CHAPTER

10

Capacity, mobility, and safety

10.1 Introduction

Capacity, mobility, and safety are the three most important transportation system performance measures. Capacity is the maximum traffic flow rate in vehicles per hour for a highway facility considering the prevailing traffic state, roadway conditions, and driver population. Mobility can be monitored by the average vehicle operating speed of the highway facility. Lastly, safety is directly measured by crash frequency and injury severity. Drivers are less affected by other road users when traffic is light, so they feel less pressured to change lanes or slow down. However, drivers' actions have more impacts on others around them when traffic is dense. Unexpected changes in speed or positioning, for example, may force other drivers to take evasive actions to avoid a collision.

The various degrees of freedom and comfort experienced by drivers are qualitatively measured on a scale of A through F, which is known as the level of service (LOS) in the highway capacity manual (HCM) (HCM, 2010). Drivers in LOS A or B have more freedom to choose their speed and maneuver as they please. As the roadway becomes more congested, comfort and convenience deteriorate. The freedom to change lanes or choose speeds becomes noticeably restricted at LOS C, and then even more at LOS D and E. Capacity, mobility, and safety result from interactions between drivers in the traffic stream, and are therefore interdependent and interrelated.

Excessive driver interactions may lead to increased traffic conflicts. The ability to identify effective conflict resolution—which warrants the smooth movement of traffic—relies on knowledge and understanding of the sources of conflict. The transformation of undisturbed passages into disruptive traffic conditions manifests through the change of traffic state variables such as speed, flow, and density. The study of traffic conditions

that precede a crash has strong merit since intuitively, those conditions are more relevant than aggregated traffic measures such as annual average daily traffic (AADT). A crash prediction model based on temporally and spatially proximal measurements (e.g., 100 m upstream within the most recent 5 min) can substantially complement existing crash count models. Moreover, collecting, archiving, and processing traffic data in real-time presents opportunities for developing real-time and proactive safety management strategies.

This chapter offers a perceptive account of one of the fastest-developing fields in highway safety analysis, involving traffic flow theory, driver behavior models, and statistical methods. The chapter begins with a theoretical car-following model to demonstrate the safety aspects of a classic driver behavior model. Then, it introduces the modeling of relationships between crashes and traffic volume and the mapping of crash typologies to a variety of traffic regimes characterized by traffic variables. Next, it presents the use of Bayesian theory to predict crash probability and logistic regression to develop real-time crash prediction models (RTCPM) given a real-time traffic input. Finally, it describes the motivation and methodology for developing RTCPM from simulated traffic data in the event that actual traffic data are not available.

10.2 Modeling space between vehicles

Forbes and Pipes were the first to study theories that describe how one vehicle follows another. When modeling car-following behavior, Forbes presented a model that required a minimum driver perception reaction time, or the time it takes the driver in the following vehicle to perceive the need to brake (Forbes et al., 1958). Pipes used the California motor vehicle code to characterize the motion of vehicles in the traffic stream as "*following another vehicle at a safe distance is to allow yourself at least the length of vehicle between your vehicle and the vehicle ahead for every ten miles* per *hour of the speed at which you are traveling*" (Pipes, 1953). In both models, the precaution taken for safe driving is explicitly measured either by the distance of the following vehicle or the time it takes the driver of the following vehicle to respond to the lead vehicle's change(s). Hence, the decision of the driver in the following vehicle is one of the key indicators of driver behavior.

Subsequent models from General Motors (GM) quantified the driver's response (i.e., the acceleration or deceleration of the following vehicle) as the function of sensitivity and stimuli (Gazis et al., 1961). Specifically, sensitivity is represented by variables such as the relative distance

between the lead and the following vehicle. The operating speed of the following vehicle and stimuli are represented by the relative speed of the lead and the following vehicle. All car-following vehicles are based on the principle of safe operating vehicles.

The space between vehicles is of particular importance in car-following models. The driver must maintain a minimum distance to avoid colliding with another vehicle. Drivers intend to operate at a lower speed when vehicles are close to one another, so any gains in capacity due to increased traffic density may be offset by the losses of capacity due to decreased speed. These phenomena are reflected in changes of capacity and traveling speed on the roadway. A simple car-following model is illustrated in Fig. 10.1, and its equation is expressed in Eq. (10.1).

$$x_l = \frac{v^2}{2d_l} \tag{10.1a}$$

$$x_f = vt + \frac{v^2}{2d_f} \tag{10.1b}$$

$$s = x_f + L + x_0 - x_l = vt + \frac{v^2}{2d_f} + L + x_0 - \frac{v^2}{2d_l} \tag{10.1c}$$

where x_l is the distance traveled by the lead vehicle with a deceleration rate of d_l; x_f is the distance traveled by the following vehicle, including the stopping distance with a deceleration rate of d_f and the distance traveled during the perception and reaction time; t is the perception reaction time; v is the vehicle velocity; L is the vehicle length; x_0 is the minimum safe space between vehicles; and s is the safe space maintained between the front bumper of the lead vehicle and the front bumper of the following vehicle.

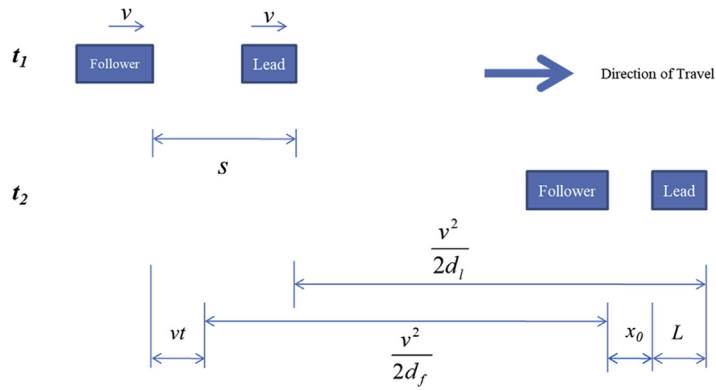

FIGURE 10.1 Car-following model.

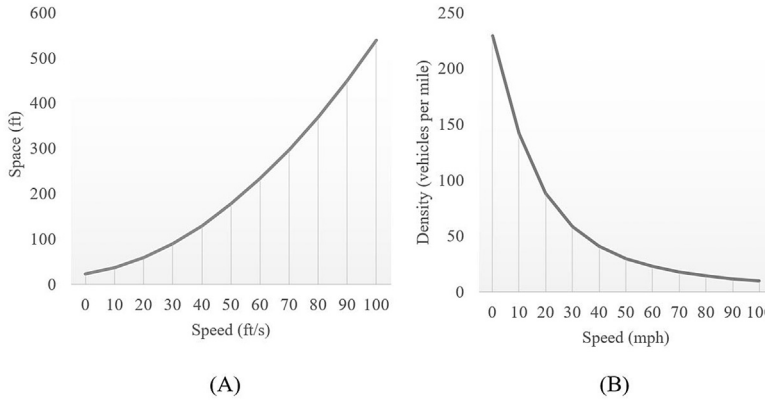

FIGURE 10.2 Traffic stream models. (A) Space versus speed. (B) Density versus speed.

A safe level of operation is defined as the space between vehicles that allows the following vehicle to apply normal deceleration and stop safely regardless of the deceleration rate of the lead vehicle. The relationship between speed and vehicle space is plotted in Fig. 10.2A assuming the following: $L = 20$ ft; $x_0 = 3$ ft; $t = 1$ s; $d_f = 8$ ft/s^2; $d_l = 24$ ft/s^2. On a more macroscopic scale, average vehicle speed and traffic density, or inversed average space is plotted in Fig. 10.2B.

The car-following models by Pipes, Forbes, GM, and others can be converted into traffic flow models pertaining to the fundamental relationships between flow speed, flow rate, and traffic density. The diagram that depicts a relationship between any two of the three macroscopic flow characteristics is called the fundamental diagram (FD). Each point along the FD curve represents a traffic state, some of which are safer than others. Exploring crashes as a function of traffic flow helps to identify what characteristics result in a safe traveling environment. The next section describes the methodologies for identifying the traffic flow patterns attributable to crashes.

10.3 Safety as a function of traffic flow

Drivers are human beings, and thus tend to make mistakes or engage in conflicts unintentionally when traffic patterns are changing. Variations in traffic flow are strong indicators of the safety status of a traveling environment. One of the common approaches to exploring the relationship between crashes and corresponding traffic flow characteristics is to model crashes as a function of prevailing traffic parameters. The safety impact of changes in flow parameters such as speed, flow rate, and density can be identified by studying the shape of a crash functional form.

An abundance of safety literature has investigated the relationship between crashes and highway traffic volumes. Volume is the most important factor in measuring traffic exposure, either in an aggregated fashion such as AADT or a disaggregated fashion such as hourly volume (Gwynn, 1967; Cedar and Livneh, 1982; Persaud and Dzbik, 1993; Mensah and Hauer, 1998; Qin et al., 2004, 2005, 2006). The HSM (AASHTO, 2010) shows how AADT can help predict the average number of crashes per year at a location using a mathematical equation called the safety performance function (SPF). Other traffic flow characteristics such as volume/capacity (v/c) ratios (Frantzeskakis and Iordanis, 1987; Zhou and Sisiopiku, 1997), vehicle density or occupancy (Brodsky and Hakkert, 1983), speed distribution (Garber and Gadiraju, 1989), level of service (Xu et al., 2014), and combination of density and v/c ratio (Lord et al., 2005) have also been studied intently.

The extensive examination of traffic flow characteristics has led to the concept of the level of safety service (LOSS). Kononov and Allery (2003) defined four LOSS levels as follows:

- LOSS-I: low potential for crash reduction;
- LOSS-II: better than expected safety performance;
- LOSS-III: less than expected safety performance; and
- LOSS-IV: high potential for crash reduction.

Kononov et al. (2011) reported a nonlinear relationship between crashes per mile per year and AADT and identified five critical points along the SPF:

- A is crash rate $\leq$ 0.64 crashes/MVMT;
- B is crash rate between 0.64 and 0.85 crashes/MVMT;
- C is crash rate between 0.85 and 1.32 crashes/MVMT;
- D is crash rate between 1.32 and 1.56 crashes/MVMT; and
- E is crash rate $\geq$ 1.4 crashes/MVMT.

Here, MVMT stands for million vehicle miles traveled. The authors further called traffic density of 24 pc/mi/ln a critical density, beyond which crashes increase at a faster pace. The portion of the SPF to the left of B is called the subcritical zone; the portion between B and D is called the transitional zone; and the portion beyond D is called the supercritical zone. Traffic density and speed are calculated for each zone boundary so that critical density (24 pc/mi/ln) and supercritical density (45 pc/mi/ln[1]) can be determined. The varying trend of crashes versus AADT suggests that crashes increase moderately when traffic is at a low density, while they accelerate when the critical density is exceeded. Both LOSS

[1] 45 pc/mi/ln is the maximum traffic density at the capacity for a freeway segment (HCM, 2010, TRB).

I–IV and critical points A–E illustrate the use of SPF to classify and measure safety performance as a function of traffic flow.

Harwood et al. (2013) also developed SPFs for describing a safety–congestion relationship. The relationships between crashes per MVMT and traffic density exhibit a U-shape. As traffic density is used as the LOS measure for capacity (HCM, 2010), it is convenient to associate safety benefits with highway capacity improvement through the change of traffic density. The equations for crashes by injury severity level are presented in Eqs. (10.2)–(10.4).

Total number of crashes:

$$Total\ per\ MVMT = 2.636 - 0.2143 \times D + 0.00708 \times D^2 - 4.80 \times 10^{-5} \times D^3 \tag{10.2}$$

Fatal and injury crashes:

$$FI\ per\ MVMT = 1.022 - 0.0842 \times D + 0.00264 \times D^2 - 1.79 \times 10^{-5} \times D^3 \tag{10.3}$$

Property damage only crashes:

$$PDO\ per\ MVMT = 1.614 - 0.1301 \times D + 0.00444 \times D^2 - 3.01 \times 10^{-5} \times D^3 \tag{10.4}$$

Where D is the traffic density. In summary, traffic flow characteristics such as traffic volume and traffic density have a direct impact on the likelihood of crashes. The safety performance of a highway facility can be measured by SPFs for different levels of AADT or other traffic flow characteristics. In this way, mobility performance can be effectively associated with safety performance when planning and designing a highway. However, AADT, hourly flow rate, vehicle density, or other flow characteristics are either average values over a long period of time or estimates from other sources, and do not necessarily reflect the exact traffic conditions at the time of the crash. This limitation has led to the use of real-time traffic data to study crashes.

10.4 Characterizing crashes by real-time traffic

Real-time crash studies use traffic data from immediately before a crash to help characterize the instantaneous traffic conditions under which crashes are more likely to occur. The primary interests of these studies are to detect disruptive traffic conditions and measure their impacts on safety. The results are instrumental in developing real-time and active traffic operational strategies. This more granular view of traffic is expected to disclose more tangible information regarding crash causes.

Traffic involves many individual travelers with different decision-making processes and varying driving behaviors. Exploratory data

analyses, which are detailed in Chapter 5—*Exploratory Analyses of Safety Data*, are regularly performed as a preliminary step to screen traffic flow patterns that have a strong association with crashes. Changes in flow rate, speed, traffic density, or combinations of them can be measured by small time steps such as 1-min intervals. Table 10.1 shows a sample of 1-min traffic loop detector data, where each record is timestamped by Date and Time.

Each crash record has its date and time and location description or GPS coordinates like latitude and longitude. With a known location and time, crashes can be associated with traffic sensors located in their proximity, making it possible to link crash data with real-time traffic data. Crash-related traffic patterns are summarized by statistics such as the mean, median, variance, difference in percentile values, as well as the trends within extended intervals (i.e., 5 min). Measurements may be taken from different sensor locations or by travel lane, depending on the data availability. Monitoring two or more consecutive sensors in the same travel direction at the same time allows analysts to classify the traffic state as congested, free-flow, traffic bottleneck, or back-of-the-queue. Lane-by-lane measurements may be necessary if a large variation exists between lanes or if frequent lane changes are observed. Table 10.2 lists representative traffic data measurements.

The cluster analysis is a useful technique for creating a subpopulation with more homogenous characteristics. Though they vary by assumptions and methodologies, all clustering algorithms intend to minimize the within group variance and maximize the between group variance. The popular cluster analysis methods include partitioning algorithms (e.g., *K*-mean), hierarchy algorithms (e.g., regression tree), correlation clustering (e.g., nonlinear canonical correlation analysis), and density-based algorithms (e.g., kernel principal component analysis).

The *K*-means cluster analysis is probably the most well-known clustering technique. It is a two-step procedure: determining optimal clusters

TABLE 10.1 Sample of 1-min loop detector data.

Region	Detector ID	Date	Time	Volume (pcpl)	Speed (mph)	Occupancy (%)
SE	5513	4/16/2014	9:00	20	67.17	5.67
SE	5513	4/16/2014	9:01	20	62.85	7.17
SE	5513	4/16/2014	9:02	14	65	4.33
SE	5513	4/16/2014	9:03	14	68.9	3.83
SE	5513	4/16/2014	9:04	15	67.67	5.83

TABLE 10.2 Input traffic flow variables for statistical measures.

Performance measures	Variable
Central tendency	Mean volume/speed/occupancy
	Median volume/speed/occupancy
Variation	Standard deviation of volume/speed/occupancy
	Difference between 90th and 50th percentiles of volume/speed/occupancy
Temporal trend	Time series measures between data points for volume/speed/occupancy
Spatial consistency	Speed and density differentials measured at the same time.

(the value of K) and then creating K subgroups of the data. Several K-means algorithms are available,[2] but the Hartigan-Wong algorithm is the standard one. The Hartigan−Wong algorithm defines the total within-cluster variation as the sum of squared Euclidean distances between data points and the corresponding centroid; it then searches for the K-value in such a way that the total within-cluster variation is minimized (Hartigan and Wong, 1979). The Hartigan-Wong algorithm is expressed as Eq. (10.5).

$$Minimize\left(\sum_{k=1}^{K} W(C_k)\right) = \sum_{k=1}^{K}\sum_{x_i \in C_k} (x_i - u_k)^2 \tag{10.5}$$

where C_k is the kth cluster and $W(C_k)$ is the within-cluster variation. x_i is ith observation and u_k is the centroid of kth cluster.

As the within-cluster variation always decreases as K increases, the practical algorithm seeks an optimal value where the marginal reduction of within-cluster variation becomes insignificant. The most extreme situation happens when each data point is its own cluster. The elbow method is a visualization method for finding the optimal K that is visible from the plot of the within-cluster variation of different K values.

After the value of K is determined, the K-means algorithm randomly selects K points from the data set to serve as the initial centroids for the clusters. Next, each of the remaining points is assigned to its closest

[2] There are other methods for determining an optimal number of clusters including the elbow method, silhouette method, and gap statistics. Details are referred to https://uc-r.github.io/kmeans_clustering#optimal.

centroid based on the Euclidean distance between the point and the centroid in what is called the "cluster assignment step." After the cluster assignment step, the algorithm computes the new mean value of each cluster in a step called the "centroid update." Then, all points are reassigned using the updated cluster means. The cluster assignment and centroid update steps are iteratively repeated until the cluster assignments converge. The *K*-means algorithm can be summarized as follows (Algorithm 10.1):

Algorithm 10.1

K-means algorithm

1. Specify the number of clusters (*K*).
2. Randomly select *k* observations from the data set as the initial "means."
3. Assign each observation to its closest mean based on the Euclidean distance.
4. Update the cluster centroid for each of the *k* clusters by calculating the new mean of all observations in the cluster.
5. Iteratively minimize the total within the sum of square (Eq. 10.5) by repeating steps 3 and 4 until either the cluster assignments stop changing or the maximum number of iterations is reached. Note that the R software uses 10 as the default value for the maximum number of iterations.

Exercise 10.1

Use the Inductive Loop Detector Dataset and the cluster Analysis method to characterize the high crash risk traffic conditions.

Inductive loop detectors collected 2163 cases on a freeway corridor between 2012 and 2014; 113 cases involved crashes, while the other 2050 were noncrash cases. Each case contains five consecutive 1-min traffic data. Each crash site is linked to two detectors, one located upstream of the crash and the other located downstream. Traffic data in 1-min intervals were collected from both detectors for 5 min before a crash.

continued

Exercise 10.1 (*cont'd*)

Noncrash cases were randomly selected from any 5-min records that are not within 2 hours of a crash. Four speed-related parameters are used to characterize the travel safety level: average 1-min speed within 5-min interval at upstream detector (AvgSpd_U); average 1-min speed within 5-min interval at downstream detector (AvgSpd_D); standard deviation of 1-min speed within 5-min interval at upstream detector (StdSpd_U); and standard deviation of 1-min speed within 5-min interval at downstream detector (StdSpd_U).

We can use the *K*-means cluster analysis method to: (1) create clusters of crash cases as the combinations of speed parameters, (2) describe the clusters, (3) compare speed characteristics near crash location, and (4) identify the traffic conditions with high crash risk. The following steps are taken to solve the problem.

First, the Elbow Method in R with the R function is applied (fviz_nbclust). Fig. 10.3 is a scree plot that shows the reduction in the within-groups sum of squares becomes marginal after the number of clusters increases to five. Thus, the optimal number of clusters is five.

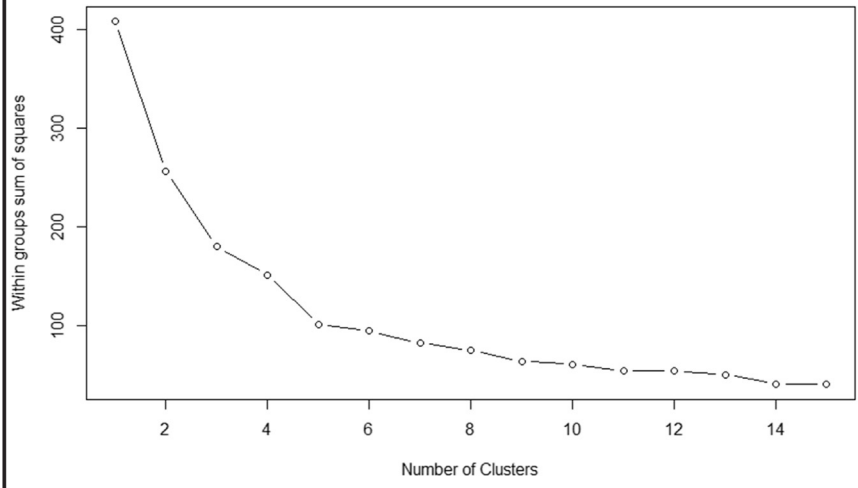

FIGURE 10.3 The Elbow method.

Exercise 10.1 (*cont'd*)

Next, the R "*k*means"[3] function is used to determine five traffic regimes characterized by the four speed-related parameters. Fig. 10.4 illustrates the five traffic regimes with a radar chart using the R function (radar chart).[4]

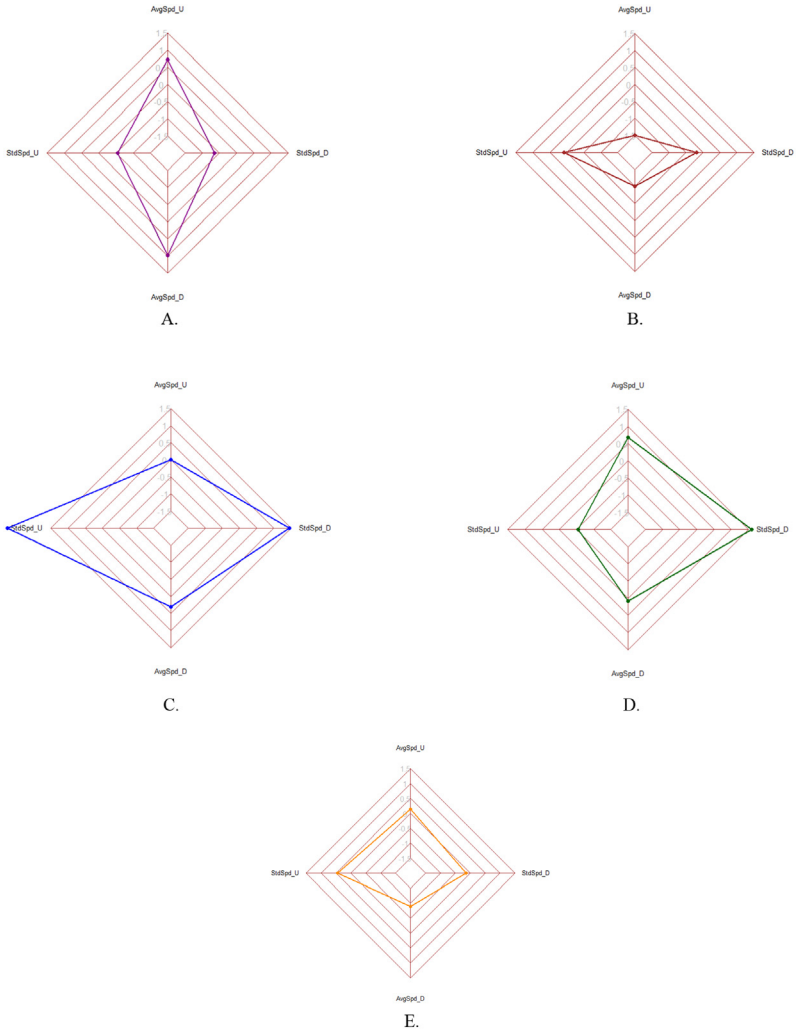

FIGURE 10.4 Traffic regimes characterized by speed data. (A) High speed and low speed variation at both upstream and downstream. (B) Low speed and low speed variation at both upstream and downstream. (C) Average speed and high speed variation at both upstream and downstream. (D) Average speed with low speed variation at upstream and average speed with high speed variation at downstream. (E) Average speed with moderately speed variation at upstream and low speed and speed variation at downstream.

continued

Exercise 10.1 (*cont'd*)

Furthermore, comparing the traffic speed and speed variation from the upstream and downstream detectors on a freeway section allows us to generalize these conditions as one of five combinations of traffic states:

- Free flow (F−F): high speed with low speed variation at both upstream and downstream locations;
- Congested (C−C): low speed and low speed variation at both upstream and downstream locations;
- Transitional (T−T): average speed with high speed variation at both upstream and downstream locations;
- Free flow to transitional (F−T): average speed with low speed variation at the upstream location, and average speed with high speed variation at the downstream location, and
- Transitional to congested (T−C): average speed with moderately high speed variation at the upstream location and low speed with low speed variation at the downstream location.

A traffic state group that has homogenous flow characteristics is also called a traffic regime. The five traffic regimes represent various risk levels associated with the turbulence of vehicle speed. A free flow to transitional condition implies a possible backward forming shockwave that originates from the downstream location, while a transitional to congested state suggests a rapid-moving, backward-forming shockwave. In either situation, drivers encounter slowing or stopping traffic ahead, which poses an elevated risk for rear-end or other crash types.

[3] *k*means, https://www.rdocumentation.org/packages/stats/versions/3.5.3/topics/kmeans.
[4] radarchart, https://www.rdocumentation.org/packages/fmsb/versions/0.6.3/topics/radarchart.

10.5 Predicting imminent crash likelihood

Characterizing the level of safety based on traffic flow variables is instrumental in comparing and identifying traffic circumstances that are more prone to crashes. Crash propensity can be calculated for these circumstances based on their appearance and duration in the traffic stream. Such information is valuable to agencies tasked with assessing transportation system safety performance, evaluating highway design and operational alternatives, and identifying and mitigating crash-prone traffic conditions.

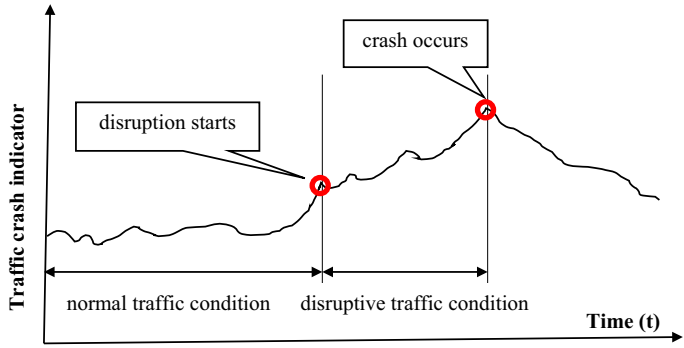

FIGURE 10.5 Traffic dynamics and crashes.

In a simple way, traffic conditions can be categorized as normal or disruptive. Normal traffic conditions are stable and highly desirable, such as LOS A to C. Disruptive conditions are unstable and undesirable, which are represented by high temporal and spatial variations in traffic parameters and often observed in LOS D or worse situations. Fig. 10.5 shows that traffic dynamics become more and more unstable over time and eventually result in a crash. High crash risk is often associated with disruptive conditions in the traffic stream. If such a pattern is recognizable and identifiable, we can forecast crashes by monitoring disruptive features in traffic.

The Bayesian method can help predict future crash likelihood through the observation of the current traffic state and the historical traffic profile. According to the Bayes' theorem, the probability of event 1 occurring given event 2 is a function of the probability of observing event 2 given event 1 and the probability of 1 and 2 occurring independently of each other. If we assume event 1 to be a crash C and event 2 to be an observable traffic event X (e.g., speed variation within 5 min), the conditional probability of C given the speed variation measurement X can be formulated in Eq. (10.6).

$$P(C|X) = \frac{P(X|C)P(C)}{P(X|C)P(C) + P(X|N)P(N)} = \frac{P(C) \times f_{crash}(X)}{P(C) \times f_{crash}(X) + P(N) \times f_{non-crash}(X)} \tag{10.6}$$

where $P(C|X)$ is the posterior probability that the given traffic measurement (e.g., speed variation) would lead to a crash; $P(C)$ is the prior probability that the given traffic measurement associated with crash cases; $P(N) = 1 - P(C)$ is the prior probability that the given traffic measurement associated with noncrash cases; $f_{crash}(X)$ is the probability

density function of the traffic measurement leading to a crash; and, $f_{non-crash}(X)$is the probability density function of the traffic measurement not related to a crash.

In a study by Oh et al. (2005), crash data and corresponding traffic data were collected from February 16 to March 19, 1993 within a 15.3 km-long freeway section on Interstate 880 (I-880) in Hayward, California. The authors defined the normal traffic conditions as a 5-min period 30 min before a crash, and disruptive traffic conditions as a 5-min period preceding a crash. Flow, density and speed data were collected from upstream detector stations every 10 s during each 5-min period. 52 crashes were matched with real-time traffic data, and 4787 periods of traffic conditions were crash-free.

Next, the probability density function of standard deviation of speed was constructed for both normal and disruptive traffic conditions, respectively. The prior probability of a crash $P(C)$ and posterior distribution of crashes given the standard deviation of speed $P(C \mid X)$ were also calculated.

The probability density function was estimated with the nonparametric kernel smoothing technique: the distribution of the 52-sample 5-min speed variation data with crashes $(f_{crash}(X))$ and the distribution of the 4787-sample 5-min speed variation data that are crash-free $(f_{non-crash}(X))$. The prior probability $(P(C))$ that given speed variation belongs to crash cases can be approximated as: $P(C) = \frac{\text{number of 5-minute intervals of crash cases}}{\text{total number of 5-minute intervals}}$. Then, the probability of crash occurrence $P(C|X)$ calculated with Eq. (10.6) within the next 5-min with respect to the standard deviation of speed.

10.6 Real-time predictive analysis of crashes

The previous section presents a convenient means of forecasting crash occurrence based on the observation of a single traffic parameter (i.e., speed variation). Though methodologically sound, determining which traffic variable is the most reliable and accurate crash indicator may not be practical. Additionally, it can be difficult to construct the probability density function for normal and disruptive traffic conditions. Furthermore, constructing the probability of given traffic measurements for crash cases can be challenging. Therefore, new studies have proposed the use of statistical models to construct meaningful relationships between historical crash data and traffic flow variables, quantify their impact on crash occurrence, and predict future crash occurrence when new traffic observations are available. Roshandel et al. (2015) conducted a systematic literature review to identify current knowledge, challenges, and gaps in real-time traffic impact on crash occurrence. The underpinning assumption of real-time predictive crash analysis is that a certain combination of

traffic conditions is likely to lead to a crash. Lee et al. (2002) referred to these traffic flow characteristics as "crash precursors."

Future predictions depend on what and how to observe: when should the observation begin, and how long will it last? Analysts must consider the lead time and the length of the observation during which traffic conditions are collected for modeling. Several studies have proposed different time intervals (Abdel-Aty et al., 2004; Pande et al., 2005; Xu et al., 2012; Chen et al., 2018). After comparing different 5-min slices (i.e., 0–5-min, 5–10-min, …, 25–30-min) and 3-min slices (i.e., 0–3-min, 3–6-min, …, 12–15-min), it was found that the 5–10-min slice is the most appropriate in terms of model performance and practicality. The 5–10-min interval before the crash occurrence has been the predominant choice in real-time crash prediction.

The next step is to choose an appropriate modeling technique, which depends on whether a study uses unmatched or matched cases for the purpose of controlling confounding factors. Controlling the impact of confounding variables ensures that the effects of traffic parameters on crash occurrence are genuine and accurate. In real-time crash prediction, a "case" refers to the traffic conditions before a crash, and a "control" represents noncrash traffic conditions. In an unmatched study design, noncrash cases are randomly selected from a large population. The unmatched design does not require the rigorous control of confounding factors, but it requires a sufficiently large sample size to ensure accurate estimation, especially when the number of variables is high. A matched case-control study requires more effort, however, so that noncrash cases similar to crash cases can be identified and confounding variables can be controlled.

Although the criteria for selecting control cases and their sizes impact the results of a case-control study, no specific guidelines are in place. For example, defining which noncrash cases are "similar" to crash cases relies on the analysts' individual interpretation. Also, the recommended case-to-control ratio is 1:4, but ratios in previous studies have ranged from 1:4 to 1:1000 because some empirical studies have shown the marginal gain in prediction beyond this ratio. The issues of defining a match and the resultant sample size are connected, as more rigorous criteria lead to fewer available control cases, and vice versa.

The binary logistic regression model and conditional logistic regression model are both salient methods for unmatched and matched RTCPM studies. The models will be introduced in the following sections.

10.6.1 Binary logistic regression model

The binary logistic regression model is a good choice for a dichotomous (binary) dependent variable such as crash versus noncrash. Eq. (10.7)

shows how the probability of a crash case can be estimated in a binary logistic regression model.

$$\text{logit}\, p_n = \log\left(\frac{p_n}{1-p_n}\right) = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \ldots + \beta_k x_{n,k} \tag{10.7}$$

where p_n represents the crash probability given $\mathbf{x}_n$ $(n = 1, N)$; $\mathbf{x}_n = (x_{n,1}, x_{n,2}, \ldots, x_{n,k})$ is a set of k explanatory variables for sample n; βs are the coefficients for $(x_{n,1}, x_{n,2}, \ldots, x_{n,k})$; k is the number of parameters; and N is the number of cases.

Eq. (10.8) shows how the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$ can be estimated by maximizing the log-likelihood function:

$$lnL(\boldsymbol{\beta}, \boldsymbol{x_n}) = \sum_{n=1}^{N}\left[(\beta_0 + \beta_1 x_{n,1} + \ldots + \beta_k x_{n,k}) - \ln(1 + e^{-(\beta_0+\beta_1 x_{n,1}+\ldots+\beta_k x_{n,k})}\right] \tag{10.8}$$

The estimated coefficients are usually presented as the log odds ratio between the probability of a crash case and a noncrash case given the change of an independent variable (Chapter 5—Exploratory Analysis of Safety Data provides details about the odds ratio). A one-unit change occurs if the independent variable is continuous, and a 0 to 1 change occurs if the independent variable is binary. The odds ratio can be estimated by taking the exponential of the estimated coefficient.

Exercise 10.2

Develop a binary logit model to estimate the safety impact of traffic variables using the same dataset as in Exercise 10.1.

Several traffic variables are assumed to have an effect on crash occurrence in this exercise. All traffic variables are collected from traffic detectors located immediately downstream of the crash location. The sample includes 103 crash cases and 2050 noncrash cases. Modeling results are shown in Table 10.3.

Four traffic variables are found to be statistically significant to crash occurrence at a 5% level of significance: average volume, volume variation (std. deviation of volume), average speed, and speed variation (std. deviation of speed). The exponential of coefficient estimates is explained as the odds ratios of a crash. With other variables being held constant, if average traffic density increases by one unit, the odds ratio of a crash is 0.943 (or $e^{-0.05875}$). In other words, the predicted odds of a crash are 5.7% lower than they are before the one-unit increase. However, if std. deviation of traffic density increases one unit when everything else being held constant, the odds ratio of a crash is 1.068 (or 6.8% higher).

Table 10.3 indicates that an increase in downstream density may decrease the likelihood of a crash, while an increase in other variables

Exercise 10.2 (*cont'd*)

may increase the likelihood of a crash. It is plausible that a higher traffic volume would lead to more interactions between vehicles, therefore increasing crash risk. A higher volume variation or speed variation indicates an unstable traffic state in which drivers are frequently required to accelerate and decelerate and are therefore more prone to crashes.

TABLE 10.3 Modeling results.

Variable	Estimate	Std. error	*P* value	Odds ratio
Intercept	−1.159	0.533	0.030	–
Average volume	0.000179	0.000090	0.047	1.000
std. dev. volume	0.001182	0.000497	0.017	1.001
Average density	−0.058750	0.006554	<0.001	0.943
std. dev. density	0.066090	0.018720	<0.001	1.068

10.6.2 Conditional logistic regression model

A matched case-control study focuses on variables of interest by controlling for nuisance factors. Abdel-Aty et al. (2004) defined a "case" as the representative traffic conditions occurring right before a crash, while "control" represents the noncrash traffic conditions. Each crash case involves several noncrash events that are selected as controllers such that the nontraffic-flow variables (e.g., location, time, season) of noncrash cases match the corresponding variables of crash cases. Each case and its controllers constitute a stratum. The controlled, nontraffic-flow variables are the same within each stratum, but are different across strata.

One example of a matched case-control design involves a crash that occurred at 11:01 a.m. on 01/17/2012. The design uses traffic measurements from 10:56 to 11:01 a.m. (0–5-min period before the crash) on the same day from the inductive loop detector stations immediately upstream and downstream of the crash. Traffic measurements are also collected from the same sites during the same time period in 2012 but on crash-free days. The matched case-control design is expected to increase the accuracy of variable estimates in real-time crash models by controlling the confounding bias while greatly reducing the required size of noncrash cases.

The crash probability $p_{nj} = \mathbf{p}(\mathbf{Y_{nj}} = \mathbf{1}|\mathbf{X}_{nj})$ in a matched case-control logistic regression can be expressed in Eq. (10.9).

$$\text{logit}(p_{nj}) = \propto_i + \sum_{k=1}^{K} \beta_k x_{njk} \tag{10.9}$$

where x_{njk} is the kth traffic flow variable for the case ($j = 0$) or the jth control in the nth stratum; $n = 1, 2, \ldots, N$ $j = 0, 1, \ldots, J$; and $k = 1, 2, \ldots, K$. N is the number of strata, J denotes the number of controls, and K represents the number of explanatory variables.

Each stratum has its own intercept parameter, meaning the number of parameters increases as more strata are added. If stratum-specific parameters $\propto_i$s are treated as nuisance parameters, a conditional likelihood of $\boldsymbol{\beta}$ could be created. The maximum likelihood estimators are expressed in Eq. (10.10) (Hosmer and Lemeshow, 2004).

$$l_n(\boldsymbol{\beta}) = \frac{\exp\left(\sum_{k=1}^{K} \beta_k x_{n0k}\right)}{\sum_{j=0}^{J} \exp\left(\sum_{k=1}^{K} \beta_k x_{njk}\right)} \tag{10.10}$$

The full conditional likelihood is the product of the $l_n(\boldsymbol{\beta})$ over N strata (Eq. 10.11).

$$l(\boldsymbol{\beta}) = \prod_{n=1}^{N} l_n(\boldsymbol{\beta}) \tag{10.11}$$

The full conditional likelihood is independent of stratum-specific parameters, $\propto_i$, and thus, $\propto_i$s cannot be estimated. So, the purpose of the conditional logistic model is not to calculate the crash probability of a specific case, but to estimate the effects of variables of interest on a crash case through the slope coefficients $\boldsymbol{\beta}$.

Consider two observation vectors $\boldsymbol{x_{i1}} = \{x_{i01}, x_{i02}, \cdots, x_{i0k}\}$ for a crash case and $\boldsymbol{x_{i2}} = \{x_{i11}, x_{i12}, \cdots, x_{i1k}\}$ for the first noncrash case from the ith strata on the k traffic flow variables. The log-odds ratio of crash occurrence specified in Eq. (10.12) can be expressed as:

$$\begin{aligned} log\left\{\frac{p(x_{i0})/[1-p(x_{i0})]}{p(x_{i1})/[1-p(x_{i1})]}\right\} &= \beta_1(x_{i01} - x_{i11}) + \beta_2(x_{i02} - x_{i12}) + \cdots \\ &+ \beta_k(x_{i0k} - x_{i1k}) \end{aligned} \tag{10.12}$$

If $\overline{x}_i = \{\overline{x}_{i1}, \overline{x}_{i2}, \cdots, \overline{x}_{ik}\}$ denote the vector of mean values of noncrash cases of the k variables within the ith stratum, then the approximate log odds ratio of crash to noncrash is expressed as follows:

$$\begin{aligned} log\left\{\frac{p(x_{i0})/[1-p(x_{i0})]}{p(\overline{x}_i)/[1-p(\overline{x}_i)]}\right\} &= \beta_1(x_{i01} - \overline{x}_{i1}) + \beta_2(x_{i02} - \overline{x}_{i2}) + \cdots \\ &+ \beta_k(x_{i0k} - \overline{x}_{ik}) \end{aligned} \tag{10.13}$$

The log-odds ratio in Eq. (10.13) can be used to predict crashes by establishing an optimal threshold that yields satisfactory crash classification accuracy.

Example 10.1

Develop a conditional logistic model to estimate the safety impact of traffic variables

In a study by Abdel-Aty et al. (2004), a total of 375 crashes and loop detector data were collected from April 1, 1999, to November 30, 1999, along a 13.25-mile stretch of Interstate four in Orlando, FL. Real-time traffic data—30 min before the crash—were extracted from the detectors immediately upstream and downstream of each crash. Traffic data were also extracted on all days that corresponded to the day of a crash. For example, if a crash happened at Station 40 on Tuesday, June 1, 1999 at 4:00 p.m., data were extracted for all Tuesdays of the same season from 3:30 to 4:00 p.m. Average $\overline{x}$, standard deviation s, and coefficient of variation CV = (s/$\overline{x}$,) for speed, volume and occupancy were calculated at each loop detector over the 5-min intervals. The nine main effects under investigation are AS, SS, CVS (average, standard deviation, and coefficient of variation, respectively) for speed; AV, SV, CVV for volume; and AO, SO, CVO for occupancy. In a matched crash−noncrash study, the nontraffic flow variables (i.e., location, time of day, and day of the week) associated with each crash are selected as matching factors.

Only two of the five variables are significantly based on the *P*-values: LOG (CVS) and AO. Table 10.4 shows the hazard ratios for the two variables for Station A.

The hazard ratio (i.e., $exp(\beta)$), as called odds ratio, is an estimate of the expected change in the risk ratio of having a crash versus noncrash per unit change in the corresponding factor. A value of β greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the factor increase, the odds of having a crash increase. For example, a hazard ratio of 2.5 corresponding to CVS means the risk for a crash increases 2.5 times for each unit increase in CVS.

TABLE 10.4 Model results for real-time freeway crash prediction[a].

Variable		AO		Log (CVS)	
Station	**Time slice**	$P_r>\chi^2$	**Hazard ratio**	$P_r>\chi^2$	**Hazard ratio**
A	1	0.008	1.021	0.007	1.862
A	2	0.008	1.021	0.024	1.682
A	3	0.005	1.022	0.015	1.740
A	4	0.039	1.016	0.017	1.728
A	5	0.083	1.014	0.069	1.520
A	6	0.034	1.017	0.186	1.355

[a]*The 30-min time period before a crash is divided into six 5-min intervals called Time Slices, with Slice 1 being the 5 min before the crash.*

10.6.3 A note about binary logit and conditional logistic regression models

Both matched and unmatched designs have been applied in the prediction of real-time crash risk, but their popularity varies depending on the study's purpose and the strengths and limitations of the methodologies. In an unmatched study design, noncrash cases are randomly selected. Confounding factors do not need to be matched in an unmatched design, so it requires less effort to identify noncrash cases. However, a sufficiently large sample size is required to ensure accurate estimation, especially when the variable number is high. The unmatched design allows for the estimation of both risk factors and confounding factors and it predicts the probability of a crash given the input variables. In the matched case-control study, the conditional logistic model does not provide the estimates of the nuisance factor, and thus it cannot directly predict the crash risk of given traffic conditions. Instead, the conditional logistic model approximates the relative crash risk through the log-odds ratio of crash occurrence. In other words, the unmatched case-control study handles the effects of confounding factors in the modeling stage by including them as variables in the regression equation; while the matched case control study tackles confounding factors in the data sampling stage. Another issue with matched case-control studies is that a randomly chosen "control" is very likely to share similar traits with crash-prone conditions because crash cases constitute only a very small fraction of all data.

Xu et al. (2016) compared the predictability of the two designs and found that given a predefined specificity (the proportion of crash cases that are correctly classified), the predictability of the RTCPM developed with unmatched data always outperformed that of the matched case-controlled data.

10.7 Using traffic simulation to predict crashes

Travel conditions can shift rapidly. Intuitively, the traffic that a driver experiences immediately before or at the time of a crash is more relevant than traffic occurring far earlier. However, the relationships between crashes and instantaneous real-time traffic flow variables may not be as useful as one would think, as the deployment and activation of crash preventive operations takes time. In practice, traffic data from the 5 to 10 minutes before a crash occurrence are normally considered as the input for predicting crashes. The lead time, or the time lapse between the current and future traffic status involving a crash, affects the practicality and accuracy of crash modeling and prediction.

Crash prediction accuracy also depends on the distance between the crash and the nearest traffic detectors, as the data from those detectors are assumed to match those of the crash site. The spacing of loop detectors can lead to a lack of consistency in the input variables, as spacing can vary substantially from location to location. The most desirable layout is that of inductive loop detectors, shown in Fig. 10.6, which are closely and uniformly spaced. However, in the real world, unevenly spaced or distant loop detectors are more prevalent.

The presence of spatial-temporal discrepancies among loop detectors will undermine the accuracy of crash prediction that depends on quality traffic flow data from loop detector stations. Simulated traffic data may be used as an alternative data source in the event that discrepancies become a concern. The macroscopic traffic flow model is a popular consideration because it formulates the relationship among traffic flow characteristics like flow, density, and speed of a traffic steam. The cell transmission model (CTM) is a popular macroscopic traffic flow model that can capture many important traffic flow behaviors (i.e., queue formation and dissipation and shockwave propagation) on a given corridor (Daganzo, 1994). In this section, CTM is first introduced as the tool to simulate spatial and temporal traffic during the time period just before a crash. Then, crash occurrence probability is estimated with simulated traffic conditions using a regression model such as binary logistic regression.

10.7.1 Cell transmission model

CTM is a macroscopic traffic simulation model that operates sufficiently with aggregated traffic data from detector stations. CTM is more computationally efficient and easier to configure and calibrate than microscopic simulation models. For a better representation of real-world situations, we consider a phenomenon called "capacity drop" that represents the discharge flow rate dropping below capacity after congestion

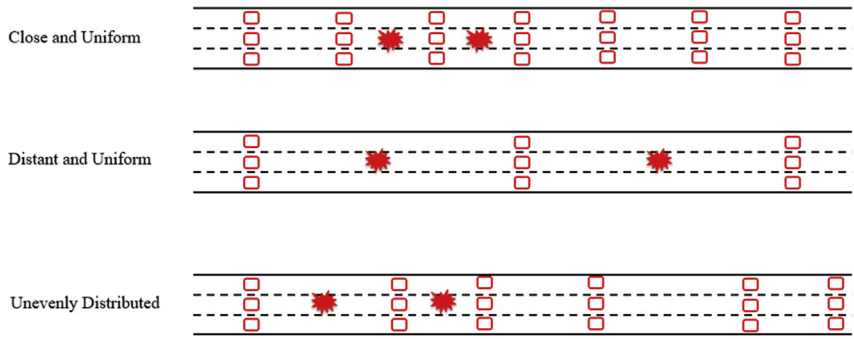

FIGURE 10.6 Spatial distribution of loop detectors.

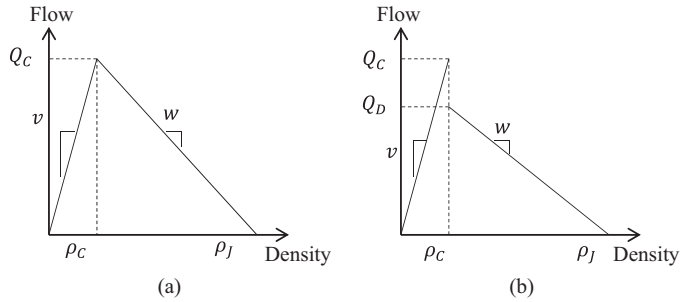

FIGURE 10.7 (A) Triangular fundamental diagram; (B) Fundamental diagram with capacity drop.

forms (Hall and Agyemang-Duah, 1991; Cassidy and Rudjanakanoknad, 2005). Fig. 10.7 shows the FD for developing a CTM with and without a capacity drop. Fig. 10.7A, where Q_C is the capacity flow, ρ_C is the critical density, ρ_J is the jam density, v is the free-flow speed, and w is the shockwave speed. Capacity drop is accounted for by adopting the FD in Fig. 10.7B where Q_D is added to the triangular FD.

A schematic diagram plotted in Fig. 10.8 illustrates the virtual loop detectors constructed for CTM along a highway corridor. If a crash occurs in Cell k, two virtual stations $k-1$ and k located upstream of the crash, and two virtual stations $k+1$ and $k+2$ located downstream can be used as candidate stations. Simulated traffic data are collected from these virtual detector stations to develop crash prediction models.

The density of each cell follows the conservation law of vehicles. The density for Cell i without on- or off-ramps is determined as follows:

$$\rho_i(k+1) = \rho_i(k) + \frac{T}{l_i}(q_i(k+1) - q_i(k)) \tag{10.14}$$

where k is the time step index; $\rho_i(k)$ is the density of Cell i during the kth time step; T is the length of the time step; l_i is the length of Cell i; and, $q_i(k)$ is the flow rate into Cell i during the kth time step.

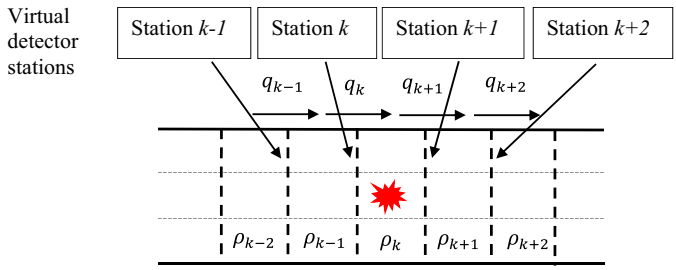

FIGURE 10.8 Layout of virtual loop detector stations.

The flow rate is determined by the sending and receiving functions. For Cell i, the sending function $S_i(k)$ represents the maximum flow that can be supplied during the kth time step, and the receiving function $R_i(k)$ represents the maximum flow that can be received. The two functions are determined in Eqs. (10.15) and (10.16), respectively, and the flow rate, $q_i(k)$ is determined by Eq. (10.17):

$$S_i(k) = \min\left(v_i\rho_i(k), Q_{C,i}\right) \tag{10.15}$$

$$R_i(k) = \min\left(Q_{C,i}, w_i\left(\rho_{J,i} - \rho_i(k)\right)\right) \tag{10.16}$$

$$q_i(k) = \min(S_{i-1}(k), R_i(k)) \tag{10.17}$$

The sending and receiving functions are modified when the capacity drops from Q_C to Q_D at the onset of congestion. The modification and expression of these processes can be seen in Eqs. (10.18) and (10.19) respectively:

$$S_i(k) = \begin{cases} v_i\rho_i(k), & \text{if } \rho_i(k) \le \rho_{C,i} \\ Q_{D,i}, & \text{if } \rho_i(k) > \rho_{C,i} \end{cases} \tag{10.18}$$

$$R_i(k) = \begin{cases} Q_{C,i}, & \text{if } \rho_i(k) \le \rho_{C,i} \\ w_i\left(\rho_{J,i} - \rho_i(k)\right), & \text{if } \rho_i(k) > \rho_{C,i} \end{cases} \tag{10.19}$$

10.7.2 Fundamental diagram calibration

An FD is required to operate the CTM simulation. Different roadway characteristics (e.g., horizontal curves, distances to on-/off-ramps, posted speed limits) make traffic patterns to vary between cells, leading to the calibration of different FDs. The calibration provides key traffic flow parameters: free-flow speed v (mph), critical density ρ_C (veh/mi), jam density ρ_J (veh/mi), capacity Q_C (veh/h), capacity drop Q_D (veh/h), and shockwave speed w (mph). The algorithm developed by Dervisoglu et al. (2009) is modified, and the full description is summarized in Algorithm 10.2.

A real-world case adapted from Chen and Qin (2019) is presented in Fig. 10.9. A freeway corridor has three lanes, one on-ramp, and one off-ramp. The corridor consists of segments S1, S2, and S3, which are 1.77-miles, 0.79-miles, and 1.59-miles long, respectively. Segment S2 starts at the end of the off-ramp and ends at the beginning of the on-ramp. The posted speed limit is 65 mph in S1, and 55 mph in S2 and S3. Lane width and shoulder width do not vary along the corridor. The corridor has seven mainline physical loop detector stations: N_1, N_2, …, N_7, which are unevenly spaced from less than a quarter mile to one mile. CTM is employed to instrument the freeway corridor with virtual detector

Algorithm 10.2

FD Calibration via the Dervisoglu et al.'s algorithm

1. Estimate the free-flow speed, v, using the least-squared method with flow-density pairs in the free-flow conditions. Treat data points with speeds exceeding the posted speed limit as observations in free-flow conditions.
2. Find the maximum measured flow rate, q_{max}, as the capacity, Q_C. Critical density is determined by $\rho_C = \frac{Q_C}{v}$. Few and unsustainable observations with extremely high flow rates, a phenomenon of capacity overestimation, should be ignored. The formula for computing the nominal capacity (in veh/h/lane) of freeways is adopted from the 2010 HCM, as opposed to using high flow rates (Transportation Research Board, 2010):

$$\text{Capacity} = \begin{cases} 2400\ veh/h/lane, & if\ FFS \geq 70mi/h \\ 2400 - 10 \times (70 - \text{FFS})\ veh/h/lane, & if\ FFS < 70mi/h \end{cases} \tag{10.20}$$

   The capacity is determined by taking the minimum of Q_c and the nominal capacity given by Eq. (10.20).

3. Estimate the shockwave speed, w, and the jam density, ρ_J, using the least-squared method, with flow-density pairs exceeding the critical density. The flow rate after the capacity drop is set as the value on the fitted flow-density line at the critical density.

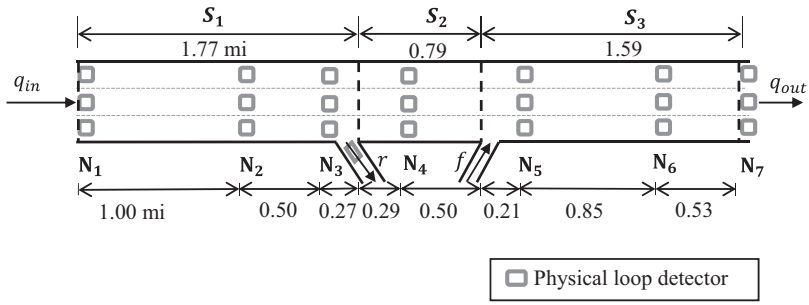

FIGURE 10.9 Layout of physical loop detector stations.

stations and measure traffic data where physical stations are not available. The highway segment is divided into 41 virtual cells, and the cell length is uniform within each of the three segments. Segment S_1 has 17 cells with a length of 0.104 mile; segment S_2 has 8 cells with a length of 0.098 mile; segment S_3 has 17 cells with a length of 0.099 mile. A virtual detector station is established at the boundaries of cells, so there are 42 virtual detector stations. The average spacing between consecutive virtual stations is 0.1 mile with a negligible variation. Virtual stations are expected to capture traffic conditions at locations closer to the crash site.

FD parameters are obtained for each physical detector station (see Table 10.5) following Algorithm 2. Note that ρ_C, ρ_J, Q_c, and Q_D are for three lanes. The magnitude of the capacity drop is from 2.0% to 6.9% for all physical stations except N_4, which has a 13.9% capacity drop rate. The set of fundamental diagram parameters calibrated for a physical station is assigned to cells near that station.

10.7.3 CTM simulation algorithm

The freeway corridor in Fig. 10.9 is used to demonstrate the CTM algorithm. The simulation time step in CTM needs to be determined to fulfill the Courant–Friedrichs–Lewy (CFL) condition (Courant et al., 1967). A vehicle cannot travel across more than 1 cell during one simulation step in the CFL condition (i.e., $v_i * \Delta t \leq l_i$ where v_i is the free-flow speed, Δt is the simulation time step, and l_i is the cell length). Thus, a 5-s time step was used ($\Delta t = 5s$) based on the lengths of cells.

The entering and exiting flows of the highway corridor are necessary to run the CTM. Four flow inputs are required for the study site: in-flow (q_{in}), out-flow (q_{out}), off-ramp flow (r), and on-ramp flow (f). The 1-min

TABLE 10.5 Fundamental diagram parameters by physical station.

Station	v (mi/h)	ρ_C (veh/mi)[a]	ρ_J (veh/mi)[a]	Q_C (veh/h)[a]	Q_D (veh/h)[a]	w (mi/h)
N_1	67.0	106.1	486.0	7111	6890	18.1
N_2	68.4	104.6	588.4	7152	6816	14.1
N_3	66.5	106.7	472.2	7095	6603	18.1
N_4	59.8	97.0	799.0	5796	4989	7.1
N_5	60.8	113.9	779.9	6924	6671	10.0
N_6	58.0	118.0	460.4	6839	6703	19.6
N_7	60.1	114.8	375.5	6903	6683	25.6

[a]*Parameters are for three lanes.*

flow data collected from the first and last physical stations(N_1 and N_7) in the 0–5 min period before a crash/noncrash are used for the in-flow and out-flow of the corridor. A linear interpolation method is applied to generate the 5-s in-flow, out-flow, on-ramp flow, and off-ramp flow data. A CTM then simulates how traffic in cells along the corridor evolves at each time step within the 5-min time interval.

In addition to the flow data, the initial densities of cells at the beginning of the simulation interval are necessary for the CTM simulation. The initial cell density is obtained from the station's density data as long as the cell has one loop detector station. Densities of cells between two such cells are interpolated using Algorithm 10.3.

10.7.4 Crash modeling

Continuing with the freeway corridor in Fig. 10.9 simulated traffic data are collected from one virtual station located upstream and one virtual station located downstream of the cell location of each crash/noncrash event in the 0–5 min period before the event. The time period of 0–5 min before an event is used to account for the temporal issue of physical station data, and the simulated traffic data for the future 5-min period

Algorithm 10.3

Interpolation algorithm for cell density

1. Compute the density change rate as the ratio of the difference in densities of 2 cells with two consecutive loop detector stations and the distance between them:

$$\nabla\rho = \frac{\rho_{d,0} - \rho_{u,0}}{x_d - x_u} \tag{10.21}$$

where $\nabla\rho$ is the density change rate; $\rho_{d,0}$ and $\rho_{u,0}$ are densities of cells having the downstream and upstream detector stations, respectively; x_d and x_u are the locations of the beginnings of the 2 cells, that is, the locations of the two detector stations.

2. Determine the initial density of 1 cell between those 2 cells by the following:

$$\rho_{i,0} = \rho_{u,0} + \nabla\rho * (x_i - x_u) \tag{10.22}$$

where x_i is the location of the beginning of one cell between the 2 cells.

would be employed for crash prediction. The total distance between the virtual upstream and downstream stations is 0.5 miles, which is not larger than the smallest spacing between physical stations.

The crash dataset consists of 113 crash cases and 2260 noncrash cases (a 20:1 noncrash to crash case ratio). The crash model includes an intercept, traffic flow variables, a series of indicator variables for traffic states representing bottleneck (BN), back-of-the-queue (BQ), congested (CT), and free flow (FF), and the interaction terms between flow state and flow variables. FF is the reference state, and therefore it does not appear in the model specification. Table 10.6 shows the model results. As can be seen, the main effects of the BN and CT states are statistically significant, while the main effect of the BQ state is not. All interaction terms are statistically significant and have the same signs. Refer to Chen and Qin (2019)for a detailed description of variables and explanations of model results. The same crash prediction models are developed from traffic data collected from physical sensors. After comparing model performance and prediction accuracy, the conclusion is that the crash prediction model performs better when it uses simulated traffic data (as opposed to physical sensor data) from uniformly and closely spaced virtual stations.

10.7.5 Crash prediction

After the crash prediction model is developed, crash-prone traffic conditions can be identified in real time following the procedures in Fig. 10.10.

The crash prediction takes real-time data from loop detectors as the input, predicts the future 5-min traffic flow rate at the existing loop

TABLE 10.6 Results of the model.

Variable	Estimate	Standard error	*P*-value
Intercept	−4.542	0.238	<.001
BN	2.126	0.524	<.001
CT	1.899	0.897	.034
FF×StdTsdDen$_d$	0.447	0.083	<.001
FF×StdTsdSpd$_d$	0.946	0.255	<.001
FF×Snow	1.168	0.494	.018
BQ×StdTsdDen$_d$	0.551	0.083	<.001
BQ×Curve	3.196	0.657	<.001
CT×AvgDen$_u$	0.00824	0.00392	.035

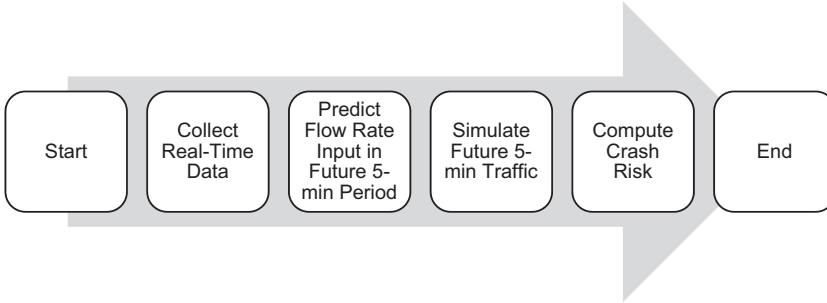

FIGURE 10.10 The procedures of predicting crashes with simulated traffic.

detector locations, then simulates the traffic in the future 5-min period for every cell using CTM and predicts the crash risk for that period based on simulated traffic data. The initial densities of all CTM cells are estimated with densities from the physical stations that are taken in the current moment. The flow rate inputs in the future 5-min period, which include in-flow (q_{in}), off-ramp flow (r), and on-ramp flow (f), can be estimated using the k-nearest neighbors (k-NN) approach (Altman, 1992). The k-NN algorithm, not to be confused with the K-means algorithm introduced earlier, is a nonparametric method for classification and regression. Algorithm 10.4 is the k-NN algorithm adapted by (Chen and Qin, 2019).

The required flow rate estimations are used to run the CTM to simulate traffic in the future 5-min period. Simulated traffic data for each cell are collected from its upstream and downstream virtual stations and are processed as the input variables to develop crash prediction models. According to Table 10.6, the predicted crash risk of Cell i can be estimated by

$$p_i = \frac{e^{\pi}}{1 + e^{\pi}} \tag{10.25}$$

Algorithm 10.4

K-NN algorithm

1. The past 30 min are considered to be the most recent time period. Flows in the recent time period are considered as the subject flow set.
2. All flow rate sets during the same time period from the last 90 days are considered as candidate flow rate sets and are matched with the subject flow rate set.

Algorithm 10.4 *(cont'd)*

K-NN algorithm

3. The ten nearest matches with the ten smallest distances are selected. The distance is determined by the following:

$$D(\boldsymbol{X^m}, \boldsymbol{Y}) = \sqrt{\sum_{i=1}^{30} \left(x_i^m - y_i\right)^2}, \mathrm{m} = 1, \ldots, 90 \tag{10.23}$$

where $\boldsymbol{X^m} = \left(x_1^m, \ldots, x_{30}^m\right)$ is the mth candidate flow rate set of 30 1-min flow rate points; $\boldsymbol{Y} = (y_1, \ldots, y_{30})$ represents the subject flow rate set.

4. The flow rate in the future 5-min period is calculated as the weighted average of flow rates in the next 5-min period for those matched flow rate sets by the following:

$$\boldsymbol{Y^F} = \frac{1}{10} \sum_{k=1}^{10} \frac{(D_k)^2}{\sum_{k=1}^{10} (D_k)^2} \boldsymbol{X^{k,F}} \tag{10.24}$$

where $\boldsymbol{Y^F} = \left(y_1^F, \ldots, y_5^F\right)$ represents the estimated flow rate set in the future 5-min period, D_k is the kth smallest distance for kth nearest matched flow rate sets among those 10 nearest matched sets, and $\boldsymbol{X^{k,F}} = \left(x_1^{k,F}, \ldots, x_5^{k,F}\right)$ is the flow rate set in the next 5-min period for kth nearest matched flow rate sets.

where

$$\begin{aligned}\pi = &-4.406 + 1.990 * BN + 1.764 * CT + 0.452 * (FF \times StdTsdSpd_d) \\ &+0.903 * (FF \times StdTsdSpd_d) - 1.049 * (FF \times OnRamp) + 1.146 * (FF \times Snow) \\ &+0.530 * (BQ \times StdTsdDen_d) + 3.111 * (BQ \times Curve) + 0.00824 * (CT \times AvgDen_u)\end{aligned}$$

Crash-prone traffic conditions are detected as long as the predicted crash probability exceeds a predetermined threshold. Chen and Qin (2019) set the threshold as 0.0427. The testing results show that 104 of the 113 cases exhibit crash-prone conditions. The application of crash prediction with simulated traffic holds the promise of using traffic simulation techniques to perform more safety-related activities such as evaluating the safety performance of new roadway designs and novel traffic control strategies.

References

AASHTO, 2010. Highway Safety Manual. American Association of State Highway and Transportation Officials (AASHTO), Washington, D.C.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F., Hsia, L., 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. Transport. Res. Rec. J. Transp. Res. Board 1897, 88–95.

Altman, N.S., 1992. An introduction to kernel and nearest-neighbor nonparametric regression. Am. Statistician 46 (3), 175–185.

Brodsky, H., Hakkert, A.S., 1983. Highway accident rates and rural travel densities. Accid. Anal. Prev. 15 (1), 73–84.

Cassidy, M.J., Rudjanakanoknad, J., 2005. Increasing the capacity of an isolated merge by metering its on-ramp. Transp. Res. Part B Methodol. 39 (10), 896–913.

Cedar, A., Livneh, M., 1982. Relationship between road accidents and hourly traffic flows-I and II. Accid. Anal. Prev. 14 (1), 19–44.

Chen, Z., Qin, X., 2019. A novel method for imminent crash prediction and prevention. Accid. Anal. Prev. 125, 320–329.

Chen, Z., Qin, X., Shaon, M.R.R., 2018. Modeling lane-change-related crashes with lane-specific real-time traffic and weather data. J. Intell. Transp. Syst. 22 (4), 291–300.

Courant, R., Friedrichs, K., Lewy, H., 1967. On the partial difference equations of mathematical physics. IBM J. Res. & Devel. 11 (2), 215–234.

Daganzo, C.F., 1994. The cell transmission model: network traffic. Transp. Res. Part B Methodol. 29 (2), 79–93.

Dervisoglu, G., Gomes, G., Kwon, J., Horowitz, R., Varaiya, P., 2009. Automatic calibration of the fundamental diagram and empirical observations on capacity. In: Transportation Research Board 88th Annual Meeting Proceedings.

Forbes, T.W., Zagorski, H.J., Holshouser, E.L., Deterline, W.A., 1958. Measurement of driver reactions to tunnel conditions. Highw. Res. Board Proc. 37, 345–357.

Frantzeskakis, J.M., Iordanis, D.I., 1987. Volume-to-capacity ratio and traffic accidents on interurban four-lane highways in Greece. Transport. Res. Rec. J. Transp. Res. Board 1112, 29–38.

Garber, N.J., Gadiraju, R., 1989. Factors affecting speed variance and its influence on accidents. Transport. Res. Rec. 1213, 64–71.

Gazis, D.C., Herman, R., Rothery, R.W., 1961. Nonlinear follow-the- leader models of traffic flow. Oper. Res. 9, 545–567.

Gwynn, D.W., 1967. Relationship of Accident Rates and Accident Involvements with Hourly Volumes. Traffic Quarter, pp. 407–418.

Hall, F.L., Agyemang-Duah, K., 1991. Freeway capacity drop and the definition of capacity. Transport. Res. Rec. J. Transp. Res. Board 1320, 91–98.

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136 A K-Means Clustering Algorithm. Appl. Stat. 28, 100–108.

Harwood, D.W., Bauer, K.,M., Potts, I.B., 2013. Development of relationships between safety and congestion for urban freeways, transportation research record. J. Transp. Res. Board 2398, 28–36.

HCM, 2010. Highway Capacity Manual. Transportation Research Board (TRB), Washington, D.C.

Hosmer Jr., D.W., Lemeshow, S., 2004. Applied Logistic in Applied Logistic Regression, first ed.

Kononov, J., Allery, B., 2003. Level of service of safety: conceptual blueprint and analytical framework. Transport. Res. Rec. 1840 (1), 57–66.

Kononov, J., Lyon, C., Allery, B., 2011. Relation of flow, speed, and density of urban freeways to functional form of a safety performance function. Transport. Res. Rec. J. Transp. Res. Board 2236, 11–19.

Lee, C., Saccomanno, F., Hellinga, B., 2002. Analysis of crash precursors on instrumented freeways. Transport. Res. Rec. J. Transp. Res. Board 1784, 1–8.

Lord, D., Manar, A., Vizioli, A., 2005. Modeling crash-flow-density and crash-flow-V/C ratio for rural and urban freeway segments. Accid. Anal. Prev. 37 (No. 1), 185–199.

Mensah, A., Hauer, E., 1998. Two problems of averaging arising in the estimation of the relationship between accidents and traffic flow. Transport. Res. Rec. J. Transp. Res. Board 1635, 37–43.

Oh, J.S., Oh, C., Ritchie, S.G., Chang, M., 2005. Real-time estimation of accident likelihood for safety enhancement. J. Transport. Eng. 131 (5), 358–363.

Pande, A., Abdel-Aty, M., Hsia, L., 2005. Spatiotemporal variation of risk preceding crashes on freeways. Transport. Res. Rec. J. Transp. Res. Board 1908 (1), 26–36.

Persaud, B.N., Dzbik, L., 1993. Accident prediction models for freeways. Transport. Res. Rec. J. Transp. Res. Board 1401, 55–60.

Pipes, L.A., 1953. An operational analysis of traffic dynamics. J. Appl. Phys. 24 (No 3), 274–287.

Qin, X., Ivan, J.N., Ravishanker, N., 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments. Accid. Anal. Prev. 36 (2), 183–191.

Qin, X., Ivan, J.N., Ravishanker, N., Liu, J., 2005. Hierarchical Bayesian estimation of safety performance functions for two-lane highways using Markov chain Monte Carlo modeling. J. Transport. Eng. 131 (5), 345–351.

Qin, X., Ivan, J.N., Ravishanker, N., Liu, J., Tepas, D., 2006. Bayesian estimation of hourly exposure functions by crash type and time of day. Accid. Anal. Prev. 38 (6), 1071–1080.

Roshandel, S., Zheng, Z., Washington, S., 2015. Impact of real-time traffic characteristics on freeway crash occurrence: systematic review and meta-analysis. Accid. Anal. Prev. 79, 198–211.

Xu, C., Liu, P., Wang, W., 2016. Evaluation of the predictability of real-time crash risk models. Accid. Anal. Prev. 94, 207–215.

Xu, C., Liu, P., Wang, W., Li, Z., 2012. Evaluation of the impacts of traffic states on crash risks on freeways. Accid. Anal. Prev. 47, 162–171.

Xu, C., Liu, P., Wang, W., Li, Z., 2014. Identification of freeway crash-prone traffic conditions for traffic flow at different levels of service. Transport. Res. Pol. Pract. 69, 58–70.

Zhou, M., Sisiopiku, V.P., 1997. Relationship between volume-to-capacity ratios and accident rates. Transport. Res. Rec. J. Transp. Res. Board 1581, 47–52.