Part III

Analytical techniques

Chapter 6

Capacity and level-of-service analysis

Summary of the chapter. When predicting the performance of a traffic facility, an important question is how much traffic the facility can carry. The fact that the first capacity studies of highway facilities date back as far as 1920 indicates that the issue has been of interest to builders and operators alike for many years. Recently, the field of capacity analysis has been extended to include level-of-service. That is, current analysis represents the trade-off between the quantity of traffic a facility can carry and the resulting level-of-service offered to the user of the facility.

This chapter focusses on capacity of uninterrupted and interrupted freeway sections, and is a summary of the relevant chapter of [36], which in turn describes largely the approach adopted in the HCM [4]. In the Netherlands, a different approach has been used. The basis for the Dutch approach is the application of the microscopic simulation model FOSIM to a large number of bottle-neck situations (on-ramps, weaving sections, etc.) for a large variety of flow conditions. On the contrary to the HCM approach, the Dutch approach does not explicitly consider level-of-service (LOS). Capacity analysis of signalised intersections is an important issue, but is beyond the scope of this course. In the ensuing chapters, we will see the importance of the notion of capacity to estimate travel times and delays using queuing models.

List of symbols

c	veh/s	capacity
c_j	veh/s	lane capacity under ideal conditions; design speed level j
N	-	number of lanes
f	-	capacity reduction factors
h_i	s	time headways
SF_j	veh/s	service flow rate at LOS i
γ	-	weaving influence factor
u	m/s	speeds
v	veh/s	traffic volume (intensity)

6.1 Capacities and level-of-service

Capacity is usually defined as follows [36]

Definition 47 *The maximum hourly rate at which persons or vehicles can reasonably be expected to traverse a point or uniform section of a lane or roadway during a given time period (usually 15 minutes) under prevailing roadway, traffic, and control conditions.*

Although we adopt this definition, it is stressed that several aspects make a practical single definition of capacity complicated. These complications are among other things due to the capacity drop phenomenon, the differences between the capacity of a motorway link (or multilane

facility, basic motorway segment), a motorway bottle-neck (on-ramps, off-ramps, weaving sections), and the stochastic nature of the capacity. The capacity drop has already been discussed in section 4.2.2.

In the US, typical values of the capacity of a freeway with a design speed of 60 or 70 miles/h is 2000 veh/h/lane under ideal conditions; in Europe and especially in the Netherlands, capacities under ideal circumstances are much higher, around 2400 veh/h. Ideal conditions in this case imply 12-foot lanes and adequate lateral clearances; no trucks, buses, or recreational vehicles in the traffic stream; and weekday or commuter traffic. When ideal conditions do not exist, the capacity is reduced. The *Highway Capacity Manual* [4], [5] proposes using the following example relation to express the influence of non-ideal conditions

$$c = c_j N f_w f_{HV} f_p \tag{6.1}$$

where

c	=	capacity (veh/h)
c_j	=	lane capacity under ideal conditions with design speed of j
N	=	number of lanes
f_w	=	lane width and lateral clearance factor
f_{HV}	=	heavy vehicle factor
f_p	=	driver population factor

For example, if ideal conditions existed along a three lane directional motorway having a design speed of 70 mph, the capacity would be

$$c = 2000 \cdot 3 \cdot 1 \cdot 1 \cdot 1 = 6000$$

However, normally ideal conditions do not exist, and in the US typical lane capacities are around 1800 veh/h. In Europe, and in particular in the Netherlands, much higher capacities are encountered. Furthermore, several studies have shown that other factors (such a weather and ambient conditions) also influence the capacity significantly. For instance, section 4.4 showed that the effect of rain on the capacity yields a factor of $f_{weather} = 0.85$; the effects of ambient conditions f_{light} are shown in Tab. 4.2 on page 97; the effect of rain for different road surfaces is shown in Tab. 4.1; other factors are discussed in Sec. 4.1.3.

Capacity is a measure of maximum route productivity that does not address the traffic flow quality or the level-of-service to the users. The *level-of-service* (LOS) reflects the flow quality as perceived by the road users. These flow quality aspects for drivers on the motorway is closely related to the experienced travel times (or travel speeds), the predictability of future traffic conditions (e.g. travel speed, waiting times), and experienced comfrot of the trip (number of stops, required accelerations and decelerations, ability to drive at the desired speed).

To include the user-related traffic flow quality aspects, the concept of *service volume* has been introduced. The service volume SF has a definition which is exactly like capacity except that a phrase is added at the end: *"while maintaining a designated level-of-service"*. In the HCM [4], [5], six service levels ranging from service level A to F are distinguished. Table 6.1 (and Fig. 6.1) shows the definitions of these levels of services. In illustration, if one wishes to operate this particular section of freeway at LOS C, the volume-capacity ratio should be limited to 0.77, and speeds over 54 mph and lane densities of less than 30 veh/mile per lane should results. Speed characteristics, density characteristics, and the relation between these characteristics have been and will be discussed elsewhere in this syllabus. Note that the values in the Netherlands are very different from the values shown in table 6.1. Moreover, the concepts are not only applicable to freeway traffic flow operations, but for instance also in the analysis of pedestrian walking facilities, such as railway stations, sport stadiums, etc.

In the remainder of this chapter, we will focus on different types of motorway facilities, starting with uninterupted multilane roads. Note that for other facilities, the analytical tools are similar.

LOS	Flow conditions	v/c limit	Service volume	Speed	Density
			(veh/h/lane)	(miles/h)	(veh/mile)
A	Free	0.35	700	≥ 60	≤ 12
B	Stable	0.54	1100	≥ 57	≤ 20
C	Stable	0.77	1550	≥ 54	≤ 30
D	High density	0.93	1850	≥ 46	≤ 40
E	Near capacity	1	2000	≥ 30	≤ 67
F	Breakdown	Unstable		< 30	> 67

Table 6.1: Level of service for basic freeway sections for 70 km/h design speed.

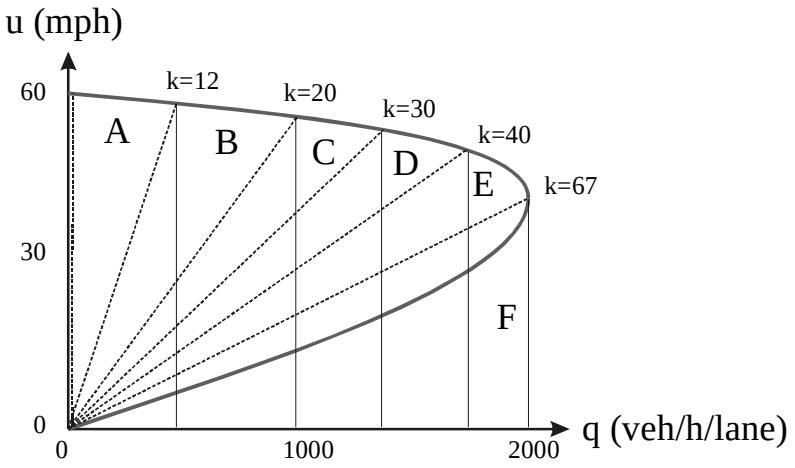

Figure 6.1: Speed-flow relation for a multilane facility for 70mph design speed (from [36])

6.2 Capacity and driver behaviour

Before discussing how the notion of capacity can be applied to basic motorway segments and bottle-necks, let us first describe how the capacity relates to the characteristics of the traffic flow or rather of the driver vehicle combinations in the flow. Recall that for a single lane of the roadway, the flow q can be determined from the headways h_i as follows

$$q = \frac{1}{\frac{1}{n}\sum_i h_i} \tag{6.2}$$

When a roadway lane operates at capacity, this thus implies that most drivers follow each other at the *minimum time headway* (empty zone), say h_i^*. Thus, we have for the capacity of a lane

$$c = \frac{1}{\frac{1}{n}\sum_i h_i^*} \tag{6.3}$$

Note that this relation indicates clearly that the capacity is related to driver behaviour, which explains how aspects like the vehicle fleet composition, lane width and lateral clearance factor, weather conditions, etc., will affect capacity, namely by (changing) the behaviour of drivers. For instance, trucks drivers generally need a larger headway with respect to their leader, due to the length of the truck, as well as larger safety margins for safe and comportable driving.

For multilane facilities, besides the car-following behaviour, the distribution of traffic over the roadway lanes will determine the capacity. Ideally, during capacity operations, all lanes of the roadway are utilised fully, that is, all driver-vehicle combinations are following their leader at the respective minimal headway h_i^*. In practise however, this is not necessarily the case, since the lane distribution will depend on the lane demands and overtaking opportunities upstream of the bottle-neck.

6.3 Multilane facilities

By definition, multilane facilities have two or more lanes available for use (for each direction of travel). The key is that multilane facilities provide uninterrupted flow conditions away from the influence of ramps or intersections. They are often referred to as basic motorway segments. In the approach proposed by the HCM, first capacity analysis under ideal conditions is performed, followed by capacity analysis under non-ideal circumstances. Ideal conditions satisfy the following criteria [4]:

- Essentially level and straight roadway
- Divided motorway with opposing flows not influencing each other
- Full access control
- Design speed of 50 mph or higher
- Twelve-foot minimum lane widths
- Six-foot minimum lateral clearance between the edge of the travel lanes and the nearest obstacle or object
- Only passenger cars in the traffic stream
- Drivers are regular users of such facilities

6.3.1 Capacity analysis under ideal conditions

The speed-flow relationships for multilane facilities have been discussed in chapter 4. These diagrams relate our three scales (flow, density, speed) that are important in LOS analysis. The average speed is an indication of the LOS provided to the users. Traffic flow is an indication of the quantity of traffic that can use the facility. The density is an indication for the freedom of movement of the users. It is noted that the upper density boundary of LOS E (of 67 veh/mile/lane) occurs at the capacity value. Only one congested state in considered in the 1985 HCM LOS classification.

For multilane facilities, the basic equation needed for capacity (and LOS) analysis under ideal conditions is

$$SF_i = \left(\frac{v}{c_j}\right)_i (c_j N) \tag{6.4}$$

where

SF_i = maximum service flow rate for level of service i
j = design speed
c_j = lane capacity under ideal conditions with design speed of j
N = number of directional lanes
$(v/c_j)_i$ = maximum volume-to-capacity ratio for LOS i

The eqn. (6.4) can be used in three ways: 1) by solving for SF_i, the maximum service flow can be determined for a given designed multilane facility under specified LOS requirement; 2) by solving for $(v/c_j)_i$, the LOS can be detemrined for a given designed multilane facility carrying a specific service flow rate. Finally, 3) by solving for $(c_j N)$, the design of a multilane facility can be determined when the LOS and the service flow are specified.

6.3.2 Capacity analysis under non-indeal conditions

The starting point for capacity and LOS analysis for multilane facilities under less than ideal conditions is to go back to eqn. (6.4). Clearly, the factor $(c_j N)$ should be reduced by some factor or a series of factors (compare eqn. (6.1)). Each factor would represent one non-ideal

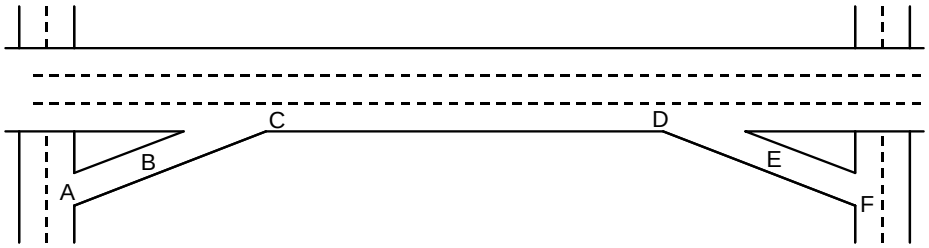

Figure 6.2: Typical motorway configuration (from [36])

condition listed in section 6.3, and multiplied to get a composite reduction factor. It should be noted that in multiplying these factors, we implicitly assume that these factors and independent and that their combined independent effects are multiplicable. In any case, eqn. (6.4) becomes

$$SF_i = \left(\frac{v}{c_j}\right)_i (c_j N)\,(f_1 \times f_2 \times ... \times f_n) \tag{6.5}$$

where $f_1, ..., f_n$ are reduction factors for non-ideal conditions. In the 1985 HCM, four reduction factors are proposed for multilane facilities, namely

1. the width reduction factor f_W, describing the reduction in capacity due to less than ideal lane widths and side clearances,
2. the heavy-vehicle reduction factor f_{HV}, describing the reduction in the capacity (in veh/h/lane!) due to the presence of heavy vehicles *under different vertical alignment conditons*,
3. the driver population factor f_P reflects the reduction in capacity due to the presence of non-regular users, and
4. the environment factor f_E to consider the reduction in capacity due to the lack of a median and/or the lack of access control.

6.4 Ramps

Ramps are sections of roadway that provide connections from one motorway facility to another motorway facility or to another non-motorway facility. Entering and exiting traffic causes disturbances to the traffic on the multilane facilities and can thus affect the capacity and the LOS of the basis motorway segments. Fig. 6.2 shows a typical (schematised) motorway configuration where an on-ramp is followed by an off-ramp. On each ramp, three locations must be carefully studied.

Location A is the entrance to the on-ramp and is affected by the ramp itself and/or by the at-grade intersection. Since the dimensions and the geometrics at location A are (normally) better or at least as good as that of location B, the effect of the physical on-ramp will be studied further at location B. Normally, the at-grade intersection controls the entrance to the on-ramp, and the potential restrictions this causes will not be studied in this course.

Location B is on the on-ramp itself and its capacity is affected by the number and the width of the lanes, as well as the length and the grade of the on-ramp. As long as the on-ramp demand is smaller than the on-ramp capacity, LOS is not really a concern. The reason for this is the relative short length of the ramps.

Locations E and F are "mirror images" of locations A and B in an analytical sense. Location E is the off-ramp itself; similar to the on-ramp, the LOS is not really a concern for the off-ramp. Location F is at the exit of the off-ramp where it connect to a crossing arterial at an at-grade

LOS	Merge flow rate	Diverge flow rate
A	≤ 600	≤ 650
B	≤ 1000	≤ 1050
C	≤ 1450	≤ 1500
D	≤ 1750	≤ 1800
E	≤ 2000	≤ 2000
F	–	–

Table 6.2: Allowable service flow rates for merging and diverging areas (passenger cars per hour) from 1985 HCM.

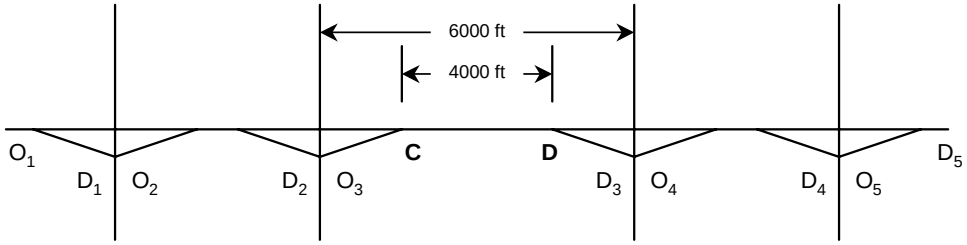

Figure 6.3: Extended typical motorway ramp configuration from [36].

intersection. An important difference between locations A and F is the location of the queues if they exit. At location A, any queues will extend into the at-grade intersection, whereas at location F, any queues will extend up the off-ramp and - if serious enough - into the multilane facility.

Locations C and D are the merge and the diverge areas and require special analytical procedures. The concept and basic principles presented by [36] are straightforward: the substance of the analytical procedure is to compare the actual demands in the merge and the diverge areas with the allowable service flow rates. This comparison is then used to determine the resulting LOS.

Table 6.2 shows an example of allowable service flow rates for merging and diverging areas for ideal conditions for various levels of service. Note that the upper limit of LOS E corresponds to the capacity of the rightmost lane under ideal conditions, which in this case equals 2000 passenger-cars per hour. As noted in table 6.2, the LOS of merge and diverge areas diminish as traffic demands in the rightmost lane increase. These allowable service flow rates should be reduced when non-ideal conditions are considered, using the reduction factors employed for basic multilane facilities. If the capacities and levels-of-service of the basic multilane motorway segment between the merge and the diverge area have been computed, the multilane service flow rates divided by the number of lanes in the basic segment can be used as the allowable lane service flow rates in the merge and diverge analysis.

The major difficulty is in estimating the traffic demands in the rightmost lane. The key to he solution is to consider that traffic in the rightmost lane is made up of subgroups of traffic each having an unique origin and destination along the multilane facility. Fig. 6.3 can be used to illustrate this approach. Table 6.3 shows all possible OD movements. Note that not all OD movements will pass through the merge and diverge areas and can thus be ignored in our analysis. The remaining nine OD movements can be combined into four groups: through, entering, exiting and local. Each will now be addressed in order to determine its share of the traffic demand in the rightmost lane in the vicinity of the merge and diverge areas is question. For demonstration purposes, the distance between the on-ramp nose and the off-ramp gore is assumed to be 4000 feet and its share of traffic in the rightmost lane will be calculated at 1000-foot intervals.

OD	D_1	D_2	D_3	D_4	D_5
O_1	-	-	Exiting	Through	Through
O_2	-	-	Exiting	Through	Through
O_3	-	-	Local	Entering	Entering
O_4	-	-	-	-	-
O_5	-	-	-	-	-

Table 6.3: Possible motorway OD movements.

Through traffic demand	Motorway lanes		
veh/h	8	6	4
≥ 6500	10	-	-
6000-6499	10	-	-
5500-5999	10	-	-
5000-5499	9	-	-
4500-4999	9	18	-
4000-4499	8	14	-
3500-3999	8	10	-
3000-3499	8	6	40
2500-2999	8	6	35
2000-2499	8	6	30
1500-1999	8	6	25
≤ 1499	8	6	20

Table 6.4: Possible motorway OD movements.

Through traffic is traffic that enters the motorway at least 4000 feet upstream of the merge area and exits the freeway at least 4000 feet downstream of the diverge area. Tab. 6.3 shows which OD movements are combined and classified as through traffic, assuming interchange spacing on the order of 1 mile. The question is now to determine how much of this traffic will be in the rightmost lane. May [36] proposes using tables that describe the perceptage of traffic in the rightmost lane for n-lane motorway facilities. Tab. 6.4 shows an example from [36]. The percentages shown in this table are assumed to be constant between the on-ramp and the off-ramp.

Entering traffic is that traffic that enters the motorway in the merge area (location C) and has a destination that is beyond the diverge area (location D); see Tab. 6.3. All entering traffic is in the rightmost lane in the merge area and as the traffic moves farther and farther downstream, a smaller and small proportion remains in the rightmost lane. Fig. 6.4a shows some figures describing the percentage of entering traffic on the rightmost lane. Fig. 6.4b shows the percentage of exiting traffic on the rightmost lane. Finally, local traffic is traffic that enters in the merge area (location C) and exits in the diverge area (location D). Generally, it is assumed that local traffic remains in the rightmost lane.

In sum, the total traffic in the rightmost lane at various locations can be determined by the sum of through, entering, exiting and through traffic. The demand on the rightmost lane is subsequently compared with the allowable service flow rates (such as those given in Tab. 6.2). The highest demand in the vicinity of the merge area is compared with the allowable merge service flow rates, and the highest demand in the vicinity of the diverge area is compared with the allowable diverge service flow rates. The resulting level of service can then be determined.

Although the principles set forth earlier for capacity and LOS analysis of merging and diverging areas are straightforward, their applications can be complicated and tedious. The complications can be caused by unusual ramp geometrics and are particulary difficult at near capacity or oversaturated situations. The 1985 HCM [4] contains many monographs that can

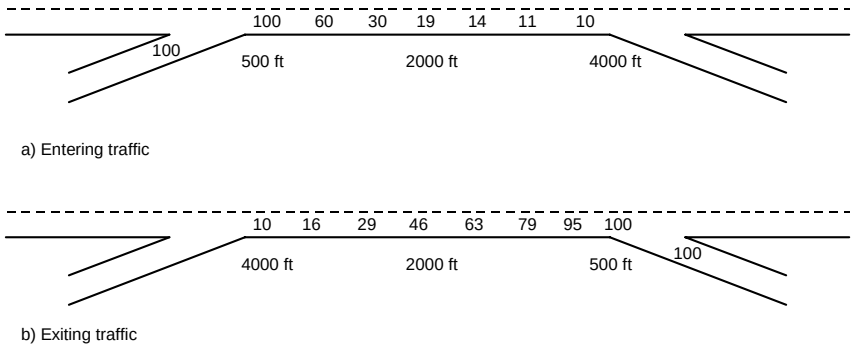

Figure 6.4: Percentage of entering and exiting traffic in rightmost lane, from [36].

be used to estimate the LOS provided in the merge and diverge areas under a wide variety of geometric configurations.

6.5 Weaving sections

Traffic entering and leaving multilane facilities can also interrupt the normal flow of basic motorwau segments by creating weaving sections.

Definition 48 *Weaving is defined as the crossing of two (or more) traffic streams traveling in the same direction along a significant length of motorway without the aid of traffic control devices.*

Weaving vehicles that are required to change lanes cause "turbulance" in the traffic flow and by so doing reduce the capacity and the LOS of weaving sections. Thus, analytical techniques are needed to evaluate this reduction.

6.5.1 HCM-1965 approach

A variety of weaving analysis techniques are available and are being used. Still, it has been recognised that further research on the capacity and LOS of weaving sections is very important. In this section, we show one specific approach to analyse a weaving area in order to show the important factors and arising complications. We will only consider one specific type of weaving section, namely the one shown in Fig. 6.5. Here v_{o1} denotes the heavy flow from A to C (outer flow 1), and v_{o2} denotes the light flow from B to D (outer flow 2). Neither of these flows is a weaving movement; their sum $v_{nw} = v_{o1} + v_{o2}$ is referred to as the total non-weaving flow. The flow from B to C and A to D cross each other's path over a certain distance and are referred to as weaving flows. The higher weaving flow is indicated by v_{w1}; the lower weaving flow is referred to as v_{w2}; their sum $v_w = v_{w1} + v_{w2}$ is referred to as the total weaving flow. The length of the weaving section is denoted by L.

In the approach of the HCM-1965, one first needs to determine if the weaving causes more than the normal amount of lane changing. For instance, when the weaving length L is large and the total weaving flow is small, then only the normal amount of lane changing is expected and the roadway section is "out of the realm of weaving". In the 1965 HCM, the following equation expresses the service flow rate for a specific weaving section

$$SF = \frac{v_{w1} + \gamma v_{w2} + v_{o1} + v_{o2}}{N} \tag{6.6}$$

where

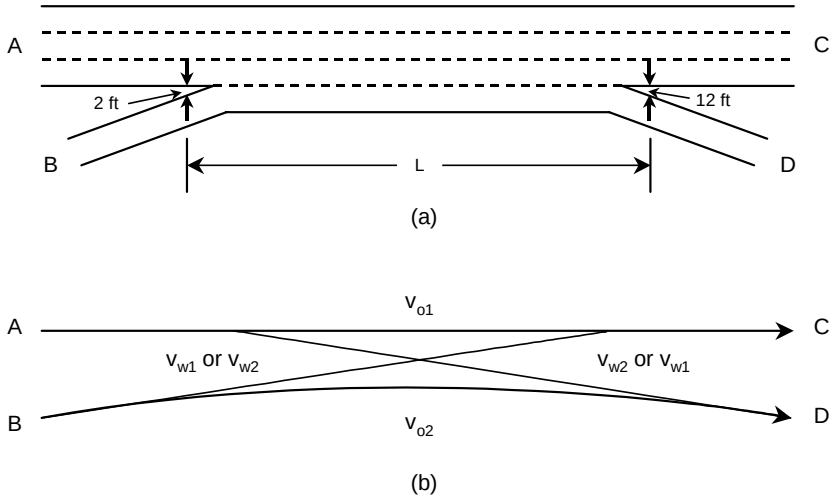

Figure 6.5: Typical simple weaving section (from [36]).

SF = service flow rate
N = number of lanes in the weaving section
γ = weaving influence factor

The weaving infleunce factor γ is a function of the total weaving traffic demand v_w and the length of the weaving section L (see example Fig. 6.6).

6.5.2 HCM 1985 approach

In the 1985 HCM [4], a more comprehensive approach to analyse the LOS is presented. In the approach, three types of weaving sections are distinguished (A, B, and C), as well as the distinction between unconstrained and constrained operations. based on field study results, 12 multiple regression equations where proposed predicting the speed of weaving and non-weaving vehicles. Using these speed predictions, the LOS can be determined.

Weaving sections types

The weaving sections are distinguished based on the required lane changing maneouvres of the weaving vehicles. Type A weaving sections (see Fig. 6.7) require that each weaving vehicle is required to make one lane changing movement, although more than one lane change may be required is weaving vehicles on are not in the correct lane at the start of the weaving section. The minimum number of lane changes equals $v_{w1} + v_{w2}$; the minimum rate of lane changes is equal to $(v_{w1} + v_{w2})/L$.

Type B weaving sections (see Fig. 6.8) require that one waving movement may be accomplished without making any lane changes, while the other movement requires one lane change. This design can be very effective if the minor weaving flow v_{w2} is relatively small. The minimum number of lane changes equals v_{w2}; the minimum lane changing rate equals v_{w2}/L.

Type C weaving sections (see Fig. 6.9) require that one waving movement may be accomplished without making a lane change, and the other waving movement requires at least two or more lane changes. This can be an effective design if the second weaving flow is small, but it can havy very adverse effects if the second weaving flow is too large, the number of lange changes is large, and the weaving length is too short. The minimum number of lane changes equals $2v_{w2}$ (or more if more than two lane changes are required); the minimum lane changing rate is equal to $2v_{w2}/L$.

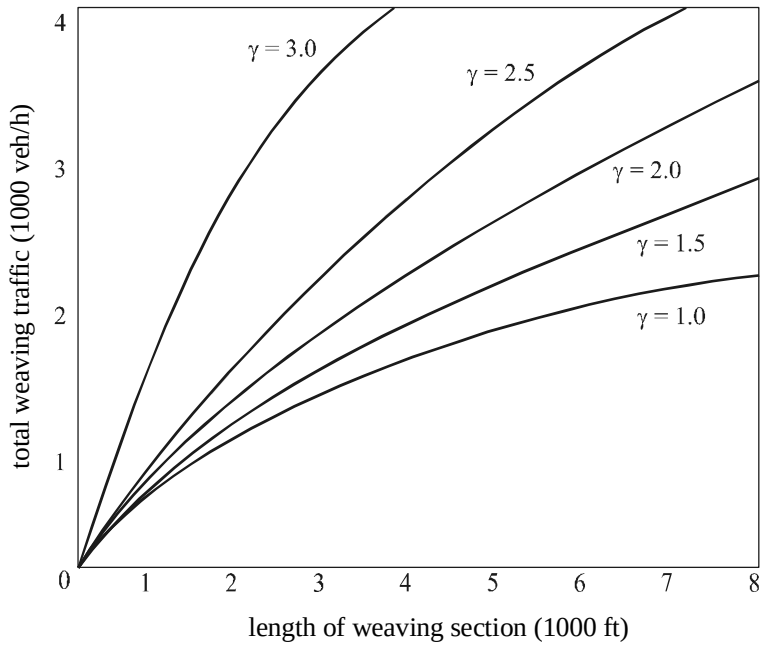

Figure 6.6: Weaving influence factor γ as a function of the length of the weaving area and the amount of weaving traffic

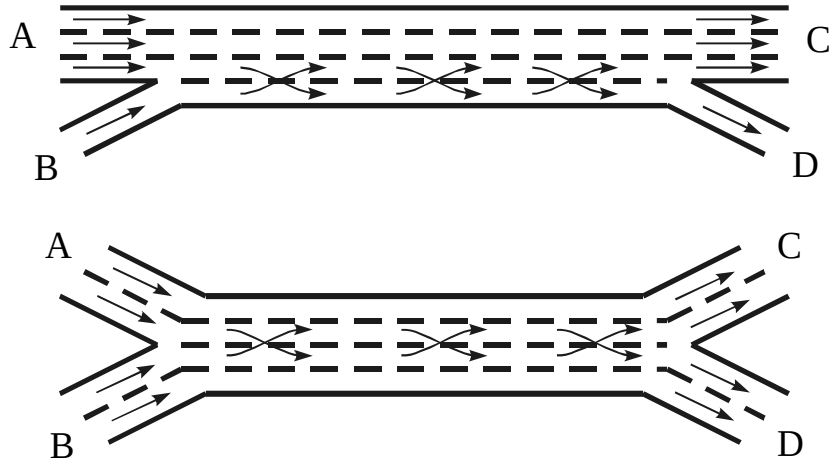

Figure 6.7: Examples of weaving area configuration A.

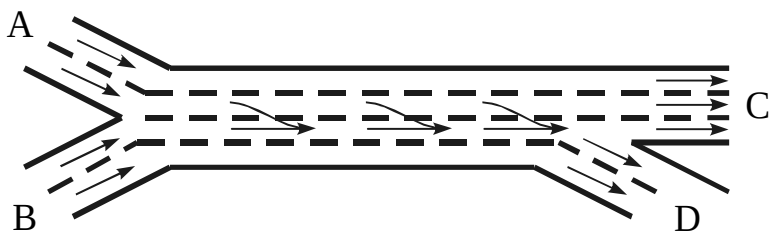

Figure 6.8: Example of weaving area configuration B.

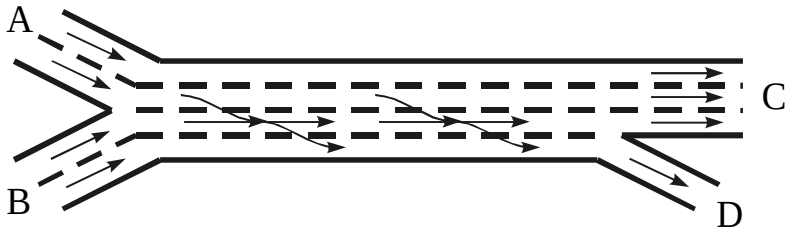

Figure 6.9: Example of weaving area configuration C.

Conf.	N_w	N_w (max)
Type A	$2.19N \cdot \left(\frac{v_w}{v_w+v_{nw}}\right)^{0.571} \left(\frac{(100L)^{0.234}}{S_W^{0.438}}\right)$	1.4
Type B	$N\left(0.085 + 0.703\frac{v_w}{v_w+v_{nw}} + \left(\frac{234.8}{L}\right) - 0.018\left(S_{nw} - S_w\right)\right)$	3.5
Type C	$N\left(0.761 - 1.1L_H - 0.005\left(S_{nw} - S_w\right) + 0.047\frac{v_w}{v_w+v_{nw}}\right)$	3.0

Table 6.5: Criteria for unconstrained and constrained operations of weaving sections. S_{nw} and S_w respectively denote the speed of the non-weaving and weaving vehicles.

Constrained and unconstrained operations

The 1985 HCM approach also distinguishes constrained and unconstrained operations. If the weaving configuration in combination with the traffic demand patterns permits the weaving and non-weaving vehicles to spread out evenly across the lanes in the weaving section, the flows will be somewhat balanced between lanes and the operation is more effective and is classified as unconstrained. On the contrary, if the configuration and demand limit the ability of weaving vehicles to occupy their proportion of available lanes to maintain balances operations, the operations is less effective and is classified as constrained. Consider for instance the weaving section shown in Fig. 6.5: if the flow from A to C is relatively light and the other flows are relatively heavy, the lanes on the left side of the weaving section will be underutilised and the lanes on the right side will be overutilised. Such imbalanced or constrained operations will result in weaving vehicles travelling at lower speed (hence lower LOS) and non-weaving vehicles travelling at a higher speed.

Determination of the type of operation is done by comparing two variables, namely N_w (number of lanes that must be used by weaving vehicles in order to achieve balanced or unconstrained operations) and N_w (max) (maximum number of lanes that may be used by weaving vehicles for a given configuration). If $N_w < N_w$ (max), the operation is defined as unconstrained, while if $N_w > N_w$ (max) the operation is defined as constrained. Based on empirical observations, procedures to compute these variables are shown in Tab. 6.5.

The next step in the analysis is to select appropriate multiregression type equations for prediction weaving and non-weaving speeds based on types of weaving configurations and types of operations. Again, empirically derived equations have been determined and can be found in the 1985 HCM. These are listed in Tab. 6.6. These parameters can subsequently be substituted in the following equation

$$S_w \text{ (or } S_{nw}) = 15 + \frac{50}{1 + a\left(1 + \frac{v_w}{v_w+v_{nw}}\right)^b \left(\frac{v}{N}\right)^c / L^d} \tag{6.7}$$

Note that the speeds of weaving and non-weaving vehicles are also required to decide between constrained and non-constrained operations, yet these speeds have not yet been determined. The suggested approach is to first assume unconstrained operations, calculate weaving and non-weaving speeds and then use the equations in Tab. 6.5 to see if the assumption of unconstrained

Conf.	Parameter values for S_w				Parameter values for S_{nw}			
and operation	a	b	c	d	a	b	c	d
A - unconstrained	0.226	2.2	1.00	0.9	0.020	4.0	1.30	1.00
A - constrained	0.280	2.2	1.00	0.90	0.020	4.0	0.88	0.60
B - unconstrained	0.100	1.2	0.77	0.50	0.020	2.0	1.42	0.95
B - constrained	0.160	1.2	0.77	0.50	0.015	2.0	1.30	0.90
C - unconstrained	0.100	1.8	0.80	0.50	0.015	1.8	1.10	0.50
C - constrained	0.100	2.0	0.85	0.50	0.013	1.6	1.00	0.50

Table 6.6: 1985 HCM parameter values for determination of speeds of weaving and non-weaving traffic.

LOS	Minimum S_w	Minimum S_{nw}
A	55	60
B	50	54
C	45	48
D	40	42
E	35	35
F	30	30

Table 6.7: Level of service criteria for weaving sections.

operations is correct. If not, the process is repeated assumig constrained operations. The final step in determining the LOS of the weaving section is to enter Tab. 6.7 with the predicted weaving and non-weaving speeds.

6.6 Dutch approach to motorway capacity analysis

The Dutch design procedure focuses on the achievable capacity of a weaving section. The level of service is not estimated: as a rule of thumb, the weaving segment design should be based on a demand-capacity ratio equal or below 0.8. This value is accepted by roadway designers as a good design value for freeway facilities in the Netherlands.

The Level of Service of a freeway facility or network can be determined by additional procedures that take into account the probability of congestion over a year. The background of the method is a cost-benefit analysis, given the distribution of capacity and expected traffic demand pattern over a year. Currently, new guidelines for the Level of Service of freeways are being developed using the average travel speed over a 10-20 km traject as variable. However, these Dutch approaches for level of service calculation cannot be compared to the LOS definition in the HCM, as they relate to a facility or network, rather than the segment based HCM approach. Capacity is defined similarly as in the HCM, except that a different analysis time interval is used. Dutch guidelines consider a 5-minute time interval, because this value is used in the simulations using the Dutch freeway trafic flow simulator FOSIM. Capacity refers to a *pre-queue capacity value*.

As an example, let us consider determination of the capacity of a weaving section. The notation used in describing weaving sections is straightforward. Two types of weaving sections are distinguished: symmetrical and asymmetrical types:

- A symmetrical weaving section '2+1' indicates that leg A has 2 incoming lanes, and leg B has 1 lane. The section is said to be symmetrical, because the number of lanes in legs C and D are equal to that of A and B respectively.
- A weaving section '2+2 $\rightarrow$ 3+1' indicates that legs A and B have 2 incoming lanes each, and the weaving section geometry is asymmetrical, with leg C having 3 lanes and leg D

having 1 lane.

By means of microscopic simulation, several tables have been established giving capacity values as function of:

- Weaving configuration types: symmetrical 1+1, 2+1, 3+1, 4+1, 2+2, 3+2, 4+2 (which are similar to the HCM weaving section type A), and several asymmetrical designs.
- Free flow speed (120 km/h or 75 mph)
- Truck proportion (with a range of 5-15%);
- Length of weaving segment (the range depends on the type of weaving segment – lenghts considered are: 100 – 1000 m)
- Weaving ratio of the leg with the smallest incoming flow (LR) (range depends on the type of weaving segment)

The capacity values in the tables are expressed in vehicles/hour. This value can be converted into pcph by using a pcu-truck value of 1.5 according to the Dutch guidelines. The overview table is published in the Dutch guidelines for freeway capacity, the so-called *CIA-manual*. In 2001 the asymmetrical types (which can be compared with HCM type C) were also studied using the microscopic simulation model FOSIM. For these types also capacity tables have been established, and are available. The calibrated and validated microscopic simulation model FOSIM was used to calculate capacity values for a wide range of weaving segment designs. The following procedure was repeated for every scenario (weaving designs varying in weaving section geometry, truck proportion, weaving segment length) in order to determine the capacity values:

1. Run simulation for a specific scenario until congestion occurs upstream or on weaving segment; Stop simulation and determine maximum 5-minute flow rate.
2. Repeat step 1 50-100 times applying different random number seed values. This results in a distribution of capacity values.
3. Determine median capacity value of the performed simulation runs. This value is denoted as 'capacity'. An important aspect to be considered is the approach followed in the simulation of a scenario. In all the simulations the weaving flow rate is equal for both legs. Thus the flow rate on the origin-destinations AD and BC was nearly equal.

6.7 Stochastic nature of motorway capacity

Maximum flows (maximum free flows of queue discharge rates) are not constant values, and vary under the influence of several factors, as was disussed in this and previous chapters. Factors influencing that capacity are among other things the composition of the vehicle fleet, the composition of traffic with respect to trip purpose, weather-, road-, and ambient conditions, etc. These factors affect the behaviour of driver vehicle combinations and thus the maximum number of vehicles that can pass a cross-section during a given time period. Some of these factors can be observed and their effect can be quantified. Some factors can however not be observed directly. Furthermore, differences exist between drivers implying that some drivers will need a larger minimum time headway than other drivers, even if drivers belong to the same class of users. As a result, the minimum headways h_i^* will not be constant values but follow a distribution function (in fact, the empty zone distribution $p_{fol}(h)$), as was explained in chapter 2. As a result, capacity will also be a random variable following a specific distribution. The shape of this distribution depends on among other things the capacity definition and measurement method / period. In most cases, a Normal distribution will can be used to describe the capacity.

6.8 Capacity drop

In section 4.2.2, we have discussed the existence of two different maximum flow rates, namely *pre-queue capacity* and *queue discharge rate* respectively. Each of these have their own maximum flow distribution.

Definition 49 *We define the pre-queue maximum flow as the maximum flow rate observed at the downstream location just before the on-set of congestion (a queue) upstream. These maximum flows are characterised by the absence of queues or congestion upstream the bottle-neck, high speeds, instability leading to congestion on-set within a short period, maximum flows showing a large variance [37].*

Definition 50 *The queue discharge flow is the maximum flow rate observed at the downstream location as long as congestion exists. These maximum flow rates are characterised by the presence of a queue upstream the bottle-neck, lower speeds and densities, a constant outflow with a small variance which can sustain for a long period, however with lower flow rates than in the pre-queue flow state [37].*

Both capacities can only be measured *downstream* of the bottle-neck location. Differences between the two capacities (so-called *capacity drop*) are in the range of -1% to -15%. Different explanations for the capacity drop can be and have been given. Dijker [16] argues that the main reason is the preference for larger headways if drivers experience congested conditions. Differences between acceleration and deceleration behaviour may also contribute to this phenomenon.

6.9 Capacity estimation approaches

To determine the capacity of a bottle-neck or a basic motorway segment, appropriate capacity estimation techniques are required. These techniques can be classified in techniques that do not require capacity observations and those who do. The former methods, which are based on free flow traffic *and* constrained traffic measurements are generally less reliable than methods using capacity measurements. If the capacity state has not been reached and a capacity estimation must be performed, the following approaches are applicable:

1. *Headway distribution method.* The observed headway distribution is used to determine the distribution of the minimum headway h_i^*, which in turn is used to estimate a single capacity value (no distinction between pre-queue capacity and queue-discharge rate). See chapter 2.

2. *Fundamental diagram method.* This approach uses the relationship between speed and density or flor rate to estimate the capacity value. A functional form needs to be selected and assumptions about the critical density must be made. See section 4.3.3.

Methods using explicitly capacity flows sometimes use additional flow measurements in order to get an improved capacity estimate. Some methods do not distinguish between queue and pre-queue maximum flows.

1. *Selected maxima method.* Measured flow rate maxima are used to estimate a capacity value or distribution. The capacity state must be reached during each maxima selection period. Approach should be applied over a long period.

2. *Bimodal distribution method.* This method may be applied if the observed frequency distributions of the flow rates exhibit a clear bimodal form. The higher flow distribution is assumed to represent capacity flows.

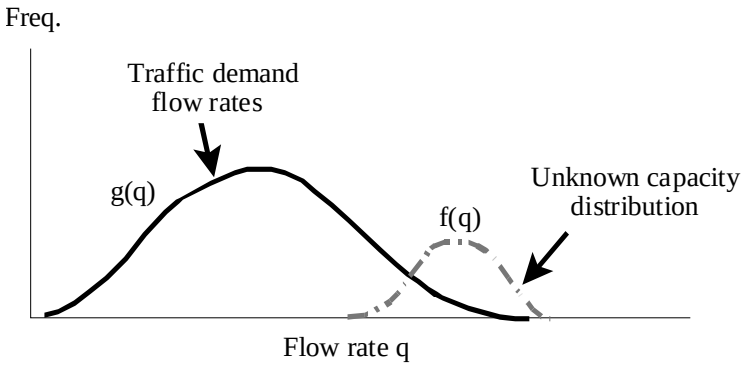

Figure 6.10: Example of probability density functions of flow rates divided into demand and capacity flows.

3. *Queue discharge distribution method.* This is a very straightforward method using queue discharge flow observations to construct a capacity distribution or a capacity value. The method requires additional observations (for instance, speeds upstream of the bottle-neck) to determine the congested state.

4. *Product-limit method.* This method uses below-capacity flows together with capacity flows to determine a capacity value distribution. Speed and / or density data is needed to distinguish the type of flow measurement at the road section upstream of the bottle-neck.

6.9.1 Product-limit method to capacity estimation

Formally, an observation is said to be 'right censored' at flow rate q if the unknown capacity value c of the observation is only known to be greater than or equal to q. This type of censoring is very common in lifetime data, and obviously apparent in capacity estimation problems. Mostly, in lifetime data analysis continuous models and data are considered.

Let $i = 1, 2, ...n$ indicate the observation period and n the total number of observed periods. A traffic flow rate in a bottle-neck section at observation period i is denoted with q_i. Simultaneous observations at the upstream section gives information over the traffic conditions. We denote δ_i as an *indicator for the type of measurement at the bottle-neck*, based on the observations at the upstream section. Two conditions are distinguished:

- $\delta_i = 0$ (capacity flow at bottle-neck in period i, i.e. congested conditions upstream);
- $\delta_i = 1$ (traffic demand flow at bottle-neck in period i, uncongested conditions upstream)

Each observation period i is assumed to have its own specific capacity value. However, we can only measure a capacity value if the bottle-necks capacity has been reached. If the bottle-necks capacity has not been reached, we are observing traffic demand. Nonetheless, this traffic demand value can be used in the Product-Limit capacity estimation procedure since it gives valuable information about the location of the capacity: the capacity value will be higher than the observed volume.

Based on this principle, we can present the maximum likelihood for a sample of observations. Firstly, let us denote functions (and their properties) which will be used in the estimation. See also Fig. 6.10.

Capacity observations are assumed to be *identically and independently distributed* with probability density function $f(q)$ and survival function $S(q)$. The survival function $S(q)$ is equal to $1 - F(q)$, where $F(q)$ is the cumulative distribution function of $f(q)$, i.e. $F(q) = \int_{-\infty}^{q} f(x)\, dx$.

The independence requirement will be met by selecting observation periods between 5 and 15 minutes.

Traffic demand observations do also have a cumulative distribution function, survival function and p.d.f., denoted respectively with $G(q)$, $K(q)$ and $g(q)$. Their shape strongly depends on the total observation time and hours of the day selected for analysis, therefore the choice of a functional form and estimation of its parameters is not relevant. The problem now is to estimate the actual – unknown – capacity distribution given the capacity survival function $S(q)$ and p.d.f. $f(q)$. Then, for a sample the likelihood is:

$$L = \prod_{i=1}^{n} f\left(q_i\right)^{1-\delta_i} S\left(q_i\right)^{\delta_i} \tag{6.8}$$

This expression can be easily understood by noticing that when an observation q_i is a capacity observation ($\delta_i = 0$), this will occur with probability $f\left(q_i\right)$; if not ($\delta_i = 1$), then the flow is less than the capacity, which occurs with probability $\Pr\left(C \geq q\right) = S\left(q\right)$. In other words, each observed capacity contributes a term $f(q)$ to the likelihood, and each below-capacity volume contributes a term $S(q)$.

Parametric Product Limit Method

Let θ denote the parameters of the capacity distribution function. Then, the likelihood function for sample q_i can be written as a function of θ. By maximizing the likelihood we can estimate the parameters θ of the capacity distribution, i.e., we can determine the parameters θ of the probability distribution function $f\left(q\right) = f\left(q;\theta\right)$ – and thus also of the survival function $S\left(q\right) = S\left(q;\theta\right)$ – as follows

$$\theta = \arg\max\left\{L\left(\theta\right)\right\} \tag{6.9}$$

In most cases, we would use the natural logaritm of the likelihood function $L\left(\theta\right)$, i.e.

$$\tilde{L}\left(\theta\right) = \ln L\left(\theta\right) = \sum_{i=1}^{n}\left\{\left(1-\delta_i\right)\ln f\left(q_i;\theta\right) + \delta_i \ln S\left(q_i;\theta\right)\right\} \tag{6.10}$$

To use this expression, one first needs to determine a good functional form of the probability density function of the capacity. In most cases, a Normal distribution with mean μ and standard deviation σ (i.e. $\theta = (\mu, \sigma)$) will be a good first approximation.

Non-parametric Product Limit Method

For roadway capacity estimation a *non-parametric form* of the capacity survival function S may be used, rather than the parameterised one presented in the previous sections. This is prefered over the parameterised one, since there is no real evidence for the choice of a particular functional form of the capacity survivor function. If there have been only capacity flow observations in a sample of size n, the empirical survival function $\hat{S}_n\left(q\right)$ is defined as:

$$\hat{S}_n\left(q\right) = \frac{1}{n}\left\{\text{Number of observations } i \text{ with } q_i \geq q\right\} \tag{6.11}$$

This cumulative frequency function decreases by $1/n$ just after each observed 'lifetime'. One may represent observed capacity values as observed lifetimes and use this equation to estimate the roadway capacity. We may choose the 50% percentile (median) as the representative capacity value, or any other percentile point.

A straightforward approach would involve only using capacity observations ($\delta_i = 0$). The resulting approach would be a special case of the generic approach described here, the Product Limit Method (PLM). When dealing with both censored and uncensored data, some modification is necessary, since the number of capacity values greater than or equal to q will not

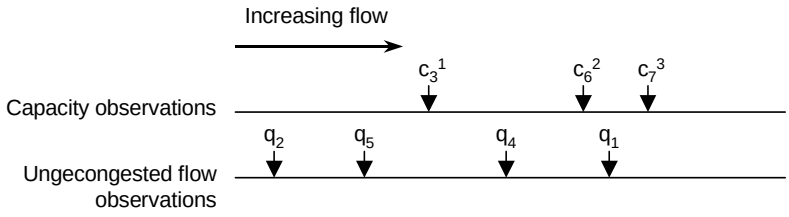

Figure 6.11: Capacity and uncongested flow observations

generally be exactly known. This modified form is called the Product-Limit estimate of the survival function, and is given by

$$\hat{S}_{PLM}(q_i) = \prod_{j=1}^{k} \frac{m_j - d_j}{m_j} \quad j = 1, ..., k \text{ and } 1 \leq k \leq n \tag{6.12}$$

In Eq. (6.12), index j indicates the ordering of the observed capacity flow rates c_i according to increasing size. Since index i indicates the observation period, we denote the size-ordered capacity values c_i with an extended index j, thus c_i^j. Obviously, since there must be at least one capacity observation to apply the method and there will not be more capacity observations than the total number of observation periods, $1 \leq j \leq n$. The total number of capacity observations is denoted with k, and is smaller than or equal to n. Furthermore in equation (6.12), m_j is the sum of

1. the number of capacity observations that *are at least as high as* c_i^j (thus including c_i^j) *and*
2. the number of uncongested observations q_i that have a higher value than c_i^j

At last, d_j represents the number of (exactly equal) capacity values observed at ordering index j, *which is mostly equal to one.* The multiplication is performed from the lowest capacity value observation $j = 1$ to the highest capacity value observation k, thus j increases from 1 to k. *This multiplication makes clear that only at observed capacity flow rates a survival probability is calculated.* The decrease at these points is not equal to $1/n$, as with the empirical distribution, but depends on the number of observations with flow rates q higher than capacity value c_i^j. This number m_j includes higher-value censored observations q_i, but higher capacity values as well.

Let us illustrate formula (6.12) with a sample of 7 observations (in Fig. 6.11 below). Four uncongested flow rates q were observed, and three capacity flows c. Thus $n = 7$ and $k = 3$. It appears that $d_j = 1$ since no equal capacity observations occurred. When applying equation (6.12) to the example shown in Fig. 6.11, the first calculation to be performed is for capacity observation c_3^1. This is the *lowest observed capacity value*, denoted with index $j = 1$. The observation was made in period 3. At this capacity value, there are *two uncongested flow rates* (q_2 and q_5) observed, and two capacity values (c_6^2 and c_7^3) having higher flow rates. Thus, including the observed capacity value, this means $m_1 = 5$. Now the probability that the capacity is at least as high as value c_3^1 can be calculated using Eq. (6.12), and equals $\frac{4}{5}$. For the next capacity value c_6^2 the same procedure applies. Now $m_2 = 3$, which would yield a survival probability of $\frac{4}{5} \cdot \frac{2}{3} = \frac{8}{15}$. The last calculation to be performed in this example is for capacity observation c_7^3. Since $m_3 = 1$ the multiplication result in a zero value, which is according to the expectations.

The motivation for the (discrete) Product-Limit estimate is essentially the same as that for the continuous approach. That is, the estimate $\hat{S}_{PLM}(c)$ is built up as a product, and each term in the product can be thought of as an estimate of the conditional probability of 'surviving' capacity flow rate c_i^j, given survival till just prior to c_i^j. It will be noted that when there is no censoring, the equation reduces to the ordinary empirical survivor estimate $\hat{S}_n$ given above.

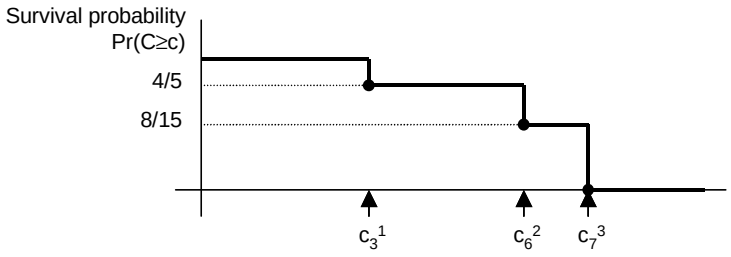

Figure 6.12: Estimated survival probabilities using PLM method.

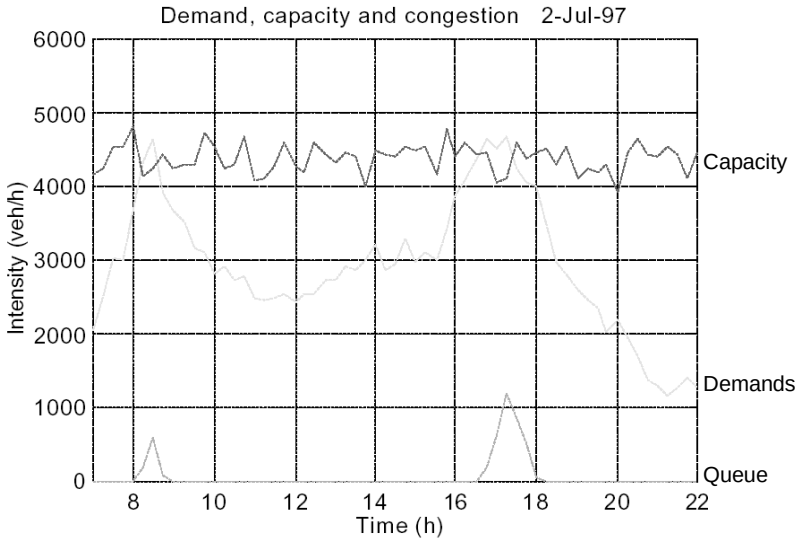

Figure 6.13: Simulation example to application of the PLM method

To effectively assess results when using PL estimates it is desirable to have an estimate of the variance. It can be shown that the variance estimate for this estimate equals

$$var\left[\hat{S}_{PLM}\right] = \hat{S}^2_{PLM} \sum_{j=1}^{k} \frac{d_j}{m_j\left(m_j - d_j\right)} \tag{6.13}$$

6.9.2 Example application of the PLM method

In order to demonstrate the validity of the PLM we are forced to use a simulated case. Fig. 6.13 shows three curves, respectively, capacity, traffic demand and congestion (queue) curve describing the flow characteristics of a sample day for the constructed case. The applied mean capacity value was set at 4400 vehicles/hour, which is a representative value for two-lane freeways in the Netherlands. A normal distribution with a standard deviation of 5% was chosen to generate stochastic road capacity values for each 15-minute period. The 15-minute interval is considered long enough to assume identically and independently distributed capacity values.

Apart from the capacity, the applied traffic demand curve has stochastic characteristics as well. However, account is taken of the correlation between the deviation from the mean demand in succeeding time intervals, smoothing the path. Queueuing directly results from the excess demand with respect to the prevailing capacity. The impact of the stochastic nature of both capacity and traffic demand on the congestion severity yields that capacity values are in the range of 3800 to 5000 veh/h. Therefore, traffic demand sometimes exceed capacity in which case congestion does occur.

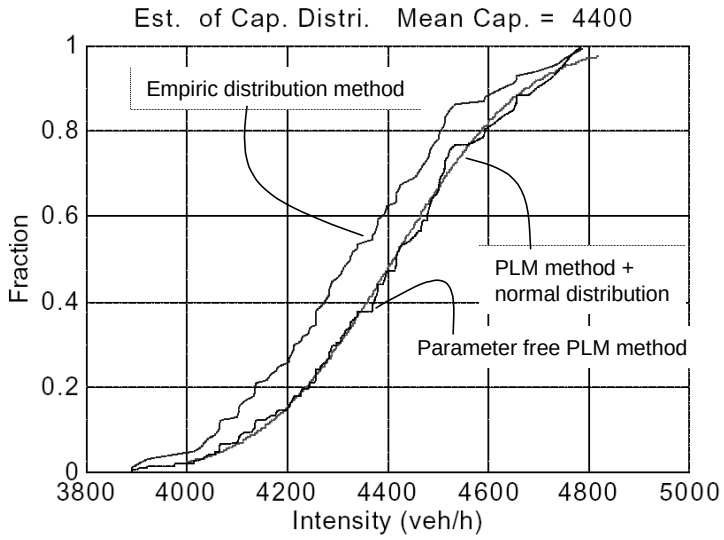

Figure 6.14: Estimates for the capacity distribution functions based different estimation methods.

The capacity value distribution was estimated based on three methods. The 'measurement' configuration is according to that of one cross-section in the bottle-neck section. Since no backward wave disturbance was simulated, it was suitable for our analysis purpose. The measurement interval equals that of the generated traffic data, that is 15-minutes.

In the simulated example, about 8% of the observed flow rates were capacity flows. The range of capacity flow values strongly overlaps with the higher range of uncongested flow values.

Three capacity distributions were calculated:

- Empirical Distribution, which is the distribution of all observed capacity flow rates;
- Estimated distribution of capacity using PLM and normally distributed capacity values;
- Estimated distribution of capacity using PLM and a parameter free distribution of capacity values;

These three estimated cumulative distributions are shown in Fig. 6.14 where one can read from the figure the differences in the capacity estimate at any desired fraction. In order to be regarded as a sound outcome, the estimated capacity values at fraction 0.5 should be close to our input value of 4400 veh/h. By looking at the three values; the empirical, PLM-normal and PLM-parameter free curves, median capacity values of approximately 4300, 4400, 4400 veh/h respectively can be found. It can be concluded that – the easily and generally applied – empirical distribution of capacity values *underestimates the 50%-percentile true value significantly.* Both forms of the Product-Limit Method, however, result in good estimations of mean capacity.

6.9.3 Other applications of the PLM method

The PLM method is not only applied for capacity estimations: it is generally applicable to problems where censored observations are present. Consider for instance the notion of *free speeds of a population of drivers*: for some drivers, we know that they are driving at their free speed, while for other drivers (constrained by the vehicle in front), we may assume that their free speed is higher than their current speed. The main problem is then to distinguish between censored (constrained drivers) and uncensored (free drivers) data, since the estimation results will be sensitive to a correct disinction between the two. A similar problem occurs in *gap acceptance analysis*: only the gaps that are accepted are monitored.

6.10 Summary

This chapter serves as an introduction to the analytical techniques for capacity and level-of-service determination of critical elements of the motorway system. Heavy emphasis has been placed on the use of the 1985 Highway Capacity Manual. However, the field of capacity analysis is not limited to motorway facilities but also includes other land transport modes as well as air and water transportation.