# Chapter 7

# Queuing analysis

*Summary of the chapter.* In this chapter the deterministic queuing model will be introduced briefly and applied to determine congestion probabilities at a single bottle-neck over a year. This is of importance because the Dutch method to assess quality of traffic operation on motorways is based on this model.

## List of symbols

| | | |
|---|---|---|
| $D$ | $veh/s$ | arrival rate |
| $C$ | $veh/s$ | capacity / queue discharge rate |
| $P_c$ | - | fraction of vehicle experiencing congestion |
| $R_t$ | $s$ | collective delays |
| $R_{mean}$ | $s/veh$ | mean delay per vehicle |
| $F$ | - | demand multiplication factors |
| $\varepsilon$ | - | errors |
| $f$ | - | capacity multiplication factors |
| $n$ | $veh$ | number of vehicles in queue |

## 7.1 Deterministic queuing theory

The queueing model to determine delay is not a realistic description of the real traffic process, the main deviation being that vehicles are stored vertically in a queue. Nevertheless with this model the delay can be calculated correctly. In chapter 2, we have introduced the notion of the cumulative vehicle plot and its applications. Let us briefly recall some of the key elements of this chapter, and introduce some new ones along the way. Recall that $\tilde{N}(x_1, t)$ denotes the arrival curve at some cross-section $x_1$. $\tilde{N}(x_2, t)$ denotes the departure curve at some other cross-section $x_2$. The arrival time of vehicle $i$ at cross-section $x_1$ is denoted by $A^{-1}(i)$; its departure time by $D^{-1}(i)$. Clearly, the travel time of vehicle $i$ equals – assuming that overtaking is not possible –

$$R_i = N^{-1}(i; x_2) - N^{-1}(i; x_1) \tag{7.1}$$

The *collective travel times* for all vehicles $i = 1, ..., N$ then equals

$$R = \sum_{i=1}^{N} R_i = \sum_{i=1}^{N} \left[ N^{-1}(i; x_2) - N^{-1}(i; x_1) \right] \tag{7.2}$$

which approximately equals the area between the arrival and the departure curve (i.e. neglecting the small errors made at the boundaries), i.e.

$$R = \int_{t=0}^{T} \left[ N(x_1, s) - N(x_2, s) \right] ds \tag{7.3}$$

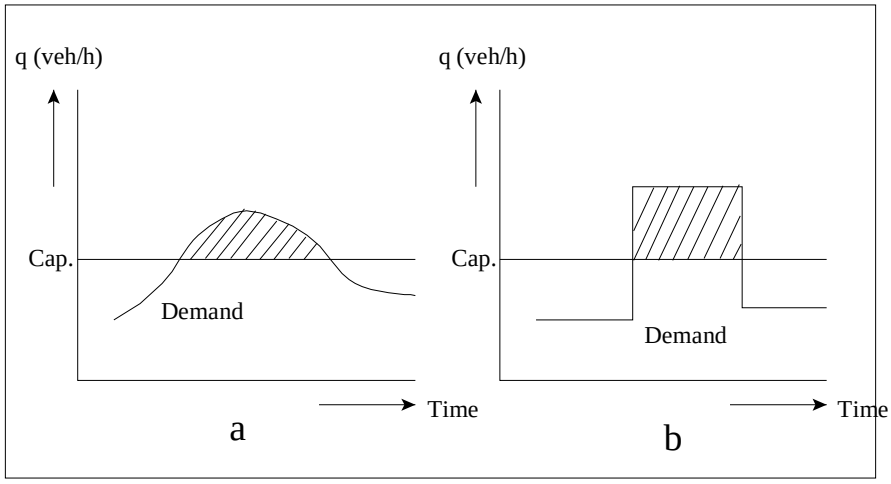


Figure 7.1: Temporarily oversaturation: (a) realistic; (b) schematised

Note that overtaking does not compromise the determination of the collective travel times $R$, but overtaking does imply that the individual travel times cannot be adequately determined. In chapter 2 we have also shown how we can determine the delays experienced by drivers / collective delays by introducting the virtual departure curve. This curve is equal to the arrival curve, translated along the $t$ axis.

## 7.2 Determination of delay with a queueing model

Let us consider a bottle-neck situation sketched in Fig. 7.2. As long as the demand flow is less than the capacity flow, we can determine the traffic flow conditions using the fundamental diagram. But what happens when the demand exceeds the capacity? Let us consider both situations in detail.

*Situation A: no oversaturation.* As long as the demand $D$ is smaller than the bottle-neck capacity $C$ there is no congestion. On the entire road, free flow conditions exists $q(x,t) = D$. In the bottle-neck, a small delay may be incurred due to the higher densities.

*Situation B: oversaturation.* When the demand $D$ exceeds the capacity $C$, vehicles can not pass the bottle-neck without experiencing delays. The next chapter on shock wave analysis with provide an in-depth discussion of the dynamics of the traffic flow conditions in this situation. However, we can also approximate the queue dynamics by using queuing theory, as is described in the remainder of this chapter.

To this end, it is assumed that all vehicles drive unhindered (without delay) to the position where the bottle-neck begins. They wait at this spot, as said before in a vertical queue, for their turn to drive through the bottle-neck. The growth of the queue depends on the difference between demand $D$ and capacity of the bottle-neck $C$. The queue starts growing when for the first time $D$ is larger than $C$. This process is usually visualised in a graph with time on the horizontal axis and the cumulative number of vehicles on the vertical axis (we have already addresses cumulative arrival curves in chapter 2). The so called 'arrival' curve is a line with demand $D$ as slope. The 'departure' curve has the capacity $C$ as slope; see Fig. 7.3 - at least when there is a queue.

The distance between the arrival curve and the departure curve at a given moment (the vertical distance) equals the length of the queue (expressed in number of vehicles). The distance between the arrival curve and the departure curve for a given arriving vehicle (the horizontal distance) equals the time the vehicle spends in the queue, assuming the queue discipline is 'First in, first out' (FIFO) implying no overtaking.

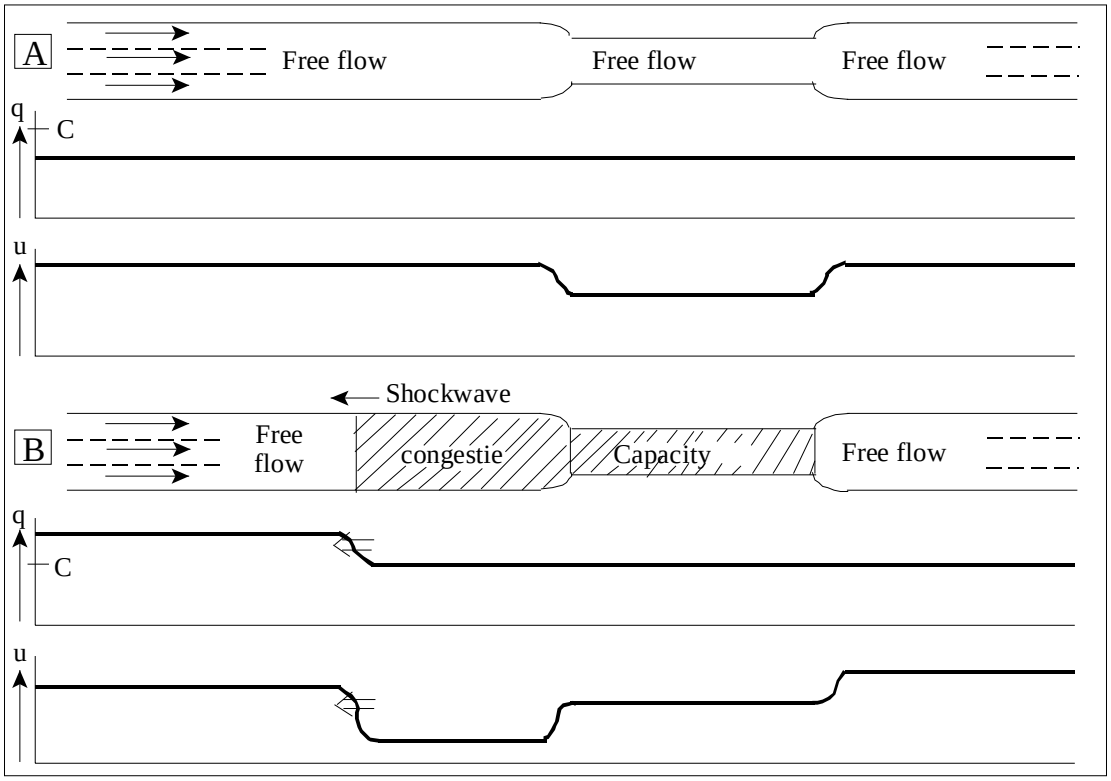


Figure 7.2: Characteristics of a traffic flow on a road with a bottleneck

**Remark 51** *Both the length of the queue and the time spent in the queue are not realistic but refer to the vertical queue.*

When $D$ becomes smaller than $C$, the slope of the arrival curve becomes smaller than the slope of the departure curve. When both curves intersect, the congestion is finished.

The delay equals the area between the arrival and the departure curve, because the area equals the sum of all horizontal distances between the arrival and the departure curve.

It is evident that congestion does not end at the moment $D$ becomes smaller than $C$; the queue that has been built still has to be broken down. The rate of decay, like the growth rate, depends on the difference $D - C$. The (vertical) queue is at its maximum at the moment $D$ becomes smaller than $C$.

This model determines the (collective) delay correctly, but not how the queue is present over the road upstream of the bottle-neck. In reality vehicles experience their delay by slowing down and possibly stopping now and then for short periods. It is important to understand that the delay only depends on the capacity of the bottle-neck and the undisturbed (by definition) demand as a function of time. This holds as long as the congestion itself does not induce extra disturbances which lead to a lower capacity.

Fig. 7.4 illustrates that the delay is independent on the form of the vehicle trajectories in congestion between cross-section $x_1$, upstream of the congestion, and cross-section $x_2$, the begin of the bottle-neck. The delay is the difference between the moment the vehicle drives through the beginning of the b-n and its virtual moment of arrival at the same cross-section. From the figure can also be concluded that overtakings in congestion do not change the collective delay. The smaller delay of the overtaker is compensated by the larger delay of the vehicle being overtaken.

**Remark 52** *A part of the delay is not accounted for by the queueing model. It is the delay due to the fact that the vehicles usually have to decelerate when they enter the b-n. However, if the queueing has any relevant size, then this extra delay can usually be neglected.*

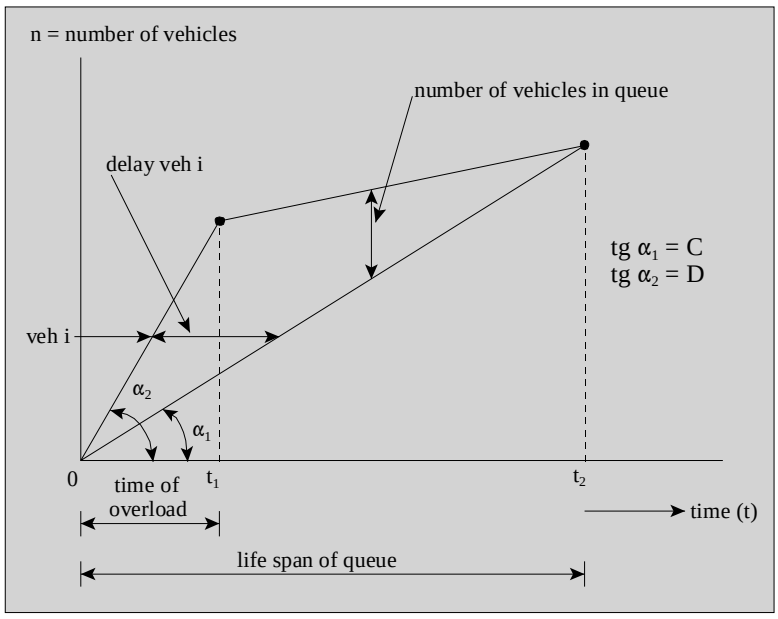


Figure 7.3: Modelling with vertical queues. $C$ denotes the capacity of the bottleneck and $D$ denotes the traffic demand during 'overload'.

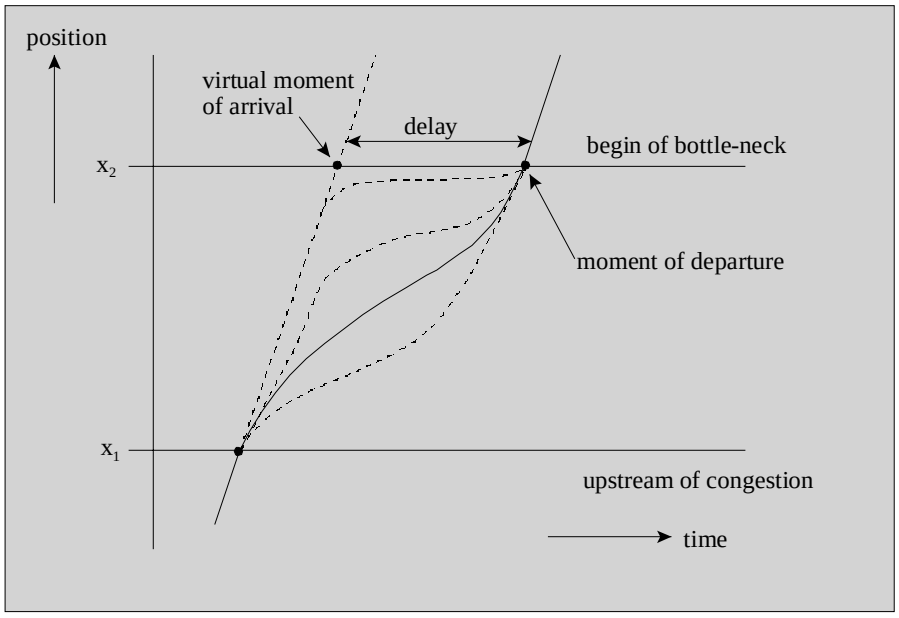


Figure 7.4: Possible vehicle trajectories in congestion.

The relatively simple queueing model has been used in a Dutch approach to take account of congestion and its fluctuation over a year in deciding how much capacity is needed. It will be discussed in the remaining part of this chapter. The main drawback of the approach is that the queue is modelled as a vertical queue, i.e. the amount of space taken up by the queue is not considered explicitly.

### 7.2.1 Computations with the queuing model

With the deterministic queuing model, we can derive several relations that provide insight into the characteristics of queuing and congestions.

*Situation:* from time $t = 0$ the capacity is exceeded. This oversaturation lasts until $t = t_1$, after which the demand is reduced to some value below $C$. This means: between 0 and $t_1$, the queue increases; after $t_1$ the queue reduces again. Let $N_D(t)$ denote the cumulative arrival curve; $q_1$ denotes the flow rate between 0 and $t_1$; $q_2$ denotes the flow rate after $t_1$. $C$ denotes the capacity. We aim to determine: the time $t_2$ at which the queue has dissappeared; the duration $T$ of congestion; the number of vehicles $M$ that experienced congestion; the collective delay $R$, the mean delay $R_{mean}$ and the maximum individual delay $R_{\max}$.

Computational steps:

1. Auxilary point (cumulative vehicle count at time $t_1$): $N(t_1) = t_1 D_1 = t_1 \tan\alpha_1 = t_1 q_1$
2. Maximum queue length $Max = N(t_1) - Ct_1$
3. Maximum delay (at time $t_1$) via

$$\tan\alpha_1 = \frac{Max}{R_{\max}} \qquad R_{\max} = \frac{Max}{C} = \frac{t_1(q_1 - C)}{C} \tag{7.4}$$

4. Duration of congestion $T$ by intersection of the cumulative demand curve

$$N_D(t) = \begin{cases} q_1 t & 0 < t < t_1 \\ q_1 t_1 + q_2(t - t_1) & t > t_1 \end{cases} \tag{7.5}$$

and supply curve

$$N_C(t) = Ct \tag{7.6}$$

yields

$$q_1 t_1 + q_2(T - t_1) = CT \rightarrow T = \frac{(q_1 - q_2)\, t_1}{C - q_2} \tag{7.7}$$

5. Total number of vehicles $M$ that has been in the queue

$$M = CT \tag{7.8}$$

6. Total collective loss $R$

$$R = \frac{Max}{2} T \tag{7.9}$$

or

$$R = \frac{M}{2} R_{\max} \tag{7.10}$$

7. Mean time loss

$$R_{mean} = \frac{R}{M} \tag{7.11}$$

or

$$R_{mean} = \frac{1}{2} R_{\max} \tag{7.12}$$

Note that $R_{mean}$ does not depend on $q_2$. Futhermore, it turns out that the collective loss $R$ increases quadratically with the duration of the oversaturation, but the individual loss $R_{mean}$ increases linearly. When $q_2$ is only slightly smaller than $C$ then

- the duration of the queue will be high
- there will be many vehicles that experience travel time delays
- the collective waiting time $R$ will be large
- the mean delays will not increase

Therefore, it can be concluded that when after congestion in the peak hour, the demand is only slightly smaller than the bottle-neck capacity, the congestion may last for a significant period of time.

## 7.3 QUAST-method to take account of variability of congestion

Congestion has existed as long as there has been traffic and transportation. Congestion occurs when the demand for the use of a system is larger than its capacity. Delay, often unexpected and unpredictable in size, is a characteristic of congestion. Because of this characteristic, congestion leads to uncertainties in the planning of activities and transportation, the use of sometimes unnecessary time margins, or to unexpected delay. Congestion on motorways is a 'relatively new phenomenon'. During the time that the main part of the motorway system was built, the missing links in the motorway network formed the bottle-necks of the national network. Even then there were days when a great deal of congestion took place on the motorways, for example on Easter Monday and Whitmonday. It was soon decided that the capacity of the network did not need to be so high as to be able to facilitate such exceptional demand patterns.

Nowadays congestion seems to attract a lot of attention. People fear grid-lock; in more objective terms they think that the accessibility of important centres (sea ports, airports, business areas in cities, etc) will degrade too much.

One of the decisions that has to be made in the design of a road is the level of service that should be offered. A commonly used rule is that a road should be able to facilitate the predicted design traffic volumes in a design year, with a certain minimum Level of Service. Since around 1970 a cost-benefit analysis in the Netherlands led to the rule: offer a Level of Service of C on inter-regional roads. This comes down to a capacity of about 4/3 of the design demand and will provide a situation in which travelling speeds are high, and freedom to manoeuver and driving comfort are also at an acceptable level.

After some time it became clear that building new roads did not (by far) keep up with the growth of the demand and congestion on motorways has become a daily returning phenomenon. This resulted in a study by working group QUAST (derived from QUAlity and STructure of the main road network) from Rijkswaterstaat (Part of Ministry of Transport) about what level of capacity should be offered . It was found that a certain amount of congestion led to an optimum in terms of costs and benefits (elements: costs of building roads, delay, accidents, and road-maintenance); see [52].

In this study the 'probability to experience congestion' was chosen as criterion for the quality of traffic operation. This is the fraction of daily road users that experiences congestion on a link. This fraction is determined as an average over all working days of a year. New in this approach is that a year is no longer represented by a single representative value. Instead, fluctuations of all kinds of factors during a year have influence; see [50].

The economic optimum appeared to be at the point of 2% probability of congestion. Later, this standard was relaxed. Whereas the 2% probability of congestion is only valid for hinterland-connections, for the rest of the main road network a 5% probability of congestion is accepted. In

addition, these are national normative values. In fact one should make a cost-benefit analysis for every section and corridor. In areas where the building costs of roads are high, the acceptable level of congestion could be higher.

### 7.3.1 QUAST-model

Needed is a calculation of the characteristics of congestion as a function of demand and capacity. This has been carried out using a deterministic queueing model with vertical queues.

The task of this model is to calculate:

- the fraction of vehicles that experience congestion ($P_c$);
- the extra travel time or delay for vehicles that experience congestion, collectively ($R_t$) as well as the mean value per vehicle ($R_{mean}$);

based on the demand pattern over a year, $D(t)$, and a capacity pattern, $C(t)$.

The queueing model with vertical queue is especially suitable to fulfill this task. It is applied to those network parts that are bottle-necks, i.e. road elements that determine the capacity of a section.

The model confronts the demand pattern and the capacity pattern with each other and calculates the resulting congestion; see Fig. 7.5. The demand pattern is represented by the fluctuations of demand over the 24 hour period of the working days of a year. Over the hours of a day the demand varies; there is a morning peak period and an evening peak period. The demand $D(t)$ also systematically varies with the day of the week (factor $F_{day}$) and the month of the year (factor $F_{month}$).

Finally there are random fluctuations $\varepsilon(t)$); given a time of the day, on a given day of the week, in a given month of the year, there is a mean value for the demand intensity. A random variable $\varepsilon_D(t)$ is added to this mean value; it has a mean value of 0 and a normal distribution with a given standard deviation. This term will further be referred to as noise.

$$D(t) = D_0(t) F_{month} F_{day} + \varepsilon_D(t) \tag{7.13}$$

Comparison of the demand patterns generated with formula (7.13), with the results of [3], showed that the demand pattern was fluctuating too wildly. This problem was solved by adding noise with interdependency (correlated noise). It is plausible that if the demand during period $i$ is higher than the average value, it is likely (probability $> 0.5$) that it is also higher than average in the next period $i+1$. This can be achieved by smoothing independent random variables $\varepsilon_0$ with the model ($\alpha$ = smoothing factor; $0 \leq \alpha \leq 1$):

$$\varepsilon_D(t) = \alpha \varepsilon_D(t-1) + (1-\alpha)\varepsilon_0(t) \tag{7.14}$$

At first glance one might think that the capacity pattern is a constant. However, this is not the case, certainly not under changing weather (factor $F_{weather}$) and light conditions (factor $F_{light}$). It is known that with rain road capacity reduces by about 10% and in darkness capacity reduces by about 5%. Rain is distributed randomly in time, whereas darkness is a systematic factor.

Similar to the demand, the capacity also has a non-systematic or random component. Under ideal and sustaining circumstances the capacity still varies. This can be explained by the different characteristics of vehicle-driver elements. However, these characteristics can not be observed and their influences are not known, but it is clear that they affect capacity; see [56].

$$C(t) = C_0 f_{weather} f_{light} + \varepsilon_C(t) \tag{7.15}$$

The random component $\varepsilon_C(t)$ has a mean value of 0 and a normal distribution with a standard deviation equal to 6 % of capacity. In contrast to the noise term of the demand, the noise of the capacity is not correlated.

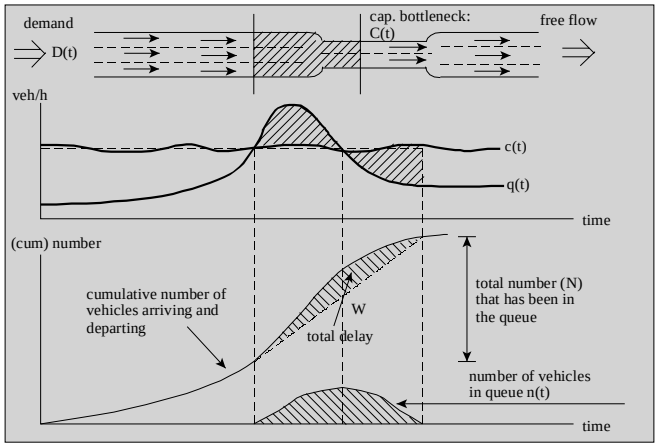


Figure 7.5: Queueing model with vertical queue

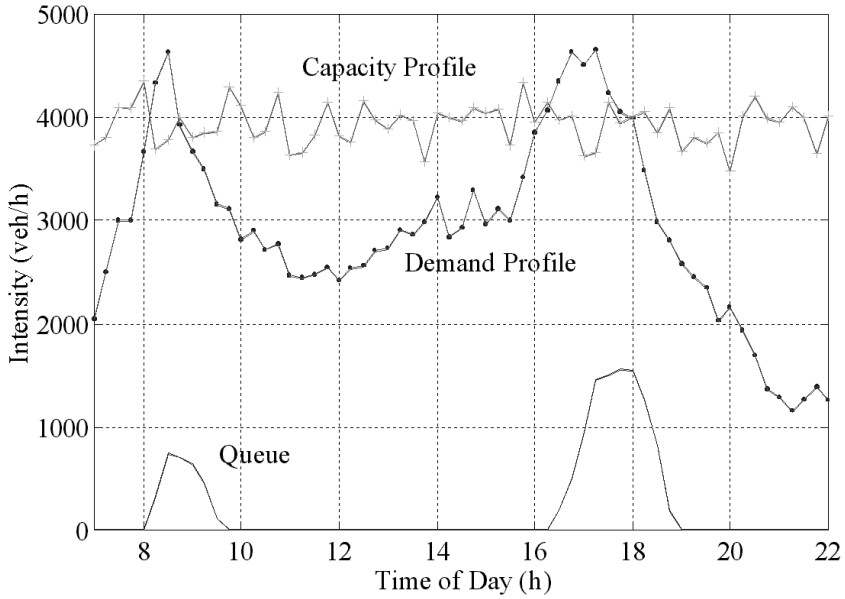


Figure 7.6: Results of a simulation for a day.

**Remark 53** *The model (7.15) has not been used literally but in a simplified way. Research in the scope of the QUAST study has led to the following rule of thumb: 'bad weather (including the effect of darkness) occurs during 10% of the peak hours and brings about a reduction of capacity of 12%'. This rule has been interpreted even more deterministically in this model. On the first Tuesday and Wednesday of every month of 20 days bad weather is assumed. Tuesday and Wednesday were chosen because the intensities on these days differ the least from the average daily value. This procedure achieves the fraction of 10% of time bad weather per year exactly. In practice the effect of bad weather and darkness will lead to more varying outcomes than this model produces.*

Fig. 7.6 shows the results of a simulation over a day. The capacity and demand are depicted as a function of the time of a day. There are short periods with demand higher than capacity, resulting in a queue indicated at the bottom line. Note that the demand and capacity are intensities and the queue is a number of vehicles; consequently the scale of the queue in Fig. 7.6 is arbitrary.

Because the pattern of the demand and capacity is now variable, the calculations have to

| Factor | Cap. increase (%) |
|---|---|
| Noise on capacity | 2.9 |
| Noise on demand | 3.7 |
| Month factor | 1.9 |
| Day factor | 1.1 |
| Bad weather | 2.6 |
| All factors | 9.9 (sum = 12.2) |

Table 7.1: Effect of different factors on roadway capacity

be carried out per time step (15 minutes), using the equation:

$$n\left(t\right) = \int_{t_1}^{t} \left(D\left(s\right) - C\left(s\right)\right) ds \tag{7.16}$$

The number of vehicles in the queue $n$ is the integral of the over saturation $D(t) - C(t)$ and the integration starts at moment $t_1$ as $D(t)$ for the first time becomes larger than $C(t)$.

The collective delay $R_t$ is the integral of $n(t)$ over time.

$$R_t\left(t\right) = \int_{t_1}^{t} n\left(s\right) ds \tag{7.17}$$

## 7.4 QUAST application results

The remainder of this chapter presents some results of application of the QUAST method.

### 7.4.1 Effect of factors on required capacity

Investigations with the model have begun with a strictly deterministic case: a fixed demand pattern and a constant capacity, chosen at a level such that $P_c$ equals 5%. The 'basic capacity' needed to obtain this $P_c$ value, given the demand pattern, is denoted as $C_0$. In next steps other sources of variation are added to the model. It turns out that the addition of every factor leads to a higher $P_c$. Increasing the capacity by a certain amount, $P_c$ can be reduced to its original value of 5 %. It is this increase, expressed as a multiplier, that has been chosen as a yard stick for the influence of the factor. Table 7.1 presents an overview of the effects of the factors investigated. It appears that the sum of the separate effects is larger than the combined effect of all factors together. Further, there are no factors significantly more or less important than others. Consequently, it can not be concluded that one or several factors can be neglected.

The calculated delay in the most extended model at the normative $P_c = 5\%$, is clearly less than the values given in publications on outcomes of QUAST (2 min against 4-5 min). Probably the explanation for this difference is that in the QUAST method of Rijkswaterstaat, the value found for $R_{mean}$ is multiplied by a factor of 2 in order to account for other factors that lead to delay.

### 7.4.2 Distribution of congestion characteristics

Besides the mean values of $P_c$, $R_{mean}$ and $R_t$, their distributions over the days of a year are also of importance. Especially the travel time reliability of the system of main roads is closely related to them. Fig. 7.7 shows the (cumulative) distributions functions where all factors are included in the model and with an overall value for $P_c$ of 5%.

It appears that all three distributions have the same character. There are many zero or nearly zero values and the tail is long. This means high values appear with a low frequency. These results are not convenient, because it means that congestion is hard to forecast. In other

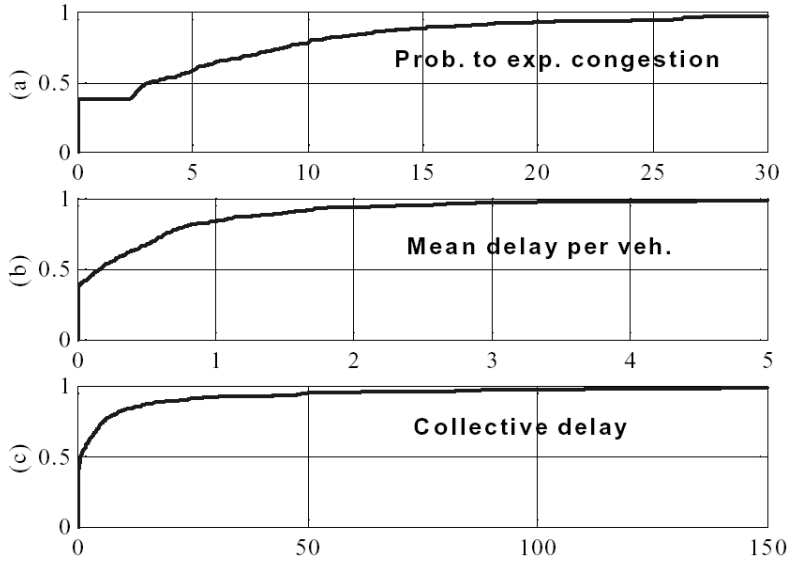


Figure 7.7: Statistical distribution functions over working days of a) $P_c$ (probability of experiencing congestion); b) $R_t$ (collective delay) and c) $R_{mean}$ (average delay per vehicle that experiences congestion).

words its unpredictability is relatively high. $P_c$ and $R_{mean}$ both show about 30% zero-values. $R_t$ even shows 50% zero-values. This means that the collective delay is nearly zero on 50% of the days.

The scale on the axis of $R_t$ requires some explanation. This scale is in minutes per percent of $AADT$; this means that the value should be multiplied by $(1/60)AADT/100$ in order to find vehicle-hours.

**Example 54** *$AADT = 50.000$ vehicles. The point on the scale at 100 corresponds to a value of 833 vehicle-hour. If the delay is valued 10 Euro per hour, then it represents an amount of Euro 8333.*

Lastly, the mutual dependence of the characteristics $P_c$, $R_t$ and $R_{mean}$ is important. The interdependence between $P_c$ and Rmean (see Fig. 7.6a) is nearly linear, but the scatter is large. The dependence between the other two (see Fig. 7.6b) is clearly not linear. $R_t$ is small for $P_c <$ 10% and beyond this point grows more rapidly than linear with $P_c$. The dependence between $R_{mean}$ and $R_t$ can easily be concluded from the other two dependencies.

## 7.5 Final remarks

Some remarks are due about the model application:

- Calculations were made with 15-minute time-steps, whereas the Ministry of Transport used 1-hour time steps.
- For the sake of simplicity calculations were made with 12 months of 20 working-days.
- Simulations were made for the time period between 7.00h and 22.00h. Congestion will not occur outside this time period in this model. The nighttime intensity has been taken into account in the 24-hour value.
- A simulation over a period of one year (240 working days) is usually not enough to get stable outcomes. One should average over a few years.

The analysis of the findings obtained with the model is far from complete. Some issues can be brought forward:

- Most analysis has focussed on the probability to experience congestion; more attention should be given to the collective and individual delay.
- The deterministic demand-pattern over the time of a day has not been varied; it is a pattern with two peak periods that differ slightly. Research should be done into the effects of this choice.
- The sensitivity of the results for change in the parameters should be determined.
- As far as the model is concerned, the factors for darkness and rain might be introduced separately.

Finally it should be born in mind that capacity in fact can only be attuned to the demand to a very limited extent. In most cases a road section capacity can only be adjusted by varying the number of lanes. The main goal of the preceding analysis has been to gain more insight into the several factors that contribute to congestion on a yearly basis and to illustrate the fluctuations in congestion. Apart from over saturation there are other factors that can cause congestion, e.g. exceptional weather conditions (dense fog, storms), road works, incidents and accidents.

Drivers can react to congestion experienced in the past by changing their travel behaviour. On a short term the change of departure time is the most likely reaction. This could lead to even more fluctuations in congestion, if the reaction is badly timed, e.g. drivers react to the congestion they experienced yesterday. In fact the process of how drivers respond to congestion is not yet very well understood.