnitude and time-of-occurrence of the bias as follows [34]

$$r_2(k) = G(k - \theta)b \tag{5.12}$$

where the function $G(k-\theta)$ describes the effect that a unit magnitude bias occurring at time θ has upon the residual at time k.

The effect of a bias of magnitude b, occurring at time θ, on the estimate at time k, $\hat{\rho}_2(k)$, is computed by propagating the bias through the filter equation beginning at time θ and continuing up to time k as follows [34].

$$\hat{\rho}_2(k) = \left\{ \sum_{j=\theta}^{k} (1-H)^j H \right\} b \triangleq F(k-\theta)b \tag{5.13}$$

Again, as a consequence of the linearity of the system, the residual equation can be decomposed into

$$r_2(k) = z_2(k) - \hat{\rho}_2(k) \tag{5.14}$$

so that

$$\left.\begin{aligned} r_2(k) &= b - F(k-\theta)b = [1-F(k-\theta)]b \\ &= G(k-\theta)b \end{aligned}\right\} \tag{5.15}$$

Thus, given the bias starting time θ and the bias magnitude, b, the signature, $r_2(k)$, can be computed using Eq. (5.15) and Eq. (5.15).

5.1.4.3 The Simplified Generalized Likelihood Ratio and the Statistics of the SGLR

Now that the nature of the residuals under H_0 and H_1 have been determined, a hypothesis test will be performed to decide between the two hypothesized conditions. A statistic will be developed on which this decision will be based.

Consider the following functions

$$c(k-\theta) = \frac{1}{\Sigma}\left[G^2(0) + G^2(1) + \ldots + G^2(k-\theta)\right] \tag{5.16}$$

$$d(k,\theta) = \frac{1}{\Sigma}\left[G(0)r(\theta) + G(1)r(\theta+1) + \ldots + G(k-\theta)r(k)\right] \tag{5.17}$$

Note that $c(k-\theta)$ is a deterministic precomputable function but that $d(k,\theta)$ is a (Gaussian) random variable.

From Eq. (5.11), Eq. (5.12) and Eq. (5.15),

$$\left.\begin{aligned} E[d(k,\theta)\,|\,H_0] &= 0 \\ E[d(k,\theta)\,|\,H_1;b,\theta] &= bc(k-\theta) \end{aligned}\right\} \tag{5.18}$$

The statistic to be used is the normalized simplified generalized likelihood ratio (SGLR) [4] which is given by

$$\ell_s(k,\theta) = \frac{d(k,\theta)}{\sqrt{c(k-\theta)}} \tag{5.19}$$

The decision rule is given by

$$\ell_s(k,\theta) \underset{H_0}{\overset{H_1}{\gtrless}} \varepsilon \tag{5.20}$$

where ε is a threshold to be determined. The decision rule of Eq. (5.20) can only detect positive biases ($b > 0$). The statistics of the statistic, $\ell_s(k,\theta)$, must examined under the two hypotheses.

First note that, because $d(k,\theta)$ is a linear combination of the residuals, which are Gaussian random variables, $d(k,\theta)$ and $\ell_s(k,\theta)$ are also Gaussian random variables. Therefore, only a mean and a variance are needed to fully characterize

the probability density function for $\ell_s(k, \theta)$.

From Eq. (5.18) and Eq. (5.19) one gets

$$\left.\begin{aligned} E[\ell_s(k,\theta)\,|\,H_0] &= 0 \\ E[\ell_s(k,\theta)\,|\,H_1;b,\theta] &= b\sqrt{c(k-\theta)} \end{aligned}\right\} \tag{5.21}$$

The residuals under H_0 will have the same variance as the residuals under H_1 since they differ only by the deterministic function $r_2(k)$. Therefore, both $d(k,\theta)$ and $\ell_s(k,\theta)$ will have the same variance under both H_0 and H_1. From Eq. (5.17)

$$E\left[d^2(k,\theta)\,|\,H_0\right] = \frac{1}{\Sigma^2}\; E\left[\sum_{j=\theta}^{k} G^2(j-\theta)r^2(j)\right] = c(k-\theta) \tag{5.22}$$

Using Eq. (5.18)

$$\mathrm{var}[d(k,\theta)\,|\,H_0] = c(k-\theta) \tag{5.23}$$

Now, Eq. (5.22), along with Eq. (5.19) and Eq. (5.21), imply

$$\mathrm{var}[\ell_s(k,\theta)\,|\,H_0] = \mathrm{var}[\ell_s(k,\theta)\,|\,H_1;b,\theta] = 1 \tag{5.24}$$

Thus, Eq. (5.21) and Eq. (5.24) completely describe the probability density function, $f_{\ell_s}(\ell)$, of the random variable $\ell_s(k,\theta)$ under the two hypotheses. (See Figure 5.1.)

5.1.4.4 Selection of the Threshold and Detection Performance

In this section, the selection of the threshold, ε, used in the decision rule of Eq. (5.20) is discussed. It is evident from Figure 5.1 that the GLR bias detection system's missed detection probability, γ, is dependent upon the threshold, ε, as well as $c(k-\theta)$ and b, while the false alarm probability, β, depends only upon ε. Since the performance of the bias detection system is certainly of interest, and is indicated by β and γ, some insight must be gained into the values of b, $c(k-\theta)$ and ε. The threshold, ε, is the only one of the three values which we have the freedom to choose. The bias magnitude, b, is an unknown quantity, while $c(k-\theta)$ is a known function of $k-\theta$. In this section, the function $c(k-\theta)$ is examined, a strategy for selecting ε is presented and γ is determined as a function of b for various values of β.

From Eq. (5.16), 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 sudden

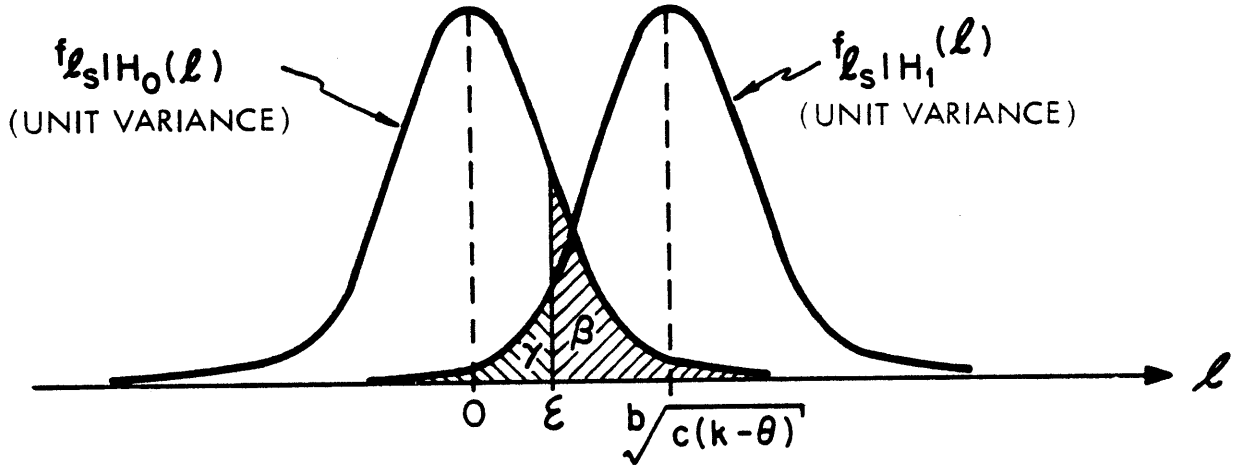


Figure 5.1 PROBABILITY DENSITY FUNCTION OF SGLR UNDER H_0 AND H_1

bias in the observations will cause a gradual bias in the filter estimate. This implies that the signature of the residuals, $G(k-\theta)$, will tend toward zero as $k-\theta$ becomes large. Consequently, from Eq. (5.16), $c(k-\theta)$ will reach a limiting value and will not diverge.

As $c(k-\theta)$ increases, the system false alarm rate, γ, decreases for fixed b and ε, because the two distributions in Figure 5.1 separate. Thus, the limiting value of $c(k-\theta)$ indicates the distinguishability of the two hypothesized conditions. The computation of this value is as follows.

From Eq. (5.13) and Eq. (5.15)

$$\left.\begin{aligned} G(j) &= 1-H(1 + (1-H) + (1-H)^2 + \ldots + (1-H)^{j-1}) \\ &= 1-H\ \frac{1 - (1-H)^j}{1 - (1-H)} \quad = \quad (1 - H)^j \end{aligned}\right\} \tag{5.25}$$

Recall that H is the steady-state Kalman gain given by Eq. (5.7). (Note that the manipulation done in Eq. (5.25) assumes that $0 < H < 1$. However, this conditions is guaranteed by Eq. (5.7) because $Q > 0$ and $R > 0$.) Then, from Eq. (5.16)

$$\left.\begin{aligned} c_\infty = \lim_{n\to\infty} c(n) &= \lim_{n\to\infty} \frac{1}{\Sigma} \sum_{j=0}^{n} G^2(j) = \\ &= \lim_{n\to\infty} \frac{1}{\Sigma} \sum_{j=0}^{n} (1-H)^{2j} = \frac{1}{\Sigma} \cdot \frac{1}{1-(1-H)^2} \end{aligned}\right\} \tag{5.26}$$

where Σ is given by Eq. (5.8). Thus

$$c_\infty = \frac{1-H}{RH(2-H)} \tag{5.27}$$

Equation (5.27) can be expressed entirely in terms of Q and R using Eq. (5.7).

From Table 5.1, we see that a typical value of Q is .10. From the discussion in Section 5.1.3, the value of R is taken to be 100. This results in $H = .031$ and $c_\infty = .16$. Figure 5.2 is a plot of c(n) versus n. It is evident that well over 25 timesteps (i.e., $n > 25$, or 125 seconds) are required before c(n) reaches its limiting behavior. Of course, one does not want to wait a large number of timesteps before making a detection. (See Section 5.1.4.7.) Thus, c(n) will, in practice, never reach its limiting value.

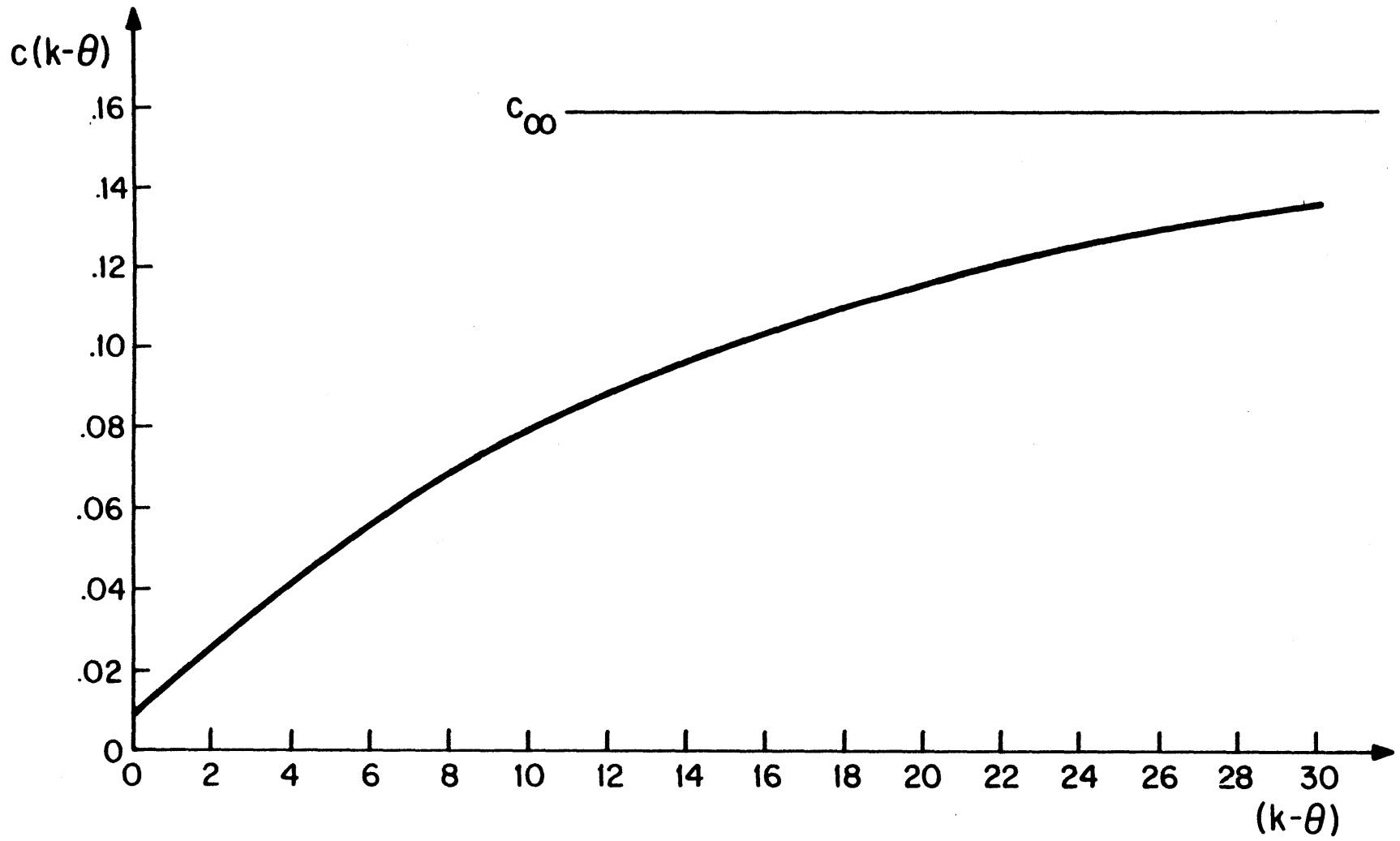


Figure 5.2 PLOT OF c(k-θ) VERSUS k-θ (Q=.1, R=100)

In simulations of this sytem, bias detections are only made for $k-\theta$ in the range of 9 to 13 (see Section 5.1.4.7). (This range on $k-\theta$ corresponds to a time-to-detect between 45 and 65 seconds.) From Figure 5.2 we see that $c(k-\theta)$ does not change appreciably over this range of $k-\theta$. Thus, we now simplify the analysis by assuming $c(k-\theta)$ is approximately constant, given that detections are made within the indicated time window. That is, we assume $c(k-\theta) \sim .08$ for $9 \leq (K-\theta) \leq 13$. The mean time-to-detect is assumed to be 50 seconds.

The threshold, ε, is selected by choosing a desired false alarm probability, β. Figure 5.3 is a plot of β versus ε. Then with c fixed at 0.8 as just described, the missed detection probability, γ, is only a function of the bias magnitude b. (A discussion of typical values of b is presented in Section 5.1.5.). Figure 5.4 is a plot of γ versus b for different values of β.

5.1.4.5 The Estimation of the Bias Magnitude, b, and the Time-of-Occurrence of the Bias, θ

The bias magnitude, b, and the time-of-occurrence of the bias, θ, are crucial unknown parameters that describe incidents. In this section, the estimation of b and θ is presented so that the GLR system can be implemented.

Assuming that a bias does indeed exist, let $\hat{\theta}$ and $\hat{b}$ denote the mazimum likelihood estimates of θ and b, respectively. It is easily shown [34] that $\hat{\theta}$ is the particular value of θ that results in the largest value of $\ell_s(k,\theta)$ for fixed k. That is

$$\ell_s(k,\hat{\theta}) \geq \ell_s(k,\theta), \ \theta \leq k, \ k \text{ fixed} \tag{5.28}$$

The maximum likelihood estimate of b is found as follows. Recall that $\ell_s(k,\theta)$, under H_1, is a Gaussian random variable with unit variance and a mean of $b\sqrt{c(k-\theta)}$. The probability density function $f_{\ell_s|H_1}(\ell)$ is, therefore,

$$f_{\ell_s|H_1}(\ell) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\ell - b\sqrt{c(k-\theta)}\right)^2\right]. \tag{5.29}$$

Because the maximum of a Gaussian distribution is at its expected value, one can solve for the maximum likelihood estimate of b as follows.

$$\ell - \hat{b}\sqrt{c(k-\theta)} = 0. \tag{5.30}$$

But, from the definition of the statistic, ℓ, in Eq. (5.19), we get $\hat{b}$ in terms of the computable quantities $\bar{c}(k-\theta)$ and $d(k,\theta)$. That is,

$$\hat{b}(\theta) = \frac{\frac{d(k,\theta)}{\sqrt{c(k-\theta)}}}{\sqrt{c(k-\theta)}} = \frac{d(k,\theta)}{c(k-\theta)} \tag{5.31}$$

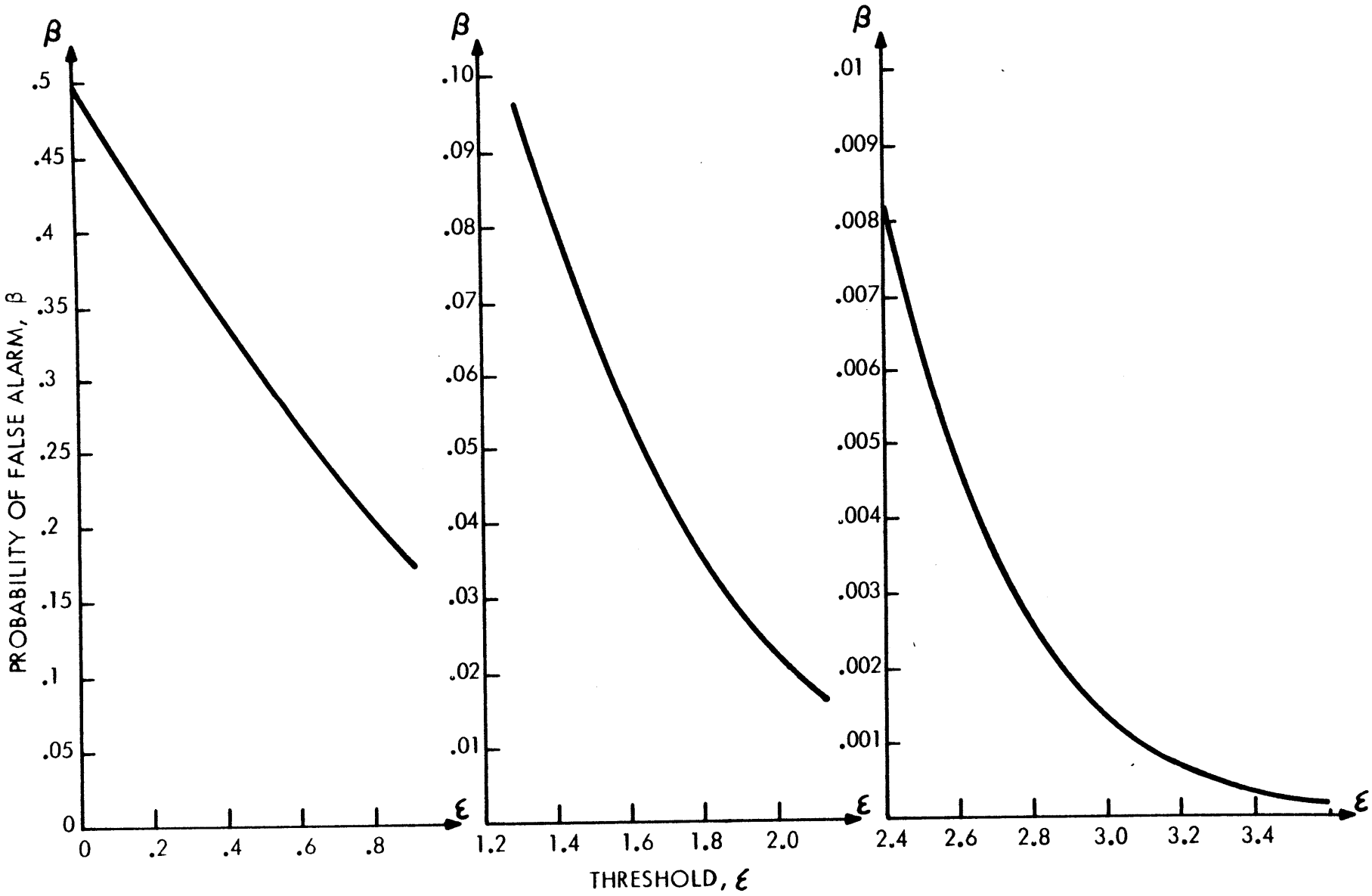


Figure 5.3 PROBABILITY OF FALSE ALARM VERSUS THRESHOLD

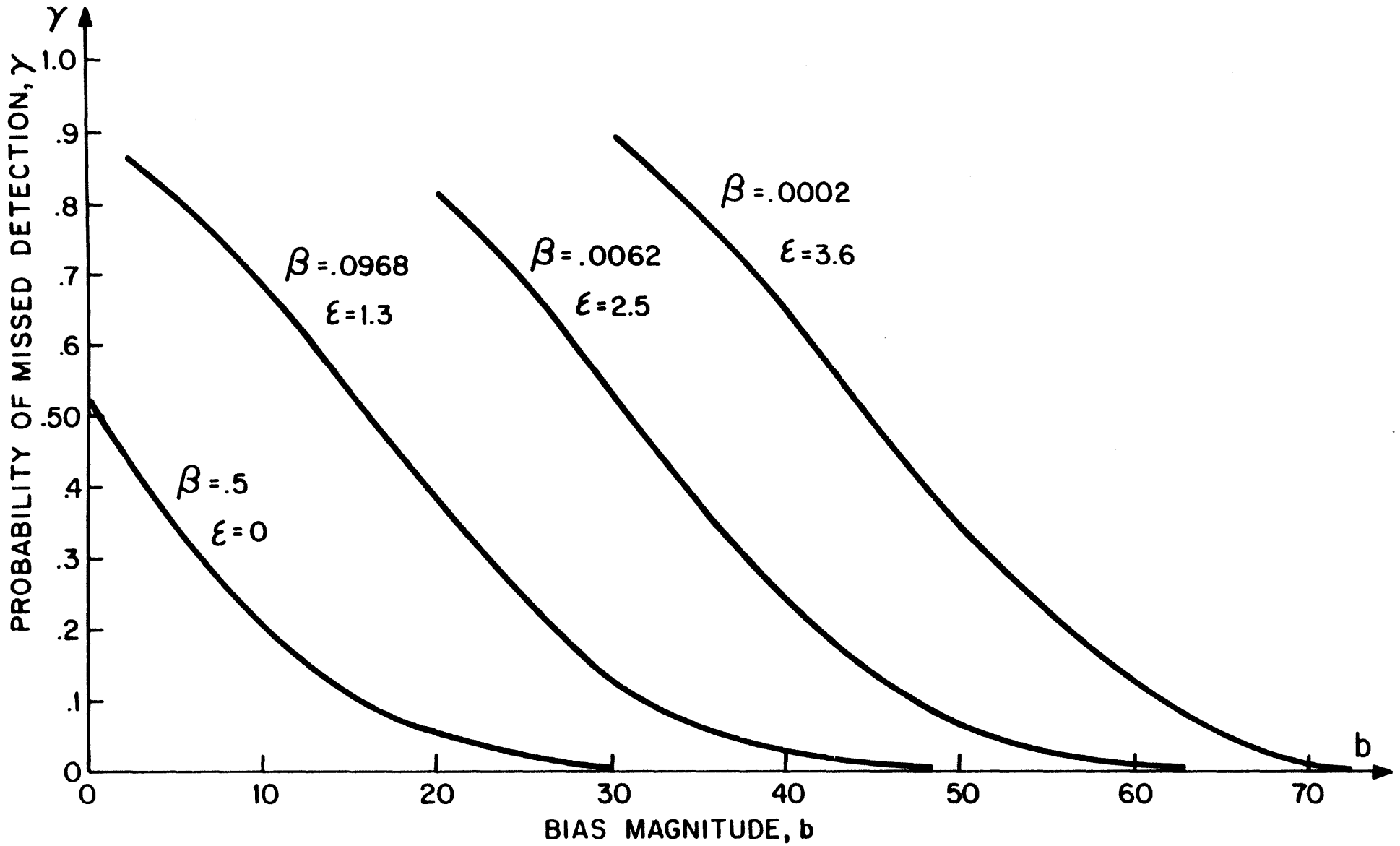


Figure 5.4 GLR DETECTION PERFORMANCE (TIME-TO-DETECT = 50 SEC.)

Equation (5.31) gives the maximum likelihood estimate of b for each θ. However, $\theta = \hat{\theta}$ is the only value of θ of concern here. Therefore, in estimating b, first θ is estimated using Eq. (5.28) and b is then estimated using Eq. (5.31) with $\theta = \hat{\theta}$.

This completes the development of the GLR bias detection system. The compensation system is described in Section 5.1.4.6.

5.1.4.6 Compensation

Suppose a bias of estimated magnitude, $\hat{b}$, was estimated to have occurred at time $\hat{\theta}$. From the filter equation, it is clear that this unmodelled bias will have caused the density estimates to also develop a bias. An estimate of how far off the current estimate is from the actual density is given by $\hat{\rho}_2(k)$ in Eq. (5.13), using $\theta = \hat{\theta}$ and $b = \hat{b}$. Compensation for this error is achieved by correcting the current estimate by this amount. In order to avoid another bias detection on the timestep following the initial detection, the residuals from the present time back to the estimated incident time are recomputed by subtracting out the signature. That is, in Eq. (5.10), $r_1(k)$ is obtained from $r(k)$ by subtracting $r_2(k)$ from it. Then $r_1(k)$, a zero-mean process, is used as the residual process. All future observations are altered by subtracting the bias from the observations. Thus, for a bias, $\hat{b}$, occurring $\hat{\theta}$ timesteps ago, the compensation system is a three step process as follows.

(1) Correct for the error in the current density estimate as follows

$$\hat{\rho}_{new}(k) = \hat{\rho}_{old}(k) - \hat{\rho}_2(k) \quad \text{where } \hat{\rho}_2(k) = \hat{b} \sum_{j=\hat{\theta}}^{k} H(1-H)^j \tag{5.32a}$$

(2) The residuals have contained a signature, for the past θ timesteps. Subtract it out as follows: $r_{new}(j) = r_{old}(j) - r_2(j)$,
$j = k-\hat{\theta}, k-\hat{\theta}+1, \ldots, k$ where $r_2(j) = b(1-H)^{j-\hat{\theta}}$. (5.32b)

(3) Change the measurement equation to

$$z(j) = \alpha\left[\frac{occup(j) + occdown(j)}{2}\right] + \hat{b} \qquad \forall\, j > k \tag{5.32}$$

This will allow the Kalman filter to subtract out the bias from all subsequent measurements.

It should be realized that this system is also able to signal the disappearance of a bias by detecting a negative bias. That is, the decision rule, Eq. (5.20), is altered to the following form.

$$\left.\begin{aligned} &\ell_s(k,\hat{\theta}) > \varepsilon \implies \text{bias occurred} \\ &\ell_s(k,\hat{\theta}) < -\varepsilon \implies \text{bias is over} \\ &-\varepsilon < \ell_s(k,\hat{\theta}) < \varepsilon \implies \text{no bias condition} \end{aligned}\right\} \tag{5.33}$$

It is clear from Figure 5.1, that if a bias occurs and then disappears (corresponding, perhaps, to an incident occurring and then clearing), the probability distribution, $f_{\ell_s|H_1}(\ell)$, will shift from the right to the left side of the $f_{\ell_s|H_0}(\ell)$ distribution. Thus, the need for a negative threshold in Eq. (5.33) is evident. Insight into the behavior of the bias magnitude, b, and its relation to incidents is given in Section 5.1.5.

5.1.4.7 Computation Issues: A Time Window

Computationally, this GLR system requires 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, as time goes on. However, it makes no sense to try to detect an incident when $k-\theta$ is, say, 120 (i.e., 10 minutes) since the $c(k-\theta)$ curve has essentially levelled off by this time. Therefore, waiting longer will not result in a detection. Similarly, one should not declare an incident to have occurred when $k-\theta$ is only 2 (i.e., 10 seconds) because it is not necessary to respond this quickly. Besides, a small amount of additional delay greatly reduces the false alarm probability. This implies that detections should be restricted to a sliding time window. In the version of this system simulated, detections are only made after 9 timesteps but before 13. This corresponds to a time-to-detect between 45 and 65 seconds.

5.1.4.8 Conclusions

In Section 5.1.4 we have developed a system which monitors the residuals of the filter of Section 5.1.3 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 estimated. The error in the current estimate due to the bias is corrected for, and the system continues to estimate section density. Figure 5.5 is a block diagram of the entire density estimation system.

5.1.5 Incident Detection

In Section 5.1.3, it was noted that the biases which occur in the observations are due to stationary, spatial inhomogeneities in the traffic. The GLR bias detection system of Section 5.1.4 is able to detect the occurrence of a bias, or, equivalently, of a stationary, spatial inhomogeneity. However, most

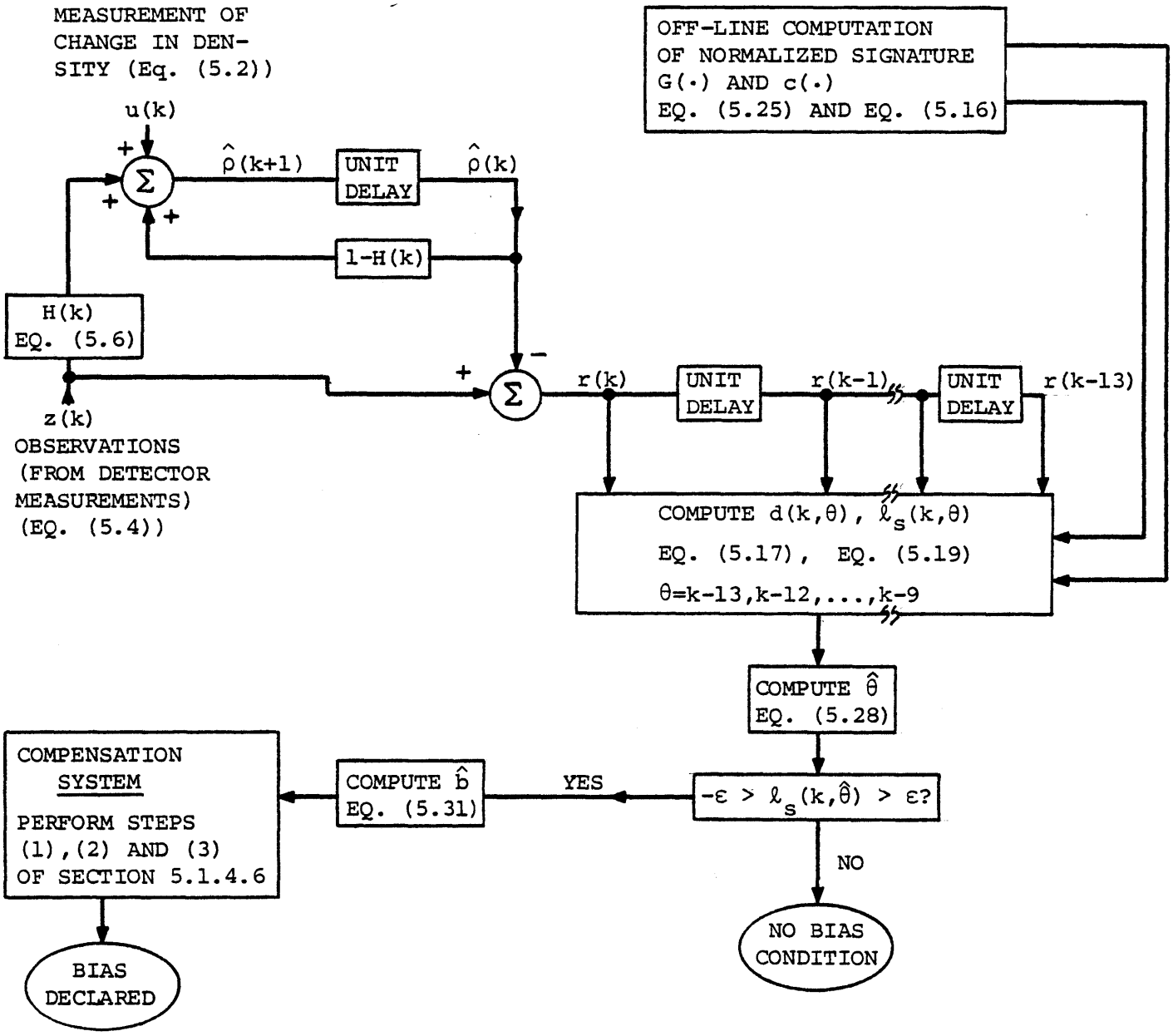


FIGURE 5.5 BLOCK DIAGRAM OF THE DENSITY ESTIMATION SYSTEM

of these bias-producing inhomogeneities are accidents or similar road blockages. Hence, the detection of a bias is, oftentimes, the detection of an accident. As an example of a non-accident traffic condition which causes a bias, consider a section of road just upstream of a curve in the road. On 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 deterministic 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. Non-topological sources of stationary spatial inhomogeneities, at present, cannot be distinguished from accidents.

In an accident at low or moderate flow levels, the associated band of congestion grows and reaches a steady-state length. (see Figure 4.5.). As the band grows, the section density increases proportionately, while the observations from detector stations (Eq. (5.4)) may not. (See Figure 4.7.) 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 time required for the bias to reach its final value is only about fifteen seconds, so that the 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 band of congestion as well as its location relative to the detector stations.

In an accident in heavy flow, the band of congestion grows endlessly in time. As the band grows, the section density increases until the section is totally congested on the upstream side of the accident. The section density then remains approximately constant. Again, the bias magnitude grows and reaches its final value when the section density becomes 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 system can determine, approximately, the magni-

tude of the bias that an accident would produce, if one were to occur, by identifying the level of traffic flow. That is, for the accidents simulated for this report, the bias is dependent mainly upon the flow level. For example, if recent density estimates are, say, around 15 veh/mile/lane, then if an accident were to occur, a bias of around 8 would be expected. (This value was obtained from simulation results.) 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 sytem, 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 purposely elapsed before a bias will be declared present. In a heavy flow accident, when the bias requires more than a minute to grow to its final value, more than one bias detection will 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.

At this point, an incident cannot be distinguished from a stationary non-incident inhomogeneity. Thus, another test must be performed. One possibility is to activate the dynamic model-based incident detection system discussed by E. Chow [4] and Greene[3]only after a bias detection by the density estimation system and let these more sophisticated systems make the distinction. Because these systems are based on a dynamic traffic model, they are capable of distinguishing between incident conditions and normal traffic dynamics. This possibility seems very promising, assuming that <u>all</u> the incidents get detected by the GLR bias detection system. If some are missed, then the

dynamic model based systems will not be activated and will not detect them either. Although no incidents have gone undetected in the simulations thus far, it is conceivable that an incident could result in a zero bias. This could happen if the incident is situated on the link such that the average of the densities at the link endpoints is approximately equal to the link density throughout the incident duration.

It was just mentioned that no accidents were missed by the GLR bias detection system in simulation studies. However, Figure 5.4 apparently indicates that, using a threshold between 2.5 and 3.6, as we did, there is a very high probability of missed detection (especially for small biases). 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 5.4. Suppose that one has selected a threshold (i.e., false alarm probability) and a bias of magnitude b suddenly appears. Figure 5.4 gives the probability that this bias will not be detected exactly 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 found as follows:

$$\gamma_s = \text{Prob[missed at t=45]} \cdot \text{Prob[missed at t=50} \mid \text{missed at t=45]} \cdot \text{Prob[missed at t=55} \mid \text{missed at 45 and 50]} \cdot \text{Prob[missed at t=60} \mid \text{missed at t=45, 50 and 55]} \cdot \text{Prob[missed at t=65} \mid \text{missed at t=45, 50, 55 and 60]}. \quad (5.34)$$

Thus, γ_s is much less than the γ given in Figure 5.4. The calculation of γ_s is difficult due to the correlation between terms in Eq. (5.34). (See [34].)

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

5.1.6 Estimation Performance

In this section the results of simulations of the density estimation

system described in Section 5.1.3 and 5.1.4 and shown in Figure 5.5 are presented. The scenarios selected span a wide variety of traffic conditions. Shown graphically in this section is the estimation performance in incident and non-incident conditions and over a wide range of flow levels. The detections made by the GLR system 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 2.

All graphs in this section plot the actual and estimated link density versus time. Also plotted are the observations of density obtained from occupancy measurements via Eq. (5.4).

The estimated and actual link density on Link 3 of Simulation 29 (see Table 4.2) are shown in Figure 5.6. 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 5.7 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 drivers clogging up traffic (as described in Table 4.2).

The estimated and actual link densities on Link 4 of Simulation 21 are shown in Figures 5.8 and 5.9. The traffic is initially very heavy. An incident occurs at t=180 (see Figure 5.8) and the incident clears at t=540 (see Figure 5.9).

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 5.6 and 5.7. 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

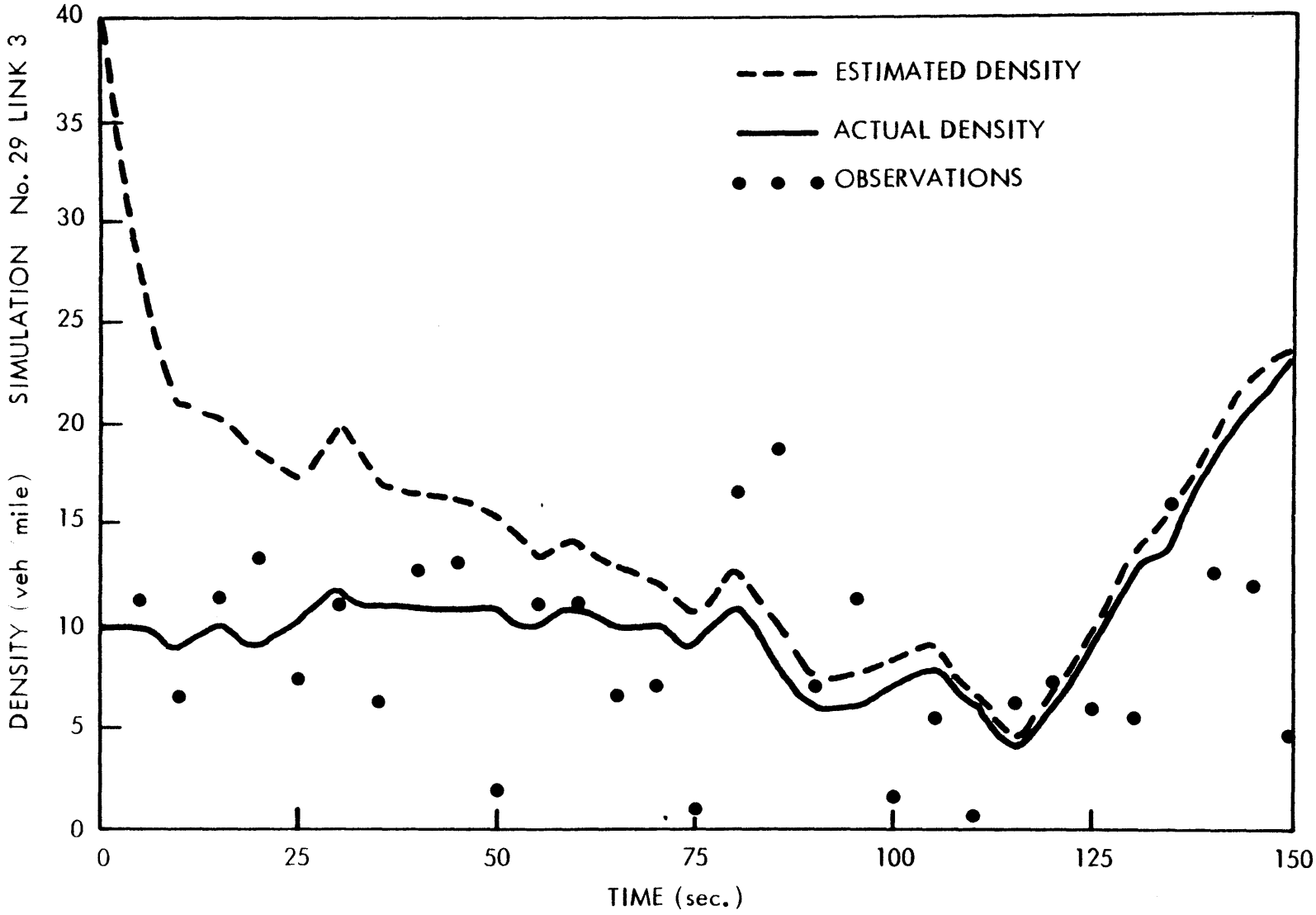


Figure 5.6 ESTIMATION PERFORMANCE: LINK 3, SIMULATION 29

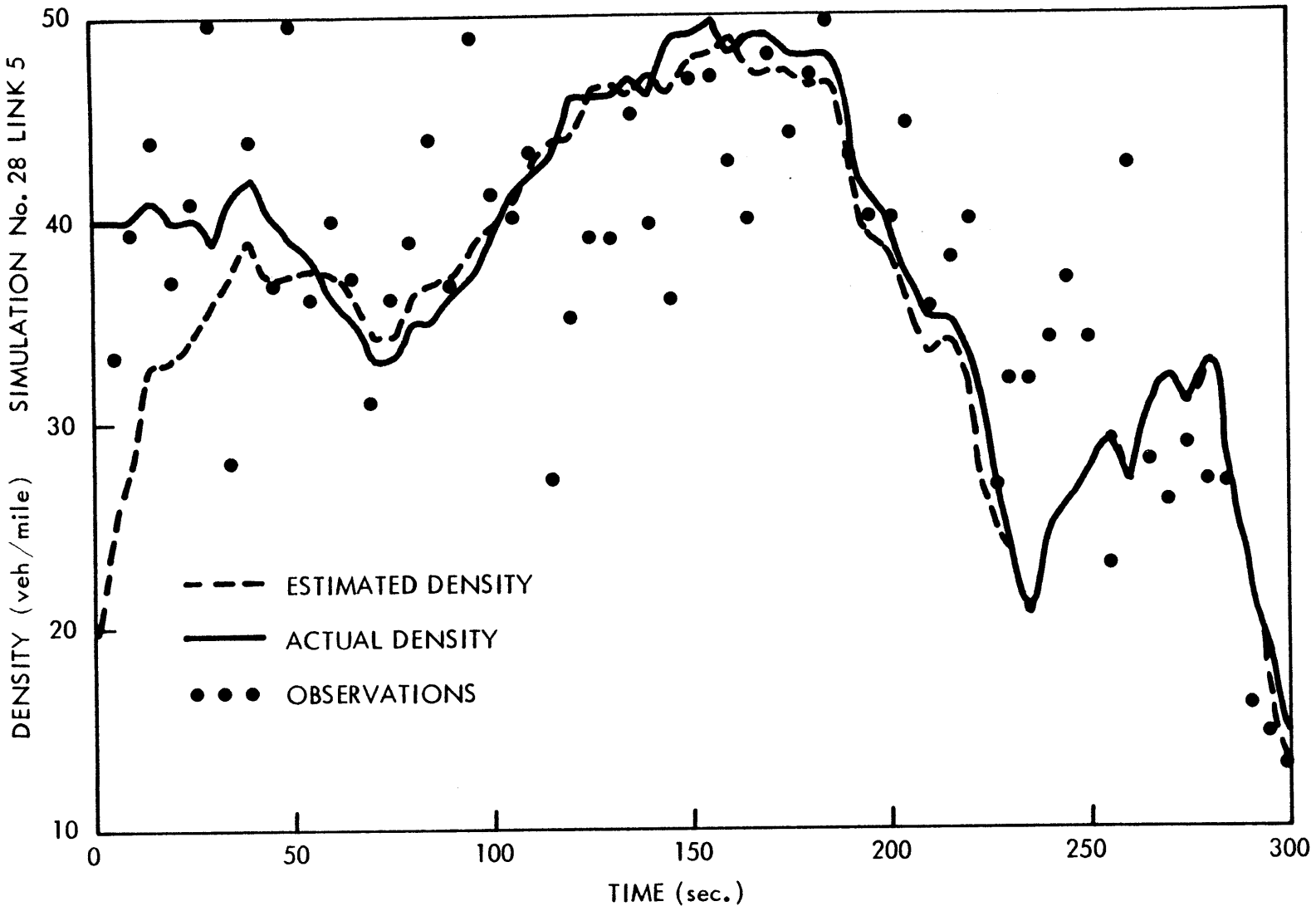


Figure 5.7 ESTIMATION PERFORMANCE: LINK 5, SIMULATION 28

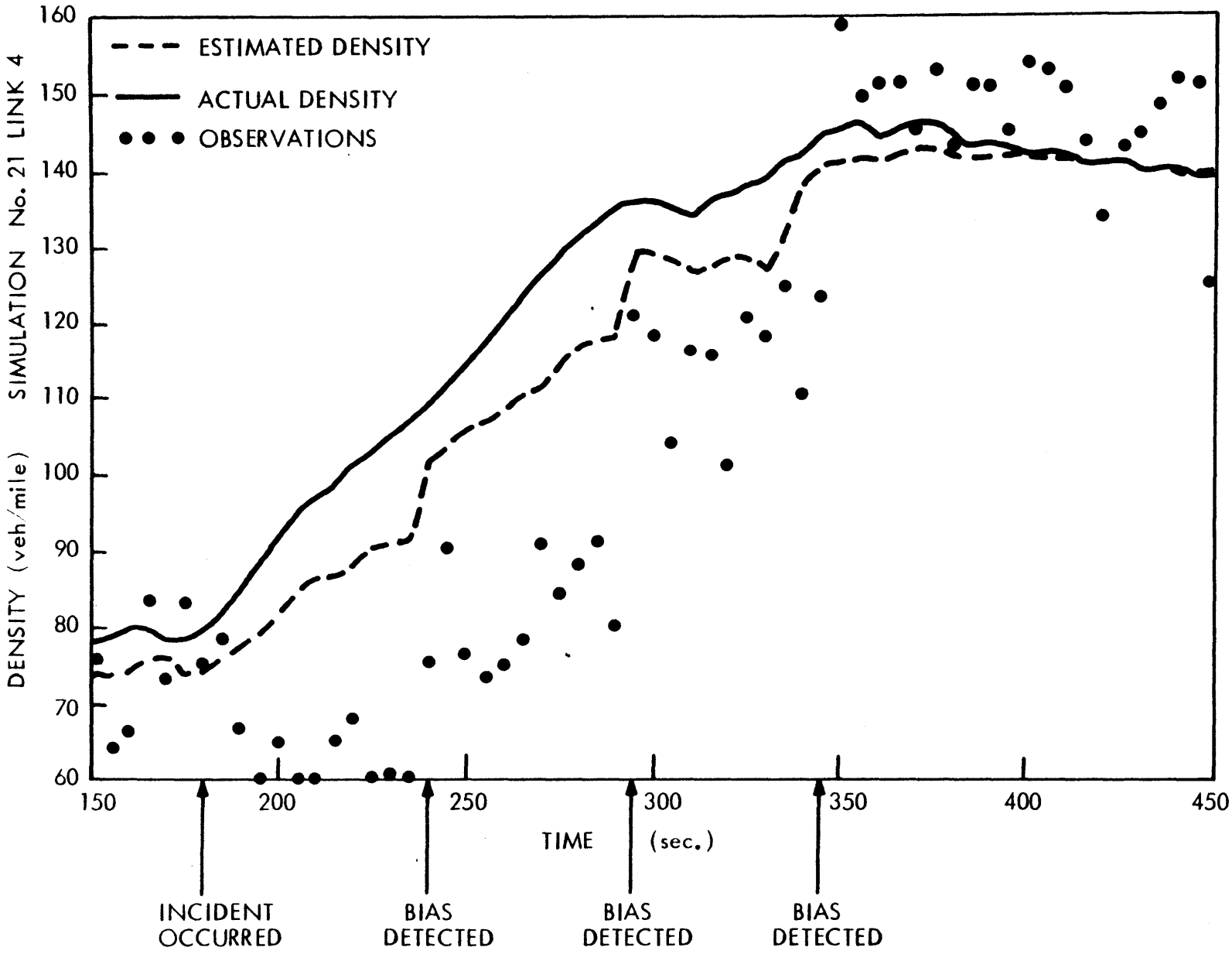


Figure 5.8 ESTIMATION PERFORMANCE: LINK 4, SIMULATION 21

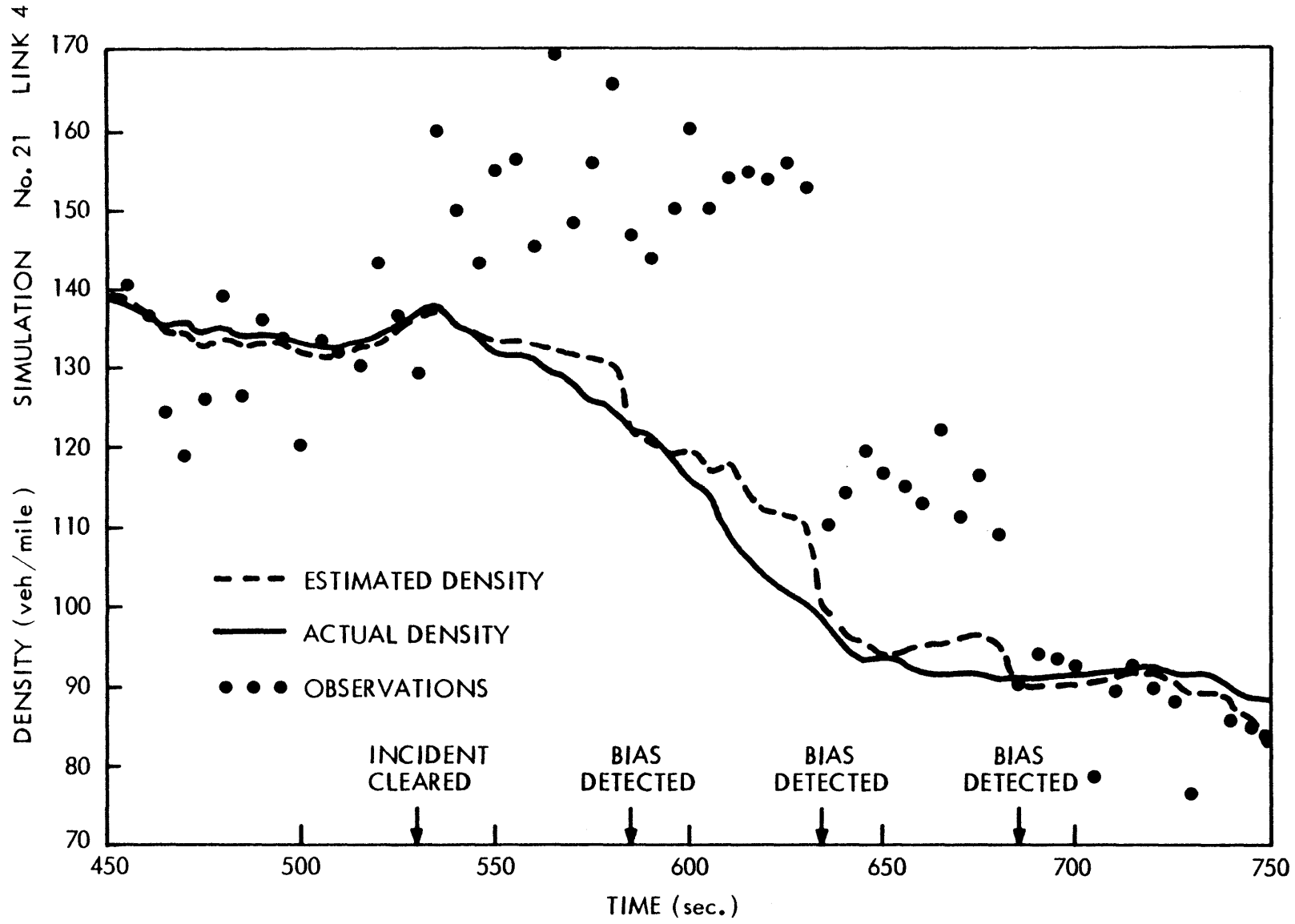


Figure 5.9 ESTIMATION PERFORMANCE: LINK 4, SIMULATION 21 (CONTINUED)

incident occurrence time, θ, was estimated to be 190. The compensation to the estimate was 10.5 and is clearly evident in Figure 5.8. 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 added into the observations at each detection, as discussed in Section 5.1.4.6, which accounts for its step-like rise with time. After the incident cleared, the observations became biased in the other direction and the detections and compensations made are shown in Figure 5.9. Thus, the end of the incident was signalled.

Figure 5.10 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 resulted in a bias in the estimated density. If the bias, b, is accurately estimated, then the observations will become zero mean around the actual density and the bias in the estimate will disappear with time.

Table 5.2 shows the error in the estimates for the simulations of Table 4.2. It is evident from Table 5.2 that this density estimation system provides very good estimates in a wide range of flow conditions and in homogeneous as well as inhomogeneous conditions.

5.1.7 Conclusions

The dynamic model based incident detection systems discussed by E. Chow [4] and Greene [3] uses the density estimates obtained from the system developed in this section. The density estimation system provides excellent estimates under all conditions. The system also shows promise as an incident detection system in itself.

5.2 SPACE-MEAN SPEED ESTIMATION

5.2.1 Introduction

Payne [1] found that the better incident detection algorithms were based upon occupancy or simple functions of occupancy, as opposed to flow measure-

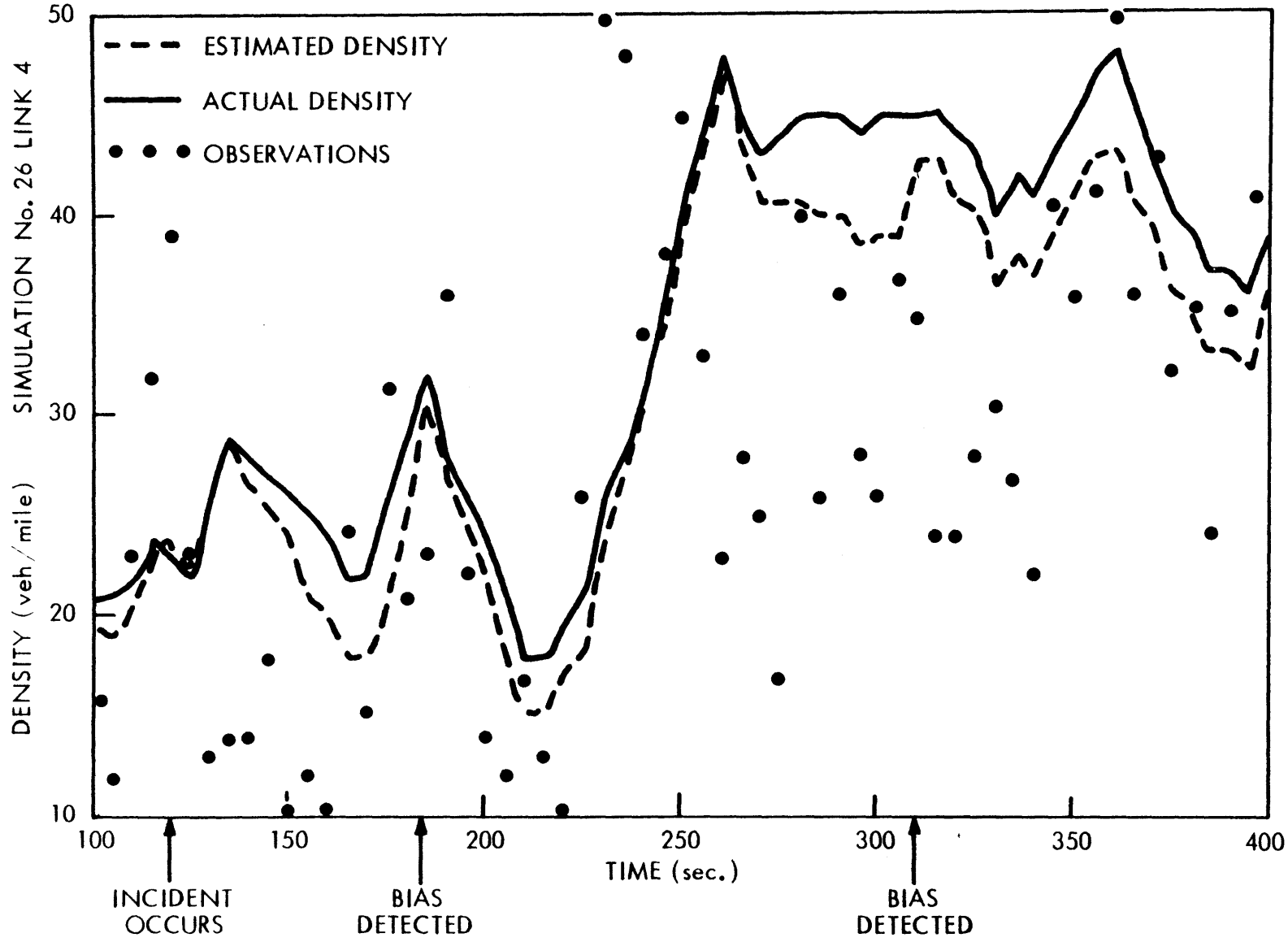


Figure 5.10 ESTIMATION PERFORMANCE: LINK 4, SIMULATION 26

TABLE 5.2

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

ments. Because occupancy is conditionally related to density, this result is consistent with the results of Mitchell [21]. He found section density to be the measurement most crucial to the detection of incidents. However, in addition to a density estimate, the incident detection systems of Volumes III and IV require a space-mean speed estimate. Space-mean speed estimation is the concern of this section.

The space-mean speed estimation system described in this section performs very poorly, relative to the density estimation system of Section 5.1. It is included here only so that readers of the reports by Chow and Greene will be aware of the quality of the space-mean speed estimates used by the dynamic model-based systems and how these estimates were derived from presence detector data.

In Section 5.2.2, other efforts at estimating $\overline{V}_s$ are discussed. Section 5.2.3 presents the method used here. The performance of this new method is described in Section 5.2.4, Section 5.2.5 concludes Section 5.2.

5.2.2 Other Efforts

The major contributions to space-mean speed estimation have come from Mikhalkin [19], [20] and Nahi [15], [16]. Mikhalkin's method consists of first estimating vehicle speeds as they cross the detector and then using a harmonic average of these speeds (see Eq. (3.9) and Appendix B) to obtain a space-mean speed estimate. This method has two weaknesses; (1) the method is complicated because the individual vehicle speed estimation scheme is complex and (2) the harmonic averaging is only valid under space-time homogeneous traffic conditions.

Nahi's method models the space-mean speed as a function of time as a first-order, stationary random process. The model also requires estimates of vehicle speeds as they cross detectors. The only result presented showed excellent space-mean speed estimation in homogeneous traffic conditions. It is not clear how the estimator would perform in incident conditions.

5.2.3 Space-Mean Speed Estimator

A simple new approach to space-mean speed estimation is presented here. The method derives an estimate of $\bar{v}_s$ from the density estimation system of Section 5.1 and flow information from the detector, thus eliminating the need for knowledge of vehicle speeds as they pass over detectors.

The estimator is

$$\hat{\bar{v}}_s(k) = \frac{\hat{\phi}(k)}{\hat{\rho}(k)} \tag{5.35}$$

where $\hat{\rho}(k)$ is the density estimate obtained from the system described in Section 5.1. $\hat{\phi}(k)$ is an estimate of the average flow on the link at time k.

At present $\hat{\phi}(k)$ is obtained by simply averaging the flow rates past each link end point over the k^{th} time interval as measured by the detectors.

Note that Eq. (5.35) is exactly Eq. (3.7) with $\rho=\hat{\rho}$, $\phi=\hat{\phi}$ and $\bar{v}_s=\hat{\bar{v}}_s$. Recall that Eq. (3.7) was derived under space-time homogeneous traffic conditions. Thus the estimator, Eq. (5.35) has two weaknesses; (1) it is valid only under space-time homogeneous conditions and (2) the flow estimate, using 5 second timesteps, is very noisy and may not track the actual average flow on the entire link. The advantages of using Eq. (5.35) to estimate space-mean speed are (1) it is simple and (2) it does not require that the speeds of vehicles be estimated as they cross detectors.

5.2.4 Estimation Performance

The estimation of space-mean speed using Eq. (5.35) is poor. In Table 5.3 the statistics of the estimation error for the simulation of Table 4.2 are presented. From this table it is evident that the variance of the estimate is quite large.

5.2.5 Conclusions

Eq. (5.35) was used to provide space-mean speed estimates to the dynamic model based incident detection systems discussed by Greene [3] and Chow [4]. The estimates are very poor relative to the quality of the density estimates provided by the system of Section 5.1.

TABLE 5.3

SPACE-MEAN SPEED ESTIMATION ERROR STATISTICS

Simulation Identification Number	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	-7.6	.74	-.5	2.4	0.83	4.9	1662.	1055.	1200.	563.	438.	239.
28	-1.7	-3.1	-4.5	-.5	-.44	1.4	379.	257.	170.	101.	109.	80.
26	-4.4	-6.3	-1.5	-13.0	-2.5	-1.0	507.	563.	296.	1014.	485.	405.
22	-13.6	-6.3	-35.	-10.4	-7.5	-2.3	1129.	845.	21069.	970.	799.	532.
21	-6.7	-10.3	-0.2	0.6	-0.6	-5.9	2672.	684.	10.9	8.5	10.6	53.8

6. DISCUSSION AND DIRECTION FOR FUTURE WORK

In this thesis we have discussed a wide variety of topics. Presence detectors were modelled, the occupancy-density relationship was explored, the California Algorithm #7 was examined and methods for estimation density and space-mean speed were developed. Incident detection was the underlying concern of each of these topics and provided continuity to the thesis.

The presence detector model which was developed was a direct result of the findings of Mikhalkin [19]. The model itself is exceedingly simple but was the best attainable from the available literature. It is suggested that a more accurate model be developed for simulating presence detector signals. The recent results of Houpt and Olesik [35] and of Mills [36] may be of interest in this cause.

Using this simple model in conjunction with the lane changing model in the microscopic traffic simulation program, results were stated concerning the frequency that errors are made in counting vehicles at detector stations. These results are believed to be new to the literature and are considered to be an important contribution to this report. The precise error frequency results stated should not be considered accurate results in a real world sense, due to the inevitability of modelling errors. That is, these results should be empirically verified by actually testing at a freeway detector station before interpreting them to be accurate. Nevertheless, the general results that detector stations almost never miss a vehicle and occasionally count a vehicle more than once, due to lane changing in the vicinity of a detector station are certainly accurate statements, based on the results of Mikhalkin[19]. This vehicle count error issue must be dealt with in designing a density estimation system in the fashion of Nahi. However, it seems to have been ignored in the literature.

The study of the relationship between occupancy and density in Chapter 3 resulted in Eq. (3.13). The relationship defined by Eq. (3.13) was found to be accurate in smoothly flowing conditions and, essentially, inaccurate otherwise. The use of this result requires that $E[1/\ell+d]$ be known, but, this may not be readily attainable. Imperfect knowledge of this term will result in a bias in the observations used in the density

estimation system. One can, initially, check for this bias and adjust the value until one gets zero bias under homogeneous conditions. No experimental work has been done with Eq. (3.13) and, again, it is highly recommended that some be done before this result be considered an accurate relation in real-world traffic conditions. It should be realized that many approximations were made in arriving at Eq. (3.13), but that they are all realistic assumptions. The general result that Eq. (3.13) estimates the local density rather noisily but with zero-mean is reliable and very important. The technique of thinking of space-time homogeneity in terms of density maps was crucial to the understanding of the relationship and, in general, is a valuable technique is studying the complex traffic flow process. All these results were developed independent of any information found in the literature and are believed to be new and of value.

The study of California Algorithm #7 in Chapter 4 very specifically identified its fundamental limitations. These results are general in the sense that they also apply to the twenty second version of California Algorithm #7 and other versions of the California Algorithm. Basically, the algorithm has very serious deficiencies at low, or even moderate, flow levels. This result is certainly not obvious at first glance, although it is pointed out in [5] that this algorithm cannot be expected to detect incidents at low flow levels. The biggest emphasis in [5] was on compression waves which cannot, as yet, be realistically simulated. These compression waves were found to be the largest source of false alarms in the algorithm and it is advised that some simulation method be developed.

It is clear from the study of CA-7 in Chapter 4 that an algorithm of this type is simply not sufficient for incident detection purposes. Presence detector data must be converted into spatial quantities (e.g. density and space-mean speed) because local quantities (e.g. occupancy and flow) do not adequately reflect the occurence of an incident due to the random spatial location of the incident. These spatial variables might be used to detect incidents using a CA-7 type of method (i.e. without the aid of a dynamic model describing their evolution). Such a system can be expected to work better than CA-7 because an incident at any point along the roadway will be reflected in the variables measured.

An investigation into this type of incident detection system is recommended.

It is very important that naturally occurring stationary spatial inhomogeneities be distinguishable from actual traffic accidents. A dynamic model describing the spatial and temporal evolution of the traffic variables in accident and non-accident conditions allows the possibility of making such a distinction. The Payne-Isaksen model has been used sucessfully in this regard by Greene [3] and Chow [4]. The results of a comparison between CA-7 and the MM and GLR systems can be found in [1]. Essentially, these new systems displayed the ability to identify a variety of incident types correctly. CA-7, on the other hand, only is capable of detecting accidents. False alarms result in CA-7 from naturally occurring stationary spatial inhomogenieties which are detected as such by the MM and GLR systems.

The MM and GLR systems can detect incidents at flow levels as low as 800 vehicles/hour per lane. This is far below the lowest level that CA-7 can detect at (i.e. about 1200 vehicles/hour per lane) and, in fact, is lower than that of any other incident detection system in existence. The The reason for this improvement over existing methods in detecting incidents at low levels is attributed to the new approach used by these systems. That is, the MM and GLR systems compare the observations (i.e. density and space-mean speed estimates) with those predicted by the Paysne-Isaksen model. If the observations do not fit the model, then an accident may be declared. Because the model describes traffic flow in a wide range of flow conditions, the systems are able to detect incidents in a wide range of flow conditions. Existing methods cannot detect over a broad range of flow levels.

When the density estimates obtained using the method described in Chapter 5 were incorporated into the MM and GLR systems of Greene [3] and Chow [4], the resulting system displayed detection capabilities which significantly exceeded those of CA-7 or any other existing system. Of course, it is recommended that these systems gain some actual experience in real world situations. However, the goal of the overall research effort was indeed reached.

The density estimation system presented in Section 5.1 is regarded as the major contribution of this report. It is believed that this system exceeds the performance of all existing methods and, in additon, is

exceedingly simple. Not only that, it works even in accident conditions and can be used to detect spatially inhomogeneous conditions. No initial knowledge of the freeway conditions are needed to start the system. However, it should be noted that we have assumed that no incident occurs when the system is in its start-up transient. This is a fundamental limitation of this system but is not necessarily a serious problem. That is, the system requires approximately one minute of data in order to lock on to the current density level. Several minutes would be preferable. If an accident were to occur during this period, the density estimates would not be meaningful until the accident congestion reached either a steady state length or fills up the entire link. However, the probability of this event occurring is small, and, even if it does, in fact, occur, the system will merely take longer to lock on to the true density level.

The observations for this system were 5 second occupancy measurements converted to density estimates using the scale factor of Eq. (3.13). It is possible that one can slightly low pass filter these observations and have them still be zero-mean, effectively white and less noisy (i.e. lower variance). This can be expected to aid in detecting smaller biases since H and consequently C_{∞} would rise. Unfortunately, the filtering would smooth out the signature of the incident in the observations (i.e. the sudden bias) and would lead to a longer mean time-to-detect.

The true capabilities of this system cannot be ascertained until it has been experimentally tested. This should be done. It is believed, however, that an experimental implementation of this system would be successful.

The success of this system can be attributed to the fact that the system uses independent measurements of density. That is, car counts are used (as in Nahi's mentod) and occupancy is used (as in Eq. (3.13)). Together, the two provide two different perspectives at which to view the actual density. This is one problem with CA-7. That is, CA-7 uses occupancy measurements only and all flow information is effectively ignored. The independent measurement characteristic of this system is not present in any previously developed density estimation method.

The incident detection capabilities of this system are deemed

exceedingly interesting and are worthy of further development. It is believed that this system can detect incidents at very low flow levels (i.e. around 700 vehicles/hour per lane) if the band of congestion associated with the incident is located between, and not on, detector stations. If the congestion is on top of a station, the resulting bias in the observations will not be as significant as if the incident were between stations, and thus, may not be detectable. However, a CA-7 type of algorithm, with appropriately selected thresholds, performs best when the incident congestion actually covers a detector station. Therefore, a possible realization of an incident detection system might be two systems operating in parallel; (1) the density estimation system/incident detection system and (2) a CA-7 type of algorithm. In very low flow conditions, the CA-7 type of algorithm could detect incidents if the incident congestion actually covered a detector station. If the congestion were between stations, the density estimation system would detect it. In moderate or heavy flow conditions, thedensity estimation system would suffice. Such a simple system would be able to detect low flow incidents and is believed worthy of future work.

There are two major areas which require further development. The first is the model for the bias used in the GLR detection system. Instead of modelling the bias as a sudden event, it should be modelled as a ramp. The rise time would be parameterized by the current flow level. Such a model more reasonably describes the physical situation and can be expected to produce even better results.

Secondly, considerable effort needs to be expended towards the development of a space-mean speed estimation system. The system presented in Section 5.2 is, by no means, considered a contribution to the literature due to its poor performance. However, the system is simple and does not require that speeds of individual vehicles be known as they cross presence detectors. From a practical standpoint, these characteristics should be preserved in the development of a space-mean speed estimation system. This is because the estimation of individual speeds is not a simple matter. Mikhalkin [19] had reasonable sucess and the work of Olesik [35] may become useful in this regard. Olesik showed that the vehicle

type can be identified from the analog signal in the presence detector. This information, along with presence time information, can be combined to yield a speed estimate. However, as pointed out earlier, given a set of individual speeds of vehicles, it is not clear how to combine them into a space-mean speed estimate which is accurate in all types of traffic conditions. Existing methods do not appear to be adequate.

APPENDIX A: SIMULATION OF VEHICLE AND DRIVER TYPES

The microscopic traffic simulation program requires that a distribution of vehicle and driver types be specified. There are six vehicle types and nine driver types modelled into the program. The vehicle types are defined by the vehicle's length and width and the vehicle's acceleration capabilities.

Each of the nine driver types require the specification of 27 parameters. Fast, medium and slow drivers are each modelled by a set of 26 of these parameters. The parameter values were taken from the results of a study (see Mitchell [21]). The 27^{th} parameter is a risk level indicator (either high, medium or low). This quantity indicates the driver's willingness to maintain a small headway and make lane changes. Fast, medium and slow drivers can each be high, medium or low risk level drivers, thus accounting for nine driver types in all.

The user of the simulation specifies the percent of each vehicle type and driver type. The driver types are distributed uniformly among the vehicle types.

A study has been made (see Mitchell [21]) to decide realistic choices for these percentages. The result, referred to as a <u>standard mix</u> of traffic is presented in Table A.1. The standard mix was used in all simulations reported here.

TABLE A.1

STANDARD MIX OF TRAFFIC

Vehicle Types				
Description	Acceleration	Length (ft)	Width (ft)	Percent
car	good	18	6	43.750
car	medium	18	6	39.375
car	poor	18	6	4.375
van	-	20	7	6.250
truck	-	40	8	3.750
truck	-	60	8	2.500

Driver Types		
Driver	Risk Level	Percent
slow	low	1
slow	medium	3
slow	high	1
medium	low	9
medium	medium	27
medium	high	9
fast	low	10
fast	medium	30
fast	high	10

APPENDIX B: HARMONIC AVERAGING OF VEHICLE SPEEDS

Under conditions of space-time homogeneity on a section $[x, x+\Delta x]$ over a time interval $[t, t+T]$, Eq. (3.9) can be used to determine the space-mean speed on the road from the speeds of vehicles as they cross a detector station. For convenience Eq. (3.9) is rewritten here.

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

where $N(t,T)$ is the number of vehicles to cross a fixed point, x_0, $x_0 \in [x, x+\Delta x]$, during the time interval $[t, t+T]$ and the v_j, $j = 1,2,\ldots N(t,T)$, are the speeds of successive vehicles.

Averaging vehicle speeds according to Eq. (B.1) is called harmonic averaging. The purpose of this appendix is to give some intuition into the reason for a harmonic average and to explain why the common arithmetic average is incorrect. An example may help clarify these issues.

Assume that there is an equal number of vehicles travelling at 30 miles/hr and 60 miles/hr on a road. That is, suppose at any instant, there are 10 vehicles travelling at 30 mph and 10 vehicles travelling at 60 mph on a one mile section of road. Let us calculate the rates at which these vehicles cross a detector at the downstream end. Assume that the vehicles at each speed are uniformly spaced along the section. After one minute all 10 60 mph vehicles will have reached the downstream end while only 5 30 mph vehicles will have reached the end. Thus, an observer at the end of the section would see twice as many 60 mph vehicles as 30 mph vehicles pass by when in fact there are equal numbers of each on the road. Clearly an arithmetic average of the speeds of successive vehicles passing the observer is incorrect. Such an average would yield 2/3(60) + 1/3(30) = 45 mph as the space mean speed on the road.

The observer should make his space-mean speed estimate as follows. He observed 15 veh/min (or 900 veh/hr) pass him by. 1/3 of the flow (5 vehicles) was from 30 mph vehicles and the rest (10 vehicles) was from 60 mph vehicles. Assuming space-time homogeneity, the density of 30 mph vehicles is

$$\frac{1/3(900 \text{ veh/hr})}{30 \text{ mph}} = 10 \text{ veh/mile} \qquad \text{(B.2)}$$

and that of 60 mph vehicles is

$$\frac{2/3\ (900 \text{ veh/hr})}{60 \text{ mph}} = 10 \text{ veh/hr} \qquad \text{(B.3)}$$

averaging these two spatial results yields the correct result. Note that the computation of Eq. (B.2) and Eq. (B.3) is the same as Eq. (B.1). The only difference is that flow rates, and not the number of vehicles, is used in Eq. (B.2) and (B.3).

In this example, the arithmetic average was larger than the harmonic average. It can be shown (see Wardrop [27]) that this is always the case.

REFERENCES

[1] A.S. Willsky, et. al., "Dynamic Detection and Identification of Incidents of Freeways Volume I: Summary", MIT Electronic Systems Laboratory, Report ESL-R-764, September 1977.

[2] A.L. Kurkjian, S.B. Gershwin, P.K. Houpt, A.S. Willsky, "Dynamic Detection and Identification of Incidents on Freeways Volume II: Approaches to Incident Detection Using Presence Detectors", MIT Electronic Systems Laboratory, Report ESL-R-765, September 1977.

[3] 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, Report ESL-R-766, September 1977.

[4] E.Y. Chow, A.S. Willsky, S.B Gershwin, "Dynamic Detection and Identification of Incidents on Freeways Volume IV: Generalized Likelihood Ratio", MIT Electronic Systems Laboratory, Report ESL-R-767, September 1977.

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

[6] L.C. Edie, "Flow Theories", Chapter 1 of Traffic Science, D.C. Gazis, Editor, Wiley, 1974.

[7] I. Prigogine, "A Boltzmann-like Approach to the Statistical Theory of Traffic Flow", in Theory of Traffic Flow, Edited by R. Herman, Amsterdam: Elsevier, 1961.

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

[9] M.C. Dunne, R.W. Rothery and R.B. Potts, "A Discrete Markov Model of Vehicular Traffic", Transportation Science, 2, No. 3, p. 233-251 (1968).

[10] A.J. Miller, "A Queueing Model for Traffic Flow," J. Royal Statistical Society, B-23, p. 64-75 (1961).

[11] M.J. Lighthill and G.B. Whitham, "On Kinematic Waves II. A Theory of Traffic Flow on Long Crowded Roads," Proc. Royal Society (London) 229A, p. 317-345 (1955).

[12] D.C. Gazis, R. Herman and R.B. Potts, "Car-Following Theory of Steady State Traffic Flow," Operations Research, 7, p. 499-505 (1959).

[13] W.E. Wilhelm and J.W. Schmidt, "Review of Car-Following Theory", Transportation Engineering Journal, p. 923-933, Nov. 1973.

[14] H.J. Payne, "Models of Freeway Traffic and Control", Simulation Councils Proceedings Series, Vol. I, No. I,Mathematical Models of Public Systems, January 1971.

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

[16] 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.

[17] 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, p. 283-301, August 1971.

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

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

[20] B. Mikhalkin, "Estimation of Roadway Behavior Using Occupancy Detectors", Ph.D. Dissertation, Dept. of Freeway Operations, Los Angèles, California.

[21] 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.

[22] A.D. St. John, "Study of Traffic Phenomena Througn Digital Simulation", Final Report of the Public Health Service, Dept. Health, Education and Welfare, Research Grant AC-000106, 1966.

[23] M. Singleton and J. Ward, "Dynamic Stochastic Control of Freeway Corridor Systems, Volume V: A Comprehensive Study of Various Types of Vehicle Detectors", MIT Electronic Systems Laboratory, Report ESL-R-612, Cambridge, October, 1975.

[24] Discussions with Dr. Paul Houpt, MIT Electronic Systems Laboratory, Cambridge.

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

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

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

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

[29] L. Breiman, "Time Scales, Fluctuations and Constant Flow Periods in Uni-Directional Traffic", Transportation Research, Vol. 7, p. 77-105, Pergamon Press, 1973.

[30] Observations of traffic behavior in the traffic microsimulation program.

[31] 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.

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

[33] 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.

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

[35] J.T. Olesik, "Use of Loop Detector Signatures for Vehicle Speed and Type Classification", E.E. Thesis, MIT, September 1976.

[36] M.K. Mills, "An Investigation of the Driving Point Impedance Change of a Rectangular Inductive Loop Due to a Parralel, Rectangular Wire Grid Simulating a Vehicle", M.S. Thesis, Catholic University, 1975.