CHAPTER 9

Intelligent Transportation and Information Technology

Information Technology (IT) has had a dramatic impact on society and transportation. This chapter discusses applications of IT to transportation systems, which is also called the Intelligent Transportation Systems (ITS) program.

ITS refers to the application of information technologies such as computer software, hardware, communications technologies, navigational devices, and electronics to improve the efficiency and safety of the transportation system. ITS offers a modern approach to meeting the challenges of increasing travel demand that replaces the physical construction of additional capacity with optimization of existing capacity. The benefits of ITS include improving traffic flow, reducing delay, and minimizing congestion. ITS improves the level of service and safety by providing timely information, advanced warnings, and efficient commercial vehicle operations.

Under the umbrella of Intelligent Transportation Systems, several applications can be identified. There are many applications of ITS; some are designed to improve the safety and efficiency of passenger transportation, and others focus on freight transportation. ITS applications may reside within the transportation infrastructure and others within the vehicles themselves. Thus, ITS applications may be referred to as Intelligent (or Smart) Roads or Intelligent (or Smart) Vehicles.

This chapter describes infrastructure-based technologies designed to improve the safety and mobility of passenger transportation. Among the areas covered are

1. Freeway and incident management systems (FIMS)
2. Advanced arterial traffic control (AATC)
3. Advanced public transportation systems (APTS)

4. Multimodal traveler information systems (MTIS)
5. Advanced technologies for rail

The topics covered should furnish an understanding of successful ITS applications and issues related to them. For each area, the operational concept will be described and illustrated by a brief description of real-world examples. The chapter will also discuss the modeling and analysis tools that can be used to aid in the planning, design, and analysis of ITS applications.

FREEWAY AND INCIDENT MANAGEMENT SYSTEMS

Freeway and incident management systems (FIMS) are designed to improve the flow of people and goods on limited access facilities. They combine field equipment (such as traffic detectors, variable message signs, and ramp meters), communications networks, and traffic operations centers and operating personnel. These elements assist FIMS to control and manage traffic efficiently and safely by reducing congestion. Figure 9.1 depicts a typical traffic control center.

Congestion on a freeway occurs when demand exceeds capacity. There are generally two types of congestion: recurrent congestion and nonrecurrent congestion. Recurrent congestion occurs on a regular basis usually during peak hours. Nonrecurrent congestion is less predictable since it is caused by occurrences such as traffic accidents, adverse weather conditions, and short-term construction work. These events result in a reduction in the capacity of a freeway segment and an increase in congestion. FIMS can deal with both types of congestion but are more effective in dealing with nonrecurrent congestion.

FIGURE 9.1
Traffic operations center.

FIMS Objectives and Functions

The objectives commonly identified for FIMS are to

- Continuously monitor the status of traffic flow, and to implement appropriate traffic control actions that reduce congestion;
- Minimize the duration and severity of nonrecurrent congestion by restoring capacity to its normal level;
- Reduce the frequency of recurrent congestion and mitigate its adverse impacts;
- Maximize freeway efficiency and improve safety; and
- Provide real-time travel information concerning the status of traffic that assists drivers to alter route plans.

The functions of FIMS are traffic surveillance; incident detection and management; ramp metering; information dissemination and dynamic route guidance; and lane management. Each of these functions is described next.

Traffic Surveillance

Traffic surveillance is the continuous monitoring of the status of the transportation system. This function provides the basis for all the other functions and applications of ITS, because ITS applications rely on the use of real-time information about the state of the system. The traffic surveillance system collects several types of data, among the most important of which are data on the status of traffic operations. Traffic operations are evaluated on the basis of the three fundamental measures of traffic previously mentioned in Chapter 4: flow, speed, and density. These measures constitute an essential component of the data collected by modern surveillance systems, but additional types of data are also captured by surveillance technologies. These are video images of transportation system operations, queue length, travel time between given origin and destination pairs, location of emergency response vehicles, bus or transit vehicle location, and environmental data, including pavement temperature, wind speed, road surface condition information, emission levels, and air quality.

Traffic Surveillance System Components and Technologies

A surveillance system consists of four components: detection methods, hardware, computer software, and communications. Detection methods use technologies such as inductive loops, nonintrusive detection devices, closed circuit TV cameras, probe vehicles, reports from police or citizens, and environmental sensors to monitor weather conditions. Hardware elements include computers, monitors, controllers, and displays. Computer software is used to convert data collected by detection devices and to interface and communicate with the field devices. The communications system connects the components located at the control center with the field devices.

Detection Methods Detection methods include inductive loop and nonintrusive detectors such as microwave, infrared, ultrasonic, and acoustic sensors; closed circuit TV; video image processing (VIP); vehicle probes, police and citizens' reporting, and environmental sensors. These technologies are described next in terms of characteristics, applications, advantages, and disadvantages.

Inductive loop detectors (ILDs) are widely used for vehicle detection. Their main use is at intersections with advanced signal traffic control systems and on freeways for incident detection and traffic monitoring. ILDs are made of insulated wire embedded in the pavement. The loop is connected with lead-in cable to the detector unit, which senses changes in the inductance within the embedded wire when a vehicle passes over the loop (Figure 9.2).

ILDs can operate in either the pulse mode or the presence mode. In the pulse mode, the loop sends a short signal (typically in the order of 0.125 sec) to the detector unit, and the pulse mode serves to detect volume counts. In the presence mode, the signal persists as long as the vehicle occupies the detection area. The presence mode provides volume counts and the time occupied by the vehicle. ILDs also measure speeds (by installing two pulse loops a short distance apart) and can determine vehicle classification. The ILDs, however, are not always reliable and may fail to operate when damaged by heavy traffic. Furthermore, installation and maintenance of loops require lane closure and modifications to the pavement.

Occupancy calculations where presence-type inductive loops provide occupancy measurements, defined as the proportion of time that a detector is "occupied" or covered by a vehicle during a given time period. Occupancy measurements can be used to calculate traffic density, one of the fundamental

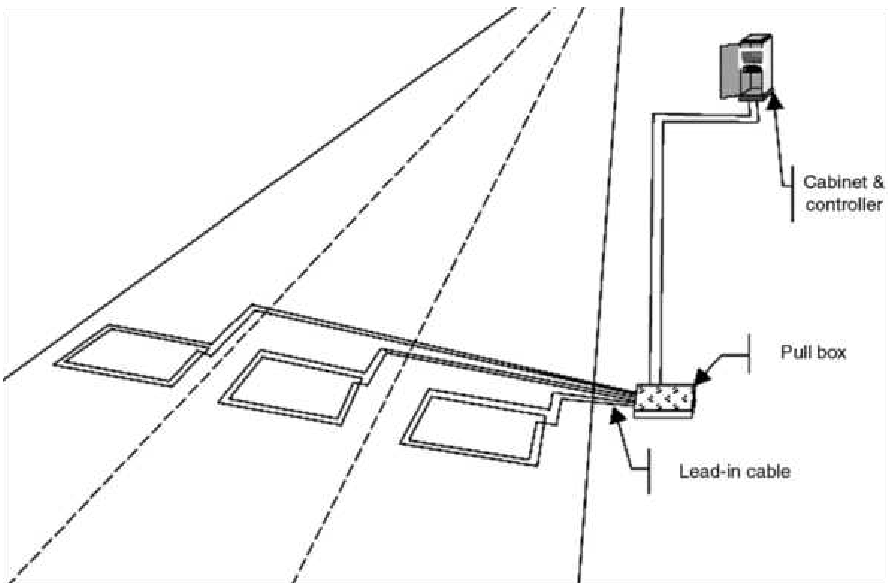


FIGURE 9.2
Inductive loop detectors.

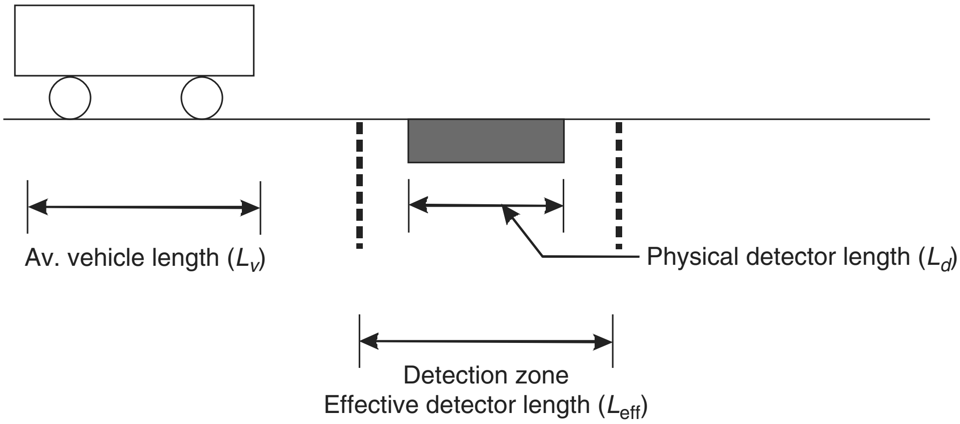


FIGURE 9.3
Occupancy measurements.

measures of traffic flow discussed in Chapter 4, by using an estimate of the average length of vehicles in the traffic stream (L_v) and the effective length of the detector (L_{eff}). The value of L_{eff} is typically greater than the physical length of the loop since vehicles are detected prior to and after they are within the loop (see Figure 9.3).

If the average length of a vehicle (L_v) is known, the following equation can be used to estimate traffic density from occupancy measurements:

$$D = \frac{10 \times Occ}{L_v + L_{eff}} \tag{9.1}$$

where

D = traffic density in veh/km/lane
Occ = occupancy measurement (percent time occupied)
L_v = average vehicle length in meter
L_{eff} = effective detector length in meter

Note that the vehicle and the detector length are added, since the detector is activated as the front bumper enters the detection zone and is deactivated when the rear bumper clears the zone. Since the occupancy measurement is for a single detector in a designated lane, the density value applies only to that lane.

EXAMPLE 9.1

Calculating Traffic Density from ILD Occupancy Measurements

A freeway detection station in one direction of a six-lane highway (3 lanes/direction) provides the occupancy measurements shown in Table 9.1. The average vehicle length of vehicles is 6 m for lane 1, 5.5 m for lane 2, and 5 m for lane 3. The effective length of each loop detector is 2.5 m. Determine the traffic density for (a) each lane, and (b) for the freeway.

TABLE 9.1
Occupancy Measurements

Lane No.	*Occupancy (%)*
Lane 1	22
Lane 2	15
Lane 3	12

Solution

(a) Estimate the density for each lane using Equation 9.1. The results are shown in Table 9.2.

TABLE 9.2
Per Lane Density

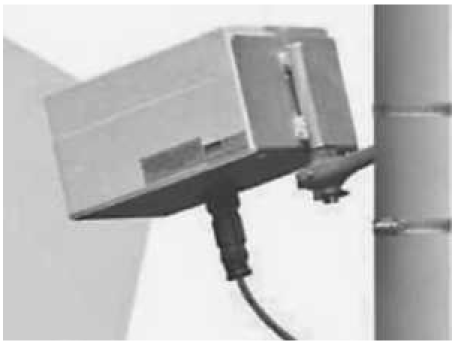
Lane No.	*Occupancy (%)*	*Av. Veh. Length (ft)*	*Density (veh/km/ln)*
Lane 1	22.00	6	25.9
Lane 2	15.00	5.5	18.8
Lane 3	12.00	5	16.0

(b) The overall density for the freeway direction measured is the sum of the per lane density for each lane:

Overall density = 25.9 + 18.8 + 16.0 = 60.7 veh/km

Microwave radar detectors are nonintrusive devices whose installation and maintenance does not require lane closure and pavement modifications since they are mounted on a structure over or to the side of the road (Figure 9.4). The types of data collected by the sensor depend on the form of electromagnetic wave transmitted. Sensors that transmit a continuous wave are designed to sense vehicle speeds by measuring the Doppler shift in the returned wave. They cannot detect stationary vehicles and thus cannot function as a presence-type detector. Microwave sensors that use a frequency-modulated continuous wave (FMCW) can measure speed and detect vehicles. The presence of a vehicle is detected by measuring the change in range when a vehicle enters the field of detection.

FIGURE 9.4
Nonintrusive traffic detector.



A major advantage of microwave sensors is their ability to function under all weather conditions. Since these sensors are installed above the pavement surface, they are not subject to the effects of ice and snow plowing. Microwave sensors can be expected to function adequately under rain, fog, snow, and wind.

Infrared sensors are nonintrusive detectors that can be either passive or active. Passive sensors do not transmit energy but detect the energy that is emitted or reflected from vehicles, road surfaces, and other objects. The amount of transmitted energy is a function of surface temperature, size, and structure. When a vehicle enters a detection zone, it creates an increase in the transmitted energy compared with a static road surface. Passive infrared detectors can measure speed, vehicle length, vehicle volume, and occupancy. Since their accuracy is affected by adverse weather conditions, they are not always reliable.

Active infrared detectors are similar to microwave radar detectors since they direct a narrow beam of energy toward the roadway surface. The beam is then directed back to the sensors, and vehicles are detected by noting changes in the round-trip propagation time of the infrared beam.

Active infrared detectors measure vehicle passage, presence, and speed information. Speed is measured by noting the time it takes for a vehicle to cross two infrared beams that are scanned across the road surface a known distance apart. Some active detectors have the ability to classify vehicles by measuring and identifying their profiles. Accuracy may be compromised by weather conditions such as fog and precipitation.

Ultrasonic detectors are similar to microwave detectors in that they actively transmit pressure waves, at frequencies above the human audible range. The waves can either be continuous or pulse. Detectors that use continuous waves sense vehicles by using the Doppler principle and measure volume, occupancy, and speed. Pulse wave detectors can also determine classification and presence. Since ultrasonic sensors are sensitive to environmental conditions, they require a high level of maintenance.

Acoustic detectors measure acoustic energy or audible sound from a variety of sources both within the vehicle and from the interaction of the tires with the road surface. Acoustic detectors use an array of acoustic microphones to detect these sounds from a single lane on a roadway. When a vehicle passes through the detection zone, a signal-processing algorithm detects an increase in sound energy and a vehicle presence signal is generated. When the vehicle leaves the detection zone, the sound energy decreases below the detection threshold and the vehicle presence signal is terminated.

Acoustic sensors can be used to measure speed, volume, occupancy, and presence. Vehicle classification can also be obtained by matching the sonic signature of a vehicle against a database of sonic signatures for different vehicles. Speed is measured by using an array of microphones such that the time delay of sound arrival will vary for each microphone. The advantage of acoustic sensors is their ability to function under all lighting and weather conditions.

FIGURE 9.5
Integrated video camera/image processing system.

Source: Autoscope Web site, http://www.autoscope.com/.

Video image processing (VIP) is a traffic detection technique that purports to meet the needs of traffic management and control. VIP detectors identify vehicles and traffic flow parameters by analyzing imagery supplied by video cameras. The images are digitized and passed through a series of algorithms that identify changes in the image background. New designs include an integrated machine vision processor, color camera, and zoom lens (Figure 9.5).

An advantage of VIP systems stems from their ability to provide wide area detection across several lanes and in multiple zones within the lane. The user can modify the detection zones through the graphical interface, without the need for pavement excavation or traffic lane closures. The performance of VIP systems can be compromised by poor lighting, shadows, and inclement weather.

VIP and closed circuit TV (CCTV) can be combined to provide an excellent detection tool, particularly for incident detection and verification purposes. When an incident occurs, the user can switch from the VIP mode to the standard CCTV mode and verify the occurrence of the incident via pan/tilt/zoom controls.

Vehicle probes involve tracking vehicles, using positioning and communications technologies, and communicating the location of the vehicle to a central computer where data from different sources are combined to determine the status of traffic flow over the measured transportation system. Vehicle probes can provide useful information not available from other detection techniques. These include link travel times, average speeds, and origin–destination information. Three different technologies that use vehicles as probes are automatic vehicle identification (AVI); automatic vehicle location (AVL); and anonymous mobile call sampling (AMCS).

Automatic vehicle identification (AVI) can identify vehicles as they pass through a detection zone. A transponder (or tag) mounted on the vehicle is read by a roadside device, using dedicated short-range communications (DSRG). The information is then transmitted to a central computer. The most common application of AVI technologies is in conjunction with automatic toll collection systems. The toll charge is automatically deducted from a driver's account when a vehicle enters the toll plaza. This technology can also be used for detection purposes by determining the average travel time on freeways between roadside readers.

Automatic vehicle location (AVL) determines the location of vehicles as they travel over the network. AVL can locate and dispatch emergency vehicles, locate buses in real time, and determine their expected arrival time at bus stops. Several technologies are used for AVL, including dead reckoning, ground-based radio, signpost and odometer, and the global positioning system (GPS), which is the currently most commonly used technology for location identification and navigation. To operate, GPS relies on signals transmitted from 24 satellites orbiting the earth at an altitude of 20,200 km. GPS receivers calculate the location of a point using the time it takes for those electromagnetic signals to travel from the satellites to the GPS receiver.

Anonymous mobile call sampling uses triangulation techniques to determine a vehicle's position by measuring signals emanating from a cellular phone within the vehicle. This concept provides a wealth of information at a relatively low cost. Anonymous mobile call sampling requires two elements: a geo-location control system (GCS), and a traffic information center. The GCS provides the latitude and longitude of cellular probes, which is communicated to the traffic information center, where the information is fused and analyzed. This concept was first tested in the Washington, D.C., area in the mid-1990s.

Mobile reports constitute another source of freeway surveillance information. In many cases, reports of incidents from citizens and the police can provide system monitoring information, at a lower cost than other surveillance technologies. Mobile reports provide event information at unpredictable intervals that could be useful for traffic management purposes. In particular, mobile reports are effective for incident detection. Examples of mobile reporting methods include cellular phones and freeway service patrols.

Cellular phones can serve as an effective tool for incident detection. Many jurisdictions around the country have established an incident reporting hotline to encourage citizens to report traffic incidents. This method has the advantage of low start-up costs. A freeway service patrol consists of a team of trained drivers who are responsible for covering a given segment of the freeway. The freeway service patrol vehicle is equipped to help stranded motorists and to clear an incident site. Examples of the supplies used by a service patrol vehicle include gasoline, water, jumper cables, vehicle repair tools, first-aid kit, push bumpers, and warning lights. Service patrols can then detect incidents and perform the entire incident management process of detecting and clearing an incident.

Environmental sensors are used to detect adverse weather conditions, such as icy or slippery conditions. This information can then be used to alert drivers via variable message signs (VMS) and can be used by maintenance personnel for optimizing maintenance operations. Environmental sensors can be divided into road condition sensors, which measure surface temperature, surface moisture,

FIGURE 9.6
Environmental station.
Source: Nu-Metrics Web site.

and presence of snow accumulations; visibility sensors, which detect fog, smog, heavy rain, and snowstorms; and thermal mapping sensors, which can be used to detect the presence of ice. In addition, many manufacturers currently provide complete weather stations that are capable of monitoring a wide range of environmental and surface conditions. Figure 9.6 shows one example of a weather station.

Hardware Computer hardware is the second component of a traffic surveillance system. Computers receive information from field devices and sensors; communicate data from the control center to field devices (e.g., control data to provide for pan/tilt/zoom of a field CCTV camera); process data to derive meaningful traffic parameters from real-time data collected by sensors; and archive the data collected.

In addition to computer hardware, a surveillance system typically includes graphical displays in the control center to provide for a visual description of the transportation system operations as obtained from the field cameras. Graphics can be provided on the workstations' monitors or on large-screen graphics displays. These displays take the form of an array of video screens (Figure 9.1).

Software Computer software constitutes the third component of a traffic surveillance system. Examples of traffic surveillance system software include incident detection algorithms, decision support systems (DSS) for incident management, and software for controlling field devices. Following sections discuss incident detection algorithms and decision support systems for incident management.

Communications System The communications system is needed to provide communications among the components of a control center and between the control center and the field devices. Communications within the center is accomplished via a local area network (LAN). Between the center and the field devices, a wire-line communication system (e.g., fiber optic, coaxial cable, twisted pair) or wireless system is used. The choice of the communications medium (for example, fiber optic versus coaxial cable) depends on the bandwidth requirements of the data transmitted. For example, video requires a wide bandwidth that can best be met using fiber optic cables.

Incident Management

The second function provided by a FIMS is incident management. Congestion on freeways can be either recurrent or nonrecurrent. Incident management systems are designed primarily to deal with nonrecurrent congestion conditions. Incident management is defined as "a coordinated and planned approach for restoring traffic to its normal operations after an incident has occurred." An incident can be a random event (a freeway accident or a stalled vehicle) or a planned or scheduled event (such as work zone lane closure). In both cases, the goal is to systematically utilize both human and mechanical resources to achieve the following:

- Quickly detect and verify the occurrence of an incident.
- Assess the severity of the situation and identify the resources needed to deal with the situation.
- Determine the most appropriate response plan that will restore the facility to normal operation.

The process of incident management can be conceptually viewed as consisting of the following four sequential stages:

- Detection and verification
- Response
- Clearance
- Recovery

The goal of the incident management process is to reduce the time needed to complete each stage and to restore normal operations. A brief discussion of these four stages follows:

Incident Detection and Verification

Incident detection is the identification of an incident. Verification is the acquisition of information about the incident, such as its location, severity, and extent. Verification provides the information used to devise an appropriate response plan. Incident detection and verification have always been the responsibility of state and local police. Technologies now available for incident detection and verification augment these functions and are either nonautomated or automated. Nonautomated detection techniques include cellular phone calls to a 911 or 511 number, an incident reporting hotline, freeway service patrols, citizen-band radio monitoring, motorist call boxes, and fleet operators. These techniques often serve an important role in the incident management process as a supplement to automated surveillance technologies. Automated incident detection methods will be discussed in a following section.

Incident Response

With an incident detected and verified, the next step in the incident management process is that of incident response. Incident response involves the activation, coordination, and management of personnel and equipment to clear the incident. Traffic incident response can be divided into two stages. Stage 1 is concerned with identifying the closest incident response agencies required to clear the incident, communicating with those agencies, coordinating their activities, and suggesting the required resources to deal with the incident effectively. Stage 2 involves traffic management and control activities aimed at reducing the adverse impacts of the incident. These include informing the public about the incident using variable message signs (VMS) or other information dissemination devices, implementing ramp metering and traffic diversion strategies, and coordinating corridor-wide traffic control strategies.

The primary goal of incident response technologies is to optimize resource allocation and minimize response time. The three elements of response time are verification of the occurrence and location of an incident; dispatch of an incident response team; and travel time of the incident response team. A number of techniques and technologies are available to reduce response time, including incident response manuals, tow truck contracts, techniques to improve emergency vehicle access, and improved traffic flow through alternative route planning.

Incident Clearance

Incident clearance refers to the safe and timely removal of an incident. There are several technologies for improving the efficiency of incident clearance. Inflatable air bag systems are one example of such technologies. The main purpose of these systems is to restore an overturned vehicle to an upright position. The system consists of rubber inflatable cylinders having various heights. These cylinders are placed underneath the overturned vehicle and inflated until the task is complete.

Incident Recovery

This stage refers to the time taken by traffic to return to normal flow conditions after the incident has been cleared. The goal is to use appropriate traffic management techniques to restore normal operations and to prevent the effect of congestion from spreading elsewhere.

Automatic Incident Detection Methods

Automatic incident detection (AID) uses algorithms to detect incidents in real time using data collected from traffic detectors. The development of AID algorithms began in the 1970s, and since then many algorithms have been developed. Evaluation of AID algorithms is based on the detection rate (DR); the false alarm rate (FAR); and time to detect (TTD).

Detection rate (DR) is a measure of how effective an AID algorithm detects incidents. It is the ratio of the number of incidents that the algorithm detects and the total number of incidents that occurred. Values range from 0 to 100% and the closer to 100, the more effective the algorithm.

False alarm rate (FAR) is the ratio of the number of false detections and the total number of observations. Most algorithms observe incidents at regular time intervals, such as every 30 seconds or every minute. The FAR results are a percentage for each detector station or simply the total number of false reports over the time period observed.

Time to detect (TTD) is the time difference between the moment an incident was detected and when the incident occurred. The *mean time to detect* (MTTD) is the average TTD over a specified number of incidents.

The three parameters are correlated. For example, increasing the value of the DR results in a corresponding increase in the FAR. If the algorithm time to detection is increased (TDD), both the DR and FAR values would improve. Experience with deployed AID has not always been favorable. In many cases, the number of false alarms that AID algorithms produce has become so unreliable that several traffic operations centers have stopped using them.

EXAMPLE 9.2

Calculating the Detection and False Alarm Rates for an AID Algorithm

A certain automatic incident detection (AID) algorithm is used in a traffic management center. The algorithm is applied every 30 s. To evaluate the performance of the algorithm, traffic was observed over a period of 30 days in which a total of 57 incidents took place. Of this number, the algorithm correctly detected a total of 49 incidents. The algorithm provided 1000 false alarms during the observation period. Determine (a) the DR and (b) FAR for this algorithm.

Solution

(a) DR:

The DR is the ratio of the number of incidents detected to the total number of incidents occurring. Thus

$$DR = \frac{49}{57} \times 100 = 86\%$$

(b) FAR is the ratio of the number of incorrect detections to the total number of times that the algorithm was applied. Therefore, it is necessary to first determine the number of times the algorithm was applied. Since the algorithm is applied every 30 s, and the observation period was 30 days long, the number of times the algorithm was applied is

$$30 \text{ days} \times 24 \text{ hours} \times 60 \text{ minutes} \times 2 \text{ applications/minute} = 86{,}400 \text{ times}$$

Thus

$$\text{FAR} = \frac{1000}{86400} \times 100 = 1.16\%$$

Although the FAR rate is relatively low (only about 1% in this problem), the absolute number of false alarms (1000) is high, which could become quite annoying to operators in traffic centers.

Comparing the Performance of Incident Detection Algorithms

The performance index, PI, is a measure that is used to compare different AID algorithms. This measure can also be used to calibrate the algorithms for a given location. The performance index is defined in Equation 9.2, with lower values of PI indicating better performance of the algorithm:

$$PI = \left[\frac{(100 - DR)}{100}\right]^m \times FAR^n \times MTTD^p \tag{9.2}$$

where

DR and FAR = detection and false alarm rates, respectively
$MTTD$ = mean time to detect in minutes
m, n and p = coefficients that can be used to emphasize or weigh how the three performance measures are used in evaluating an algorithm's performance (e.g., using higher values for the coefficient m compared to n and p would emphasize the role of the detection rate for the algorithm in judging its performance)

EXAMPLE 9.3

Comparing the Performance of AID Algorithms

The performance of seven AID algorithms was evaluated by recording the DR, FAR, and MTTD for each algorithm. The results are shown in Table 9.3. Values of the coefficients m, n, and p are all equal to 1.0. Using Equation 9.2, determine how each AID algorithm performs. Which of these is preferred?

TABLE 9.3
Comparing the Performance of AID Algorithms

	DR (%)	*FAR (%)*	*MTTD (min)*
AID1	82	1.73	0.85
AID2	67	0.134	2.91
AID3	68	0.177	3.04
AID4	86	0.05	2.5
AID5	80	0.3	4
AID6	92	1.5	0.4
AID7	92	1.87	0.7

Solution

To solve this problem, the PI is computed for each AID. The results are shown in Table 9.4. From Table 9.4, it can be seen that AID4 has the lowest PI value and is therefore the algorithm with the best performance.

TABLE 9.4
PI Calculations for $m = 1; n = 1;$ and $p = 1$

	DR (%)	*FAR (%)*	*MTTD (min)*	*PI*
AID1	82	1.73	0.85	0.265
AID2	67	0.134	2.91	0.129
AID3	68	0.177	3.04	0.172
AID4	**86**	**0.05**	**2.5**	**0.018**
AID5	80	0.3	4	0.240
AID6	92	1.5	0.4	0.048
AID7	92	1.87	0.7	0.105

EXAMPLE 9.4

Detection Time of AID Algorithms

In the previous example, the traffic engineer is interested in emphasizing the importance of the quick detection of incidents. Given this, the engineer decides to rerun the analysis but doubling the value for the p coefficient (i.e., $p = 2$). Which algorithm would be judged to have the best performance in this case?

Solution

The analysis is rerun with $m = 1$, $n = 1$, and $p = 2$. The results are shown in Table 9.5.

In this case, AID6, which has a mean detection time of only 0.4 min, is judged to be the best algorithm.

TABLE 9.5
PI Calculations

	DR (%)	*FAR (%)*	*MTTD (min)*	*PI*
AID1	82	1.73	0.85	0.225
AID2	67	0.134	2.91	0.374
AID3	68	0.177	3.04	0.523
AID4	86	0.05	2.5	0.044
AID5	80	0.3	4	0.960
AID6	**92**	**1.5**	**0.4**	**0.019**
AID7	92	1.87	0.7	0.073

Types of AID Algorithms

AID algorithms can be broadly divided into four groups based upon the principle behind the algorithm's operation: (1) comparative-type or pattern recognition algorithms; (2) catastrophe theory algorithms; (3) statistical-based algorithms; and (4) artificial intelligence–based algorithms. This section describes the comparative-type or pattern recognition algorithms as these form the basis for other applications. Other algorithms are briefly discussed.

Comparative-type or pattern recognition algorithms These are among the most commonly used AID algorithms. These algorithms are based on the premise that the occurrence of an incident results in an increase in the density of upstream traffic detectors and in a decrease in density for the downstream detectors. Figure 9.7 illustrates this phenomenon.

Comparative-type algorithms attempt to distinguish between "usual" and "unusual" traffic patterns by comparing the values of the traffic volumes, densities, and speeds at upstream and downstream detector stations with preestablished thresholds. If the field-observed values exceed established thresholds, an alarm is triggered indicating that an incident may have occurred. The most challenging part in implementing comparative-type algorithms involves establishing values for preestablished thresholds since values differ for specific freeway locations. A few examples of comparative-type algorithms are given below.

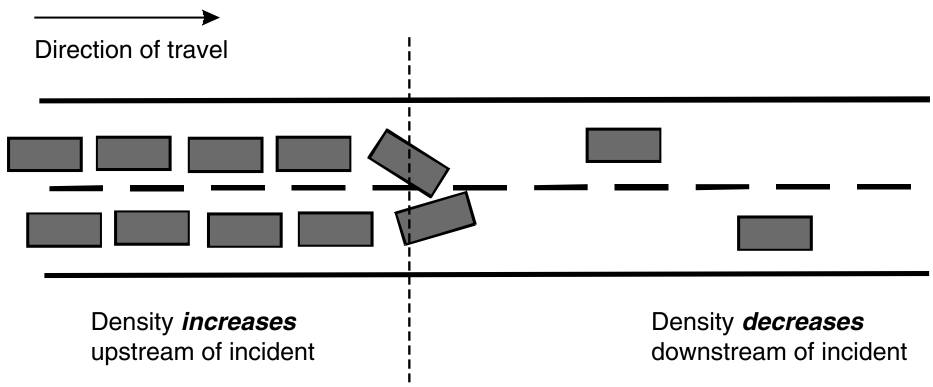


FIGURE 9.7
Occupancy changes as a result of an incident.

The California algorithm is one of the earliest comparative-type AID algorithms to be developed and is often used for comparisons and benchmarking. The algorithm tests for an incident by comparing occupancy (density) values from two adjacent detection stations, according to the following logic:

Step 1: The difference between the upstream station occupancy (OCC_{up}) and the downstream station occupancy (OCC_{down}) is compared against threshold T_1. If the threshold value is exceeded, then the algorithm proceeds to step 2.

Step 2: The ratio of the difference in the upstream and downstream occupancies to the upstream station occupancy ($OCC_{up} - OCC_{down}$)/OCC_{up}) is checked against threshold T_2. If this threshold is exceeded, the algorithm proceeds to step 3.

Step 3: The ratio of the difference in the upstream and downstream occupancies to the downstream station occupancy ($OCC_{up} - OCC_{down}$)/OCC_{down} is checked against threshold T_3. If this threshold is exceeded, a potential incident is indicated. No alarm is indicated, yet step 2 is repeated for the following time interval. If thresholds T_2 and T_3 are again exceeded, a potential incident is assumed.

An incident state is terminated when threshold T_2 is no longer exceeded. The thresholds are calibrated from empirical data. The application of the California algorithm is quite straightforward, yet it is challenging to determine appropriate values of the algorithm's thresholds (T_1, T_2, and T_3) for each location.

EXAMPLE 9.5

Applying the California Algorithm for Incident Detection

Table 9.6 provides occupancy readings for two detection stations along a freeway equipped with a California-type AID algorithm. The algorithm is applied at regular time intervals of 30 s. Based on off-line calibration, three threshold values T_1, T_2, and T_3 were determined to be equal to 20, 0.25, and 0.50. Apply the California algorithm logic to determine the time step when an incident alarm would be triggered, and the time step when the incident state would be terminated.

Solution

For each time step, compute the values for the following three quantities:

Step 1: $(Occ_{up} - Occ_{down})$
Step 2: $(Occ_{up} - Occ_{down})/Occ_{up}$
Step 3: $(Occ_{up} - Occ_{down})/Occ_{down}$

The calculations are shown in Table 9.7 in columns 4 through 6.

TABLE 9.6
Detection Stations Readings

Time Step	*Occ_{up} (%)*	*Occ_{down} (%)*
1	60	10
2	62	15
3	59	17
4	65	14
5	67	22
6	64	19
7	59	22
8	48	27
9	37	29
10	32	29
11	30	28
12	32	31

TABLE 9.7
California Algorithm Calculations

Column [1]	*Column [2]*	*Column [3]*	*Column [4]*	*Column [5]*	*Column [6]*
Time Step	*Occ_{up} (%)*	*Occ_{down} (%)*			
1	60	10	50	0.83	5.00
2	62	15	47	0.76	3.13
3	59	17	42	0.71	2.47
4	65	14	51	0.78	3.64
5	67	22	45	0.67	2.05
6	64	19	45	0.70	2.37
7	59	22	37	0.63	1.68
8	48	27	21	0.44	0.78
9	37	29	8	0.22	0.28
10	32	29	3	0.09	0.10
11	30	28	2	0.07	0.07
12	32	31	1	0.03	0.03

Column [4] = $(Occ_{up} - Occ_{down})$ = Column [2] − Column [3]
Column [5] = $(Occ_{up} - Occ_{down})/Occ_{up}$ = (Column [2] − Column [3])/Column [2]
Column [6] = $(Occ_{up} - Occ_{down})/Occ_{down}$ = (Column [2] − Column [3])/Column [3]

The values in columns 4 through 6 are then compared to the three threshold values, T_1, T_2, and T_3, respectively, to determine whether the threshold values are exceeded. The results are shown in Table 9.8.

It can be seen that an alarm would be triggered after time step 2, since the algorithm needs two time steps where the thresholds are exceeded before an alarm is triggered. The incident state would then be terminated after time step 9.

TABLE 9.8
California Algorithm Results

Column [1] Time Step	*Column [2] Occ_{up} (%)*	*Column [3] Occ_{down} (%)*	*[4] > T_1*	*[5] > T_2*	*[6] > T_3*
1	60	10	YES	YES	YES
2	***62***	***15***	***YES***	***YES***	***YES***
3	59	17	YES	YES	YES
4	65	14	YES	YES	YES
5	67	22	YES	YES	YES
6	64	19	YES	YES	YES
7	59	22	YES	YES	YES
8	48	27	YES	YES	YES
9	***37***	***29***	***NO***	***NO***	***NO***
10	32	29	NO	NO	NO
11	30	28	NO	NO	NO
12	32	31	NO	NO	NO

Since the original California algorithm was first developed, refinements have been made on its performance. At least 10 new algorithms have been produced, of which algorithms 7 and 8 are the most successful. The TSC 7 algorithm represents an attempt to reduce the false alarm rate of the original algorithm. To do this, the algorithm requires that traffic discontinuities continue for a specified period of time before an incident is declared. The TSC 8 algorithm provides a repetitive test for the propagation of congestion effects upstream of the incident. It also categorizes traffic volumes into different states, which require that more parameters be calibrated. The TSC 8 algorithm can be regarded as the most complex algorithm to emerge from the modified California series, but also the best performer.

Catastrophe Theory Algorithms Catastrophe theory derives its name from sudden changes that take place in one variable that is being monitored, while other related variables under investigation show smooth and continuous changes. For incident detection, catastrophe theory algorithms monitor the three fundamental variables of traffic flow—namely, speed, flow, and occupancy (density). When the algorithm detects a drastic drop in speed, without an immediate corresponding change in occupancy and flow, this is an indication that an incident has probably occurred. This is because incidents typically cause a queue to form suddenly. The advantage of catastrophe theory algorithms as compared to the comparative-type is that catastrophe theory algorithms use multiple variables and compare these to previous trends of the data, whereas the comparative type typically uses a single variable and compares it to a preestablished threshold. By using more than one variable, catastrophe theory algorithms are better at distinguishing between nonrecurrent and recurrent congestion. The McMaster algorithm, developed at McMaster University in Canada, is a good example of an algorithm that is based on this idea.

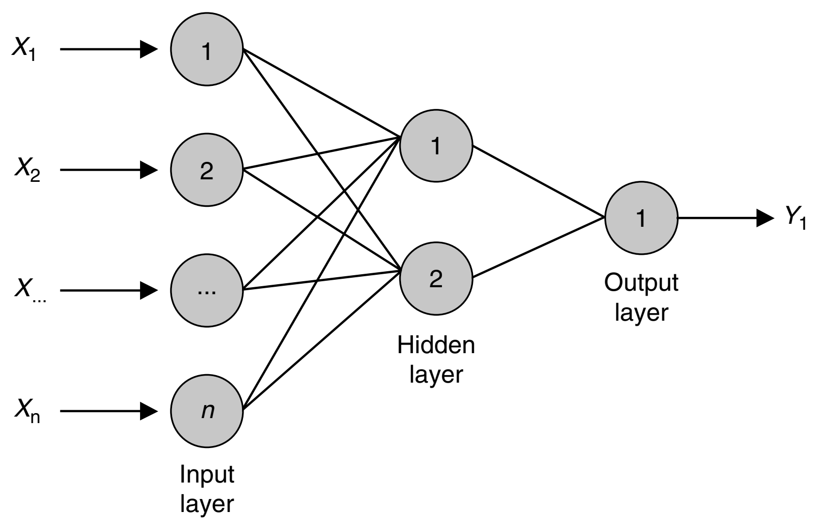
Statistical-Based Algorithms The idea behind these algorithms is the use of statistical and time series methods to forecast future traffic states or conditions. By comparing real-time observed traffic data with data forecasts, unexpected changes are then classified as incidents. An example of these algorithms is the auto-regressive integrated moving-average (ARIMA) time series algorithm. In this algorithm, ARIMA, a time series technique, is used to provide short-term forecasts of traffic occupancies based upon observed data from three previous time intervals. The algorithm also computes the 95% confidence intervals. If observations fall outside the 95% range as predicted by the model, an incident is assumed to have occurred.

Artificial Intelligence–Based Algorithms A number of the computational paradigms using artificial intelligence (AI) have been applied to problems in transportation engineering and planning. Automatic incident detection is one application. The incident detection problem is a good example of a group of problems called pattern recognition or classification problems. Several AI paradigms are available to solve classification problems, of which neural networks (NNs) are among the most effective.

NNs are biologically inspired systems consisting of a massively connected network of computational "neurons," organized in layers (Figure 9.8). By adjusting the weights of the network connections, connecting the neurons in the different layers of the network, NNs can be "trained" to approximate virtually any nonlinear function to a required degree of accuracy. NNs typically learn by providing the network with a set of input and output exemplars. A learning algorithm would then be used to adjust the weights of the network so that the network would give the desired output, in a type of learning commonly called supervised learning. Once trained, the NN can be used to predict the likely output for new cases.

Over the years, several NN types and architectures were developed. The type used for incident detection is the multilayer perceptron (MLP) neural network,

FIGURE 9.8
A multilayer perceptron neural network.



which is among the most widely used of NN architectures. As seen in Figure 9.8, MLPs typically consist of three layers: (1) the input layer; (2) the hidden layer(s); and (3) the output layer. The input layer takes data from loop detectors, the intermediate layer processes data, and the output layer gives an incident or an incident-free signal. Training is performed by presenting the network with a set of both incident and incident-free traffic scenarios. The training helps the network adjust its weights so as to be able to distinguish between traffic states that are incident-free and those that point to the occurrence of an incident.

Estimating the Benefits of Incident Management Systems

A major benefit of incident management systems is the reduction in the duration of an incident. The components of incident duration reduction are reductions in the time to detect and verify the occurrence of an incident; respond to the incident; and clear the incident. Incident management systems have been known to reduce the duration of an incident by up to 55%. The reduction in an incident duration resulting from the deployment of incident management systems can be used to estimate the likely benefits of their deployment, as illustrated by the following example.

EXAMPLE 9.6

Estimating the Benefits of Incident Management Systems Deployment

A six-lane freeway system (three lanes in each direction) carries an estimated 4200 veh/h during the peak hour in the peak direction. The capacity of the freeway is 2000 veh/h/lane. An incident occurs of 60 min duration that blocks 50% of the freeway capacity. Determine the time savings possible if an incident management system is implemented such that the duration is reduced to 30 minutes.

Solution

We first calculate the total vehicle delay for the case when the incident duration is one hour. To do this, the cumulative plot method is used as shown in Figure 9.9 (a similar problem was solved in Chapter 2, Example 2.4).

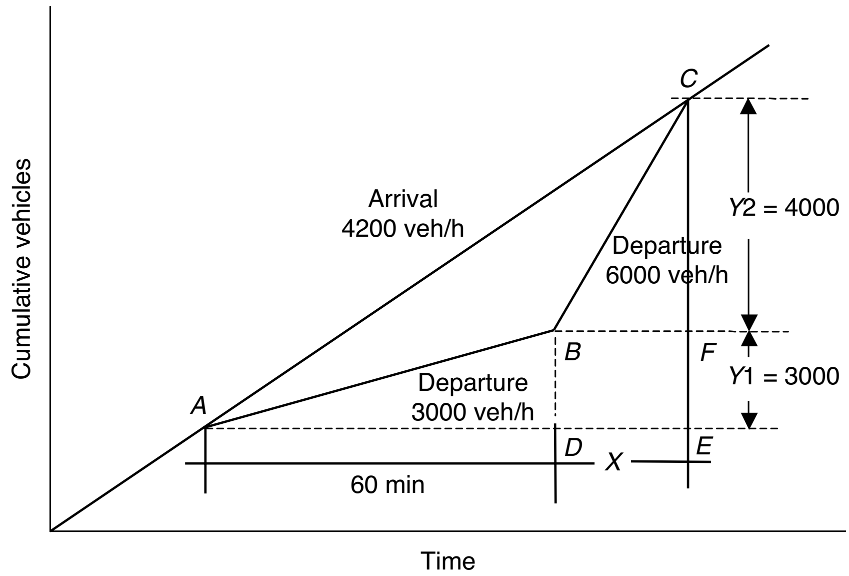
The arrival rate is 4200 veh/h. The departure rate is 3000 veh/h for a duration of 60 min and, when the incident clears, departure is 6000 veh/h. The total vehicle delay is computed as the area of the triangle between the arrival and departure curves.

(a) Solve for X, the time required for the queue to dissipate:

$$(3000)(1) + (6000)(X) = 4200(1 + X)$$

$$3000 + 6000X = 4200 + 4200X$$

FIGURE 9.9
Cumulative vehicle arrivals and departures.



$1800X = 1200$

$X = 0.667$ h

(b) Determine the cumulative number of vehicles, indicated by the vertical distances $Y1$ and $Y2$, as follows:

$Y1 = 3000 \times 1 = 3000$ vehicles

$Y2 = 6000 \times 0.6667 = 4000$ vehicles

(c) Determine the total delay in veh.hours due to congestion by calculating the area of the triangle, ABC, between the arrival and departure curves. The area of triangle ABC is determined by first calculating the area of triangle AEC and then subtracting the area of triangle ABD, the area of rectangle $BDEF$, and the area of triangle CBF, as follows:

$$(\tfrac{1}{2})(7000)(1.67) - (\tfrac{1}{2})(1)(3000) - (3000)(0.67) - (\tfrac{1}{2})(0.67)(4000)$$
$$= 1000 \text{ veh.hours}$$

Next, we consider the case when the incident duration is reduced to 30 min. Figure 9.10 develops the cumulative plot for this case. The total delay is calculated in a manner similar to that described previously, as follows:

(a) Solve for X, the time required for the queue to dissipate:

$(3000)(0.5) + (6000)(X) = (4200)(0.5 + X)$

$1500 + 6000X = 2100 + 4200X$

$1800X = 600$

$X = 0.333$ h

(b) Determine the cumulative vehicles $Y1$ and $Y2$:

$$Y1 = 3000 \times 0.5 = 1500 \text{ vehicles}$$

$$Y2 = 6000 \times 0.333 = 2000 \text{ vehicles}$$

(c) Determine the total delay in veh.hours due to congestion:

$$(\tfrac{1}{2})(3500)(0.833) - (\tfrac{1}{2})(0.5)(1500) + (1500)(0.33) + (\tfrac{1}{2})(0.33)(2000) = 250 \text{ veh.hours}$$

The deployment of the incident management system has reduced delay from 1000 to 250 veh.hours or a reduction of 75%.

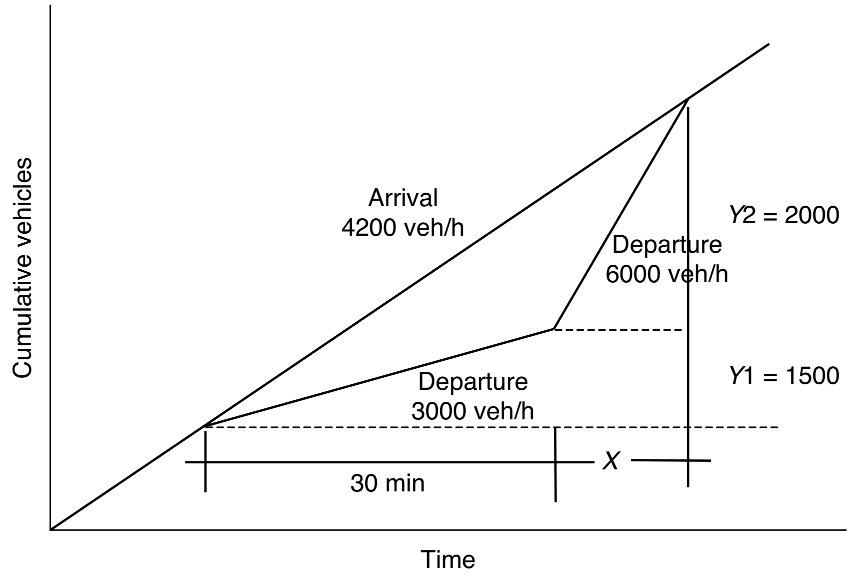


FIGURE 9.10 Cumulative vehicles arrivals and departures.

Ramp Metering

Ramp metering is the third function of a FIMS. It involves the regulation of vehicle entry to a freeway system by traffic signals at the entrance ramps. Ramp metering systems are intended to reduce recurrent congestion during peak hour periods as well as to improve safety when some geometric deficiencies exist.

Ramp metering is not a new strategy and existed in the early 1950s and 1960s. Ramp control systems operate in many areas, including Minneapolis/St. Paul, Minnesota, Seattle, Washington, and Austin, Texas. Most of these systems have achieved the goals of reducing delay and improving safety. The focus of this section is on various types of ramp control systems and their operational concepts. The main system components and technologies used by ramp control systems are described followed by examples of successful ramp metering projects.

Ramp Metering Control Philosophy

As was discussed in Chapter 4, as traffic flow (q) increases, traffic density (k) increases, reaching an optimum k_o at maximum capacity (q_{max}). At density levels greater than k_o, traffic flow conditions deteriorate and flow conditions change from stable to unstable. Ramp metering is designed to prevent unstable flow. Ramp metering controls the amount of traffic entering the freeway in an attempt to maintain densities at or below optimal (k_o) and to assure that traffic does not transition to an unstable or congested condition.

Benefits of Ramp Metering

Ramp metering is designed to achieve the following improvements in traffic operations:

Improved system operation: The main objective of ramp metering is to reduce congestion on a facility by controlling the number of entering vehicles. It is important, however, to make sure that congestion is not moved to surface streets. Vehicle queues on the ramps should not exceed the ramp length. Ramp metering systems can also minimize turbulence caused by merging at the junction of the ramp and the main line lanes, shortening platoons of entering vehicles such that they join the main line flow stream one or two vehicles at a time.

Improved safety: Many freeway accidents occur near on-ramps as the intensity of the merge maneuver increases and large platoons of vehicles arrive. By breaking up platoons of merging vehicles and smoothing the merge operation, ramp metering systems improve the safety of traffic operations. Also, by reducing stop-and-go conditions, ramp metering systems improve safety of traffic operations on a freeway facility.

Reduction in vehicle emissions and fuel consumption: There is a direct relationship between improved traffic operations and the reduction in harmful vehicle emissions and fuel consumption. Thus, ramp metering can improve air quality and energy consumption.

Promotion of demand management strategies: Ramp metering can be designed to encourage demand management and reduction strategies. For example, ramp metering systems can be designed to provide high-occupancy and transit vehicles with preferential treatment by adding a separate lane at the ramp entrance that allows these vehicles to bypass the ramp metering signal. Thus, ramp metering can contribute to strategies aimed at reducing single occupancy vehicles.

Classification of Ramp Metering Strategies

Ramp metering strategies can be classified as restrictive and nonrestrictive metering, and local versus systemwide metering.

Restrictive and nonrestrictive metering: Restrictive ramp metering sets the metering rate at a level lower than the nonmetered ramp volume. As a result, restrictive metering results in queue buildup on ramps and causes drivers to use alternative surface streets. Nonrestrictive metering sets the metering rate equal to or even greater than the average arriving volume. As a result, smaller queues form and diversion to surface streets is reduced. Nonrestrictive metering is often used for the purposes of improving the safety of operations at the vicinity of the ramp by breaking up ramp platoons. It also helps to delay the onset of congestion by smoothing the merge process.

Local versus systemwide ramp metering: Local ramp metering rates are determined based on traffic conditions in the vicinity of the ramp. Local metering is used when traffic congestion can be reduced by metering a single ramp or when several nonmetered ramps are near metered ramps. Systemwide ramp metering rates are implemented at more than one ramp along a freeway section in an integrated fashion and are more effective overall than local ramp metering.

Metering Rate Strategies

Ramp metering success depends on the metering rate selected that permits vehicles to enter the system. The metering rate for single lane ramps is between 240 and 900 veh/h. Ramp metering rates are either pretimed or traffic responsive. Pretimed strategies maintain a constant metering rate for a given time period regardless of the actual traffic volumes on the freeway. Traffic-responsive strategies vary the metering rates based on actual traffic volumes. Traffic-responsive strategies may be *local*, based upon the local traffic conditions detected at the vicinity of the ramp, or *systemwide*, where several ramps are controlled together as a part of an integrated system and metering rates are determined based upon traffic measurements over a large segment of the freeway. Types of ramp metering strategies are described as follows:

Pretimed Metering Pretimed metering rates are determined based on historical observations. Rates are specified for different time periods within a typical day. The metering rate selected depends on the objective to be achieved; that is, whether metering is designed to reduce congestion or to improve safety.

If the system is intended to relieve congestion, the rates are determined to ensure that main line traffic flow is less than the capacity. Thus, the metering rate will be a function of the upstream traffic flow, the ramp volume, and the downstream capacity. The metering should satisfy Equation 9.3 as depicted in Figure 9.11:

$$\text{Metering rate} + \text{upstream volume} \leq \text{downstream capacity} \tag{9.3}$$

Other factors to be considered in setting the metering rate are the availability of adequate storage on the ramp to accommodate queuing and adequate capacity throughout the corridor to accommodate vehicles that may be diverted.

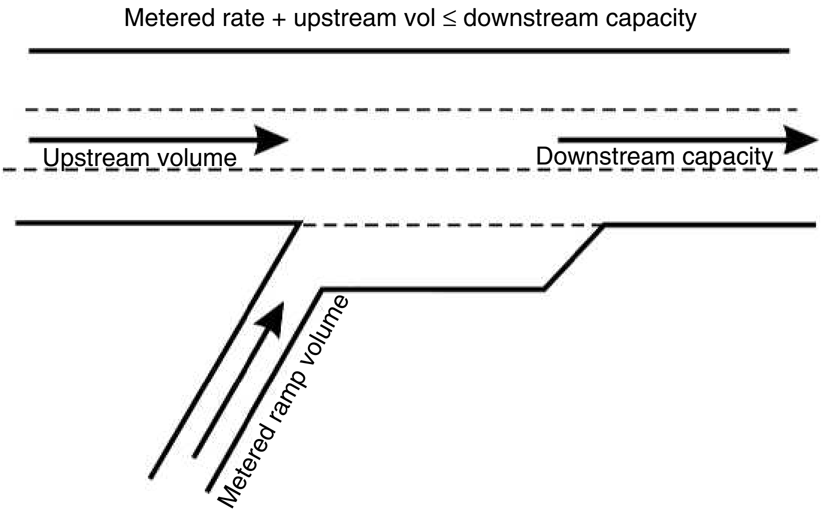


FIGURE 9.11
Pretimed ramp metering.

If the system is intended to improve safety, the metering rate is selected based on merging conditions at the end of the ramp. At ramps and junctions, rear-end and lane-change collisions can occur when vehicle platoons attempt to merge with main line traffic. Ramp metering can alleviate this situation by reducing the number of vehicles in a platoon. The metering rate depends on the geometrics of the ramp and the availability of acceptable gaps in the freeway traffic stream.

With pretimed ramp control, the ramp signal operates according to a predefined plan for the period under consideration. The timing of the red, yellow, and green intervals differs for single-entry, platoon metering, or two-abreast metering, as discussed next.

Single Entry Only one vehicle is allowed per green interval. The green interval (or the green-plus-yellow) is thus typically in the order of 1.5–2.0 s to ensure that only one vehicle enters per green interval. The red interval duration depends on the metering rate in effect.

EXAMPLE 9.7

Designing a Pretimed Single-Entry Ramp Meter

Design a pretimed single-entry ramp metering system on a four-lane freeway. The upstream traffic volume is equal to 3400 veh/h/direction, and the freeway capacity is 2000 veh/h/lane. Green interval is 2 s.

Solution

The downstream capacity in one direction is calculated as (2)(2000) = 4000 veh/h.

The ramp metering rate can be calculated using Equation 9.3:

Metering rate + 3400 = 4000

Metering rate = 4000 − 3400 = 600 veh/h

Since the green interval is 2 sec, the red interval is (cycle length) − (2.0):

Cycle length = 3600/600 = 6 s

Thus, the red interval is (6.0 − 2.0) = 4.0 s/cycle, and the ramp meter signal cycle is green for 2 s and red for 4 s.

PLATOON METERING For metering rates greater than 900 veh/hr, platoon metering is used, where two or more vehicles per cycle enter the freeway. The minimum length of the green interval must be sufficient to allow the platoon to pass.

EXAMPLE 9.8

Designing a Pretimed Platoon Metering System

Design a signal plan for a ramp metering system, given the following information:

Upstream volume = 4800 veh/h

Number of lanes/direction = 3 lanes

Capacity = 2000 veh/h/lane

Solution

Calculate the metering rate using Equation 9.3 using a downstream capacity of $3 \times 2000 = 6000$ veh/h.

Metering rate = 6000 − 4800 = 1200 veh/h

Since the metering rate is greater than 900 veh/h, platoon metering is required. The metering rate is 1200/60 = 20 veh/min.

If two vehicles enter on the green signal, 10 cycles/min are required (i.e., (2)(10) = 20). The cycle length is 60/10 = 6 s and the green interval is 4 s for 2 s per vehicle. The red interval is 6 – 4 = 2 s.

TWO-ABREAST METERING Two vehicles are released side by side (on a two-lane ramp). Vehicles are released alternately, and the green interval is set to allow the release of one veh/cycle. With two-abreast metering, as many as 1700 veh/h can be accommodated.

Local Traffic Responsive Metering Traffic-responsive metering rates are not prefixed. Rather, they are determined in real time, based on traffic measurements. Local traffic responsive metering rates are selected based on real-time measurements of traffic conditions in the vicinity of the ramp. Traffic-responsive ramp metering systems use traffic flow models that include the variables of flow rate (q), speed (u), and density (k); see Figure 4.2 of Chapter 4. The basic strategy of traffic-responsive metering is to

- Obtain real-time measurements of current traffic flow parameters.
- Determine the current state of traffic flow from traffic flow models.
- Determine the maximum ramp metering rate that would ensure that the flow is kept within the uncongested portion of the fundamental traffic flow diagram (see Figure 9.12).

Ramp metering strategies differ from one another based upon which traffic flow parameters they use in order to determine the appropriate ramp metering rate. Two of the most widely used traffic-responsive ramp metering strategies are demand-capacity control and occupancy control.

DEMAND-CAPACITY CONTROL Metering rates are based on real-time comparisons of upstream traffic *volumes* against downstream capacity. The upstream volume is measured in real time; the downstream capacity is determined either on the basis of historical data or computed in real time based upon downstream volume measurements. The ramp metering rate for the next control period (typically 1 min) is computed as the difference between the downstream capacity and the upstream volume to ensure that downstream capacity is not exceeded. For example, if at a certain control interval the upstream volume equals 3000 veh/h (i.e., 50 veh/min) and the downstream capacity equals 3600 veh/h (i.e., 60 veh/min), a metering rate up to $(60 - 50) = 10$ veh/min could be accommodated.

A problem with using *volume* alone as the traffic flow performance measure is that low volume values may be associated with free-flow conditions as well as with congested conditions, depending upon whether traffic density is less or greater than the density at capacity. As can be seen from Figure 9.12, corresponding to a volume value, $V1$, there are two possible density values, one corresponding to uncongested conditions and the other to congested conditions. To overcome this problem and be able to distinguish between congested and uncongested conditions, occupancy measurements are required.

OCCUPANCY CONTROL Metering rates are selected based upon *occupancy* measurements upstream. There are two types of occupancy control: open-loop occupancy control and closed-loop occupancy control.

Open-loop occupancy control provides a schedule of metering rates. Based upon occupancy measurements upstream the metered ramp, one out of several predefined metering rates is selected for the next control period. The predefined

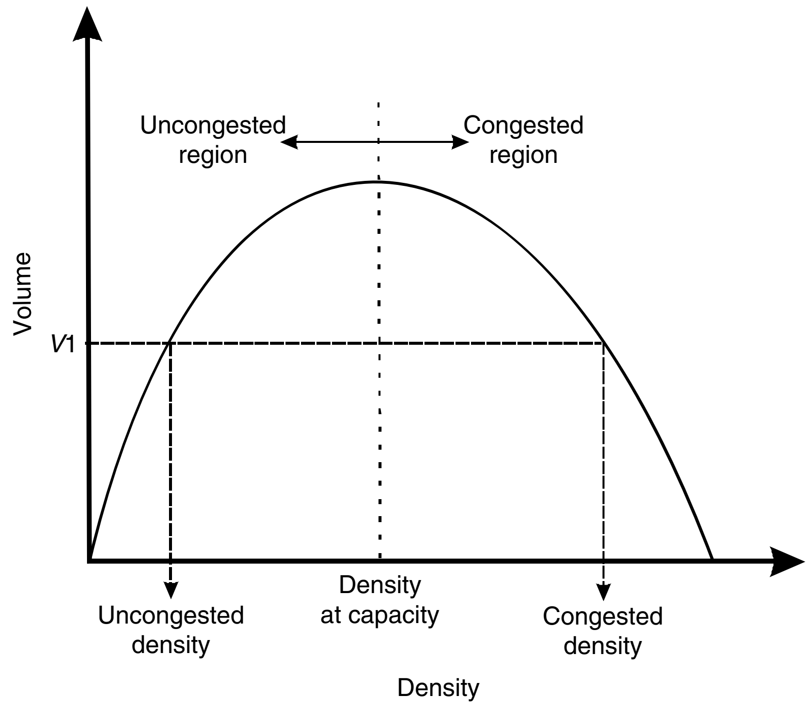


FIGURE 9.12
A typical volume–density plot.

metering rates are determined from a study of a plot of the relationship between volumes and occupancy for the facility of interest. Using this plot, for each level of occupancy, a metering rate can be established that corresponds to the difference between the predetermined estimate of capacity and the real-time estimate of the volume that corresponds to the occupancy measured. The volume can be estimated using a volume–occupancy plot as shown in Figure 9.13, which determines an approximate relationship between occupancy and volume.

Table 9.9 can be used to determine appropriate local traffic-responsive metering rates as a function of the measured mainline upstream capacity. As can

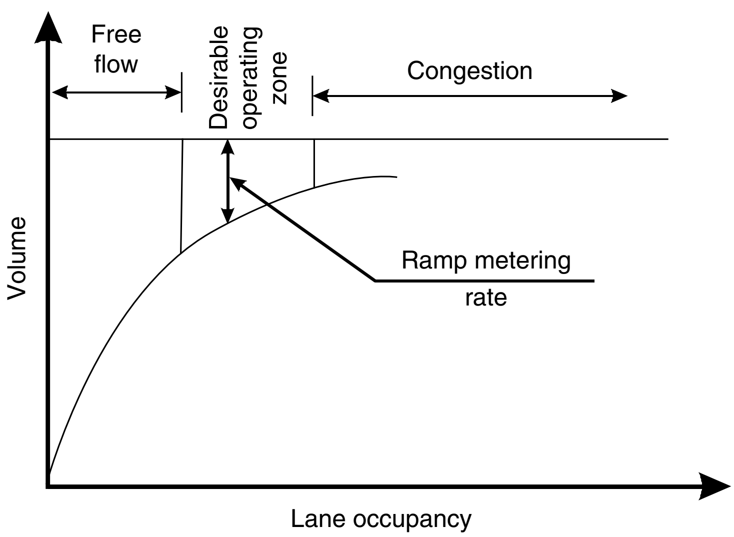


FIGURE 9.13
Calculating ramp metering rates based upon volume–occupancy plots.

TABLE 9.9
Metering Rates as a Function of Upstream Occupancy

Occupancy (%)	*Metering Rate (Veh/Min)*
≤10	12
11–16	10
17–22	8
23–28	6
29–34	4
>34	3

be seen in Table 9.9, if the measured occupancy exceeds the preset capacity occupancy (i.e., 34% in this case), a minimum metering rate of 3 veh/min is selected.

The type of control described is termed *open loop* because it controls the flow rate based upon preset values and does not check the impact of the control action on the controlled environment. It does not control the flow rate to explicitly achieve a parameter value sensed by detectors, such as the downstream occupancy, as does closed-loop control.

EXAMPLE 9.9

Determining Metering Rates for an Open-Loop Ramp Meter

Given the occupancy measurements at a local traffic-responsive ramp metering site shown in Table 9.10, determine the metering rates for the different control periods.

TABLE 9.10
Data for Example 9.9

Control Period	*Measured Occupancy (%)*
1	23%
2	25%
3	29%
4	21%
5	18%

Solution

This problem can be solved using Table 9.9 to determine the appropriate metering rates for each occupancy level. The solution is given in Table 9.11.

TABLE 9.11
Metering Rates for Example 9.9

Control Period	*Measured Occupancy (%)*	*Metering Rate (Veh/min)*
1	23	6
2	25	6
3	29	4
4	21	8
5	18	8

Closed-loop occupancy control explicitly controls the downstream occupancy to conform with the desired occupancy value. Occupancy values measured downstream of the ramp are fed back to the controller in order to determine the metering rate that would bring the downstream capacity to its desired value. One of the more well-known closed-loop ramp metering control algorithms is called ALINEA. The algorithm is designed to operate with a main line detector station that measures occupancy values downstream of the ramp. The metering rate for a given control period, i, is then calculated using the following equation:

$$r(i) = r(i-1) + K_R(o_s - o_{\text{out}}(i)) \tag{9.4}$$

where

$r(i)$ = metering rate for the control interval i
$r(i-1)$ = metering rate during the previous control interval $(i-1)$
o_s = preset or desired value for the downstream occupancy
$o_{\text{out}}(i)$ = measured downstream occupancy for the control interval i

K_R is a coefficient that is commonly referred to as the gain coefficient. The value of K_R affects the sensitivity of the controller and how quickly it reacts to changes in the controller input. The higher the value of K_R, the faster the controller reacts to changes. At the same time, however, high values of K_R tend to make the control more oscillatory and more sensitive to errors in the measured occupancy.

For inductive loop detectors, the occupancy set point (o_s) is typically established in order to ensure that the level of service (LOS) on the freeway does not fall below a specified LOS (e.g., LOS D or E). The calculation proceeds by first looking up the upper value density for the specified LOS from the *Highway Capacity Manual* (HCM) curves or tables. With this determined, Equation 9.1, which relates the density and occupancy values, can be used to calculate the corresponding occupancy set point. The following example illustrates the procedure.

EXAMPLE 9.10

Determining the Occupancy Set Point for a Traffic-Responsive Ramp Meter

A closed-loop, occupancy control ramp meter works by measuring the downstream occupancy using an inductive loop detector and then determining the metering rate using the ALINEA algorithm. It is desired to establish the occupancy set point for the ALINEA algorithm so that the LOS on the freeway will be LOS E. Determine this set point, given the following information:

Average passenger car length = 5.4 m

Average commercial vehicle length = 8.1 m

Percentage of commercial vehicles in traffic stream = 4%

Effective length of the detector = 2.4 m

Upper density level corresponding to LOS E = 28 passenger cars/km/lane

Solution

Calculate the average vehicle length with 4% commercial vehicles (i.e., 96% passenger cars). The average vehicle length, L_v, is

$$L_v = 5.4 \times (1 - 0.04) + 8.1 \times (0.04) = 5.51 \text{ m}$$

The occupancy set point is calculated using Equation 9.1:

$$D = \frac{10 \times Occ}{L_v + L_{\text{eff}}}$$

$$28 = \frac{10 \times Occ}{5.51 + 2.4}$$

$$\text{Occupancy} = \frac{28 \times 7.91}{10} = 22.15\%$$

Systemwide Traffic-Responsive Metering This is the application of metering strategies to a series of ramps. For each control interval, real-time traffic measurements are made of parameters such as volume and/or occupancy, which define demand capacity conditions at each ramp. Ramp metering rates are determined for the entire system as well as for individual ramp meters. Appropriate algorithms will include stored pretimed metering rates. The system will typically use the more restrictive of the traffic responsive and pretimed rates.

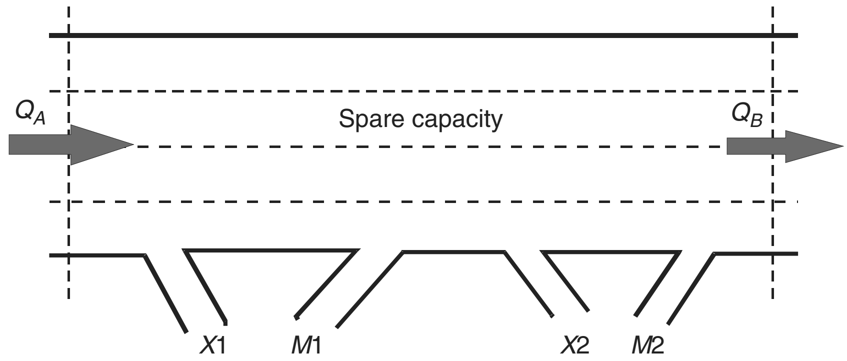


FIGURE 9.14
Volumes entering and leaving a freeway zone.

Most systemwide traffic-responsive metering algorithms start by dividing the freeway into a number of zones. For each zone, the algorithm calculates the number of excess vehicles based on direct mainline measurements. The metering rates of the ramps within that zone are then established based upon the number of excess vehicles.

The Minnesota algorithm serves to illustrate the process. The algorithm regulates traffic within freeway zones by ensuring that the total number of vehicles leaving each zone is greater than the number entering. As shown in Figure 9.14, each freeway zone has three input variables (representing vehicles entering the zone) and three output variables (representing vehicles leaving the zone).
The input variables are

Q_A = main line upstream volume entering the zone, determined from an upstream detection station

M = total ramp volume entering the zone through the metered on-ramps. In Figure 9.14, $M = M1 + M2$

U = total unmetered ramp volume entering the zone

The output variables are

Q_B = main line downstream volume leaving the zone

X = total volume exiting through the zone's off-ramps. In Figure 9.14, $X = X1 + X2$;

S = spare capacity or the additional volume that can enter the zone without causing congestion; calculated based on measured main line speed and volume data

The Minnesota algorithm can be expressed as follows:

$$Q_B + X + S \geq Q_A + M + U \tag{9.5}$$

Therefore

$$M \leq Q_B + X + S - Q_A - U \tag{9.6}$$

Equation 9.6 is the maximum number of vehicles that can pass through all the ramp meters in a given freeway zone. The volume M is then dispersed throughout the zone in proportion to the demand (D) on the metered entrance ramps, using Equation 9.7:

$$R_n = M \times (D_n/D) \tag{9.7}$$

where

R_n = metering rate for on-ramp, n
D_n = demand at ramp, n
D = total demand on all metered ramps within the zone

EXAMPLE 9.11

Determining Metering Rates for a Systemwide Ramp Metering System

Determine appropriate metering rates for on-ramps A and B of the freeway zone shown in Figure 9.15. The projected demand rates for ramps A and B are D_A = 550 and D_B = 700 veh/h. Traffic is flowing smoothly within the zone, and spare capacity is 1000 veh/h.

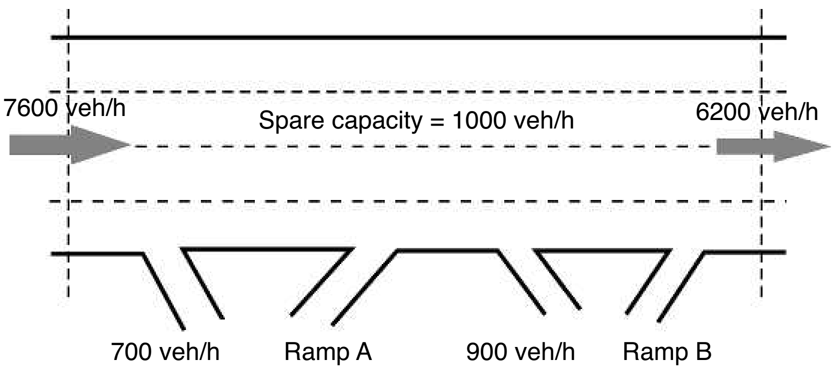


FIGURE 9.15
Traffic volumes for Example 9.11.

Solution

Calculate the total number of vehicles that can pass through the metered ramps A and B. Use Equation 9.6:

$$M = Q_B + X + S - Q_A - U$$
$$= 6200 + (700 + 900) + 1000 - 7600 - 0 = 1200 \text{ veh/h}$$

(Note that U is equal to 0, since all on-ramps within the zone are metered.)

The metering rate for ramps A and B can then be determined using Equation 9.7 as follows:

$$R_1 = 1200 \times \frac{550}{(550 + 700)} = 528 \text{ veh/h (answer)}$$

$$R_2 = 1200 \times \frac{700}{(550 + 700)} = 672 \text{ veh/h (answer)}$$

Ramp Metering System Layout

The typical components of a metering system are shown in Figure 9.16. The system consists of the following elements:

- Ramp metering signal, either a red-yellow-green signal or just a red-green signal;
- Local controller, similar to those used at signalized intersections;
- Advance ramp control sign to inform drivers that the ramp is being metered;
- Vehicle detectors, devices that establish conditions within the ramp area. There are five types of detectors at ramp metering systems, as described next.

  Check-in detectors: Ramp signal remains red until a vehicle is detected. A minimum metering rate, however, is used to avoid problems caused by possible detector failure or a vehicle not stopping near enough to the line to actuate the detector.

  Check-out detectors: These ensure single-vehicle entry. When a vehicle is allowed to pass the ramp, it is detected by the check-out detector and the green is terminated. This ensures that the green interval is sufficient for the passage of one vehicle only.

  Queue detectors: Queue detectors sense the backing or spillback of ramp traffic onto the surface roads. When a queue is sensed, the ramp metering rate may be increased to allow the queue to shorten.

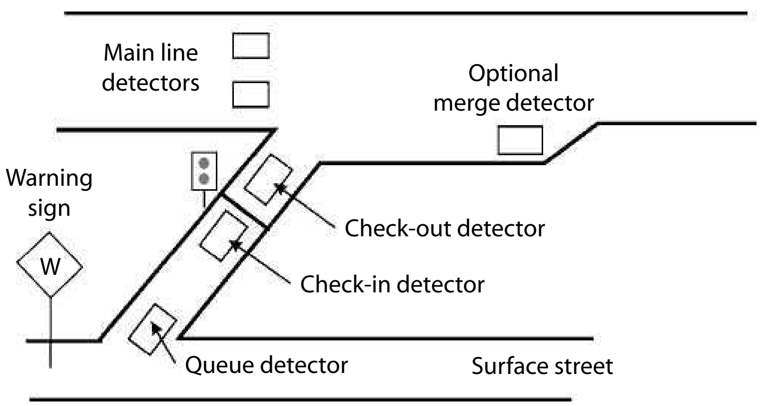


FIGURE 9.16

Ramp metering system layout.

Merge detectors: Merge detectors may be used to detect the presence of vehicles in the merge area. When a vehicle is blocking the merge area, the ramp signal remains red until the detected vehicle merges with freeway traffic.

Main line detectors: Main line detectors sense traffic volumes upstream of the merge area and may be either single or multilane. These detectors provide the input information for the ramp meter control algorithm.

Ramp Storage Requirements

Adequate storage space is required at ramps in order to avoid queues from the ramp backing onto the street network. Storage requirements for ramps can be calculated using the principles of queuing theory as described in Chapter 2.

As discussed in Chapter 2, queuing systems are classified based on the way in which customers arrive and depart. For ramp meters, both the interarrival as well as the service times are best described by the negative exponential distribution. Thus, the $M/M/1$ queuing model, discussed in Chapter 2, can be used to solve ramp metering queuing problems. The following example illustrates the procedure.

EXAMPLE 9.12

Ramp Meters Queue Analysis

Traffic is to be regulated on an on-ramp leading to a freeway using a traffic-responsive ramp meter. The ramp has adequate storage space for eight vehicles. During the peak hour, it is estimated that the metering rate will not exceed 600 veh/h. The average volume on the ramp during a typical peak hour is 480 veh/h. Using queuing theory, determine (1) the average queue length on the ramp; (2) the average delay for vehicles at the ramp meter; and (3) the probability that the ramp will be full.

Solution

(1) We first calculate the ratio of the arrival to the service flow rate, ρ, for the ramp meter described in the problem as follows:

$$\rho = \lambda/\mu = 480/600 = 0.80$$

As discussed in Chapter 2, the average queue length, $\overline{Q}$, for an $M/M/1$ queue is given by Equation 2.28 as

$$\overline{Q} = \frac{\rho^2}{(1 - \rho)}$$

Therefore,

$$\overline{Q} = \frac{0.80^2}{(1 - 0.80)} = 3.2 \text{ vehicles}$$

(2) Also from Chapter 2, the average delay, $\overline{W}$, for an $M/M/1$ queue is given by Equation 2.29 as follows:

$$\overline{W} = \frac{\lambda}{\mu(\mu - \lambda)}$$

where

λ = the arrival rate (customers/time)
μ = the service rate (customers/time)

In this example,

$\lambda = 480$ veh/h $= 480/60 = 8$ veh/min

$\mu = 600$ veh/h $= 600/60 = 10$ veh/min

Therefore

$$\overline{W} = \frac{\lambda}{\mu(\mu - \lambda)} = \frac{8}{10(10 - 8)} = 0.40 \text{ min/veh or } 24 \text{ s/veh}$$

(3) The ramp will be full when we have more than eight vehicles in the queue. For $M/M/1$ queues, the probability of having exactly n customers, p_n, in the queue is given by Equation 2.31 as follows:

$$P_n = (1 - \rho)\rho^n$$

The probability of $n > 8$ can be expressed as follows:

$$p(n > 8) = 1.0 - p(n \le 8)$$

That is,

$$p(n > 8) = 1 - p(0) - p(1) - p(2) - p(3) - p(4) - p(5) - p(6) - p(7) - p(8)$$

The calculations can be easily performed using Excel, as shown in Figure 9.17. The probability that the ramp will be full is equal to 0.1342.

FIGURE 9.17
Probability calculations for Example 9.12.

	A	B	C	D	E	F	G	H
1	ρ =	0.8						
2								
3	n	p(n)						
4	0	0.2000		=(1-B1)*B1^A4				
5	1	0.1600						
6	2	0.1280						
7	3	0.1024						
8	4	0.0819						
9	5	0.0655						
10	6	0.0524						
11	7	0.0419						
12	8	0.0336						
13				=SUM(B4:B12)				
14	p (n <= 8)	0.8658						
15	p (n > 8)	0.1342						
16				=1-B14				
17								

Information Dissemination

Information dissemination is another function that FIMS are designed to provide. Effective communication with drivers is an essential component of the freeway management process. Freeway management systems utilize several travel information dissemination devices to keep drivers informed about current as well as expected travel conditions on the freeway. Travel information dissemination is divided into pretrip information dissemination and en route information dissemination.

Pretrip travel information concerns providing travelers with information before they start their trips. Examples of pretrip travel information includes current or expected traffic conditions, current and expected weather conditions, and bus schedules and fares. The information is typically provided via devices such as cable TV and the Internet and is intended to enable travelers to make informed route/mode/time of departure decisions. Such informed decisions are likely to improve the overall level of service of the transportation network.

En route travel information dissemination involves providing travelers with information while they are en route. Dissemination devices in this case include dynamic message signs (DMS) (Figure 9.18), Highway Advisory Radio (HAR), low-power FM radio, cell phones, and in-vehicle display devices. En route travel information systems provide information about current and expected traffic and weather conditions, incidents, and alternative routes. Dynamic route guidance (DRG) uses real-time information about traffic flow conditions to reroute drivers around congested areas or incident locations.

FIGURE 9.18
Dynamic message signs (DMS).

Dynamic Route Guidance

Closely associated with the information dissemination function of a FIMS is the notion of dynamic route guidance (DRG). Travelers typically select the shortest route to their destination while taking account of congestion, if possible. For individual travelers, it is difficult to know in advance the congestion level on the route they plan to use. This is especially true in cases where unexpected incidents and accidents occur on the transportation network. The idea behind DRG is to take advantage of the information provided by the advanced monitoring and surveillance equipment of an intelligent transportation infrastructure and use this information to develop an optimal way to assign or distribute traffic on the network in real time. Routing recommendations are then communicated to drivers via DMS (Figure 9.18) or in-vehicle display devices.

While developing optimal routes, DRG algorithms consider real-time traffic and congestion levels. These algorithms are therefore called dynamic traffic assignment (DTA) algorithms as opposed to the static assignment techniques previously discussed in relation to transportation planning, which focus on average, steady-state conditions. The following section describes the difference between the dynamic and static assignment problems in some detail.

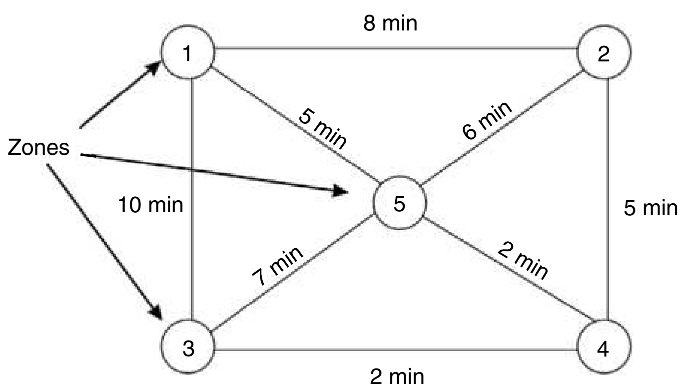
Dynamic versus Static Traffic Assignment

The general traffic assignment problem includes a network and a set of ordered pairs of points on the network where trips originate and end. For each origin–destination pair, a function $R(t)$, $0 \leq t \leq T$, where T is the planning horizon, is given that defines the rate at which vehicles leave the origin at time t, destined to a particular destination. This function yields what is called the origin–destination (O–D) matrix, as shown in Figure 9.19. In addition, the capacity of each link

FIGURE 9.19
The traffic assignment problem.

Origin–Destination Matrix

Zone	**1**	**2**	**3**	**4**	**5**
1	0	1000	2000	900	0
2	500	0	1200	1700	700
3	1200	900	0	1100	1500
4	800	700	1500	0	2000
5	1100	750	1150	1500	0



(roadway segment) in the network is provided, *Cap*(t). The assignment problem is to determine the traffic pattern or flows on the links of the network satisfying certain optimality or equilibrium conditions.

When both *R*(t) and *Cap*(t) are constant over time, the problem reduces to the static traffic assignment problem. While this assumption may be reasonable for transportation planning applications, it is not very realistic for real-time modeling and control of transportation networks. The assumption of constant demand and supply does not hold for many realistic traffic situations. Peak-hour conditions, for example, are typically characterized by variations in traffic demand. The occurrence of incidents affects the capacity (i.e., the supply side) of the network. For such conditions of variable demand and/or supply, a dynamic traffic assignment (DTA) problem formulation is needed. This is the formulation required for optimally routing drivers in real-time in the DRG problem.

Mathematical Formulation of the Dynamic Route Guidance or Traffic Assignment Problem

The DRG or DTA problem can be formulated as a mathematical program. For this problem, the decision variables are the time-varying traffic splits at each diversion point that optimize network performance (e.g., minimize total travel time). This defines how traffic should be distributed over the network. The objective function expresses the measure of the highway network's performance to be optimized (such as the total travel time for all vehicles), and the set of constraints attempt to model traffic flow in the region and ensure flow conservation at the nodes and along the links of the network. The formulated model is solved to determine the routing strategy that will optimize the objective function.

Challenges of the DRG The DRG problem is a challenging problem. For realistic transportation networks, with hundreds and even thousands of nodes, links, and alternative routes, the computation effort required to solve the problem is intensive. This is especially true given the fact that the recommended routing strategies need to be developed in real time. As soon as traffic conditions change, such as when an incident occurs, routing strategies must be revised to address the new situation. Second, the problem formulation discussed previously assumes that travel demand and travelers' origins and destinations are known. In practice, predicting travelers' origins and destinations is far from being a straightforward problem. One also needs to be able to predict how drivers will respond to the routing recommendations generated. Finally, there is the problem of missing or incomplete information, since the surveillance system will cover only a subset of the network. Moreover, sensor malfunctions are a common occurrence in the harsh freeway environment.

Real-Time Execution Requirement of the DRG Problem

Real-time execution, in the context of incident traffic flow management, refers to the immediate, online response to an incident, through the implementation of a

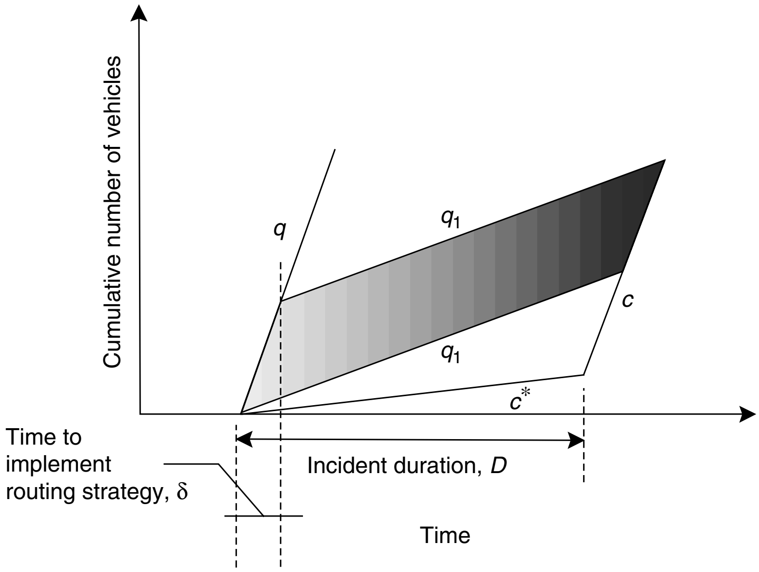


FIGURE 9.20
Incident queuing diagram.

routing strategy, so as to minimize resulting delays. Delay in the implementation of the routing strategy results in incurring additional delays. To illustrate this, we will look at an example next that uses cumulative plots. Consider Figure 9.20, which depicts a queuing diagram for cumulative vehicle arrivals and vehicle departures during a particular incident scenario of duration, D, minutes, similar to the cumulative plots we developed in Chapter 2.

The traffic arrival rate before a routing strategy is implemented is denoted by q (veh/h) and is represented by the slope of the cumulative arrival function. Similarly, q_1 (veh/h) denotes the reduced traffic arrival rate after a routing strategy is implemented. The reduced capacity of the segment caused by the occurrence of the incident is denoted by c^* (veh/h), whereas the normal capacity on the absence of incidents is denoted by c (veh/h). The capacities, c^* and c, are represented by the slope of the cumulative departure curves.

As Figure 9.20 illustrates, waiting for a period, say δ minutes, to implement the routing strategy results in incurring additional delay costs, as indicated by the shaded region of Figure 9.20. Geometrically, the area of the shaded region can be shown to be equal to

$$(q - q_1) \times \delta \times \left[\frac{\delta(q - c) + 2D(c - c^*)}{120(c - q_1)}\right] \text{veh.min} \tag{9.8}$$

where

q = traffic arrival rate before routing (veh/h)
q_1 = reduced traffic arrival rate after routing (veh/h)
δ = waiting time before implementing a routing strategy in minutes
c = normal capacity of segment without any incidents (veh/h)
c^* = reduced capacity of segment as a result of an incident (veh/h)
D = incident duration in minutes

The units for the resulting extra delay incurred will be veh.min.

EXAMPLE 9.13

Extra Delay Resulting from Waiting to Implement Routing Strategies

A six-lane freeway segment, whose unrestricted capacity is 2200 veh/h/lane, carries an average volume of 6000 veh/h. An incident occurs that results in a 60% reduction in the capacity of the segment. The incident lasts for 45 min. To relieve congestion during the incident, traffic routing is implemented which reduces the traffic volume on the segment down to 3600 veh/h. What would be the extra delay incurred if it would take 5 min to implement the routing strategy versus only 30 s?

Solution

To calculate the extra delay incurred as a result of waiting for a period of δ minutes to implement routing strategies, we use Equation 9.8.

For the 30-s case:

$q = 6000$ veh/h

$q_1 = 3600$ veh/h

$\delta = 0.5$ min

$c = 6600$ veh/h

$c^* = 0.4 \times 6600 = 2640$ veh/h

$D = 45$ min

Substituting in Equation 9.8 gives

$$\text{Extra delay} = (q - q_1) \times \delta \times \left[\frac{\delta(q - c) + 2D(c - c^*)}{120(c - q_1)}\right]$$

$$= (6000 - 3600) \times 0.5 \times \left[\frac{0.5(6000 - 6600) + 2 \times 45 \times (6600 - 2640)}{120(6600 - 3600)}\right]$$

$= 1187$ veh.min

For the 5-min case, substituting in Equation 9.8 gives

$$\text{Extra delay} = (q - q_1) \times \delta \times \left[\frac{\delta(q - c) + 2D(c - c^*)}{120(c - q_1)}\right]$$

$$= (6000 - 3600) \times 5 \times \left[\frac{5(6000 - 6600) + 2 \times 45 \times (6600 - 2640)}{120(6600 - 3600)} \right]$$

$$= 11780 \text{ veh.min}$$

Therefore, the extra delay incurred as a result of taking 5 min to implement the routing strategy versus only 30 s is equal to 11,780 − 1187 = 10,593 veh.min.

Lane Management

The lane management function of a FIMS attempts to maximize the utilization of the available lane capacity of the freeway. One major application of the lane management function involves the use of reversible-lane flows, which change the directional capacity of a freeway to accommodate peak directional traffic demands. The use of reversible lanes is warranted when traffic flow exhibits significant directional imbalance (e.g., when there is more than 70% of the two-directional traffic volume in the peak direction). In such cases, the use of reversible lanes allows for using the existing capacity in a more efficient way. Reversible lanes or contraflow lanes are also very useful during some incident management scenarios and for emergency evacuation.

The use of reversible lanes, however, raises some safety concerns, and appropriate measures would need to be implemented to ensure safe operations. These include the use of barrier gates to prevent vehicles from entering in the wrong direction; pop-up lane delimiters; video cameras for vehicle detection; and dynamic message signs (DMS) to inform motorists of the current operating direction.

Real-World Examples of Freeway and Incident Management Systems and Their Benefits

Real-world freeway and incident management systems can now be found throughout the United States, as well as around the world. Examples in the United States include systems in Atlanta, Houston, Seattle, Minneapolis–St. Paul, New York, Chicago, Milwaukee, Los Angeles, San Diego, and Northern Virginia, among others. Freeway and incident management has been proven to be quite effective in alleviating recurrent and nonrecurrent congestion. The San Antonio's TransGuide freeway management system in Texas, for example, helped reduce accidents by 15% and emergency response time by 20%. Ramp metering was shown to help increase throughput by 30% in the Minneapolis–St. Paul metro area, with peak hour speeds increasing by 60%. Ramp meters in Seattle, Washington, are credited with a 52% decrease in travel time and a 39% decrease in accidents. The evaluation of the initial operation of the Maryland CHART program showed a benefit/cost ratio of 5.6:1, with most of the benefits resulting from a 5% decrease (which amounted to around 2 million veh/h/yr) in delays from nonrecurrent congestion.

Advanced Arterial Traffic Control (AATC) Systems

Signalized intersections play a major role in determining the overall performance of arterial networks and many other types of transportation facilities. They are the points where conflicting traffic streams meet and compete for the same physical space, creating many potential conflicts. For a long time, transportation practitioners have been thinking of ways to make signalized intersections more efficient, and a key tool that they have been trying to take advantage of in this regard is information technology (IT). To a large extent, improving the performance of signalized intersections through the use of IT has centered on two simple ideas.

The first idea attempts to make the traffic signal more intelligent and more responsive to actual traffic demands. The concept is to use traffic sensors or loop detectors, similar to those described in relation to FIMS, at the approaches to the intersection. These sensors would detect the presence or passage of vehicles and would communicate this information to the signal controller. Based on this information, the signal controller would attempt to optimize the signal plan so as to minimize the vehicle's delay at the intersection. These signals are commonly referred to as actuated traffic signals.

The second idea involves controlling a group of signals that lie along a major corridor in an integrated or, to use signal control terminology, in a coordinated way. This means that the signal plans of the individual intersections would be coordinated in such a fashion that a platoon of vehicles discharged from one intersection will not be stopped right away at a downstream intersection, but instead would proceed through a sequence of coordinated intersections without stopping. In addition to actuated and coordinated signals, AATC applications include adaptive traffic control and signal preemption to allow emergency vehicles to safely reach their destination as soon as possible. The following sections will describe these applications in more detail.

Actuated Traffic Signals

Actuated signal control could be regarded as one of the very first applications of IT to transportation problems, which predates the term *ITS* by several years. As opposed to pretimed signals, actuated signals have the capability to revise their timing plans based upon actual traffic demands obtained from traffic detectors. The idea behind the use of actuated controllers is to have an adaptive type of control that is responsive to the continuously changing traffic conditions. For pretimed controllers, the implemented signal plan is only optimal for the volumes assumed in developing the "offline" plan. These volumes could be very different from the actual volumes, particularly if signal plans are not updated regularly, which is often the case. Actuated controllers are capable of optimizing the allocation of time based upon real traffic volumes.

To understand the basic concept of operations for actuated controllers, we first need to define the following three parameters:

Minimum green. Each signal phase of an actuated controller is assigned a minimum green time. This time is typically taken to be equal to the time it takes a queue of vehicles potentially stored between the stop line and the approach detector location to enter the intersection.

Passage time interval. The passage time interval is the time it takes a vehicle to travel from the detector location to the stop line. The passage time also defines the maximum gap, which is the maximum time period allowed between vehicles' arrivals at the detector for the approach to retain the green. If a time period equal to the passage time interval elapses without vehicle actuations at the detector, the green for that approach is terminated, and another approach, with a call for service waiting, gets the green. In such a case, the terminating phase is said to have "gapped out."

Maximum green time. In addition to assigning each phase a minimum green, each phase is assigned a maximum green. If the demand on one approach is sufficient to retain the green until this limit (i.e., vehicles keep arriving before the maximum gap expires), the phase is terminated after the maximum green time is exceeded. In this case, the terminating phase is said to have "maxed out."

Figure 9.21 shows the operational concept of an actuated controller. When a certain phase becomes active, the minimum green is displayed first. Following this, the green is extended by the vehicle passage time. Depending upon vehicle actuations, the minimum green is extended by the passage time interval for each vehicle actuation. If a subsequent actuation occurs within one passage time interval, another passage time interval is added (measured from the time of the new actuation, and not from the end of the previous passage time interval). Finally, the green is terminated according to one of two mechanisms: A passage time elapses without a vehicle actuation (the phase gaps out); or the maximum green time for that phase is exceeded (the phase gaps out).

Readers interested in learning the details of actuated signal controller design are referred to appropriate references in traffic and highway engineering, including *Traffic and Highway Engineering* by Garber and Hoel.

Signal Coordination

When a number of traffic signals are located close to one another along a major corridor, one simple idea to improve the efficiency of the transportation system

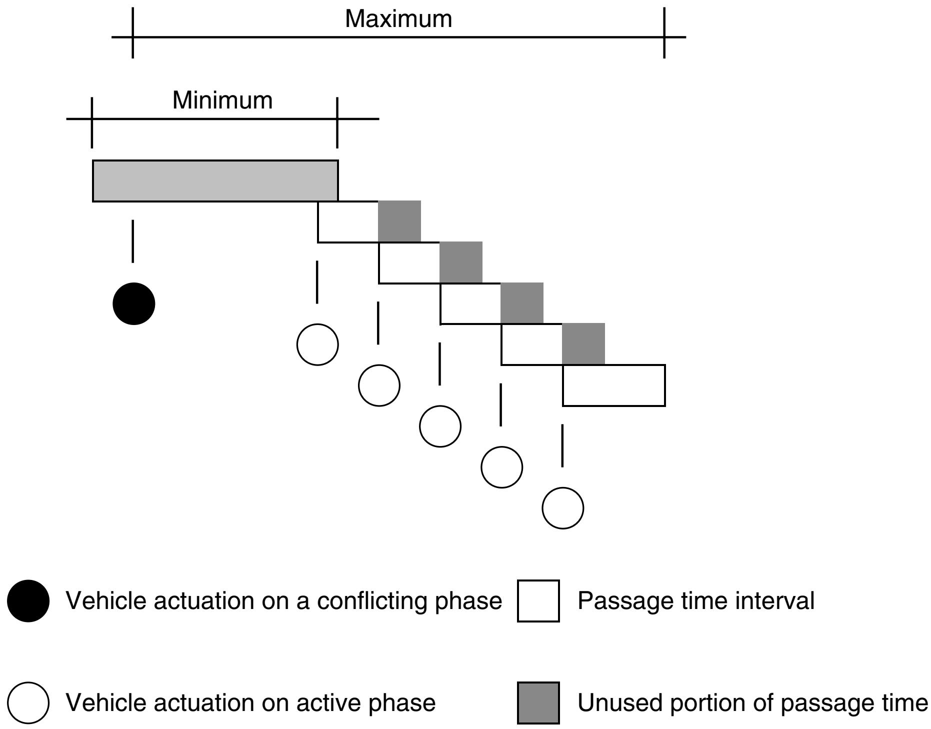


FIGURE 9.21
Actuated control operational concept.

is to coordinate the start of the green for those signals. By carefully setting the time difference between the start of green at successive intersections (i.e., this difference is commonly referred to as the signal offset, as will be explained later), it may be possible to create a "green wave" along the corridor that would allow drivers to go through those signals without having to stop at each and every intersection.

A key requirement for coordinating signals is that successive signals are close enough to one another, thereby allowing vehicles to arrive at intersections in the form of platoons (i.e., a group of vehicles closely spaced to one another). Intersections that are far apart from one another are not good candidates for coordination because vehicles, after traveling for long distances in between intersections, tend to disperse, and the platoonlike structure of the traffic stream is destroyed. In these cases, intersections can be regarded as if they were isolated intersections, and vehicles' arrival patterns at those intersections tend to become random.

To allow for coordination, all signals along a coordinated system have to have the same cycle length (in some cases, however, an intersection with exceptionally high volumes may be allowed to have double the cycle length). A common cycle length is needed so that the start of the green would occur at the same time relative to the nearby intersections. While the cycle length has to be the same, the length of the green at different intersections can vary. Given the requirement for common cycle lengths, most signals along coordinated systems are set to operate on a pretimed basis. It is still possible to coordinate traffic-actuated signals, but they would have to have a common background cycle length.

Coordinated actuated controllers are therefore often of the semiactuated type, which allow for varying the green given to the side streets from one cycle to the next.

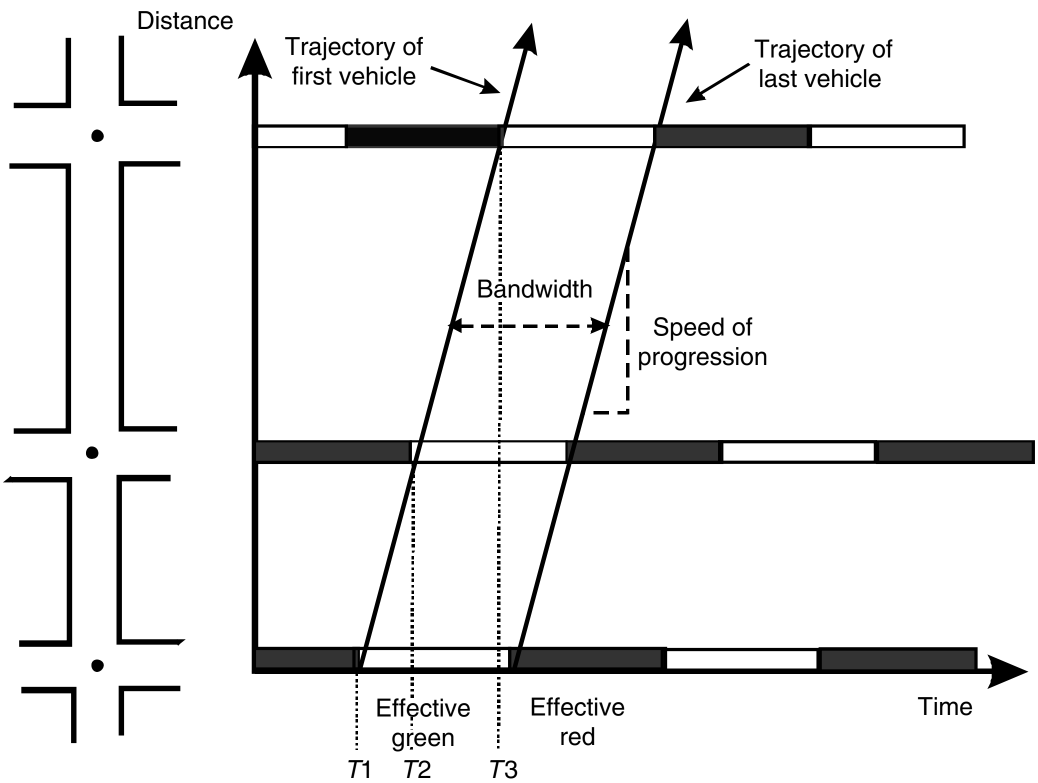
For signal coordination, individual controllers need to be interconnected in order to achieve the necessary synchronization. Typically, in a coordinated system, a master controller would send coordination pulses to all other controllers within the coordinated system (these are typically referred to as local controllers). Direct communication could be established using hard-wired cable, telephone lines, coaxial cable, fiber optic cable, or radio communications. In addition, indirect communication could be established using time-based coordinators.

Time–Space Diagram and Signal Coordination

A powerful tool that has historically been used to design coordinated signal systems is the time–space diagram introduced earlier in Chapter 2. At the present time, the use of the time–space diagram for the design of coordinated signal plans has largely been replaced by more powerful traffic simulation software and optimization algorithms. Nevertheless, the diagram is still quite useful in illustrating the concepts, factors, and challenges of signal coordination.

Figure 9.22 shows a typical time–space diagram for a signal coordination problem. To the left of the *y*-axis of the diagram, which represents the distance, we draw to scale a plan of the corridor or street along which the signals are to

FIGURE 9.22
Signal coordination on a time–space diagram.



be coordinated. We then focus on a given direction (say the NB direction in this example), and at the location of each intersection and along the x-axis, we draw a schematic representation of the phase sequences for the selected direction at that particular intersection. To make things simpler, we typically only plot the duration of the effective green (i.e., the green + the yellow) as a blank line, and the effective red as a solid line. When representing the signal plan for each intersection, it is important to correctly record the beginning of the green for each signal. Vehicles' trajectories could then be drawn and their interactions with the signal plan studied.

As can be seen from Figure 9.22, the first signal turns green at time, $T1$, followed by the second signal at time, $T2$, and the third at time $T3$. The difference between the time when an upstream signal turns green and when the downstream turns green is referred to as the signal offset. Typically the offset is defined as ($T2 - T1$) or ($T3 - T2$) and therefore is usually a positive number between 0 and the common cycle length for the coordinated system of signals. Also shown in Figure 9.22 is the concept of bandwidth. This is the amount of green that can be used by a platoon of vehicles moving through a series of intersections without having to stop at any of these intersections.

Determining "Ideal" Offsets

If we focus on one direction (such as the NB direction in Figure 9.22), determining the values for the "ideal" offsets is straightforward. If the offset for a given signal is to be related to the signal directly upstream from that signal, the ideal offset can be easily computed as follows:

$$O_{\text{ideal}} = L/S \tag{9.9}$$

where

L = distance between signalized intersections
S = average vehicle speed

The calculations are illustrated by the following example.

EXAMPLE 9.14

Calculating Ideal Offsets for Signal Coordination

It is required to coordinate signals along the one-way corridor shown in Figure 9.23. All signals shown have a common cycle length of 80 s, and the effective green for the direction to be coordinated for all signals is about 60% of the cycle length. Given that the average vehicle speed along the corridor is 55 km/h and the distances between intersections are shown in Figure 9.23, calculate the ideal offsets for the signals.

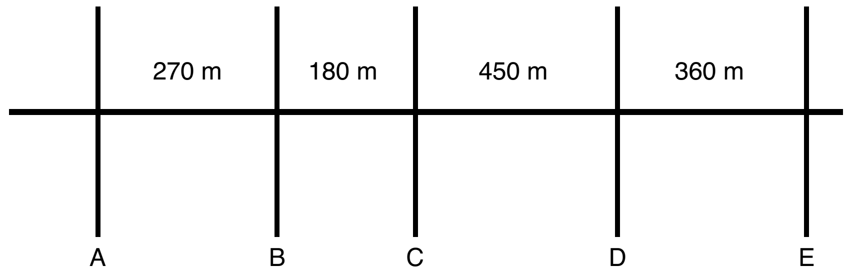


FIGURE 9.23
Calculation of ideal offsets for one-way progression.

Solution

We first convert the speed given in km/h into the equivalent value in m/s as follows:

Average speed = 55 km/h = 55 × 1000/3600 = 15.3 m/s

We then apply Equation 9.9 to calculate the offsets as shown in Table 9.12. The offset for a given signal is calculated relative to the one right upstream of that signal.

TABLE 9.12
Calculating the Ideal Offsets for the Corridor of Figure 9.23

Signal	*Offset Calculated Relative to Signal*	*Ideal Offset (s)*
B	A	270/15.3 = 17.6 s
C	B	180/15.3 = 11.8 s
D	C	450/15.3 = 29.4 s
E	D	360/15.3 = 23.5 s

Bandwidth Concept

As previously mentioned with reference to Figure 9.22, the bandwidth can be defined as the time difference, in seconds, between the trajectories of the first and last vehicle in a platoon of vehicles capable of moving through a series of intersections without having to stop at any of these intersections. *Bandwidth efficiency* gives an indication of the efficiency of the coordination scheme. It is generally defined as the ratio of the bandwidth to the cycle length, as given by Equation 9.19.

$$\text{Bandwidth efficiency} = \left(\frac{BW}{C}\right) \times 100 \tag{9.10}$$

where

BW = bandwidth, in second
C = cycle length, in second

In general, a bandwidth around 50% is regarded as an indication for good coordination.

Bandwidth capacity gives the number of veh/h that can go through the coordinated system without stopping. The bandwidth capacity can be easily computed by first determining the number of vehicles per traffic lane that go without stopping in each traffic signal cycle. This can be done by dividing the bandwidth in seconds by the discharge or saturation headway, which is typically around 2 s/veh (see Chapter 4). This number is then multiplied by the number of signal cycles/h, and by the number of traffic lanes, as shown in Equation 9.11.

$$\text{Bandwidth capacity (in veh/h)} = \frac{3600 \times BW \times N}{C \times h} \tag{9.11}$$

where

BW = bandwidth, s
N = number of through lanes in the indicated direction
C = cycle length, s
h = discharge or saturation headway, s

Determining the bandwidth for a given coordinated system can be estimated graphically from a time–space diagram similar to the one shown in Figure 9.22. Once this is determined, the efficiency and capacity of the bandwidth can be calculated. The following example illustrates the procedure.

EXAMPLE 9.15

Calculating Bandwidth, Bandwidth Efficiency and Bandwidth Capacity

Figure 9.24 shows a set of three traffic signals along an arterial with two lanes in each direction. The signals are coordinated primarily for the NB direction. The cycle length, the duration of the green for the N-S phase, and the offset for each of the three signals (A, B, and C) are as shown in Table 9.13.

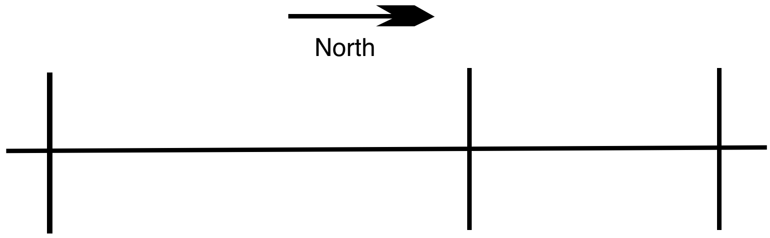
TABLE 9.13
Signal Data for Example 9.15

Signal	*Cycle Length*	*Green for N-S Phase*	*Offset Relative to Upstream Signal*
Signal A	80 s	35 s	0 s
Signal B	80 s	45 s	20 s
Signal C	80 s	40 s	15 s

Given that the average vehicle speed along the corridor is 66 km/h,

1. Draw a time–space diagram for the coordinated system.
2. Determine the bandwidth efficiency and the bandwidth capacity for the NB direction.
3. Determine the bandwidth efficiency and capacity for the SB direction.

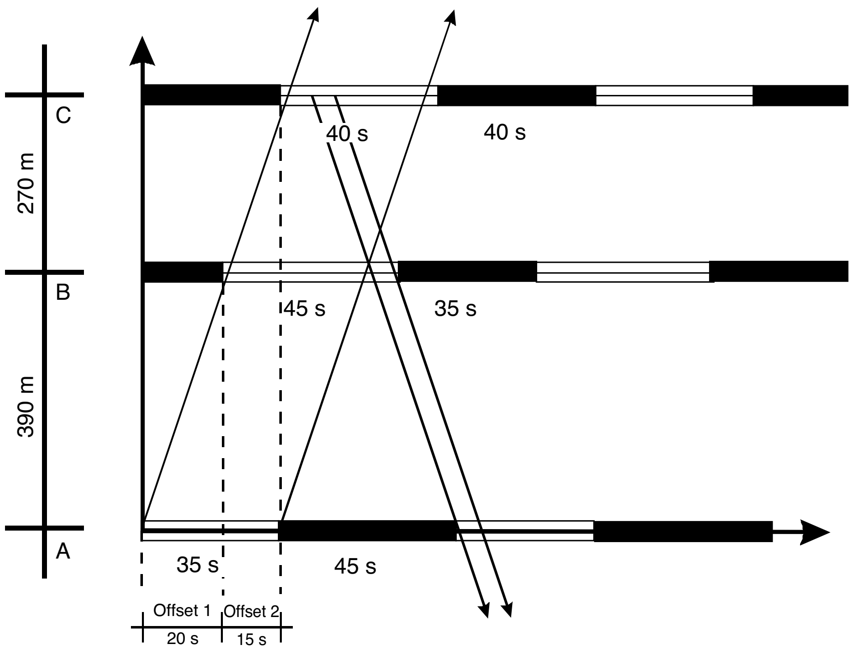
FIGURE 9.24
Roadway sketch for Example 9.15.



Solution

(1) The first step in solving this problem is to draw the time–space diagram for the coordinated system as shown in Figure 9.25.

FIGURE 9.25
Time–space diagram for the northbound direction.



The road with the three signals was first drawn to scale along the y-axis of the time–space diagram. Next, the signal timings for each of the three signals A, B, and C were sketched along the x-axis. For signal A and for the N-S direction, we have 35 s of green followed by 45 s of red (to make the 80 s cycle). Signal B has 45 s of green followed by 35 s of red. Since the offset for signal B is 20 s, the green of signal B is drawn so as to start 20 s after the start of the green for signal A. Finally, signal C gets 40 s of green and 40 s of red. The green for signal C starts 15 s after the green for signal B.

We then draw the vehicle trajectories. The average vehicle speed along the corridor is 66 km/h, which is equivalent to $66 \times 1000/3600 = 18$ m/s. Vehicle trajectories are therefore represented by straight lines having a slope of 18 m/s, as shown in Figure 9.25, for both the NB and SB directions.

(2) As can be seen from Figure 9.25, for the NB direction, the bandwidth is equal to 35 s. Given this, the bandwidth efficiency can be easily calculated from Equation 9.10 as follows:

$$\text{Bandwidth efficiency} = \left(\frac{BW}{C}\right) \times 100 = \left(\frac{35}{80}\right) \times 100$$
$$= 43.75\%$$

Bandwidth capacity can be calculated from Equation 9.11 as

$$\text{Bandwidth capacity} = \frac{3600 \times BW \times N}{C \times h} = \frac{3600 \times 35 \times 2}{80 \times 2}$$
$$= 1575 \text{ veh/h}$$

(3) For the SB direction, as can be clearly seen for Figure 9.25, the bandwidth is much smaller, only about 6 s. With the bandwidth determined, the bandwidth efficiency and safety can be easily calculated from Equations 9.10 and 9.11 as follows:

$$\text{Bandwidth efficiency} = \left(\frac{BW}{C}\right) \times 100 = \left(\frac{6}{80}\right) \times 100$$
$$= 7.5\%$$

Bandwidth capacity can be calculated from Equation 9.11 as

$$\text{Bandwidth capacity} = \frac{3600 \times BW \times N}{C \times h} = \frac{3600 \times 6 \times 2}{80 \times 2}$$
$$= 270 \text{ veh/h}$$

Challenges in Coordinating Signals

While signal coordination on one-way streets is straightforward, this is not the case for two-way streets and grid networks. The complexity arises from the fact that, for a two-way street, once the offsets are determined for a given direction (based upon the needs of that direction), the offsets for the other direction are fixed (see Figure 9.26). These offsets (for the other direction) may be inappropriate for the needs of that other direction, as Figure 9.26 and Example 9.15 illustrate.

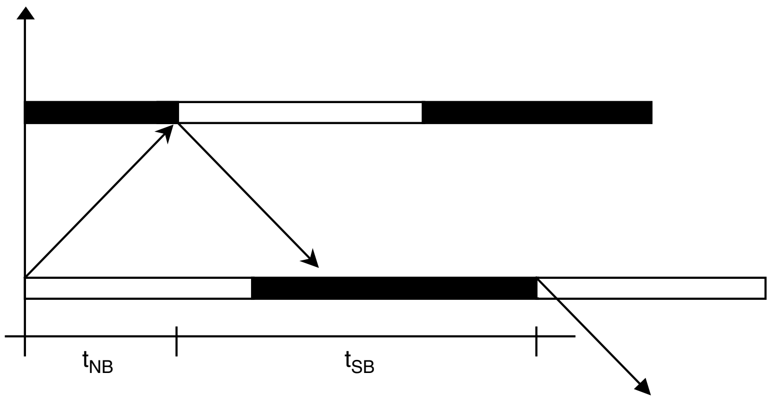


FIGURE 9.26
Relationship between offsets on two-way streets.

Determining the offsets for a two-way street starts with the realization that the offsets in the two directions add to one cycle length, or to an integer multiple of cycle lengths for longer block lengths (see Figure 9.26). Therefore, with reference to Figure 9.26, we could say that

$$t_{NB} + t_{SB} = C \tag{9.12}$$

where

t_{NB} = offset in the northbound direction
t_{SB} = offset in the southbound direction
C = cycle length

The actual offset, which has to satisfy Equation 9.12, can then be expressed as

$$t_{actual} = t_{ideal} + e \tag{9.13}$$

The goal of most signal optimization programs is to minimize the weighted sum of the difference between the actual and ideal offsets.

A number of computer programs are currently available to help in the design of optimal signal timing plans for coordinated systems. The idea behind these programs is to find a set of timing parameters (such as offsets, cycle lengths, and phase intervals) that would minimize a given performance measure (such as total delay or total number of stops) while satisfying the different constraints (such as the constraint defined in Equation 9.12). Among the most famous of these computer programs are TRANSYT-7F and SYNCHRO.

TRANSYT (TRAffic Network StudY Tool) was first developed by the U.K.'s Transport Road Research Laboratories in the late 1960s and has undergone several revisions since then. TRANSYT Version 7 was Americanized for the Federal Highway Administration (FHWA) in the late 1970s and early 1980s; hence the name TRANSYT-7F. Currently, TRANSYT-7F is one of the most widely used computer programs for developing optimal timing plans for corridors and networks. To develop optimal coordinated signal plans, TRANSYT varies the cycle length, the phase lengths, and the signal offsets until the plan that optimizes a user-defined objective function is identified.

SYNCHRO is another signal timing software that could be used to generate optimal signal plans (cycle length, phase lengths, and offsets). For optimization, SYNCHRO uses an objective function that attempts to minimize a combination of delay, number of stops, and number of vehicles in queue. A unique advantage of SYNCHRO is its ability to accurately model the operation of actuated controllers within a coordinated system.

Adaptive Traffic Control Systems

Adaptive or computer traffic control refers to the use of a digital computer to control the operation of a group or system of signals. Adaptive traffic control systems combine the concept of actuated or computer control with the concept of signal coordination. These systems can therefore be regarded as the next step in the evolution of traffic signal control systems. The idea behind adaptive or computer traffic control systems is to take advantage of the power of digital computers to control many signals, along an arterial or in a network, from one central location. Computer traffic control systems predate ITS by several decades; the first installation of these systems took place in the early 1960s. These systems, however, have undergone continuous refinement since that time. In the following paragraphs, we give some historical perspective on the development of these systems.

The most basic type of computer signal control system first appeared in the 1960s. The idea was for a computer to control a series of controllers, but with no "feedback" of information from the field detectors to the computers. In such a system, the traffic plans implemented are not responsive to the actual traffic demand. Instead, the plans are developed "offline" from historical traffic counts and implemented based upon the time of day and day of the week. While this system appears rather simplistic, it offers several advantages, including the ability to update signal plans from a central location, the ability to store a large number of signal plans, and the automatic detection of malfunctioning equipment.

The next development was to have signal control systems in which information from the field traffic detectors is fed back into the central computer. The computer would then use this information to select the signal plan to be implemented. Plan selection is conducted according to one of the following methods.

Select plan from a library of predeveloped plans. In this method, the system has access to a database (library) that stores a large number of different traffic patterns along with the "optimal" signal plans for each pattern (these plans are developed offline). Based upon information from the traffic detector, the computer matches the observed traffic pattern against the patterns stored in the library and identifies the closest one. The plan associated with the identified pattern is then implemented. This type of adaptive traffic control system is often referred to as a first-generation system. The distinguishing feature of these systems is that the plans, while responsive to traffic conditions, are still developed offline. Typically, the frequency of signal update is every 15 min. First-generation systems do not generally have traffic prediction capabilities.

Develop plan online. In this method, the "optimal" signal plan is computed and implemented in real time. This requires enough computational power to do the necessary computations online. Systems that develop plans online are classified as either second-generation or third-generation systems. These systems typically have a much shorter plan update frequency compared to first-generation systems. In addition, signal plans are computed in real time based on forecasts of traffic conditions obtained from feeding the detector's information into a short-term traffic-forecasting algorithm. For second-generation systems, the plan update frequency is every 5 min, whereas third-generation systems have an update interval ranging from 3 to 5 min. The next section will describe some examples of these systems that are in use throughout the world.

Adaptive Traffic Control Algorithms

A number of adaptive traffic control algorithms are currently available. Among the most widely accepted of these are SCOOT, and SCATS. SCOOT (Split, Cycle, Offset Optimization Technique) is an adaptive traffic control system developed by the U.K.'s Transport Research Laboratory (TRL) in the early 1980s. In 1996, the system was in operation in more than 130 towns and cities around the world. SCOOT operates by attempting to minimize a performance index (PI), which is typically taken as the sum of the average queue length and the number of stops at all approaches in the network. In order to do this, SCOOT modifies the cycle lengths, offsets, and splits at each signal in real time in response to the information provided by the vehicle detectors.

The operation of SCOOT is based upon cyclic flow profiles (CFPs), which are histograms of traffic flow variation over a cycle, measured by loops and detectors, placed midblock on every significant link in the network. Using CFPs, the offset optimizer calculates the queues at the stop line. The optimal splits and cycle length are then computed. In recent years, a number of additional features have been added to SCOOT to improve its effectiveness and flexibility. These include the ability to provide preferential treatment or signal priority for transit vehicles; the ability to automatically detect the occurrence of incidents; and the addition of an automatic traffic information database that feeds historical data into SCOOT, allowing the model to run even if there are faulty detectors.

The Sydney Co-ordinated Adaptive Traffic System (SCATS) was developed in the late 1970s by the Roads and Traffic Authority of New South Wales in Australia. For operation, SCATS requires only stop line traffic detectors, and not midblock traffic detection as does SCOOT. This is definitely an advantage since the majority of existing signal systems are equipped with sensors only at stop lines. SCATS is a distributed intelligence, hierarchical system that optimizes cycle length, phase intervals (splits), and offsets in response to detected volumes. For control, the whole signal system is divided into a large number of smaller subsystems ranging from 1 to 10 intersections each. The subsystems run individually unless traffic conditions require the "marriage" or integration of the individual subsystems.

In developing real-time signal plans, SCATS' objective is generally to equalize the saturation flow ratio of the conflicting approaches. Consequently, the system in many cases does not minimize delays on major arterials, which may actually exhibit deterioration in their level of service particularly during peak periods. This was evident in the FAST-TRAC ITS field test in Oakland County, Michigan. In that project, video detection was used to feed a SCATS system, which then developed timing plans in real time.

Signal Preemption and Priority

Advanced arterial traffic control (AATC) systems often include signal preemption and priority capabilities. These capabilities allow controllers to detect vehicles approaching signalized intersections and to provide them with some type of preferential treatment. There are several instances where such systems could be used. For example, signal preemption could be used to provide an approaching emergency vehicle with the green light, an act that could save the lives of those in an emergency situation. Signal preemption could be used at highway–rail grade crossings to prevent a vehicle from being trapped on the railway track. They could also be used to provide transit vehicles with some preferential treatment that may involve extending the green at an intersection for an approaching bus in order to allow buses to stay on their schedule.

Historically, the term *signal preemption* has been used to refer to highway–rail grade crossing systems, emergency vehicles systems, and transit systems. More recently, the use of the term *signal priority* is preferred to reflect the fact that there is a need to assign different priorities to different requests. For example, a highway–rail crossing typically is assigned the highest priority, which would typically involve instantaneous response from the controller to avoid trapping vehicles on the railway track. Emergency vehicles are typically assigned a slightly lower priority to allow a signal from a highway–rail grade crossing to override the emergency vehicle request, if a need arises to do so. Finally, transit vehicles are assigned an even lower priority. Such requests received from transit vehicles typically do not cause major disruptions of the phase sequence but may extend a green split by a prescribed amount, allowing the bus to pass the signal.

There are several control strategies that could be used to grant transit vehicles preferential treatment at signalized intersections. In this section, however, we focus primarily on emergency vehicle systems. In our discussion, we will use the term *preemption* since this is the term that is most often used today to refer to emergency vehicle systems. Signal preemption is usually designed to give the green light in the direction of the oncoming emergency vehicle while posting a red light for all other approaches (Figure 9.27). The other, less frequently used, alternative is for preemption to result in all approaches turning red. There are basically two different approaches for signal preemption—the first is based on local communication between the vehicle and the controller. In this case, the controller identifies approaching vehicles using acoustic, optic, or special loop technology.

Under the second approach, the right-of-way is granted based upon requests from an emergency management center to a traffic management center. This approach requires a highly integrated ITS system whereby the emergency management center would track its vehicles in real time, using global positioning systems technology, and would send signal preemption requests to the traffic management center. The center would then grant the emergency or transit vehicle

FIGURE 9.27
Signal preemption.
Source: 3M Web site.

the right-of-way. The approach allows for the development of more sophisticated signal coordination strategies compared to the local signal preemption approach, strategies that would anticipate the vehicle's turning movements and would minimize overall system disruption. However, it is much more complex and more expensive than the local signal preemption approach.

Advanced Arterial Traffic Control Systems Benefits

The benefits to be expected from advanced arterial traffic control systems include travel time reduction benefits; environmental benefits resulting from improved traffic flow conditions, lower emission rates, and less fuel consumption; and safety-related benefits resulting from lower accident rates under improved travel conditions. Following is a brief discussion of each of these benefits.

Travel Time Reduction Benefits

Evaluation studies conducted throughout the United States indicate that advanced signal control systems could result in travel time reduction in the range of 8 to 25%. The exact value of time reduction will depend upon a number of factors, including the variability of travel demand; the overall level of congestion; the time interval between signal timing plan modifications; and the density of traffic signals.

Environmental Benefits

Studies show that advanced signal control systems could result in a reduction in air pollutants (such as hydrocarbons and carbon monoxide) that range between 16 and 19%. They could also result in a 4 to 12% reduction in fuel consumption.

Safety Benefits

Some studies show that advanced signal control systems could also result in a reduction in the frequency of injury-related accidents ranging from 6 to 27%.

As an example of how the benefits of an advanced arterial traffic control system may be calculated, consider a segment on an arterial carrying an average annual daily traffic (AADT) of about 20,000 veh/day. Assuming that the average trip length for this segment is about 10 min, and using the conservative estimate of a 10% reduction in trip travel time (as discussed, studies show a reduction in the range 8 to 25%), the time savings resulting from deploying the ITS system could be estimated at ($0.10 \times 10 = 1.0$ min/veh/day). Assuming that the value of time is equal to \$8.90/h, the benefits could be computed as follows:

$$\begin{aligned}\text{Benefits} &= (\text{\# of vehicles}) \times (\text{time saved}) \times (\text{time value}) \times 365 \\ &= 20{,}000 \times (1.0/60) \times (\$8.90) \times 365 = \$\,1{,}082{,}833/\text{year}\end{aligned}$$

Advanced Public Transportation Systems

Advanced public transportation systems attempt to improve the efficiency, productivity, and safety of transit systems. They also strive to increase ridership levels and customer satisfaction. In this section, we describe some examples of advanced public transportation systems. The examples described can be categorized under the following four categories: automatic vehicle location (AVL) systems; transit operations software; transit information; and electronic fare payment systems.

Automatic Vehicle Location (AVL) Systems

AVL systems are designed to allow for tracking the location of transit systems in real time. These systems work by measuring the actual real-time position of each vehicle and communicating the information to a central location. This information can then be used to increase dispatching and operating efficiency; allow for quicker response to service disruptions; provide input to transit information systems; and increase driver and passenger safety and security.

While a number of technologies are available for AVL systems, including dead reckoning, ground-based radio, signpost and odometer, and GPS, most agencies now are choosing GPS-based systems. GPS is a navigational and positioning system that relies upon signals transmitted from satellites for its operation. In 1996, there were 86 transit agencies across the country operating, implementing, or planning AVL systems; over 80% of these were using GPS technologies. The following section describes some real-world implementations of AVL systems.

Real-World Examples of Transit AVL Systems

In Atlanta, Georgia, about 250 buses of the Metropolitan Atlanta Rapid Transit Authority (MARTA)'s 750-vehicle fleet are equipped with AVL. The system is linked to the Georgia Department of Transportation traffic management center. There are also electronic signs at a few bus stops for displaying information to passengers waiting at these stops. The system was shown to yield concrete benefits, including improved on-time performance as well as increased safety.

The Tri-County Metropolitan Transportation District of Oregon (Tri-Met) has recently completed the implementation of a GPS AVL system for 640 fixed-route vehicles and 140 paratransit vehicles. The AVL is being employed as part of a regional ITS system, whereby the buses will be used as probes for traffic monitoring, as discussed previously.

The Milwaukee Transit System (MTS) completed the installation of a GPS AVL on 543 buses and 60 support vehicles. Preliminary results indicate a 28% decrease in the number of buses more than one minute behind schedule.

Transit Operations Software

Transit operations software allow for automating, streamlining, and integrating many transit functions. This includes applications such as computer-aided dispatching (CAD), service monitoring, supervisory control, and data acquisition. The use of operations software can improve the effectiveness of operations dispatching, scheduling, planning, customer service, and other agency functions. Operations software is available for fixed-route bus operations as well as for paratransit or demand-responsive operations.

Operations software for demand response transit implement new scheduling and dispatching software for improved performance and increased passenger-carrying capability of the vehicles. Systems vary widely in their capabilities; the high-end systems have integrated automated scheduling and dispatching software with AVL, geographic information systems, and advanced communications systems. These systems provide dispatchers with the capability to view maps of the service area with the locations of all the vehicles in real time. Drivers have mobile data terminals displaying the next hour's pickups and dropoffs.

Real-World Examples of Transit Operations Software Implementations

Kansas City, Missouri, was able to reduce up to 10% of the equipment required for bus routes using an AVL/CAD system. This allowed Kansas City to recover its investment in the system within two years. On-time performance was improved by 12% in the first year of operating the AVL system. In *Ann Arbor, Michigan*, the city's paratransit service (A-Ride) has implemented computer-aided dispatch, automated scheduling, and advanced communications for eight AVL-equipped paratransit vehicles. This system is able to provide service 24 hours a day, with the services of a dispatcher needed only to take reservations and cancellations from callers and to confirm rides.

Transit Information Systems

Transit information systems implement traveler information systems that provide travelers with transit-related information. Three types of such systems can be identified: pretrip systems; in-terminal/wayside systems; and in-vehicle transit information systems.

Pretrip information systems provide travelers with accurate and timely information before starting their trips to allow them to make informed decisions about modes as well as routes and departure times. The information provided can cover a wide range of categories, including transit routes, maps, schedules, fares, park-and-ride locations, points of interest, and weather. In addition, these systems often support itinerary planning. Methods of obtaining pretrip information include touch-tone phones, pagers, kiosks, the Internet, fax machines, and cable TV.

In-terminal/wayside systems provide information to transit riders who are already en route. This information is typically communicated using electronic signs, interactive information kiosks, and CCTV monitors. The overall goal is to provide real time bus and train arrival and departure times, reduce waiting anxiety, and increase customer satisfaction. In-vehicle information systems provide en route information for travelers on board the vehicle. The major impetus behind such systems is to comply with the applicable provisions of the American Disabilities Act of 1991.

Real-World Examples of Transit Information Systems

A good example of pretrip transit information systems can be found in Seattle, Washington, where a key product of Seattle Metro's transit information system is a Web site whereby travelers can obtain information on transit schedules and fares, van and carpooling, ferries, and park-and-ride facilities. This site also provides assistance to transit users in planning their trips. In addition, the University of Washington has developed a Java applet to allow users to view the locations of all buses traveling throughout the Metro System. The University has also developed Web pages to help travelers predict the time of arrival of buses at different bus stops.

Examples of in-terminal and in-vehicle information systems can be found in Ann Arbor, Michigan, where a pair of 79 cm video monitors is used to display real time data generated by the AVL system to inform passengers at the downtown transit center (in-terminal) about arrival status, delays, and departure times. The Ann Arbor system also includes in-vehicle annunciators and displays from which passengers will receive next-stop and transfer information. The latter will take the form of announcements that identify valid bus transfers at upcoming stops.

Electronic Fare Payment Systems

The idea behind electronic fare payment systems is to facilitate the collection and management of transit fare payments by using electronic media rather than cash or paper transfers. These systems consist of two main components: a card and a card reader. Cards could be of the magnetic-strip type, where the reader does most of the data processing. They could also be equipped with a microprocessor (smart cards), and, in this case, data processing could occur on the card itself. Electronic fare payment systems offer a number of advantages: they offer convenience to vehicle operators by eliminating the need for any actions to be taken on their part; they eliminate the need for a passenger to worry about having exact change for the bus fare; they facilitate the collection and processing of fares; and they allow for adopting more complex and effective fare structures.

There are two types of electronic fare payment systems: (a) closed systems, and (b) open systems. Closed systems are limited to one main purpose (i.e., paying transit fares) or to a few other applications, such as paying parking fees. However, the value stored on the card cannot be used outside the defined set of activities; hence the name *closed system*. Open systems can be used outside the transit system. A prime example of an open system is a credit card, which naturally can be used with multiple merchants.

Transit AVL and Operations Software Benefits

Studies show that the deployment of transit AVL and operations software could result in both capital as well as operating cost savings to the operating agency. The default values are a reduction between 1 and 2% in fleet size and a reduction in the range of 5 to 8% in operating costs. To illustrate how the benefits from deploying such systems could be estimated, consider the case of a transit agency whose annual capital and operating costs are \$2,000,000 and \$1,500,000, respectively. Assuming a cost saving of 1.5% for capital costs and 6% saving for operating costs, the annual savings are equal to

$$2{,}000{,}000 \times 1.5/100 + 1{,}500{,}000 \times 6/100 = \$120{,}000/\text{yr}$$

Multimodal Traveler Information Systems

Multimodal traveler information systems are designed to provide static, as well as real-time, travel information over a variety of transportation modes (e.g., highways, transit, ferries, etc.). In essence, these systems integrate the traffic information dissemination functions of freeway and incident management systems with the functions of transit information systems. They then add more information from sources such as Yellow Pages, tourist organizations, and weather services. Traveler information can be provided before or during a trip (pretrip and en route traveler information). Table 9.14 shows an extensive list of data of potential interest to the traveler that could be part of the traveler information system. Information is classified as either static or real-time information.

Following the collection and processing of data, telecommunications technologies, including voice, data, and video transmission over wire-line and wireless channels, are then used to disseminate the information to the public. Among

TABLE 9.14
Potential Contents of a Multimodal Traveler Information System

Static information: Known in advance, changes infrequently	Planned construction and maintenance activities
	Special events, such as state fairs and sporting events
	Transit fares, schedules, and routes
	Intermodal connections (e.g., ferry schedules along Lake Champlain)
	Commercial vehicle regulations (such as Hazmat (hazardous material) and height and weight restrictions)
	Parking locations and costs
	Business listings, such as hotels and gas stations
	Tourist destinations
	Navigational instructions
Real-time information: Changes frequently	Roadway conditions, including congestion and incident information
	Alternate routes
	Road weather conditions, such as snow and fog
	Transit schedule adherence
	Travel time

the means of information dissemination are the Internet, cable TV, radio, phone systems, kiosks, pagers, personal digital assistants (PDAs), and in-vehicle display devices. In addition, efforts are currently underway to implement a national number, 511, which would provide travelers in the United States with real-time multimodal travel information.

Benefits of Multimodal Traveler Information Systems

Multimodal traveler information services allow users to make more informed decisions regarding time of departure, routes, and mode of travel. They have been shown to increase transit usage and reduce congestion when travelers choose to defer or postpone trips or select alternate routes. A good example of a regional multimodal traveler information system is given by the Smart Trek Model Deployment Initiative in Seattle, Washington. At the center of this system is a set of protocols and paradigms designed to collect and fuse data from a variety of sources; process the data to derive useful information; disseminate the information derived to independent information service providers; and warehouse the data for the purposes of long-range planning.

Challenges Facing Multimodal Traveler Information Systems

Although multimodal traveler information systems have the potential to yield significant benefits for both travelers and system operators, the demand for products has been slow to materialize. The size of the market for traveler information systems has been modest to date. Several reasons could be provided for the slow growth of the traveler information market. First, consumer awareness of traveler information products is currently quite low. Second, the price of some products, especially in-vehicle display devices, is still high. Finally, the quality of the information and the extent of coverage need to be increased.

Advanced Technologies for Rail

The rail industry has also been quite active in applying IT to rail transportation applications. While there are numerous examples of the application of advanced technologies to improve the safety and efficiency of rail transportation, we limit our discussion here to just two representative examples, namely, positive train control (PTC) systems, and (2) intelligent rail intersections.

Positive Train Control

Positive train control (PTC) systems are designed to allow for controlling train movements with safety, security, and efficiency. PTC systems integrate digital data communications networks, GPS navigation systems, computers on board trains, in-cab displays, and control center computers and displays. PTC systems will allow personnel in the control center to track the location of trains and maintenance crews in real time, and will control train movements so as to achieve optimum speeds and hence maximum track capacities. PTC will also allow a control center to stop a train should the locomotive crew be incapacitated, thereby providing a greater level of safety and security. Demonstration projects of PTC systems are currently underway in the United States, and wide-scale implementation is expected to begin shortly.

Intelligent Highway–Rail Intersections (HRIs)

Intelligent HRI systems are designed to eliminate accidents at highway–rail grade crossings. Active warning systems at intersections (such as flashing lights and gates) are activated as an approaching train is detected. The equipment at the HRI may also be connected to the adjacent signal system. In case a trapped vehicle on the railway track is detected, the system would immediately preempt the signal and simultaneously warn the locomotive engineer. Intelligent HRI also continuously monitors the health of the detection and warning systems and reports any detected malfunctioning to the appropriate authorities.

Summary

In this chapter, we have discussed some of the applications of information technology to improve the efficiency and safety of the transportation system. As was discussed, this effort is often referred to as Intelligent Transportation Systems or ITS applications. The chapter focused primarily on five key ITS applications: (1) freeway and incident management systems; (2) advanced traffic signal systems;

(3) advanced public transportation systems; (4) multimodal traveler information systems; and (5) advanced technologies for rail. The operational concept behind each application was discussed, as well as a description of the likely benefits of these systems. A brief mention was also made of advanced technologies applications to rail transportation. The application of advanced technologies to transportation is still an evolving field, and new ideas are proposed every day.

PROBLEMS

9.1 Select an ITS project in your state that you are familiar with, and briefly describe its basic concept of operation. What are the likely benefits of that project?

9.2 List the primary objectives of a freeway and incident management system (FIMS).

9.3 What is the difference between *recurrent* and *nonrecurrent* congestion?

9.4 Describe the four basic components of a traffic surveillance system.

9.5 Select four different methods for traffic detection, and briefly discuss the advantages and disadvantages of each method.

9.6 A freeway detection station on an eight-lane freeway gives the occupancy measurements shown below. The average length for vehicles is 6.25 m for lane 1, 5.75 m for lane 2, 5.25 m for lane 3, and 5 m for lane 4. Assuming that the effective length for the loop detectors is 2.5 m, determine the traffic density for each lane and for the freeway direction.

Lane No.	*Occupancy (%)*
Lane 1	24%
Lane 2	17%
Lane 3	14%
Lane 4	12%

9.7 In the context of traffic monitoring, explain what is meant by "probe vehicles." Discuss the different technologies that can be used to implement the concept.

9.8 Briefly discuss the four different stages of the incident management process.

9.9 Describe the parameters used to evaluate the performance of automatic incident detection (AID) algorithms.

9.10 To evaluate the performance of an AID algorithm, its performance was observed over a period of 45 days. During this period, a total of 80 incidents occurred, out of which the algorithm managed to detect 63 incidents. At the same time, the algorithm gave a total of 1300 false alarms. If the algorithm is applied every 30 seconds, determine the detection rate (DR) and the false alarm rate (FAR) for the algorithm.

9.11 A transportation agency desires to compare the performance of five different AID algorithms in order to select one algorithm for implementation in its traffic operations center. To do this, the agency looks at some historical data that give the detection rate (DR), false alarm rate (FAR), and Mean Time to Detect (MTDD) for the five algorithms. The data compiled are shown in the following table. Assuming that the agency would like to place equal emphasis on the three performance measures, which algorithm should the agency select?

	DR (%)	*FAR (%)*	*MTTD (min)*
AID1	83	0.37	2.1
AID2	92	0.86	1.2
AID3	87	0.03	0.4
AID4	95	1.24	0.7
AID5	72	0.73	0.2

9.12 In Problem 9.11, assuming that the agency would like to lay twice as much emphasis on the time to detect incidents as on either the detection rate, or false alarm rate, which algorithm should the agency choose?

9.13 Pick three different types of AID algorithms, and briefly outline how they work.

9.14 Two detection stations on a freeway equipped with a California-type AID algorithm give the following readings. The calibration process for the AID algorithm shows that the values for the algorithm's three thresholds (T_1, T_2, and T_3) are equal to 25%, 0.30, and 0.45, respectively. Determine the time step when an incident alarm would be triggered and the time step when the incident state would be terminated.

Time Step	Occ_{up} (%)	Occ_{down} (%)
1	55	18
2	67	17
3	72	15
4	70	14
5	67	13
6	69	14
7	74	9
8	65	17
9	60	24
10	42	30
11	39	34
12	37	33

9.15 A four-lane freeway system carries an estimated volume of 3400 veh/h during the peak hour in the main direction. An incident occurs resulting in a loss of about 60% of the original freeway capacity. Without an incident management system, the incident is likely to last for a period of 1 hour. Determine the time savings possible if an incident management system is implemented such that the duration is reduced to only 20 mins. Assume the capacity of a freeway lane is equal to 2200 veh/h/lane.

9.16 An accident occurs on a six-lane freeway carrying a peak hourly volume of 4800 veh/h in the main direction. The accident blocks two of the freeway's three lanes, resulting in a significant reduction of the freeway capacity to a value of only 2000 veh/h. Assuming a lane capacity of 2100 veh/h/lane, compare the maximum length of queue, the maximum delay incurred by a vehicle, and the total vehicle delay for the following two cases:

(a) Without an incident management system, the incident lasts for a duration of 75 minutes.

(b) With an incident management system in place, the incident duration is reduced to only 30 minutes.

9.17 A four-lane freeway system carries an estimated 2600 veh/lane during the peak hour in the peak direction. Studies have shown that the maximum capacity of the freeway is 2000 veh/h/lane. For an incident with a duration of 90 minutes that would block 50% of the capacity of the freeway, what are the time savings resulting from implementing an incident management system that would cut the incident duration in half?

9.18 Briefly discuss the likely benefits of ramp metering.

9.19 Distinguish between restrictive and nonrestrictive ramp metering.

9.20 Briefly describe the different types of ramp metering strategies.

9.21 Briefly describe the difference between open-loop and closed-loop control for ramp metering systems.

9.22 Design a pretimed single entry ramp metering system on a four-lane freeway. The upstream traffic volume is 3800 veh/h/direction, and the freeway lane capacity is equal to 2300 veh/h/lane. Assume the green interval is equal to 2 s.

9.23 Design a signal plan for a ramp metering system on a six-lane freeway that carries a total volume of 5220 veh/h in the peak direction. Assume the freeway lane capacity is equal to 2100 veh/h/lane.

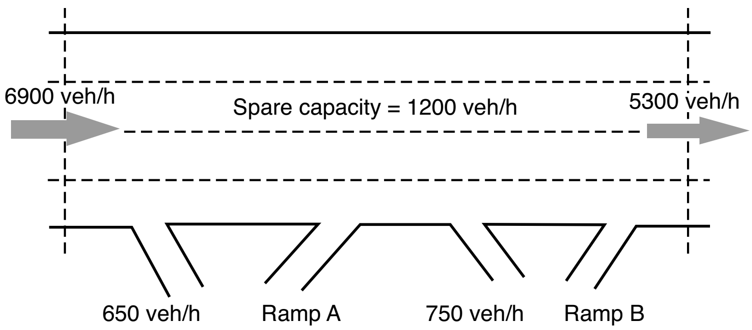
9.24 The occupancy measurements at a local traffic-responsive ramp metering station are as shown in the following table. Determine the metering rates for the different control periods.

Control Period	*Measured Occupancy (%)*
1	12
2	18
3	17
4	24

9.25 Determine the set point for an ALINEA-type control ramp meter given the following:

Average passenger car length = 5.25 m
Average truck length = 8.5 m
Percentage of trucks in traffic stream = 8%
Effective detector length = 2.5 m
Upper density level that corresponds to LOS E = 28 passenger car/km/lane

9.26 Determine the appropriate metering rates for the two on-ramps A and B shown if the projected demand for ramp A is 900 veh/h and that for ramp B is 700 veh/h.



9.27 A traffic-responsive ramp has adequate storage space for 10 vehicles. During the peak hour, it is estimated that the metering rate will not exceed 750 veh/h, whereas the average traffic demand is 620 veh/h. Using queuing theory, determine the average number of vehicles on the ramp, the average delay for the on-ramp vehicles, and the probability that the ramp will be full.

9.28 Explain the difference between pretrip and en route travel information user services.

9.29 Discuss the difference between dynamic versus static traffic assignment.

9.30 As was discussed earlier, delay in implementing dynamic traffic routing strategies results in incurring additional delays. The additional delay can be calculated using Equation 9.8 as follows:

Additional delay in veh/minutes =

$$(q - q_1) \times \delta \times \left[\frac{\delta(q - c) + 2D(c - c^*)}{120(c - q_1)} \right]$$

where

q = traffic arrival rate before routing (veh/h)
q_1 = reduced traffic arrival rate after routing (veh/h)
δ = waiting time before implementing a routing strategy in minutes
c = normal capacity of segment without any incidents (veh/h)
c^* = reduced capacity of segment as a result of an incident (veh/h)
D = incident duration in minutes

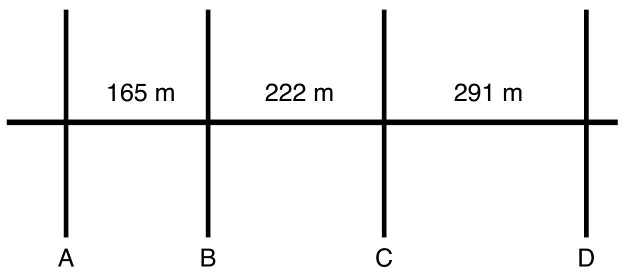
Use a cumulative plot to confirm the validity of the preceding equation.

9.31 A four-lane freeway segment carries an average volume of 3800 veh/h. The unrestricted capacity of a freeway lane can be assumed to be equal to 2300 veh/h/lane. An incident occurs that results in a 65% reduction in the capacity of the segment. The incident lasts for 60 mins. To relieve congestion, traffic routing is implemented, which reduces the traffic volume on the segment to 2200 veh/h. What would be the extra delay incurred if it took 4 mins to implement the routing strategy versus only 20 s?

9.32 Briefly define the following terms in relation to actuated signals: (1) minimum green; (2) passage time; and (3) maximum green.

9.33 Briefly describe the operational concept for an actuated controller.

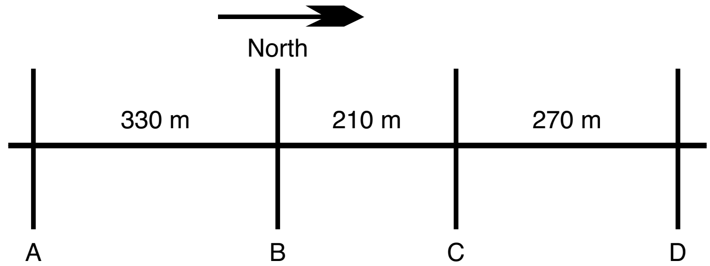
9.34 It is required to coordinate the signals along the two-lane, one-way corridor shown:



All signals have a common cycle of 90 s, and the effective green for the direction to be coordinated is 66.67% of the cycle length. Given that the average speed along the corridor is 63.4 km/h, calculate the following:

(a) The ideal offsets
(b) The bandwidth efficiency
(c) The bandwidth capacity

9.35 A north-south arterial with two lanes in each direction has four of its signals coordinated for the northbound direction, as shown. The average vehicle speed along the corridor is 55.5 km/h.



The cycle length, the duration of the green for the N-S phase, and the offset for each of the four signals (A, B, C, and D) are as shown:

Signal	*Cycle Length (s)*	*Green for N-S Phase*	*Offset Relative to Upstream Signal*
A	100	55	0
B	100	60	21
C	100	45	14
D	100	50	18

(a) Draw a time–space diagram for the coordinated system.

(b) Determine the bandwidth efficiency and bandwidth capacity for the NB direction.

(c) Determine the bandwidth efficiency and bandwidth capacity for the SB direction.

9.36 Briefly trace the development of adaptive traffic control systems since the early 1960s.

9.37 Discuss the difference between signal preemption and signal priority.

9.38 Determine the benefits to be expected from the deployment of an advanced traffic signal system on an arterial with an average ADT of 30,000 veh/day. The average trip length on the segment where the system is to be deployed is 15 mins. Assume that the value of time is equal to $10.00/h.

9.39 Give some examples for (1) real-world transit tracking systems; and (2) transit information systems.

9.40 A transit agency has an annual capital and operating cost of $3,000,000 and $2,500,000, respectively. Determine the benefits to be anticipated from deploying an AVL and transit operations software system.

9.41 Describe two applications of advanced technologies in rail transportation.

References

1. Bishop, R., *Intelligent Vehicles Technology and Trends*, Artech House, Inc., Norwood, MA, 2005.
2. Bretherton, D., "Current Developments in SCOOT: Version 3," in *Transportation Research Record 1554,* TRB, National Research Council, Washington, D.C., 1996.
3. Cambridge Systematics and ITT Industries, *ITS Deployment Analysis System User's Manual.* Cambridge, MA, 2000.
4. Chowdhury, M. A., and Sadek, A., *Fundamentals of Intelligent Transportation Systems Planning*, Artech House, Inc., Norwood, MA, 2003.
5. Cheu, R. L., and Ritchie, S. G., "Automated Detection of Lane-Blocking Freeway Incidents Using Artificial Neural Networks," *Transportation Research C,* Vol. 3(6), pp. 371–388.
6. Dailey, D. J., Smart Trek: A Model Deployment Initiative. U.S. Department of Transportation, 2001. Available at http://www.its.washington.edu/pubs/smart_trek_report.pdf
7. Garber, N. J. and Hoel, L. A., *Traffic & Highway Engineering,* Brooks/Cole, Pacific Grove, CA, 2002.
8. Hansen, B. G., Martin, P. T., and Perrin, H. Joseph, Jr., "SCOOT Real-Time Adaptive Control in a CORSIM Simulation Environment," in *Transportation Research Record 1727,* TRB, National Research Council, Washington, D.C., 2000.
9. Head, K. L., Mirchandani, P. B., and Sheppard, D., "Hierarchical Framework for Real Time Traffic Control," in *Transportation Research Record 1360,* TRB, National Research Council, Washington, D.C., 1992.
10. Institute of Transportation Engineers, *Intelligent Transportation Systems Primer*, Washington, D.C., 2001.
11. Lucas, D. E., Mirchandani, P. B., and Head, K. L., "Remote Simulation to Evaluate Real Time Traffic Control Strategies," in *Transportation Research Record 1727,* TRB, National Research Council, Washington, D.C., 2000.
12. Michalopoulos, P. G., Jacobson, R. D., Anderson, C. A., and Barbaresso, J. C., "Field Deployment of Machine Vision in the Oakland County ATMS/ATIS Project," *Proceedings,* IVHS America 1994 Annual Meeting, Atlanta, GA, April 1994, pp. 335–342.
13. Mitretek Systems, Inc., *ITS Benefits: 1999 Update*, Report No. FHWA-OP-99-012, Federal Highway Administration, U.S. Department of Transportation, Washington, D.C., 1999.
14. Neudorff, L. G., Randall, J. E., Reiss, R., and Gordon, R., *Freeway Management and Operations Handbook*, Report No. FHWA-OP-04-003, Federal Highway Administration, U.S. Department of Transportation, Washington, D.C., 2003.
15. Payne, H. J., and Tignor, S. C., "Freeway Incident Detection Algorithms Based on Decision Trees with States," in *Transportation Research Record 682*, TRB, National Research Council, Washington, D.C., 1978, pp. 30–37.
16. Persaud, B., and Hall, F. L., "Catastrophe Theory and Pattern in 30-Second Freeway Traffic Data—Implication for Incident Detection," *Transportation Research A*, Vol. 23(2), pp. 103–113, 1989.
17. Smith, Brian L., Pack, Michael L., Lovell, David J., and Sermons, M. William, "Transportation Management Applications of Anonymous Mobile Call Sampling," *Proceedings of the 11th Annual Meeting of ITS America*, Miami, FL, 2001.
18. U.S. Department of Transportation, Federal Highway Administration, *The National Intelligent Transportation Systems Architecture, Version 5.1*, 2005. Available from http://www.iteris.com/itsarch.
19. U.S. Department of Transportation, Federal Highway Administration, *Intelligent Transportation Systems, Compendium of Field Operational Test—Executive Summaries*, Washington, D.C., 1998.
20. U.S. Department of Transportation, Federal Highway Administration, *Developing Traveler Information Systems Using the National ITS Architecture*, Washington, D.C., 1998.
21. U.S. Department of Transportation, *Advanced Public Transportation Systems: The State of the Art—1998 Update*, Report No. FTA-MA-26-7007-98-1, Federal Transit Administration, Washington, D.C., 1998.
22. U.S. Department of Transportation, Federal Highway Administration, *The National Intelligent Transportation Systems Program Plan*, Washington, D.C., 1995.