

Chapter 2

Transportation Supply Models

2.1 Introduction

This chapter deals with the mathematical models simulating transportation supply systems. In broad terms a transportation supply model can be defined as a model, or rather a system of models, simulating the performances and flows resulting from user demand and the technical and organizational aspects of the physical transportation supply.

Transportation supply models combine traffic flow theory and network flow theory models. The former are used to analyze and simulate the performances of the main supply elements, the latter to represent the topological and functional structure of the system. Therefore, in Sect. 2.2 we present some of the basic results of traffic flow theory. Section 2.3 covers the constituent elements of a transportation network supply model: such elements form an abstract model of transportation supply (transportation network) which combines network flow theory with the functions that express dependence between transportation flows and costs on the network. This is followed by some general indications on the applications of network models in Sect. 2.4. Specific models for transportation systems with *continuous* services (such as road systems) are described in Sect. 2.4.1; models for *discrete* or *scheduled services* (such as bus, train, or airplane) are described in Sect. 2.4.2. Throughout this chapter, as stated in Chap. 1, it is assumed that the transportation system is intraperiod (within-day) stationary (unless otherwise stated); extensions of supply models to intraperiod dynamic systems are dealt with in Chap. 7.

2.2 Fundamentals of Traffic Flow Theory¹

Models derived from traffic flow theory simulate the effects of interactions between vehicles using the same transportation facility (or the same service) at the same time. For simplicity's sake, the models presented refer to vehicle flow, although most of them can be applied to other types of users, such as trains, planes, and pedestrians. In the sections below we describe stationary uninterrupted flow models (nonstationary models are introduced in Chap. 7), followed by models of interrupted flow, derived from queuing theory.

¹Giulio Erberto Cantarella is co-author of this section.

2.2.1 Uninterrupted Flows

Multiple vehicles using the same facility may interact with each other and the effect of their interaction will increase with the number of vehicles. This phenomenon, called *congestion*, occurs in most transportation systems, generally worsening the overall performances of the facility, such as the mean speed or travel time. Indeed, it may happen that a vehicle is forced to move at less than its desired speed if it encounters a slower vehicle. The higher the number of vehicles on the infrastructure, the more likely this condition is to happen. This circumstance may also occur in transportation systems with scheduled services: the higher the number of vehicles on the infrastructure, the more likely out-of-schedule vehicles are to cause a delay to other vehicles.

In general, stochastic models may be used to characterize in a probabilistic sense an interaction event that causes a delay. For congested systems with continuous services it is very often sufficient to adopt the aggregate deterministic models described below; they may be applied in areas far away from interruptions such as intersections and toll booths.

2.2.1.1 Fundamental Variables

Several variables can be observed in a *traffic stream*, that is, a sequence of cars moving along a road segment referred to as a link, a . In principle, although all variables should be related to link a , to simplify the notation the subscript a may be implied. The fundamental variables are as follows (see Fig. 2.1).

τ	The time at which the traffic is observed
L_a	The length of road segment corresponding to link a
s	A point along a link, or rather, its abscissa increasing (from a given origin, usually located at the beginning of the link) along the traffic direction ($s \in [0, L_a]$)
i	An index denoting an observed vehicle
$v_i(s, \tau)$	The speed of vehicle i at time τ while traversing point (abscissa) s

For traffic observed at point s during time interval $[\tau, \tau + \Delta\tau]$, several variables can be defined (see Fig. 2.1) as follows.

$h_i(s)$	The headway between vehicles i and $i - 1$ crossing point s
$m(s \tau, \tau + \Delta\tau)$	The number of vehicles traversing point s during time interval $[\tau, \tau + \Delta\tau]$
$\bar{h}(s) = \sum_{i=1, \dots, m} h_i(s) / m(s \tau, \tau + \Delta\tau)$	The mean headway, among all vehicles crossing point s during time interval $[\tau, \tau + \Delta\tau]$
$\bar{v}_\tau(s) = \sum_{i=1, \dots, m} v_i(s) / m(s \tau, \tau + \Delta\tau)$	The time mean speed, among all vehicles crossing point s during time interval $[\tau, \tau + \Delta\tau]$

Similarly, for traffic observed at time τ between points s and $s + \Delta s$, the following variables can be defined.

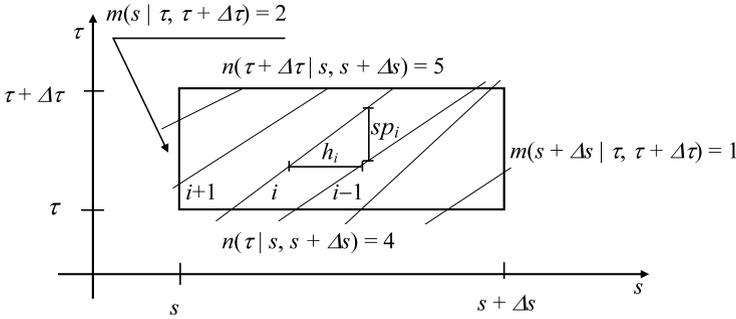


Fig. 2.1 Vehicle trajectories and traffic variables

$sp_i(\tau)$ The spacing between vehicles i and $i - 1$ at time τ

$n(\tau | s, s + \Delta s)$ The number of vehicles at time τ between points s and $s + \Delta s$

$\bar{sp}(\tau) = \sum_{i=1, \dots, n} sp_i(\tau) / n(\tau | s, s + \Delta s)$ The mean spacing, among all vehicles between points s and $s + \Delta s$ at time τ

$\bar{v}_s(\tau) = \sum_{i=1, \dots, n} v_i / n(\tau | s, s + \Delta s)$ The space mean speed, among all vehicles between points s and $s + \Delta s$ at time τ

During time interval $[\tau, \tau + \Delta\tau]$ between points s and $s + \Delta s$, a general flow conservation equation can be written:

$$\begin{aligned} \Delta n(s, s + \Delta s, \tau, \tau + \Delta\tau) + \Delta m(s, s + \Delta s, \tau, \tau + \Delta\tau) \\ = \Delta z(s, s + \Delta s, \tau, \tau + \Delta\tau) \end{aligned} \tag{2.2.1}$$

where

$\Delta n(s, s + \Delta s, \tau, \tau + \Delta\tau) = n(\tau + \Delta\tau | s, s + \Delta s) - n(\tau | s, s + \Delta s)$ is the variation in the number of vehicles between points s and $s + \Delta s$ during $\Delta\tau$

$\Delta m(s, s + \Delta s, \tau, \tau + \Delta\tau) = m(s + \Delta s | \tau, \tau + \Delta\tau) - m(s | \tau, \tau + \Delta\tau)$ is the variation in the number of vehicles during time interval $[\tau, \tau + \Delta\tau]$ over space Δs

$\Delta z(s, s + \Delta s, \tau, \tau + \Delta\tau)$ is the number of entering minus exiting vehicles (if any) during time interval $[\tau, \tau + \Delta\tau]$, due to entry/exit points (e.g., on/off ramps), between points s and $s + \Delta s$

In the example of Fig. 2.1 there are no vehicles entering/exiting in the segment Δs ; then $\Delta z = 0$ (Δn is equal to 1 and Δm is equal to -1).

With the observed quantities two relevant variables, *flow* and *density*, can be introduced:

$f(s | \tau, \tau + \Delta\tau) = m(s | \tau, \tau + \Delta\tau) / \Delta\tau$ is the flow of vehicles crossing point s during time interval $[\tau, \tau + \Delta\tau]$, measured in vehicles per unit of time

$k(\tau | s, s + \Delta s) = n(\tau | s, s + \Delta s) / \Delta s$ is the density between points s and $s + \Delta s$ at time τ , measured in vehicles per unit of length

Flow and density are related to mean headway and mean spacing through the following relations.

$$f(s | \tau, \tau + \Delta\tau) \cong 1/h(s)$$

$$k(\tau | s, s + \Delta s) \cong 1/sp(\tau)$$

Note that if observations are perfectly synchronized with vehicles, the near-equality in the previous two equations becomes a proper equality.

Moreover, if the general flow conservation equation (2.2.1) is divided by $\Delta\tau$, the following equation is obtained.

$$\Delta n/\Delta\tau + \Delta f = \Delta e \quad (2.2.2)$$

where

$\Delta f(s, s + \Delta s, \tau, \tau + \Delta\tau) = \Delta m(s, s + \Delta s, \tau, \tau + \Delta\tau)/\Delta\tau$ is the variation of the flow over space

$\Delta e(s, s + \Delta s, \tau, \tau + \Delta\tau) = \Delta z(s, s + \Delta s, \tau, \tau + \Delta\tau)/\Delta\tau$ is the (net) entering/exiting flow

Finally, dividing by Δs , we obtain a further formulation of (2.2.1) (useful for comparisons with nonstationary models based on the fluid-dynamic analogy described in Chap. 7) that expresses the role of variation in density:

$$\Delta k/\Delta\tau + \Delta f/\Delta s = \Delta e/\Delta s \quad (2.2.3)$$

where

$\Delta k(s, s + \Delta s, \tau, \tau + \Delta\tau) = \Delta n(s, s + \Delta s, \tau, \tau + \Delta\tau)/\Delta s$ is the variation of the density over time

2.2.1.2 Model Formulation

In this subsection we describe several deterministic models developed under the assumption of stationarity, formally introduced below. Extensions to nonstationarity conditions are reported in Chap. 7 (some information on stochastic models is reported in the bibliographical note). In formulating such models it is assumed that a traffic stream (a discrete sequence of vehicles) is represented as a continuous (one-dimensional) fluid.

Traffic flow is called *stationary* during a time interval $[\tau, \tau + \Delta\tau]$ between points s and $s + \Delta s$ if flow is (on average) independent of point s , and density is independent of time τ (other definitions are possible):

$$f(s | \tau, \tau + \Delta\tau) = f$$

$$k(\tau | s, s + \Delta s) = k$$

Note that this condition is chiefly theoretical and in practice can be observed only approximately for mean values in space or time. It is nevertheless useful in that it

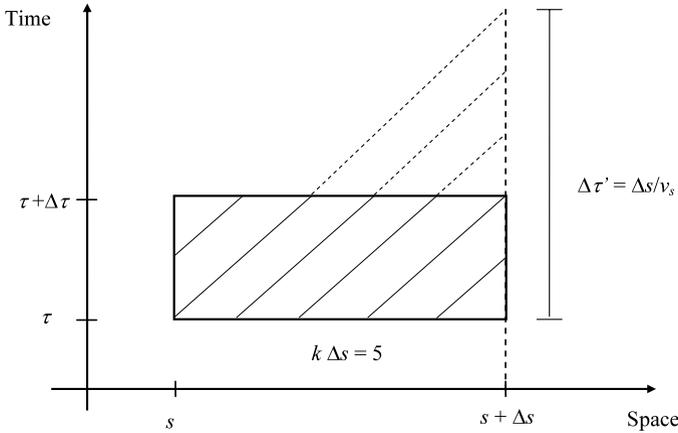


Fig. 2.2 Vehicle trajectories and traffic variables for stationary (deterministic) flows

allows effective analysis of the phenomenon. In this case, the time mean speed is independent of location and the space mean speed is independent of time:

$$\bar{v}_\tau(s) = \bar{v}_\tau$$

$$\bar{v}_s(\tau) = \bar{v}_s$$

In the case of stationarity, both terms in the left side of the conservation equation (2.2.3) are identically null, anyhow other flow conservation conditions may be formulated. Hence, let $n = k \cdot \Delta s$ be the number, time-independent due to the assumption of stationarity, of vehicles on the stretch of road between cross-sections s and $s + \Delta s$, and let \bar{v}_s be the space mean speed of these vehicles. The vehicle that at time τ is at the start of the stretch of road, cross-section s , will reach the end, cross-section $s + \Delta s$, on average at time $\tau + \Delta\tau'$, with $\Delta\tau' = \Delta s / v_s$. Due to the assumption of stationarity, the number of vehicles crossing each cross-section during time $\Delta\tau$ is equal to $f \cdot \Delta\tau$. Thus the number of vehicles contained at time τ on section $[s, s + \Delta s]$ is equal to the number of vehicles traversing cross-section $s + \Delta s$ during the time interval $[\tau, \tau + \Delta\tau']$ (see Fig. 2.2); that is, $k\Delta s = f\Delta\tau' = f\Delta s / v_s$. Hence, under stationary conditions, flow, density, and space mean speed must satisfy the *stationary flow conservation equation*:

$$f = kv \tag{2.2.4}$$

where

$v = \bar{v}_s$ is the space mean speed, simply called speed for further analysis of stationary conditions.²

²It is worth noting that the time mean speed is not less than the space mean speed, as can be shown because the two speeds are related by the equation $\bar{v}_\tau = \bar{v}_s + \sigma^2 / \bar{v}_s$, where σ^2 is the variance of speed among vehicles. In Fig. 2.2 $\sigma^2 = 0$, hence $\bar{v}_\tau = \bar{v}_s$.

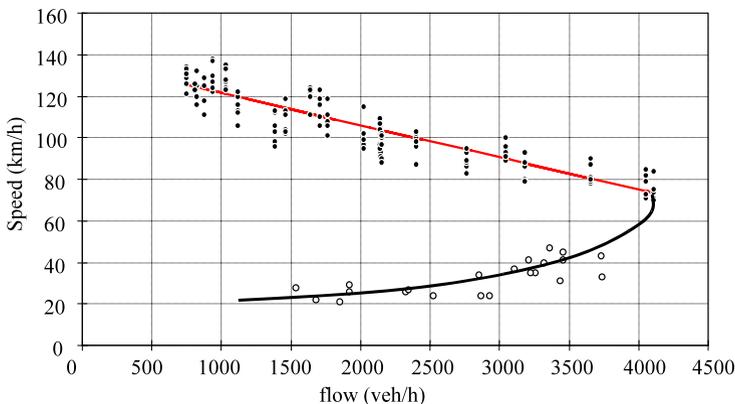


Fig. 2.3 Relationship between speed and flow

In stationary conditions, empirical relationships can be observed between each pair of variables: flow, density, and speed. In general, observations are rather scattered (see Fig. 2.3 for an example of a speed–flow empirical relationship) and various models may be adopted to describe such empirical relationships. These models are generally given the name *fundamental diagram (of traffic flow)* (see Fig. 2.4) and are specified by the following relations.

$$v = V(k) \quad (2.2.5)$$

$$f = f(k) \quad (2.2.6)$$

$$f = f(v) \quad (2.2.7)$$

Although only a model representation of empirical observations, this diagram permits some useful considerations to be made. It shows that flow may be zero under two conditions: when density is zero (no vehicles on the road) or when speed is zero (vehicles are not moving). The latter corresponds in reality to a stop-and-go condition.

In the first case the speed assumes the theoretical maximum value, *free-flow speed* v_0 , whereas in the second the density assumes the theoretical maximum value *jam density*, k_{jam} . Therefore, a traffic stream may be modeled through a *partially compressible fluid*, that is, a fluid that can be compressed up to a maximum value.

The peak of the *speed–flow* (and *density–flow*) curve occurs at the theoretical maximum flow, *capacity* Q of the facility; the corresponding speed v_c and density k_c are referred to as the *critical speed* and the *critical density*. Thus any value of flow (except the capacity) may occur under two different conditions: low speed and high density and high speed and low density. The first condition represents an unstable state for the traffic stream, where any increase in density will cause a decrease in speed and thus in flow. This action produces another increase in density and so on until traffic becomes jammed. Conversely, the second condition is a stable state because any increase in density will cause a decrease in speed and an increase in

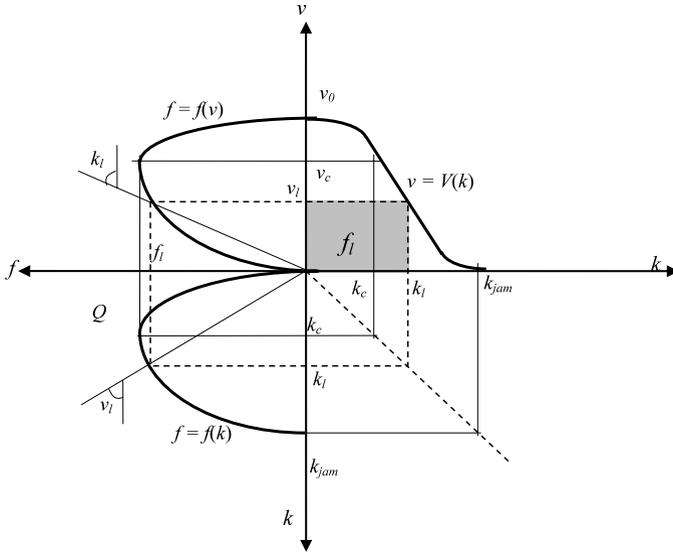


Fig. 2.4 Fundamental diagram of traffic flow

flow. At capacity (or at critical speed or density) the stream is nonstable, this being a boundary condition between the other two.

These results show that flow cannot be used as the unique parameter describing the state of a traffic stream; speed and density, instead, can univocally identify the prevailing traffic condition. For this reason the relation $v = V(k)$ is preferred to study traffic stream characteristics.

Mathematical formulations have been widely proposed for the fundamental diagram, based on single regime or multiregime functions. An example of a single regime function is Greenshields' linear model:

$$V(k) = v_0(1 - k/k_{jam})$$

or Underwood's exponential model (useful for low densities):

$$V(k) = v_0e^{-k/k_c}.$$

An example of a multiregime function is Greenberg's model:

$$V(k) = a_1 \ln(a_2/k) \quad \text{for } k > k_{min}$$

$$V(k) = a_1 \ln(a_2/k_{min}) \quad \text{for } k \leq k_{min}$$

where a_1, a_2 and $k_{min} \leq k_{jam}$ are constants to be calibrated.

Starting from the speed–density relationship, the flow–density relationship, $f = f(k)$, may be easily derived by using the flow conservation equation under station-

ary conditions, or fundamental conservation equation (2.2.4):

$$f(k) = V(k)k$$

Greenshields' linear model yields:

$$f(k) = v_0(k - k^2/k_{\text{jam}})$$

In this case the capacity is given by

$$Q = v_0 k_{\text{jam}}/4$$

Moreover the flow–speed relationship can be obtained by introducing the inverse speed–density relationship: $k = V^{-1}(v)$, thus

$$f(v) = V(k = V^{-1}(v)) \cdot V^{-1}(v) = v \cdot V^{-1}(v)$$

For example, Greenshields' linear model yields: $V^{-1}(v) = k_{\text{jam}}(1 - v/v_0)$ thus

$$f(v) = k_{\text{jam}}(v - v^2/v_0)$$

In general, the flow–speed relationship may be inverted by only considering two different relationships, one in a stable regime, $v \in [v_c, v_o]$, and the other in an unstable regime, $v \in [0, v_c]$. Greenshield's linear model leads to:

$$v_{\text{stable}}(f) = \frac{v_0}{2} \left(1 + \sqrt{1 - 4f/(v_0 k_{\text{jam}})} \right) = \frac{v_0}{2} \left(1 + \sqrt{1 - f/Q} \right)$$

$$v_{\text{unstable}}(f) = \frac{v_0}{2} \left(1 - \sqrt{1 - f/Q} \right)$$

In the particular case that one can assume the flow regime is always stable, with reference to relation $v = v_{\text{stable}}(f)$ the corresponding relationship between travel time t and flow may be defined (some examples of this type of empirical relationship may be found in Sect. 2.4):

$$t = t(f) = L/v_{\text{stable}}(f) \tag{2.2.8}$$

2.2.2 Queuing Models

The average delay experienced by vehicles that queue to cross a flow interruption point (intersections, toll barriers, merging sections, etc.) is affected by the number of vehicles waiting. This phenomenon may be analyzed with models derived from queuing theory, developed to simulate any waiting or user queue formation at a server (administrative counter, bank counter, etc.). The subject is treated below with reference to generic users, at the same time highlighting the similarities with uninterrupted flow.

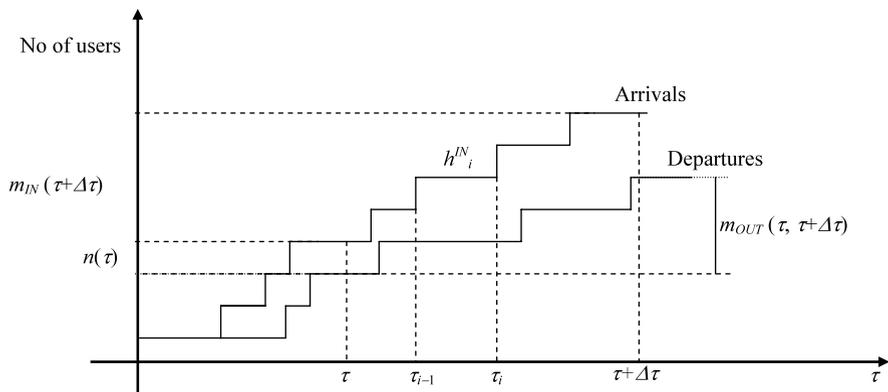


Fig. 2.5 Fundamental variables for queuing systems

2.2.2.1 Fundamental Variables

The main variables that describe queuing phenomena are:

τ The time at which the system is observed

τ_i The arrival time of user i

$h_i = \tau_i - \tau_{i-1}$ The headway between successive users i and $i - 1$ joining the queue at times τ_i and τ_{i-1}

$m_{IN}(\tau, \tau + \Delta\tau)$ Number of users joining the queue during $[\tau, \tau + \Delta\tau]$

$m_{OUT}(\tau, \tau + \Delta\tau)$ Number of users leaving the queue during $[\tau, \tau + \Delta\tau]$

$h(\tau, \tau + \Delta\tau) = \sum_{i=1, \dots, m} h_i / m_{IN}(\tau, \tau + \Delta\tau)$ Mean headway between all vehicles joining the queue in the time interval $[\tau, \tau + \Delta\tau]$

$n(\tau)$ Number of users waiting to exit (queue length) at time τ

With reference to observable quantities, flow variables may be introduced.

$u(\tau, \tau + \Delta\tau) = m_{IN}(\tau, \tau + \Delta\tau) / \Delta\tau$ arrival (entering) flow during $[\tau, \tau + \Delta\tau]$

$w(\tau, \tau + \Delta\tau) = m_{OUT}(\tau, \tau + \Delta\tau) / \Delta\tau$ exiting flow during $[\tau, \tau + \Delta\tau]$

Note that the main difference with the basic variables of running links is that space ($s, \Delta s$) is no longer explicitly referred to because it is irrelevant. Some of the above variables are shown in Fig. 2.5.

With reference to the service activity, let:

$t_{s,i}$ Be service time of user i

$t_s(\tau, \tau + \Delta\tau)$ Average service time among all users joining the queue in time interval $[\tau, \tau + \Delta\tau]$

tw_i Total waiting time (pure waiting plus service time) of user i

$tw(\tau, \tau + \Delta\tau)$ Average total waiting time among all users joining the queue in time interval $[\tau, \tau + \Delta\tau]$

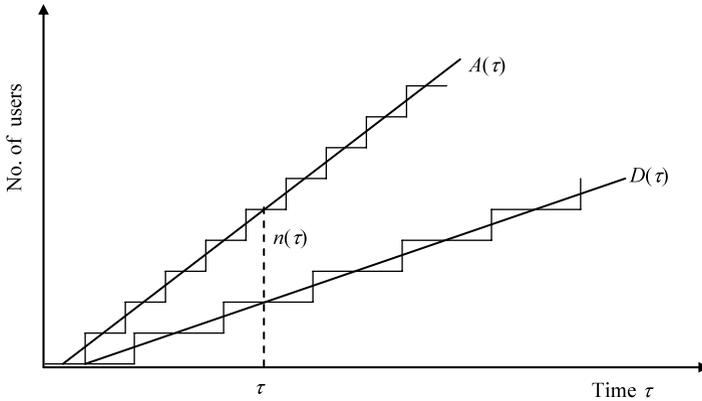


Fig. 2.6 Fluid approximation of deterministic queuing systems

$Q(\tau, \tau + \Delta\tau) = 1/t_s(\tau, \tau + \Delta\tau)$ the (transversal³) capacity or maximum exit flow, that is, the maximum number of users that may be served in the time unit, assumed constant during $[\tau, \tau + \Delta\tau]$ for simplicity's sake (otherwise $\Delta\tau$ can be redefined)

The capacity constraint on exiting flow is expressed by

$$w \leq Q.$$

A general conservation equation, similar to (2.2.1) and (2.2.2) introduced for uninterrupted flow, holds in this case:

$$n(\tau) + m_{IN}(\tau, \tau + \Delta\tau) = m_{OUT}(\tau, \tau + \Delta\tau) + n(\tau + \Delta\tau). \tag{2.2.9}$$

Moreover, dividing by $\Delta\tau$ we obtain:

$$\Delta n/\Delta\tau + [w(\tau, \tau + \Delta\tau) - u(\tau, \tau + \Delta\tau)] = 0. \tag{2.2.10}$$

In the following subsection we describe several deterministic models developed under the assumption that the headway between two consecutive vehicles and the service time are represented by deterministic variables. This is followed by a subsection on stochastic models developed using random variables. In formulating such models, as in the case of uninterrupted flow models, we assume arrival at the queue is represented as a continuous (one-dimensional) fluid.

³In some cases it is also necessary to introduce longitudinal capacity, that is, the maximum number of users that may form the queue.

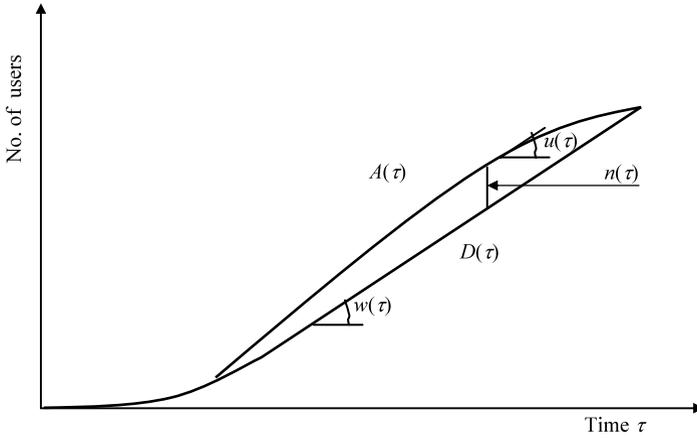


Fig. 2.7 Cumulative arrival and departure curves

2.2.2.2 Deterministic Models

Deterministic models are based on the assumptions that arrival and departure times are deterministic variables. According to the fluid approximation introduced above, the conservation equation (2.2.10) for $\Delta\tau \rightarrow 0$ becomes (see Fig. 2.6):

$$\frac{dn(\tau)}{dt} = u(\tau) - w(\tau)$$

Deterministic queuing systems can also be analyzed through the cumulative number of users that have arrived at the *server* by time τ , and the cumulative number of users that have departed from the *server* (leaving the queue) at time τ , as expressed by two functions termed *arrival curve* $A(\tau)$, and *departure curve* $D(\tau) \leq A(\tau)$, respectively; see Fig. 2.7. Queue length $n(\tau)$ at any time τ is given by:

$$n(\tau) = A(\tau) - D(\tau) \tag{2.2.11}$$

provided that the queue at time 0 is given by $n(0) = A(0) \geq 0$ with $D(0) = 0$. The arrival and departure functions are linked to entering and exiting users by the following relationships.

$$m_{IN}(\tau, \tau + \Delta\tau) = A(\tau + \Delta\tau) - A(\tau) \tag{2.2.12}$$

$$m_{OUT}(\tau, \tau + \Delta\tau) = D(\tau + \Delta\tau) - D(\tau) \tag{2.2.13}$$

The *flow conservation equation* (2.2.9) can also be obtained by subtracting member by member the relationships (2.2.12) and (2.2.13) and taking into account (2.2.11). The limit for $\Delta\tau \rightarrow 0$ of (2.2.12) and (2.2.13) leads to (see Fig. 2.7):

$$u(\tau) = \frac{dA(\tau)}{d\tau}$$

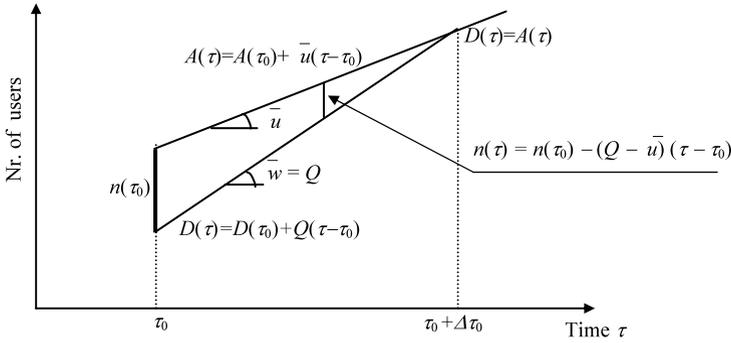


Fig. 2.8 Undersaturated queuing system

$$w(\tau) = \frac{dD(\tau)}{d\tau}$$

If during time interval $[\tau_0, \tau_0 + \Delta\tau]$ the entering flow is constant over time, $u(\tau) = \bar{u}$, then the queuing system is named (*flow-*)*stationary* and the arrival function $A(\tau)$ is linear with slope given by \bar{u} :

$$A(\tau) = A(\tau_0) + \bar{u} \cdot (\tau - \tau_0) \quad \tau \in [\tau_0, \tau_0 + \Delta\tau]$$

The exit flow may be equal to the entering flow \bar{u} , or to the capacity Q as described below.⁴

(a) Undersaturation When the arrival flow is less than capacity ($\bar{u} < Q$) the system is *undersaturated*. In this case, if there is a queue at time τ_0 , its length decreases with time and vanishes after a time $\Delta\tau_0$ defined as (see Fig. 2.8)

$$\Delta\tau_0 = n(\tau_0)/(Q - \bar{u}) \quad (2.2.14)$$

Before time $\tau_0 + \Delta\tau_0$, the queue length is linearly decreasing with τ and the exiting flow \bar{w} is equal to capacity Q :

$$\begin{aligned} n(\tau) &= n(\tau_0) - (Q - \bar{u})(\tau - \tau_0) \\ \bar{w} &= Q \\ D(\tau) &= D(\tau_0) + Q(\tau - \tau_0) \end{aligned} \quad (2.2.15)$$

After time $\tau_0 + \Delta\tau_0$ the queue length is zero and the exiting flow \bar{w} is equal to the arrival flow \bar{u} :

$$n(\tau_0 + \Delta\tau_0) = 0$$

⁴In stationary queuing models used on transportation networks, the inflow \bar{u} can be substituted with the flow f_a of the link representing the queuing system.

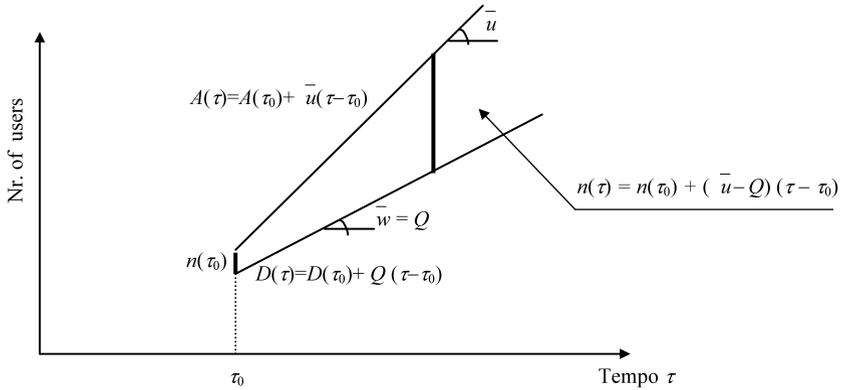


Fig. 2.9 Oversaturated queuing system

$$\bar{w} = \bar{u} \tag{2.2.16}$$

$$D(\tau) = A(\tau) = A(\tau_0) + \bar{u}(\tau - \tau_0)$$

(b) Oversaturation When the arrival flow rate is larger than capacity, $\bar{u} \geq Q$, the system is *oversaturated*. In this case queue length linearly increases with time τ and the exiting flow is equal to the capacity Q (see Fig. 2.9):

$$\begin{aligned} n(\tau) &= n(\tau_0) + (\bar{u} - Q)(\tau - \tau_0) \\ \bar{w} &= Q \end{aligned} \tag{2.2.17}$$

$$D(\tau) = D(\tau_0) + Q(\tau - \tau_0)$$

(c) General Condition By comparing (2.2.15) through (2.2.17) it is possible to formulate this general equation for calculating the queue length at generic time instant τ :

$$n(\tau) = \max\{0, (n(\tau_0) + (\bar{u} - Q)(\tau - \tau_0))\} \tag{2.2.18}$$

With the above results, any general case can be analyzed by modeling a sequence of periods during which arrival flow and capacity are constant. An important case is that of the queuing system at traffic lights which may be considered a sequence of undersaturated (green) and oversaturated (red) periods with zero capacity (see p. 73: *Application of Queuing Models*).

The *delay* can be defined as the time needed for a user to leave the system (passing the server), accounting for the time spent queuing (pure waiting). Thus the delay is the sum of two terms:

$$tw = t_s + tw_q$$

where

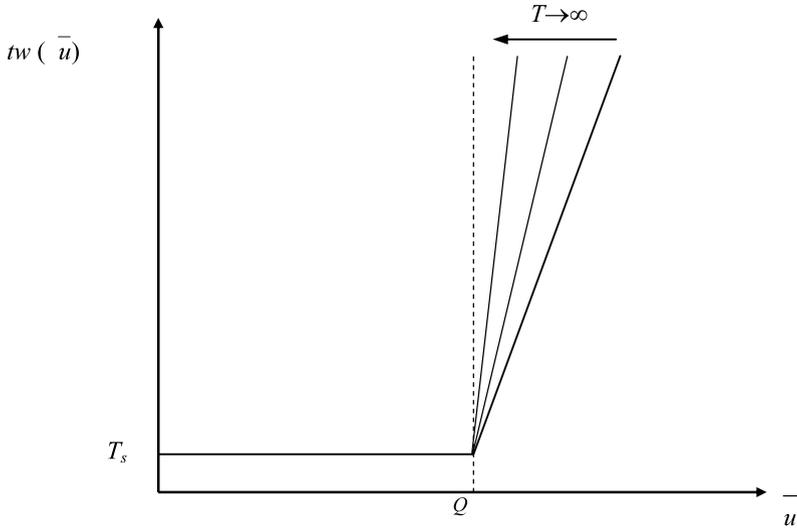


Fig. 2.10 Deterministic delay function at a server

tw is the total delay

$t_s = 1/Q$ is the average service time (time spent at the server)

tw_q is the queuing delay (time spent in the queue)

In undersaturated conditions ($\bar{u} < Q$) if the queue length at the beginning of period is zero (it remains equal to zero), the queuing delay is equal to zero, $tw_q(u) = 0$, and the total delay is equal to the average service time:

$$tw(\bar{u}) = t_s$$

In oversaturated conditions ($\bar{u} \geq Q$), the queue length, and respective delay, would tend to infinity in the theoretical case of a stationary phenomenon lasting for an infinite time. In practice, however, oversaturated conditions last only for a finite period T . If the queue length is equal to zero at the beginning of the period, it will reach a value $(\bar{u} - Q) \cdot T$ at the end of the period. Thus, the average queue over the whole period T is:

$$\bar{n} = \frac{(\bar{u} - Q)T}{2}$$

In this case the average queuing delay is \bar{x}/Q , and average total delay is (see Fig. 2.10):

$$tw(\bar{u}) = t_s + \frac{(\bar{u} - Q)T}{2Q} \tag{2.2.19}$$

2.2.2.3 Stochastic Models

Stochastic models arise when the variables of the problem (e.g., user arrivals, service times of the server, etc.) cannot be assumed deterministic, due to the observed fluctuations, as is often the case, especially in transportation systems. If the system is undersaturated, it can be analyzed through (stochastic) queuing theory which includes the particular case of the deterministic models illustrated above. Some of the results of this theory are briefly reported below, without any claim to being exhaustive.

It is particularly necessary to specify the stochastic process describing the sequence of user arrivals (arrival pattern), the stochastic process describing the sequence of service times (service pattern) and the queue discipline. Arrival and service processes are usually assumed to be stationary renewal processes, in other words with stable characteristics in time that are independent of the past: that is, headways between successive arrivals and successive service times are independently distributed random variables with time-constant parameters. Let N be a random variable describing the queue length, and n the realization of N . The characteristics of a queuing phenomenon can be redefined in the following concise notation,

$$a/b/c(d, e)$$

where

a denotes the type of arrival pattern, that is, the variable which describes time intervals between two successive arrivals:

D = Deterministic variable

M = Negative exponential random variable

E = Erlang random variable

G = General distribution random variable

b denotes the type of service pattern, such as a

c is the number of service channels: $\{1, 2, \dots\}$

d is the queue storage limit: $\{\infty, n_{\max}\}$ or longitudinal capacity

e denotes the queuing discipline:

$FIFO$ = First In–First Out (i.e., service in order of arrival)

$LIFO$ = Last In–First Out (i.e., the last user is the first served)

$SIRO$ = Service In Random Order

$HIFO$ = High In–First Out (i.e., the user with the maximum value of an *indicator* is the first served)

Fields d and e , if defined respectively by ∞ (no constraint on maximum queue length) and by $FIFO$, are generally omitted. In the following we report the main results for the $M/M/1$ ($\infty, FIFO$) and the $M/G/1$ ($\infty, FIFO$) queuing systems, which are commonly used for simulating transportation facilities, such as signalized intersections.

Some definitions or notation differ from those traditionally adopted in dealing with queuing theory (the relative symbols are in brackets) so as to be consistent with those adopted above. The parameters defining the phenomenon are as follows.

- $u, (\lambda)$ The arrival rate or the expected value of the arrival flow
 $Q = 1/t_s, (\mu)$ The service rate (or capacity) of the system, the inverse of the expected service time
 $u/Q, (\rho)$ The traffic intensity ratio or utilization factor
 n A value of the random variable N , number of users present in the system, consisting of the number of users queuing plus the user present at the server, if any (the significance of the symbol n is thus slightly different)
 tw A value of the random variable TW , the time spent in the system or overall delay, consisting of queuing time plus service time

(a) M/M/1 (∞ , FIFO) Systems In *undersaturated* conditions ($u/Q < 1$):

$$E[N] = \frac{\frac{u}{Q}}{1 - \frac{u}{Q}} = \frac{u}{Q - u} \quad (2.2.20)$$

$$\text{VAR}[N] = \frac{\frac{u}{Q}}{(1 - \frac{u}{Q})^2}$$

According to Little's formula, the expected number of users in the system $E[N]$ is the product of the average time in the system (expected value of delay) $E[TW]$ multiplied by arrival rate u :

$$E[N] = uE[TW] \quad (2.2.21)$$

from which:

$$E[TW] = \frac{1}{Q - u} \quad (2.2.22)$$

The expected time spent in the queue $E[tw_q]$ (or queuing delay) is given by the difference between the expected delay $E[tw]$ and the average service time $t_s = 1/Q$:

$$E[TW_q] = \frac{1}{Q - u} - \frac{1}{Q} = \frac{u}{Q(Q - u)}. \quad (2.2.23)$$

According to Little's second formula, the expected value of the number of users in the queue $E[N_q]$ is the product of the expected queuing delay $E[TW_q]$ multiplied by the arrival rate u :

$$E[N_q] = uE[TW_q] \quad (2.2.24)$$

and then:

$$E[N_q] = \frac{u^2}{Q(Q - u)} \quad (2.2.25)$$

(b) M/G/1 (∞ , FIFO) Systems In this case the main results are the following.

$$E[N] = \frac{u}{Q} \left[1 + \frac{u}{2(Q-u)} \right]$$

$$E[TW] = \frac{1}{Q} \left[1 + \frac{u}{2(Q-u)} \right]$$

$$E[TW_q] = \frac{u}{2Q(Q-u)}$$

2.3 Congested Network Models

This section provides a general mathematical formulation of transportation supply models, based on congested network flow models. The bases for these models are *graph models*. Next, *network models*, including link performances and costs, and *network flow models*, including link flows, are introduced. Finally, *congested network (flow) models*, modeling relationships among performances, costs, and flows, are developed.

2.3.1 Network Structure

The network structure is represented by a *graph*. The latter is defined by a set N of elements called *nodes* and by a set of pairs of nodes belonging to N , $L \subseteq N \times N$, called *links*. The graphs used to represent transportation services are generally oriented; that is, the links have a direction and the node pairs defining them are ordered pairs. A link connecting the node pair (i, j) can also be denoted by a single index, say a .

The links in a graph modeling a transportation system represent phases and/or activities of possible trips between different traffic zones. Thus, a link can represent an activity connected to a physical movement (e.g., covering a road) or an activity not connected to a physical movement (such as waiting for a train at a station). Links are chosen in such a way that physical and functional characteristics can be assumed to be homogeneous for the whole link (e.g., the same average speed). In this sense, links can be seen as the partition of trips into segments, each of which has certain characteristics; the level of detail of such a partition can clearly be very different for the same physical system according to the objectives of the analysis.

Nodes correspond to significant events delimiting the trip phases (links), that is, to the space and/or time coordinates in which events occur that they represent. In *synchronic networks*, nodes are not identified by a specific time coordinate, and the same node represents events occurring at different moments (instants) of time. For example, the different entry or exit times in a road segment, an intersection, or a station, may be associated with a single node, representing all the entry/exit events.

Centroid nodes, introduced in Sect. 1.3.1, represent the beginning or end of individual trips. In *diachronic networks*, on the other hand, nodes may have an explicit time coordinate and therefore represent an event occurring at a given instant. The graphs considered in this chapter are synchronic, because diachronic networks assume a within-period system representation; diachronic graphs for scheduled services are introduced in Chap. 7.

A trip is a sequence of several phases and, in a graph that represents transportation supply, it consists of a *path* k , defined as a succession of consecutive links connecting an initial node (path origin) to a final node (path destination). Usually, only paths connecting centroid nodes are considered in transportation graphs. On this basis, each path is unambiguously associated with one, and only one, O-D pair, whereas several paths can connect the same O-D pair. An example of a graph with different paths connecting the centroid nodes is depicted in Fig. 2.11.

A binary matrix called the *link–path incidence matrix* $\mathbf{\Delta}$, can represent the relationship between links and paths. This matrix has a number of rows equal to the number of links n_L and a number of columns equal to the number of paths n_P . The generic element δ_{ak} of the binary matrix $\mathbf{\Delta}$ is equal to one if link a belongs to path k , $a \in k$, and zero, otherwise, $a \notin k$ (see Fig. 2.11). The row of the link–path incidence matrix corresponding to the generic link a identifies all the paths including that link (columns k for which $\delta_{ak} = 1$). Moreover, the elements of a column corresponding to the generic path k identify all the links that make it up (rows a for which $\delta_{ak} = 1$).

2.3.2 Flows

A *link flow* f_a can be associated with each link a . Link flow is the average number of homogeneous units using link a (i.e., carrying out the trip phase represented by the link) in a time unit. In other words, the link flow is a random variable of mean f_a . Several link flows can be associated with a given link depending on the homogeneous unit considered. *User flows* relate to users, such as travelers or goods, possibly of different classes. *Vehicle flows* relate to the number of vehicles, perhaps of different types such as automobiles, buses, trains, and so on.

For individual modes, such as automobiles or trucks, user flows can be transformed quite straightforwardly into vehicle flows through average occupancy coefficients. For scheduled modes, such as trains, vehicle flows derive from the service schedule and are often treated as an input to the supply model.

The link flow of the generic user class or vehicle type i is denoted by f_a^i . In accordance with the results of traffic flow theory (see Sect. 2.2), link performance and cost variables are affected by user or vehicle flow. To allow for this dependence it is often worth homogenizing the various classes of users or various types of vehicles by defining *equivalent flows* associated with links. In this case the flows of different user classes or vehicle types are homogenized to a reference class or type:

$$f_a = \sum_i w_i f_a^i$$

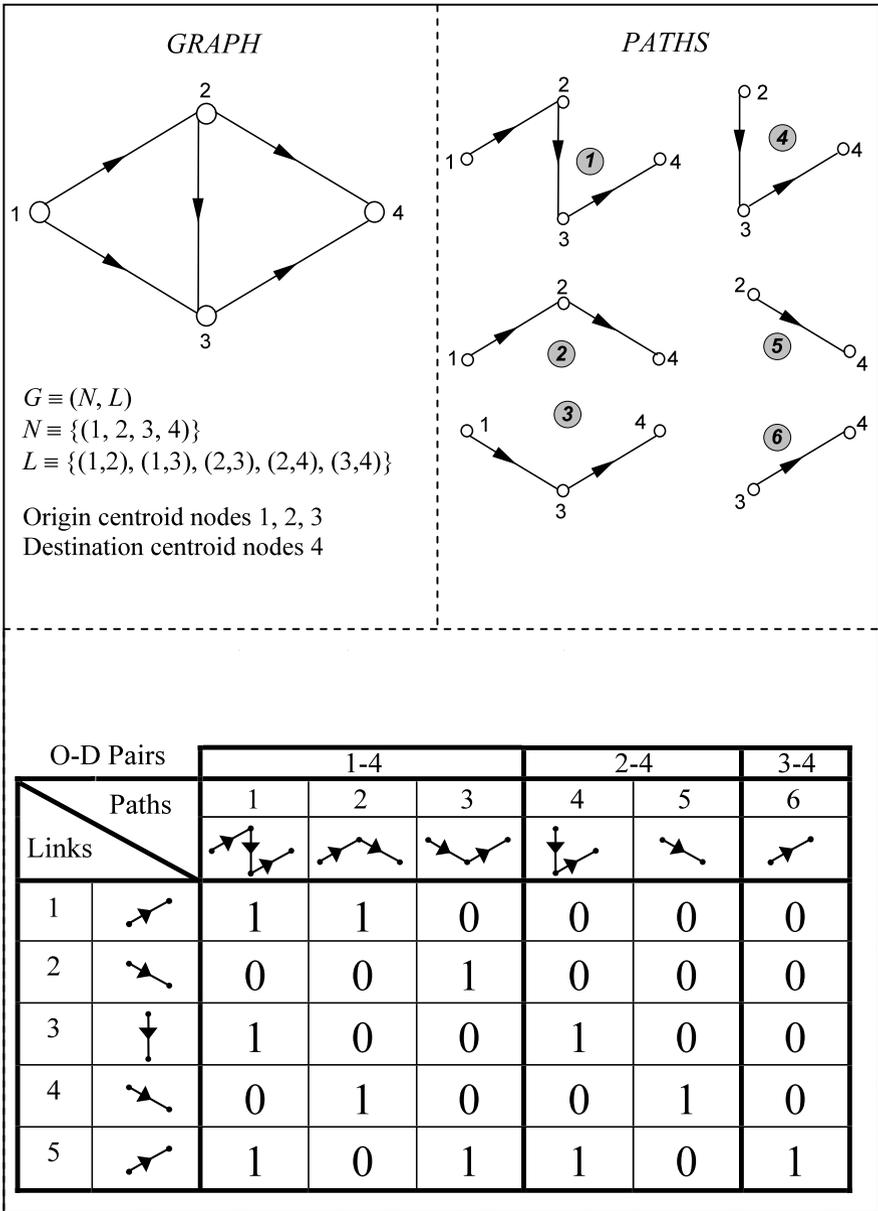


Fig. 2.11 Example of a graph and link-path incidence matrix

where w_i is the homogenization coefficient of the users of class i with respect to their influence on link performances. For example, for road flows, automobiles are usually the reference vehicle type ($w_i = 1$) and the other vehicle flows are trans-

formed into equivalent auto flows with coefficients w_i . The latter are greater than one if the contribution to congestion of these vehicles is greater than that of cars (buses, heavy vehicles, etc.), less than one in the opposite case (motorcycles, bicycles, etc.).

The *vector of link flows* \mathbf{f} has, as a generic component, the flow on link a , f_a , for each $a \in L$ (see Fig. 2.12).

Flow variables can also be associated with paths. Under the within-day stationarity hypothesis, the average number of users, who in each subinterval travel along each path, is constant. The average number of users, who in a time unit follow path k , is called the *path flow* h_k . If the users have different characteristics (i.e., they belong to different classes), path flows per class i , h_k^i , can be introduced. Path flows of different user classes or vehicle types can be homogenized by means of coefficients w_i similar to those introduced for link flows; the equivalent path flow is obtained as:

$$h_k = \sum_i w_i \cdot h_k^i$$

There is clearly a relationship between link and path flows. Indeed, the flow on each link a can be obtained as the sum of the flows on the various paths containing that link. This relationship can be expressed by using the elements δ_{ak} of the link-path incidence matrix as

$$f_a = \sum_k \delta_{ak} \cdot h_k \quad (2.3.1)$$

or in matrix terms:

$$\mathbf{f} = \mathbf{\Delta} \mathbf{h} \quad (2.3.2)$$

where \mathbf{h} is the path flow vector.

Equation (2.3.1) or (2.3.2) expresses the way in which path flows induce flows on individual links. For this reason it is referred to as the (*static*) *Network Flow Propagation (NFP)* model (see Fig. 2.11). Note that the linear algebraic structure of (2.3.1) depends crucially on the assumption of intraperiod stationarity (within-day static model); if this assumption is removed, the model loses its algebraic-linear nature as shown in Chap. 7.

2.3.3 Performance Variables and Transportation Costs

Some variables perceived by users can be associated with individual trip phases. Examples of such variables are travel times (transversal and/or waiting), monetary cost, and discomfort. These variables are referred to as *level-of-service* or *performance attributes*. In general, performance variables correspond to disutilities or costs for the users (i.e., users would be better off if the values of performance variables were reduced). The average value of the n th performance variable, related to link a , is denoted by r_{na} . The *average generalized transportation link cost*, or simply the

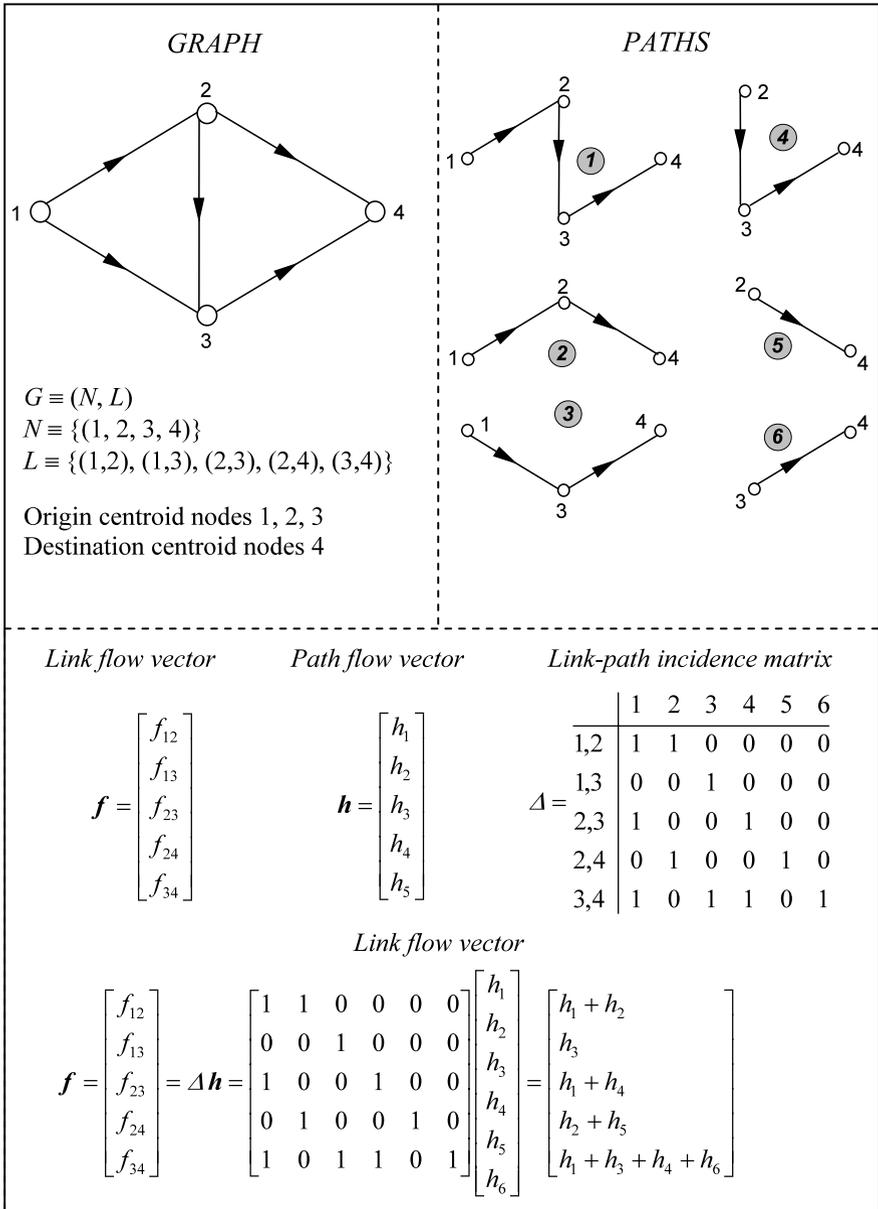


Fig. 2.12 Transportation network with link and path flows

transportation link cost, is a variable *synthesizing* (the average value of) the different performance variables borne and perceived by the users in travel-related choice and, more particularly, in path choices (see Sect. 4.3.3). Thus, the transportation link

cost reflects the average users' disutility for carrying out the activity represented by the link. Other performance variables and costs, which cannot be associated with individual links but rather to the whole trip (path), are introduced shortly.

Performance variables making up the transportation cost are usually nonhomogeneous quantities. In order to reduce the cost to a single scalar quantity, the different components can be homogenized into a generalized cost applying reciprocal substitution coefficients β , whose value can be estimated by calibrating the path choice model (see Sect. 4.3.3). For example, the generalized transportation cost c_a relative to the link a can be formulated as

$$c_a = \beta_1 \cdot t_a + \beta_2 \cdot mc_a$$

where t_a is the travel time and mc_a is the monetary cost (e.g., the toll) connected with the crossing of the link. More generally, the link transportation cost can be expressed as a function of several link performance variables as

$$c_a = \sum_n \beta_n \cdot r_{na}$$

Different users may experience and/or perceive transportation costs, which differ for the same link. For example, the travel time of a certain road section generally differs for each vehicle that covers it, even under similar external conditions. Furthermore, two users experiencing the same travel time may have different perceptions of its disutility. If we then add the fact that the analyst cannot have perfect knowledge of such costs, we realize that the perceived link cost is well represented by a random variable distributed among users, whose average value is link transportation cost c_a . There may be other "costs" both for users (e.g., accident risks or tire consumption) and for society (e.g., noise and air pollution) associated with a link. It is usually assumed that these costs are not taken into account by users in their travel-related choices and are not included in the perceived transportation cost. The transportation cost is, therefore, an internal cost, used to simulate the transportation system and, in particular, travelers' choices. The other cost items are external costs, used for project design and assessment. External costs are sometimes referred to as impacts; they are dealt with in Sect. 2.3.5.

Different groups (or classes) of users may have different average transportation costs. This may be due to different performance variables (e.g., their speeds and travel times are different or they pay different fares) or to differences in the homogenization coefficients β_n (e.g., different time/money substitution rates corresponding to different incomes). In this case a link cost c_a^i can be associated with each user class i . In what follows, for simplicity of notation, the class index i is taken as understood unless otherwise stated. Other considerations relative to users belonging to different classes are made in Chap. 6.

Link performance variables and transportation costs can be arranged in vectors. The *performance vector* \mathbf{r}_a is made up by the n th performance variable for each link, its components being r_{na} . Analogously, the vector \mathbf{c} , whose generic component c_a is the generalized transport cost on link a , is known as the *link cost vector*.

The concepts of performance variables and generalized transportation cost can be extended from links to paths. The *average performance variable* of a path k , z_{nk} , is the average value of that variable associated to a whole origin–destination trip, represented by a path in the graph. Some path performance variables are *linkwise additive*; that is, their path value can be obtained as the sum of link values for all links making up the path.

Examples of additive path variables are travel times (the total travel time of a path is the sum of travel times over individual links) or some monetary costs, which can be associated with some or all individual links. An *additive path performance variable* can be expressed as the sum of link performance variables as

$$z_{nk}^{\text{ADD}} = \sum_{a \in k} r_{na} = \sum_a \delta_{ak} r_{na}$$

or in vector notation

$$z_n^{\text{ADD}} = \mathbf{\Delta}^T \mathbf{r}_n$$

Other path performance variables are *nonadditive*; that is, they cannot be obtained as the sum of link specific values. These variables are denoted by z_{nk}^{NA} . Examples of nonadditive performance variables are monetary cost in the case of tolls that are nonlinearly proportional to the distance covered or the waiting time at stops for high-frequency transit systems, as shown below.

The *average generalized transportation cost* of a path k , g_k , is defined as a scalar quantity homogenizing in disutility units the different performance variables perceived by the users (of a given category) in making trip-related choices and, in particular, path choices.

The path cost in the most general case is made up of two parts: linkwise additive cost g_k^{ADD} and nonadditive cost, g_k^{NA} , assuming that they are homogeneous:

$$g_k = g_k^{\text{ADD}} + g_k^{\text{NA}} \quad (2.3.3)$$

The *additive path cost* is defined as the sum of the linkwise additive path performance variables:

$$g_k^{\text{ADD}} = \sum_n \beta_n \cdot z_{nk}^{\text{ADD}}$$

Under the assumption that the generalized cost depends linearly on performance variables, the additive path cost can be expressed as the sum of generalized link costs. The relationship between additive path cost and link costs can be expressed by combining all the equations previously presented:

$$g_k^{\text{ADD}} = \sum_n \beta_n z_{nk}^{\text{ADD}} = \sum_n \beta_n \sum_a \delta_{ak} r_{na} = \sum_a \delta_{lk} \sum_n \beta_n r_{na} = \sum_a \delta_{ak} c_a$$

or

$$g_k^{\text{ADD}} = \sum_a \delta_{ak} c_a \quad (2.3.4)$$

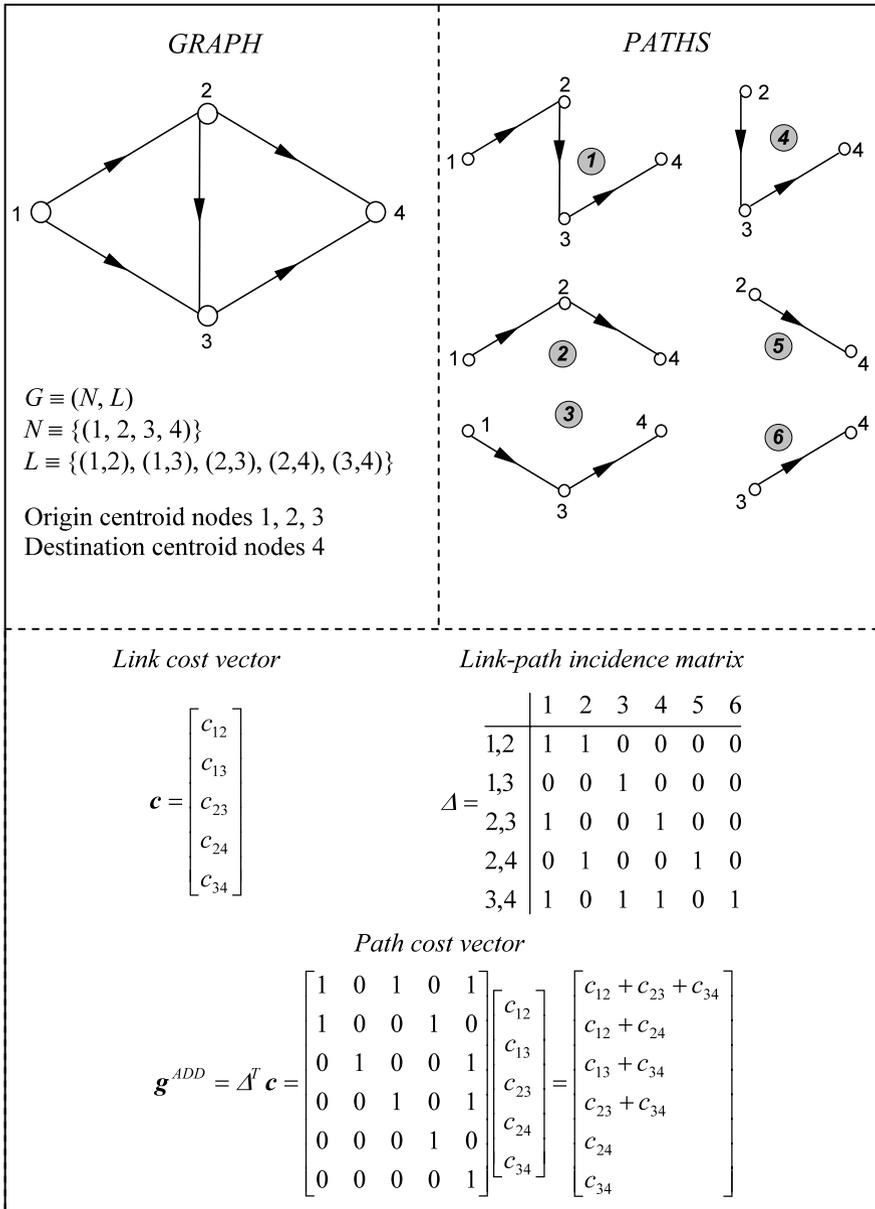


Fig. 2.13 Transportation network with link and path costs

The expression (2.3.4) can also be formulated in vector format by introducing the vector of additive path costs \mathbf{g}^{ADD} (see Fig. 2.13):

$$\mathbf{g}^{ADD} = \Delta^T \mathbf{c} \tag{2.3.5}$$

The nonadditive path cost g_k^{NA} includes nonadditive path performance variables:

$$g_k^{\text{NA}} = \sum_n \beta_n z_{nk}^{\text{NA}}$$

Finally, the path cost vector \mathbf{g} , of dimensions $(n_P \times 1)$, can be expressed as

$$\mathbf{g} = \mathbf{\Delta}^T \mathbf{c} + \mathbf{g}^{\text{NA}} \quad (2.3.6)$$

where \mathbf{g}^{NA} is the nonadditive path cost vector.

In many applications, the nonadditive path cost vector is, or is assumed to be, null. This affects the efficiency of the calculation algorithm for assignment models, as shown in Chaps. 5 and 6.

2.3.4 Link Performance and Cost Functions

Link performance attributes generally depend on the physical and functional characteristics of the facility and/or the service involved in the trip phase represented by the link itself. Typical examples are the travel time on a road section depending on its length, alignment, allowed speed, or the waiting time at a bus stop depending on the headway between successive bus arrivals. When several travelers or vehicles use the same facility, they may interact with each other, thereby influencing link performance. This phenomenon is known as *congestion* and was introduced in Sect. 2.2.1. Typically, the effects of congestion on link performance increase as the flow increases. For instance, the larger the flow of vehicles traveling along a road section, the more likely faster vehicles will be slowed by slower ones, thus increasing the average travel time. Moreover, the larger the flow arriving at an intersection, the longer is the average waiting time; the larger the number of users on the same train, the lower is the riding comfort.

In general, congestion effects are such that the performance attributes of a given link may be influenced by the flow on the link itself and by flows on other links.

Link performance functions relate the generic link performance attribute r_{na} to physical and functional characteristics of the link, arranged in a vector \mathbf{b}_{na} , and to the equivalent flow on the same link and, possibly, on other links, arranged in the vector \mathbf{f} :

$$r_{na} = r_{na}(\mathbf{f}; \mathbf{b}_{na}, \boldsymbol{\gamma}_{na})$$

where $\boldsymbol{\gamma}_{na}$ is a vector of parameters used in the function.

Because the generalized transportation cost of a link c_a is a linear combination of link performance attributes, *link cost functions*⁵ can be expressed as functions of

⁵A distinction should be made between cost functions in microeconomics and in transportation systems theory. In the first case, the cost function is a relationship connecting the production cost of a good or service to the quantity produced and the costs of individual production factors. Cost

the same parameters:

$$c_a = c_a(\mathbf{f}; \mathbf{b}_a, \boldsymbol{\gamma}_a) \quad (2.3.7)$$

where vectors \mathbf{b}_a and $\boldsymbol{\gamma}_a$ have the same meaning as above.

Link performance and cost functions may have some mathematical properties, which are used in Chaps. 5 and 6 to study the properties of supply–demand interaction models and to analyze the convergence of their solution algorithms.

Performance and cost functions can be classified as *separable* and *nonseparable* across a link. In the former case, the performances and cost variables of a link depend exclusively on the (equivalent) flow on the link itself:

$$c_a(\mathbf{f}) = c_a(f_a)$$

In the latter case, they also depend on the flow on other links. Examples of both types of function are given in the following sections.

The *cost function vector* $\mathbf{c}(\mathbf{f})$ is obtained by ordering the n_L functions of the individual network links:

$$\mathbf{c} = \mathbf{c}(\mathbf{f}) \quad (2.3.8)$$

Under the assumption that the first partial derivative of $\mathbf{c}(\mathbf{f})$ exists and is finite, the Jacobian matrix, $\mathbf{Jac}[\mathbf{c}(\mathbf{f})]$, may be defined:

$$\mathbf{Jac}[\mathbf{c}(\mathbf{f})] = \begin{bmatrix} \frac{\partial c_1}{\partial f_1} & \cdots & \frac{\partial c_1}{\partial f_{n_L}} \\ & \frac{\partial c_i}{\partial f_i} & \cdots \\ \frac{\partial c_{n_L}}{\partial f_1} & \cdots & \frac{\partial c_{n_L}}{\partial f_{n_L}} \end{bmatrix}$$

The cost functions generally have an asymmetric Jacobian. In some cases, they may have a symmetric Jacobian: $\partial c_i / \partial f_j = \partial c_j / \partial f_i$, $\forall i, j$; that is, the cost variation on link a , due to a flow variation on link j , is equal to the cost variation on link j , due to a flow variation on link i . Separable cost functions are clearly a special case, the Jacobian being a diagonal matrix: $\partial c_i / \partial f_j = 0$, $\forall i \neq j$.

In the case of uncongested networks the cost functions are independent of the flows, so the partial derivatives are all equal to zero and the Jacobian is null.

2.3.5 Impacts and Impact Functions

Design and evaluation of transportation systems, in addition to performance variables perceived by the users, require the modeling of impacts borne by the users, but not perceived in their mobility choices, and of impacts on nonusers. Examples

functions in transportation systems provide the cost perceived by users in their trips. Transportation cost is therefore a cost of use rather than of production. The cost of producing transportation services is usually indicated as the service production cost, and similarly the functions correlating it to the relevant quantities are called production cost functions.

of the first type include indirect vehicle costs (e.g., tire or lubricant, vehicle depreciation, etc.) and accident risks with their consequences (death, injury, material damage). The impacts for nonusers include those for other subjects directly involved in the transportation system, such as costs and revenues for the producers of transportation services, and impacts “external” to the transportation system (or market). Examples of externalities are the impacts on the real estate market, urban structure, or on the environment such as noise and air pollution. The mathematical functions relating these impacts to physical and functional parameters of the specific transportation systems and, in some cases, to link flows are called *impact functions*. Often these functions are named with respect to the specific impact they simulate (e.g., fuel consumption functions or pollutant emission functions). Some impacts can be associated with individual network links and depend on the flows, $e_l(f)$. Link-based impact functions are usually included in transportation supply models; see Fig. 2.1. Some impact functions may be quite elementary whereas others may require complex systems of mathematical models. Examples of link-based impact functions are those related to air and noise pollution due to vehicular traffic. Some impact functions are discussed in Chap. 10 in the context of evaluation of transportation system projects.

2.3.6 General Formulation

To summarize the above points, a *transportation network* consists of the set of nodes N , the set of links L , the vector of link costs \mathbf{c} , which depend on the vector \mathbf{r} of link performances, the vector \mathbf{g}^{NA} of nonadditive path costs and the vector \mathbf{e} of relevant impact variables: $(N, L, \mathbf{c}, \mathbf{g}^{\text{NA}}, \mathbf{e})$. For congested networks, the link cost vector is substituted by the flow-dependent cost functions $\mathbf{c}(\mathbf{f})$; the same holds for flow-dependent internal and external impacts $\mathbf{e}(\mathbf{f})$, whereas the nonadditive costs vector \mathbf{g}^{NA} is usually assumed to be independent of the flows. In this case the abstract transportation network model can be expressed as $(N, L, \mathbf{c}(\mathbf{f}), \mathbf{g}^{\text{NA}}, \mathbf{e}(\mathbf{f}))$. Performance variables and functions are not explicitly mentioned, as they are included in the generalized transportation cost functions.

The set of relationships connecting path costs to path flows is known as the *supply model*. The supply model can therefore be formally expressed combining (2.3.2), (2.3.6), and (2.3.8) into a relationship connecting path flows to path costs:

$$\mathbf{g}(\mathbf{h}) = \Delta^T \mathbf{c}(\Delta \mathbf{h}) + \mathbf{g}^{\text{NA}} \quad (2.3.9)$$

where it is assumed that nonadditive path costs, if any, are not affected by congestion. Link characteristics can be obtained through performance, cost and impact functions for the link flows corresponding to the path flow vector. Clearly the model (2.3.9) expresses the abstract congested network model described in the previous sections. The same type of models can be used to describe other systems such as electrical or hydraulic networks.

The general structure of a supply model is depicted in Fig. 2.14. The graph defines the topology of the connections allowed by the transportation system under

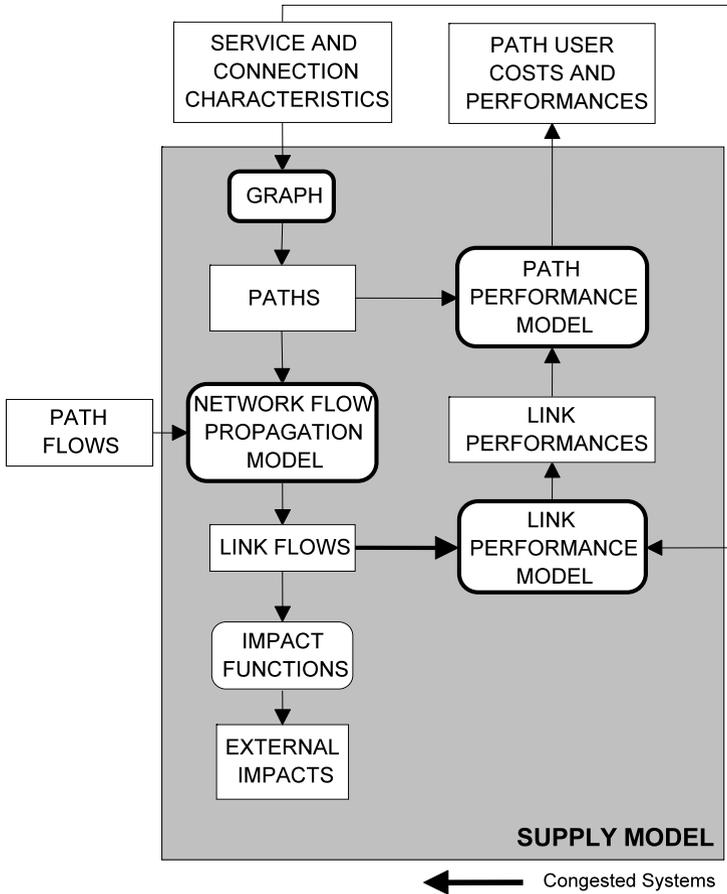


Fig. 2.14 Schematic representation of supply models

study, and the flow propagation model defines the relationship among path and link flows. The link performance model expresses for each element (link) the relationships among performances, physical and functional characteristics, and flow of users. The impact model simulates the main external impacts of the supply system. Finally, the path performance model defines the relationship between the performances of single elements (links) and those of a whole trip (path) between any origin–destination pair.

2.4 Applications of Transportation Supply Models

Network models and related algorithms are powerful tools for modeling transportation systems. A network model is a simplified mathematical description of the phys-

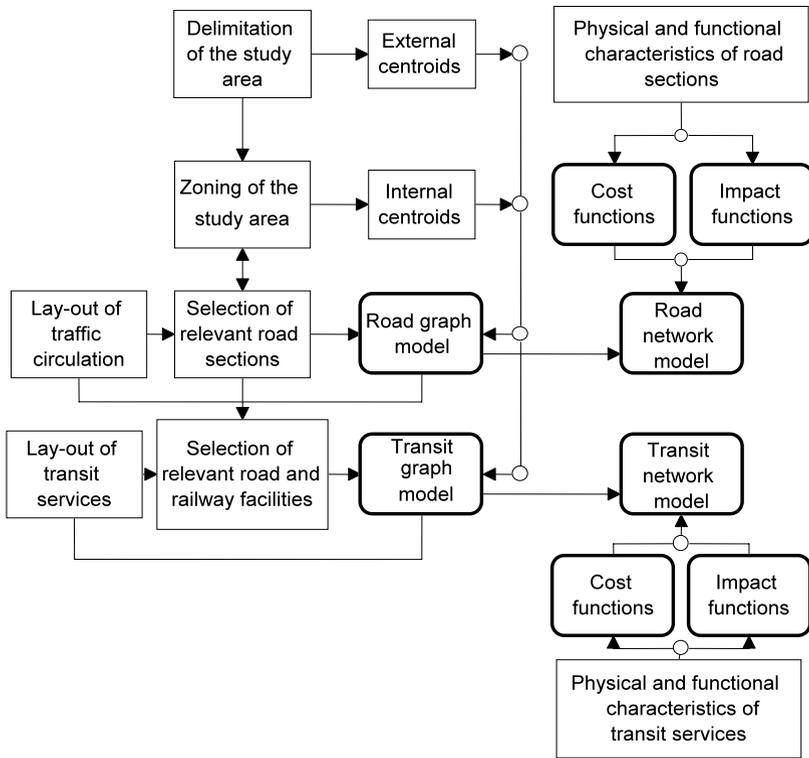


Fig. 2.15 Functional phases for the construction of an urban bimodal network model

ical phenomena relevant to the analysis, design, and evaluation of a given system. Thus transportation network models depend on the purpose for which they are used. Building a network model usually requires a sequence of operations whose general criteria are described in the following. A schematic representation of the main activities in the case of a bimodal supply system (road and transit urban systems) is depicted in Fig. 2.15.

In the most general case, a supply network model is built through the following phases.

- (a) Delimitation of the study area
- (b) Zoning
- (c) Selection of relevant supply elements (basic network)
- (d) Graph construction
- (e) Identification of performance and cost functions
- (f) Identification of impact functions

Phases (a), (b), and (c) relate to the relevant supply system definition. They are described, respectively, in Sect. 1.3.1 of Chap. 1 and are not repeated here. The rest of this section introduces some general considerations related to phases (d), (e), and

(f) for a generic system. Specific models are described separately for two different types of transportation systems: continuous services (such as road), in Sect. 2.4.1, and scheduled services (such as train or buses), in Sect. 2.4.2.

The construction of a transportation graph requires the definition of the relevant trip phases and events (links and nodes) that depend on the physical system to be represented. Important nodes in transportation graphs are the so-called *centroid nodes*. They correspond to the events of beginning and ending a trip in a given zone. As was seen in Sect. 1.3.1, the centroids can approximate the internal points within a traffic zone. In general, the zone centroid is a *fictitious node*, that is, a node which does not correspond to any specific location but which represents the set of points of the zone where a trip can start or end. Therefore, a zone centroid is placed “barycentrically” with respect to such points or to some proxy variables (e.g., the number of households or workplaces). In principle, different centroid nodes may be associated to different trip types (e.g., origin and destination centroids). In other cases, centroids represent the places of entry into or exit from the study area for the trips, which are partly undertaken within the system (*cordon centroids*). In this case they are usually associated with physical locations (road sections, airports, railway stations, etc.).

A graph usually includes *links* of different types: *real links* and *connectors*. Real links represent trip phases corresponding to “physical” components (infrastructures or services), such as traversing a road section or riding a train between two successive stations. When centroid nodes do not correspond to a physical element, *connector links* are introduced into the graph. These links represent the trip phase between the terminal point (zone centroid) and a physical element of the network. In the remainder of this section, links are referred to according to the trip phase (activity) or the infrastructure or service which allows that activity. For example, there are road links, transit line links, and waiting links at stops.

A transportation graph will have different levels of complexity, depending on the system being represented and the details required to do so. In general, short-term or operational projects, such as a road circulation plan or the design of transit lines, require a very detailed representation of the real system. By contrast, strategic or long-term projects usually require less detailed, larger-scale graphs both because of the geographical size of the area and the number of elements included in the system.

As shown shortly, different graphs can be associated with the same basic network, depending on the aim of the model. Graphs can also represent transportation infrastructures; in general, infrastructure graphs are not used directly for system models, but rather they are referred to during the construction of service graphs. User flows and supply performances depend on the transportation services using the infrastructures rather than on the infrastructures themselves.

Specification of link performance and cost functions for a transportation network requires the study of the functioning of the individual elements that comprise it. In practice, performance functions used at times derive from explicit assumptions on system behavior, following a “deductive” approach, as for queuing models for barrier systems such as motorway toll booths, road intersections, air and sea terminals, and the like (see Sect. 2.2.2). When this approach, albeit based on simplifying

assumptions, proves particularly complex, we use “descriptive” models developed according to an “inductive” approach, as in most stationary traffic flow models (see Sect. 2.2.1). Such models are made up of statistical relationships between performance attributes and the explicative variables of the phenomenon. Examples of both types of performance functions are given in the next two sections.

Both approaches use unknown parameters, vectors γ_n and γ , respectively, in expressions (2.4.11) and (2.4.12), which should be calibrated for each specific supply model. To estimate behavioral model parameters or to specify the functional form and estimate nonbehavioral model parameters, the usual methods of inferential statistics may be used. However, in many applications the cost functions calibrated in similar contexts are transferred to the system in question to save application time and costs.

2.4.1 Supply Models for Continuous Service Transportation Systems

Continuous and simultaneous services are available at every instant and can be accessed from a very large number of points. Typical examples are individual modes such as cars and pedestrians using road systems.

2.4.1.1 Graph Models

In graphs representing road systems, nodes are usually located at the intersections between road segments included in the supply model. Nodes can also be located where significant variations occur in the geometric and/or functional characteristics of a single segment (such as changes in a road cross-section and lateral friction). Intersections with secondary roads not included in the “base network,” however, are not represented by nodes. Links usually correspond to connections between nodes allowed by the circulation scheme. Therefore, a two-way road is represented by two links going in opposite directions, whereas a one-way road has a single link going in the allowed direction. Figure 2.16 shows the graph representing part of the urban road network shown in Fig. 1.3.

In applications two distinct types of links are considered: *running links*, which represent the vehicle’s real movement as the trip along a motorway or urban road section; and *waiting or queuing links*, representing queuing at intersections, toll barriers, and so on (see Fig. 2.17).

The level of detail of the road system depends on the purpose of the model. This is especially true for road intersections. In a coarse representation, a road intersection is usually represented by a single node where the access links converge. Alternatively, we can adopt a more detailed representation that distinguishes different turning movements and excludes nonpermitted turns (if any). Such a representation can be obtained by using a larger number of nodes and links. Figure 2.18 shows the two possible representations of a four-arm road intersection. Note that in the single-node representation, paths requiring a left turn (4-5-2) cannot be excluded if

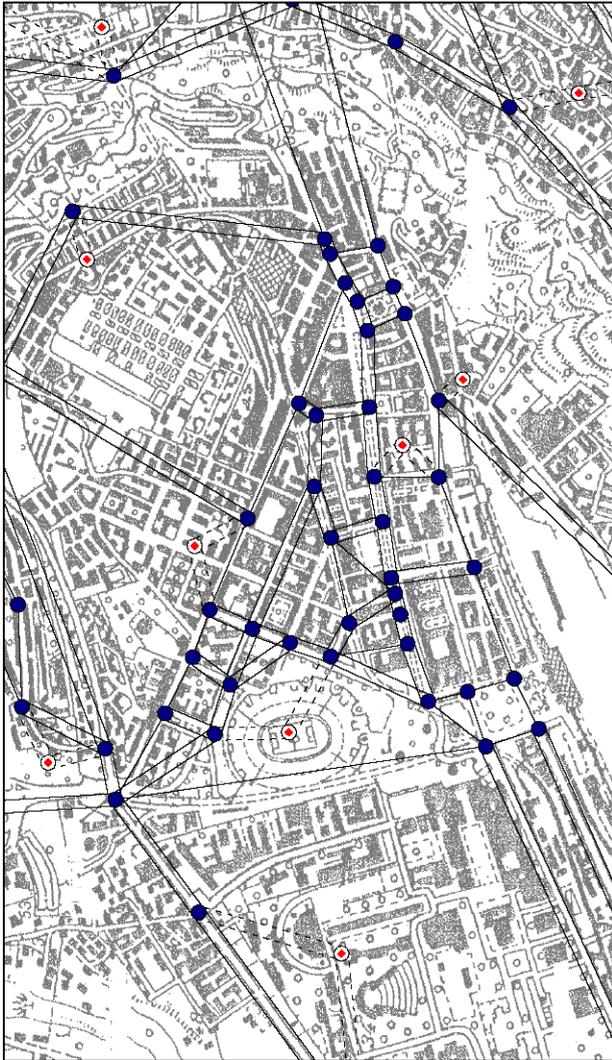
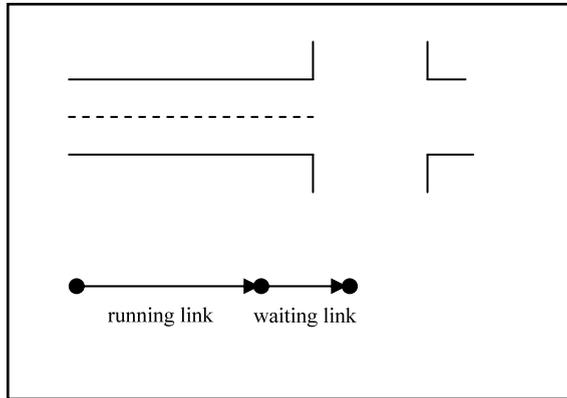


Fig. 2.16 Example of a graph representing part of an urban road system

Fig. 2.17 Representation of a road intersection with running and waiting links



this turning movement is not allowed; furthermore, different waiting times cannot be assigned to maneuvers with different green phase durations, such as right turns (4-5-3). Both of these possibilities are allowed by the detailed representation.

Parking is another element of a road system that can be represented with different levels of detail. In detailed road graphs, trip phases corresponding to parking can be represented with different links for different parking facilities available in a given zone (see Fig. 2.19). *Parking links* can be connected through pedestrian links to the centroid of the zone where they are located, and to the centroids of traffic zones within walking distance. In less detailed graphs, parking is included in connector links; in this case, however, congestion and different parking policies cannot be simulated.

2.4.1.2 Link Performance and Cost Functions

The *generalized transportation cost* of a road link is usually made up by several performance attributes. For example, three attributes can be selected: travel time along the section, waiting time (e.g., at the final intersection, at the tollbooth, etc.), and monetary cost. In this case, the cost function can be obtained as the sum of three performance functions:

$$c_a(\mathbf{f}) = \beta_1 tr_a(\mathbf{f}) + \beta_2 tw_a(\mathbf{f}) + \beta_3 mc_a(\mathbf{f}) \tag{2.4.1}$$

where

- $tr_a(\mathbf{f})$ is the function relating the running time on link a to the flow vector
- $tw_a(\mathbf{f})$ is the function relating the waiting time on link a to the flow vector
- $mc_a(\mathbf{f})$ is the function relating the monetary cost on link a to the flow vector

The dependence on physical and functional variables \mathbf{b}_a , and parameters $\boldsymbol{\gamma}$, has been omitted for simplicity's sake. Note that in (2.4.1) it has been assumed that homogenization coefficients may differ for the different time components. Furthermore, not all of the components in (2.4.1) are present for each link; for example,

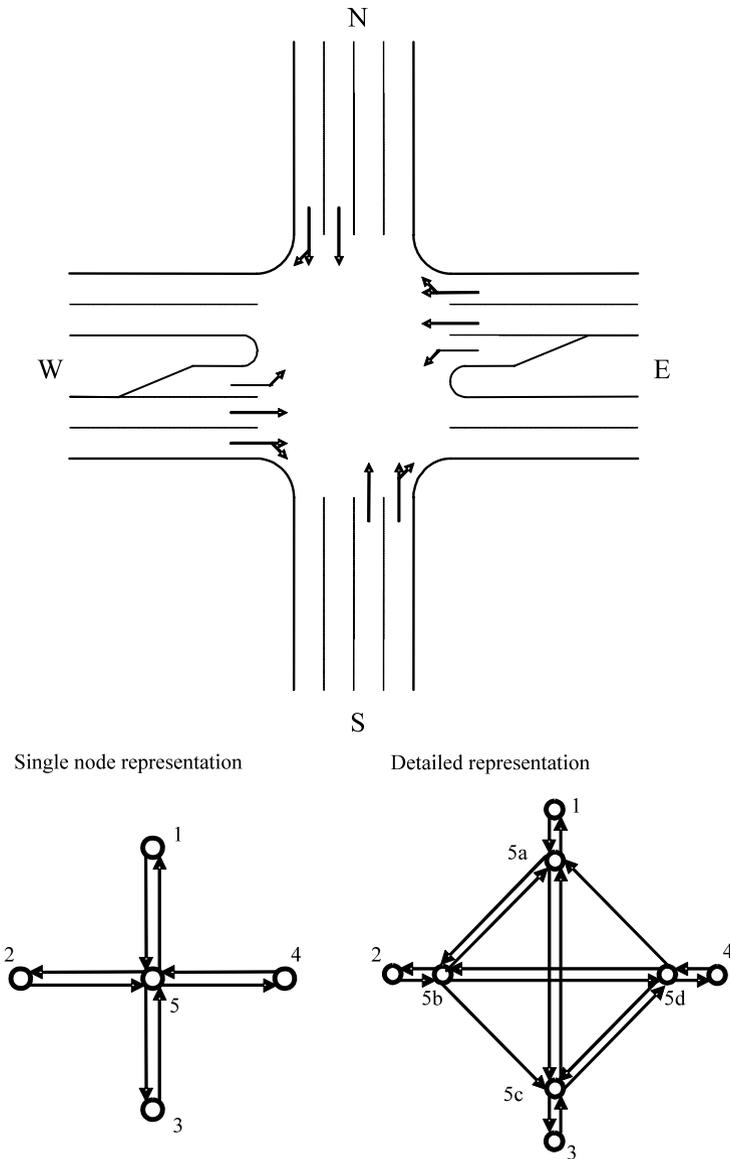


Fig. 2.18 Graphs for a road intersection

if the link represents only the waiting time for a maneuver, tr_a and mc_a are zero, and the same consideration is true for monetary costs and waiting times on most pedestrian links. If an individual link represents both the trip along a road section and queuing at the intersection, its cost function will include both travel time tr_a and queuing time tw_a .

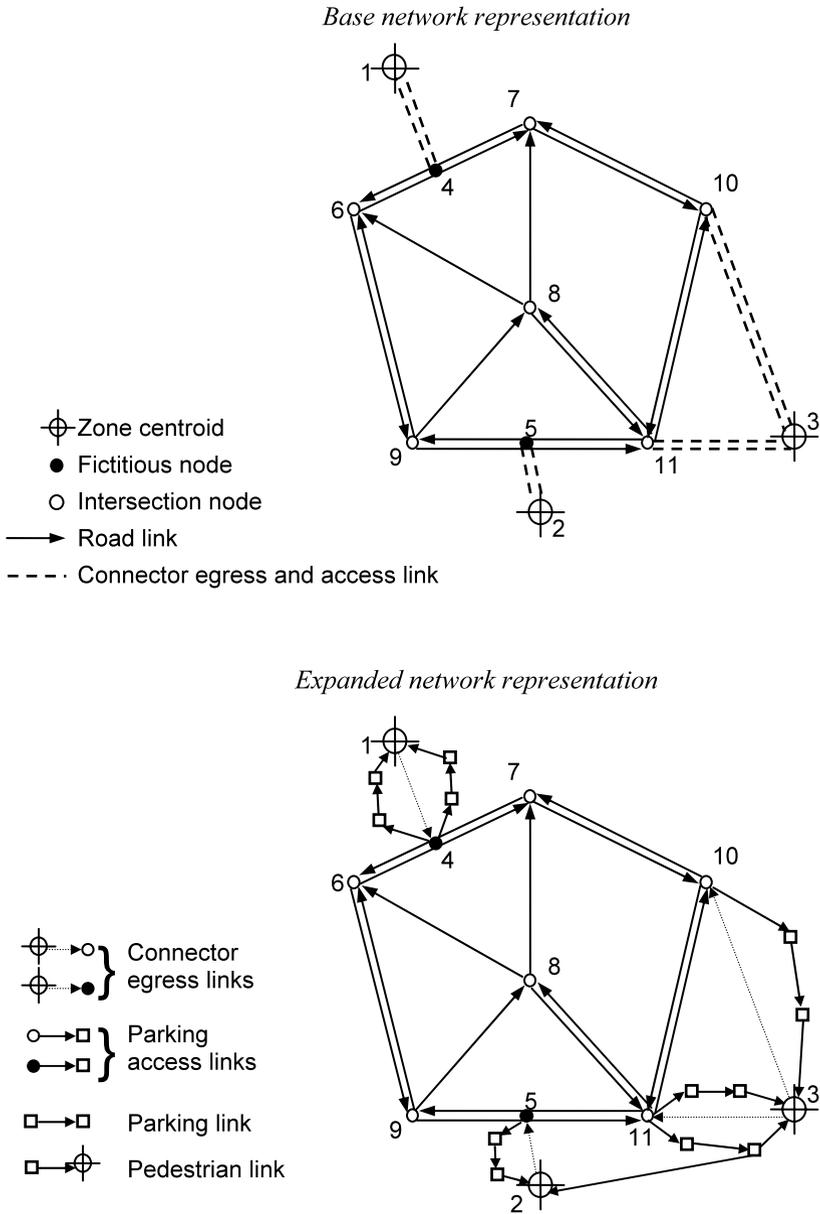


Fig. 2.19 Explicit representation of parking supply

In the most general case, the monetary cost term mc_a includes the cost items that are perceived by the user. Because users do not usually perceive other consumption (motor oil, tires, etc.), in applications monetary costs are usually identified as the

toll (if any) and fuel consumption:

$$mc_a = mc_{\text{toll}} + mc_{\text{fuel}}(\mathbf{f}).$$

The latter depends on the specific consumption (liters/km), which can vary in relation to the average speed and hence to the congestion level. In practice, these variations are sometimes ignored and the monetary cost is calculated as a function of the toll and the average unit consumption.

Performance functions for travel time and queuing time attributes are derived by following both a behavioral (deductive) and experimental (inductive) approach. For the waiting links, for example, the results of queuing theory are generally used (see Sect. 2.2.2). However, their mere implementation has not always permitted proper coverage of all situations in practice, which is why such relations often include approximated adjustment terms obtained from empirical observations.

Listing all the performance functions that can be adopted for the elements of different continuous service systems is beyond the scope of this book. In the following, we therefore present some examples of performance functions both for travel links and waiting links, following the two approaches mentioned. It should also be stressed that, consistently with the assumption of intraperiod stationarity, stationary traffic flow variables and results are used.

Running Links Starting from the (stable regime) speed–flow relationship, the (stable regime) travel time of a running link a can be calculated as a function of the flow:

$$tr_a = L_a/v_a(f_a) \quad (2.4.2)$$

where

- tr_a is the running time on link a
- f_a is the flow on link a
- L_a is the length of the running link a
- v_a is the mean speed on link a assuming a stable regime

Below we introduce the relationships between travel time tr_a and flow f_a for uninterrupted flow conditions, for various types of road infrastructures: motorways and urban and extraurban roads.

(a) Motorway Links On motorway links flow conditions are typically uninterrupted and it is assumed that the waiting time component is negligible because it occurs on those sections (ramps, tollbooths, etc.) that are usually represented by different links.

Link travel time is usually obtained through empirical statistical relationships. One of the most popular expressions, referred to as the BPR cost function, has the following specification.

$$tr_a(f_a) = \frac{L_a}{v_{oa}} + \left(\frac{L_a}{v_{ca}} - \frac{L_a}{v_{oa}} \right) \left(\frac{f_a}{Q_a} \right)^4 \quad (2.4.3)$$

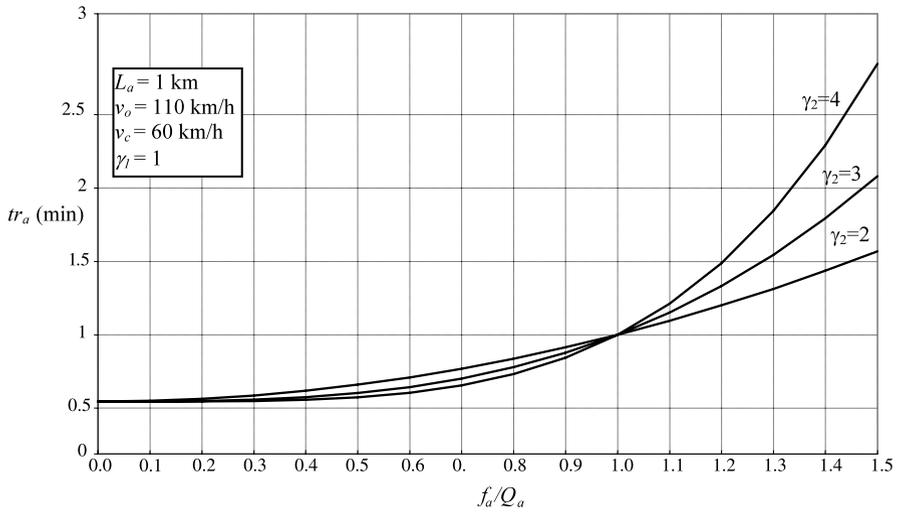


Fig. 2.20 Motorway travel time function (2.4.3) for different values of some parameters

where

- L_a is the length of link a
- v_{0a} is the free-flow average speed
- v_{ca} is the average speed with flow equal to capacity
- Q_a is link capacity, that is, the average maximum number of equivalent vehicles that can travel along the road section in a time unit. Capacity is usually obtained as the product of the number of lanes on the link a , N_a , and lane capacity, Q_{ua}

From (2.4.3) it can be noted that, in the case of motorways, cost functions are separable. The influence of flows on the performances of other links (e.g., the opposite direction or entrance/exit ramps) is significantly reduced by the characteristics of the infrastructure (divided carriageways, grade-separated intersections, etc.).

The values of v_{0a} , v_{ca} , and Q_a depend on the geometric and functional characteristics of the section (width of lanes, shoulders, and median strips; bend radiuses; longitudinal slopes; etc.). Typical values can be found in different sources; the Highway Capacity Manual (HCM) is the most complete and systematic (see Reference Notes). Parameters γ_1 and γ_2 are typically estimated on empirical data.

Figure 2.20 shows a diagram of (2.4.3) for different parameter values. Note that this function associates a travel time with the link also when flows are above link capacity (oversaturation), even though such flows are not possible in reality. However, in applications oversaturation is often allowed for reasons connected with mathematical properties and solution algorithms of static equilibrium assignment models (see Chap. 5). From a computational point of view, the oversaturation assumption should not influence the results significantly if the value of parameter γ_2 , that is, the delay penalty due to capacity overloading, is large enough.

Values of γ_2 are typically much larger than one; that is, the function is more-than-linear in flow/capacity ratios. This phenomenon is rather frequent in congested systems. It should also be noted that, if the flow is close to capacity, resulting instability challenges the within-day stationarity assumptions and the cost functions adopted. In this sense, delay functions should be considered as “penalty” functions preventing major oversaturation, rather than estimates of actual travel times.

(b) *Extraurban Road Links* Users traveling on an extraurban road behave differently according to the number of lanes available for each direction: single lane (two-lane arterial) or two or more lanes (four-lane arterial, six-lane arterial, etc.).

In the former case, the capacity and travel conditions in each direction are not influenced by the flow in the opposite direction. For this type of road, the same formula (2.4.3) described for motorway links can be used, although with different parameters. These can again be deduced from capacity manuals, such as the HCM, or from other specific empirical studies.

In the case of roads with one lane in each direction, link performances depend on the flow in both directions: because overtaking is not always possible, vehicles may reduce the average speed. In practice, it is often assumed that link capacity has a value common to both directions, and the travel time function is modified as follows.

$$tr_a(f_a, f_{a^*}) = \frac{L_a}{v_{0a}} + \gamma_a \left(\frac{L_a}{v_{ca}} - \frac{L_a}{v_{0a}} \right) \left(\frac{f_a + f_{a^*}}{Q_{aa^*}} \right)^{\gamma_2} \quad (2.4.4)$$

where, apart from the symbols introduced previously, the link in the opposite direction is denoted by a^* and overall capacity in both directions by Q_{aa^*} .

(c) *Urban Road Links* In an urban context, given the relatively short lengths of road sections, travel speed is more dependent upon road physical and functional characteristics than upon the flow traveling on them. The higher the dependence is on factors such as section bendiness or roadside parking, the lower the impact of flow.

As an example, we report the empirical relation for estimating travel speed calibrated on survey sample data from the Napoli (Italy) urban area, integrated with microscopic simulation data (see the bibliographical note):

$$v_a = 29.9 + 3.6Lu_a - 0.6P_a - 13.9T_a - 10.8D_a - 6.4S_a + 4.7Pv_a - 1.0E-04 \frac{(f_a/Lu_a)^2}{1 + T_a + D_a + S_a} \quad (2.4.5)$$

where

- Lu_a is the useful width in meters of link a
- P_a is the nonnegative slope in % of link a
- T_a is the tortuosity of link a , in values in the interval $[0, 1]$
- D_a is an index of disturbance to traffic from external factors (entry from sideroads, irregular parking, pedestrian crossings, etc.) in values in the interval $[0, 1]$

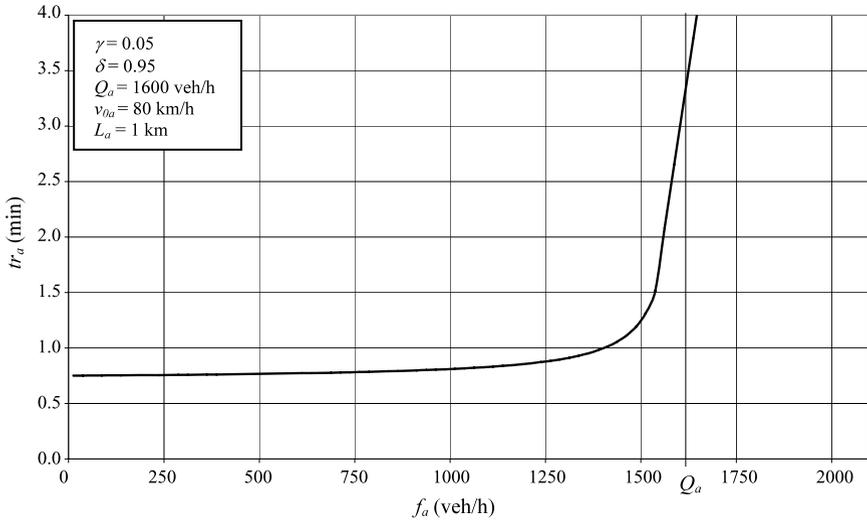


Fig. 2.21 Hyperbolic travel time cost function

- S_a is the percentage of length of a occupied by parking
- Pv_a is a dummy variable of 1 if the pavement of link a is asphalt, 0 otherwise
- f_a is the equivalent flow on link a in equiv. vehicles/hour

The travel time on link a may thus be calculated by multiplying the time obtainable from (2.4.5) by a corrective factor $c(L_a)$, which makes allowance for the effect of transient motions at the ends of the link (in the case of stopping at intersections):

$$tr_a = \frac{L_a}{v_a} \cdot c(L_a) = \frac{L_a}{v_a} \cdot \frac{1}{1 - \exp(-0.47 - 0.48E-02 \cdot L_a)} \tag{2.4.6}$$

where L_a is the road section length in km.

A further example of link travel time function is the hyperbolic expression given by Davidson, which also holds for interrupted flow (delays at intersections are thus included):

$$\begin{cases} tr_a = (L_a/v_{0a})(1 + \gamma f_a/(Q_a - f_a)) & \text{for } f_a \leq \delta Q_a \\ tr_a = \text{tangent approximation} & \text{for } f_a > \delta Q_a \end{cases} \tag{2.4.7}$$

with $\delta < 1$ and $Q_a =$ link capacity. Also see Fig. 2.21.

In this last case the tangent approximation is necessary because tr_a tends to ∞ for f_a going to Q_a . This condition is unrealistic because the oversaturated period has a finite duration.

Waiting Links

(a) *Toll-Barrier Links* In the case of links representing queuing systems, it is assumed that average waiting time is the only significant time performance variable. In

simple cases (e.g., a link corresponds to all toll lanes), the average undersaturation waiting time can be obtained by using a stochastic queuing model:

$$tw_a^u(f_a) = T_s + (T_s^2 + \sigma_s^2) \cdot \frac{f_a}{2} \cdot \frac{1}{1 - f_a/Q_a} \quad (2.4.8)$$

where

T_s is the average service time for each toll lane
 σ_s^2 is the variance of the service time at the pay-point
 $Q_a = N_a/T_s$ is the link (toll-barrier) capacity equal to the product of the number of lanes (N_a) by the capacity of each lane ($1/T_s$)

Expression (2.4.8) is derived from the assumption of a queuing system $M/G/1$ (∞ , *FIFO*) with Poisson arrivals and general service time (see Sect. 2.2.2.3).

The values of T_s and σ_s^2 depend on various factors such as the tolling structure (fixed, variable) and the payment method (manual, automatic, etc.). Note that the average waiting time obtained through (2.4.8) is larger than the average service time T_s even though the arriving flow is lower than the system's capacity. This effect derives from the presence of random fluctuations in the headways between user arrivals and service times. Hence the delay expressed by (2.4.8) is known as "stochastic delay."

Moreover, the average delay computed with (2.4.8) tends to infinity as the flow f_a tends to capacity (i.e., if f_a/Q_a tends to one). This would be the case if the arrivals flow f_a remained equal to capacity for an infinite time, which does not occur in reality. In order to avoid unrealistic waiting times and for reasons of theoretical and computational convenience, two different methods can be adopted. The first, and less precise, method assumes that (2.4.8) holds for flow values up to a fraction α of the capacity, for example, $f_a \leq 0.95Q_a$. For higher values, the curve is extended following its *linear approximation*, that is, in a straight line passing through the point of coordinates αQ_a , $tw(\alpha Q_a)$ with angular coefficient equal to the derivative of (2.4.8) computed at this point:

$$tw_a(f_a) = tw_a(\alpha Q_a) + K(f_a - \alpha Q_a) \quad (2.4.9)$$

with

$$K = \frac{T_s^2 + \sigma_s^2}{2} \cdot \frac{1}{(1 - \alpha)^2}$$

Figure 2.22 shows the relationships (2.4.8) and (2.4.9) for some values of the parameters.

A more rigorous method is based on calculating oversaturation delay using a deterministic queuing model with an arrival rate equal to f_a , deterministic service times equal to T_s and an oversaturation period equal to the reference period duration T (see Sect. 2.2.2.2). The deterministic average (oversaturation) delay tw_a^d is then equal to:

$$tw_a^d = T_s + \left(\frac{f_a}{Q_a} - 1 \right) \frac{T}{2} \quad (2.4.10)$$

which, for a *given* capacity, is a linear function of the arrivals flow f_a .

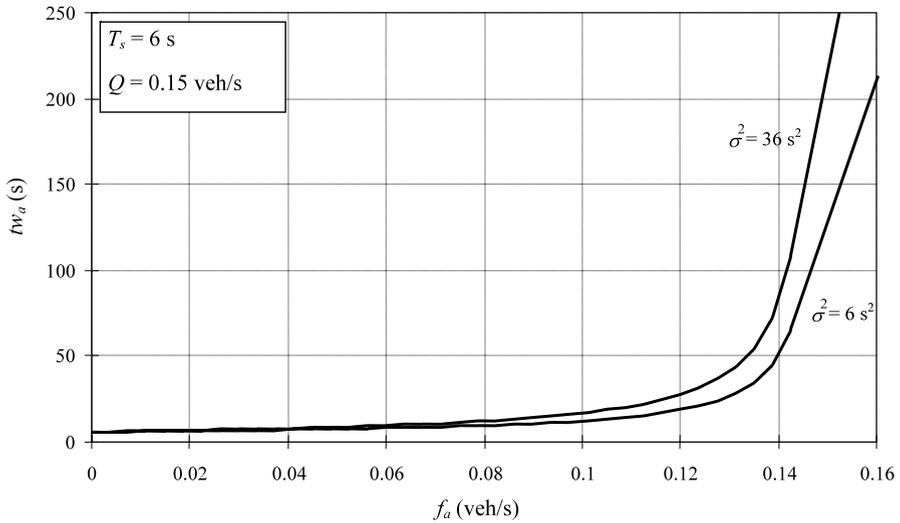


Fig. 2.22 Waiting time functions (2.4.8) and (2.4.9) at toll-barrier links

Note that in this case the assumption of intraperiod stationarity is challenged because even if the arrivals flow rate f_a and capacity $1/T_s$ are constant over the whole reference period T , the waiting time is different for users arriving in different instants of the reference period. In static models it is assumed that users perceive the average waiting time. Intraperiod dynamic models, discussed in Chap. 7, remove this assumption.

The average delay tw_a can be calculated by combining the stochastic undersaturation average delay tw_a^u expressed by (2.4.8) with the deterministic average oversaturation delay tw_a^d , expressed by (2.4.10). The combined delay function is such that the deterministic delay function is its oblique asymptote (see Fig. 2.23). The following equation results.

$$\begin{aligned}
 tw_a(f_a) = T_s + (T_s^2 + \sigma^2) \frac{f_a}{2} + \frac{T}{4} \left\{ \frac{f_a}{Q_a} - 1 \right. \\
 \left. + \left[\left(\frac{f_a}{Q} - 1 \right)^2 + \frac{4(f_a/Q_a)}{Q_a T} \right]^{1/2} \right\} \quad (2.4.11)
 \end{aligned}$$

(b) *Signal-Controlled Intersection Links* Queuing and delay phenomena at signalized intersections can be obtained from the queuing theory results reported in Sect. 2.2.2. In fact, signalized intersections are a particular case of servers for which capacity is periodically equal to zero (when the signal is red). During such times the system is necessarily oversaturated.

The simplest case is that of a *signal-controlled intersection* not interacting with adjacent ones (*isolated intersection*), without lanes reserved for right or left turns.

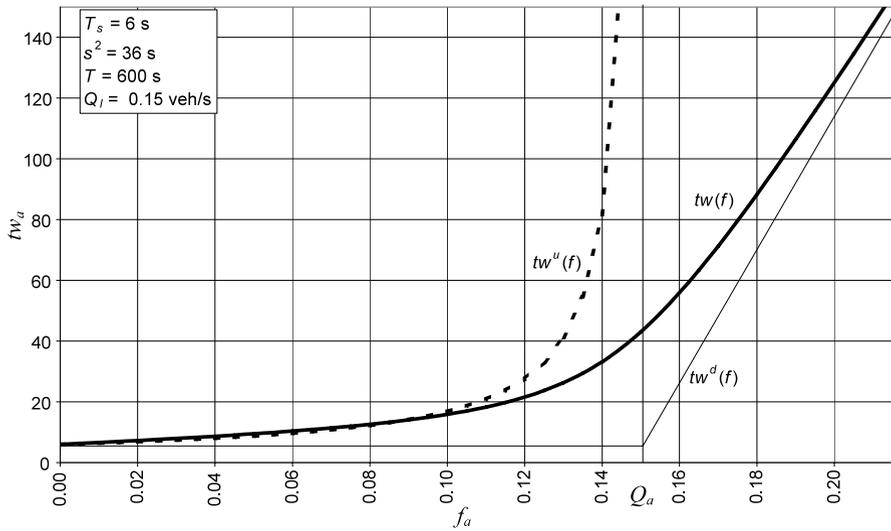


Fig. 2.23 Under- and oversaturation waiting time functions for toll barrier links

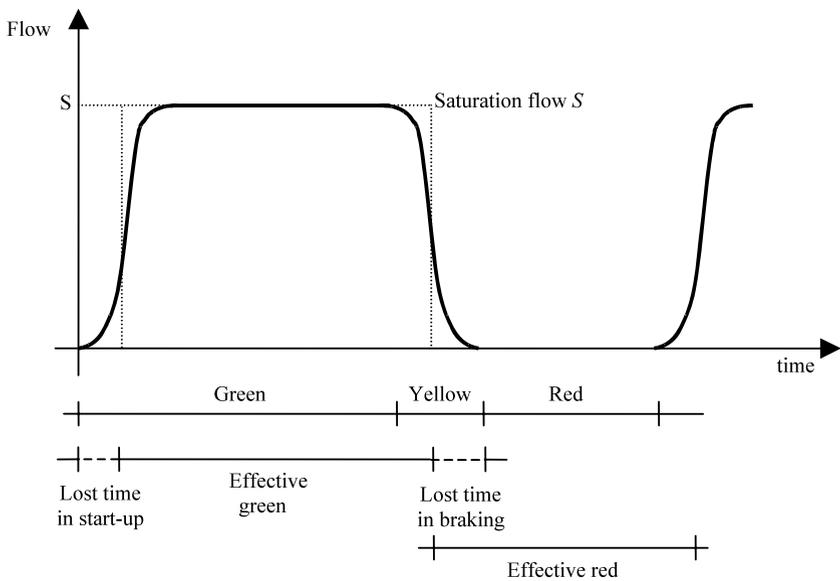


Fig. 2.24 Discharge flow from signal-controlled intersection in relation to cycle phases

Below we first introduce the assumptions and variables for each access as well as the most widely used calculation method. We then present the various models for calculating delays at intersections.

It is common to divide the cycle length into two time intervals (Fig. 2.24 illustrates the quantities associated with a traffic-light cycle). The effective green time equals the green plus yellow time minus the lost time, during which departures occur at a constant service rate, given by the inverse of saturation flow. The effective red time is the difference between cycle length and the effective green time, during which no departures occur.

Below, to simplify the notation, we omit the index of link a . Moreover, to facilitate application of the results in Sect. 2.2.2, the symbol \bar{u} instead of f is used for the arrivals flow. Let:

T_c be the cycle length for the whole intersection
 G be the effective green time for an approach
 $R = T_c - G$ be the effective red time for the approach
 $\mu = G/T_c$ be the effective green/cycle ratio for the approach

The number of vehicles arriving at the approach during time interval T_c is given by the following equation.

$$m_{IN}(\tau, \tau + T_c) = \bar{u} \cdot T_c$$

The maximum number of users that may leave the approach, during time interval T_c , is given by:

$$S \cdot G = \mu \cdot S \cdot T_c$$

where S is the saturation flow of the intersection approach, that is, the maximum number of equivalent vehicles which in the time unit could cross the intersection if the traffic lights were always green ($\mu = 1$). Alternatively, the saturation flow may be defined as the maximum discharge rate that may be sustained by a queue during the green–amber time.

Hence the actual capacity of the approach is given by:

$$Q = \frac{S \cdot G}{T_c} = \mu \cdot S$$

Thus, the approach can be defined *undersaturated* if:

$$\bar{u} \cdot T_c < \mu \cdot S \cdot T_c$$

that is:

$$\bar{u} < \mu \cdot S \tag{2.4.12}$$

On the other hand the approach is defined *oversaturated* if:

$$\bar{u} \geq \mu \cdot S \tag{2.4.13}$$

The saturation flow rate of an intersection can in principle be obtained through specific traffic surveys; in practice, however, empirical models based on average results are often used. The Highway Capacity Manual (HCM) describes one of the

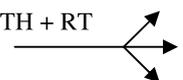
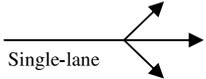
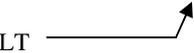
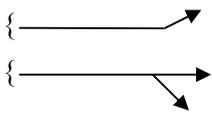
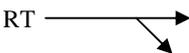
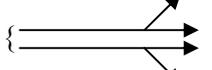
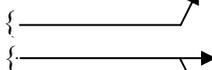
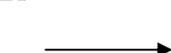
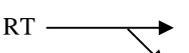
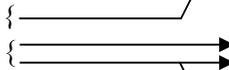
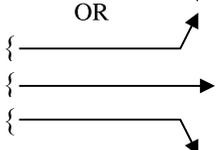
No. of Lanes	Movements by Lane	Lane Group Possibilities
1	LT + TH + RT 	1  Single-lane
2	EXC LT  TH + RT 	2 
2	LT + TH  TH + RT 	1  OR 2 
3	EXC LT  TH  TH + RT 	2  OR 3 

Fig. 2.25 Typical lane groups for the HCM method for calculating saturation flow

most popular methods. To apply this method, it is necessary to determine appropriate lane groups. A lane group is defined as one or more lanes of an intersection approach serving one or more traffic movements with which a single value of saturation flow, capacity, and delay can be associated. Both the geometry of the intersection and the distribution of traffic movements are taken into account to segment the intersection into lane groups. In general, the smallest number of lane groups that adequately describes the operation of the intersection is used. Figure 2.25 shows some common lane group schemes suggested by the HCM. The saturation flow rate of an intersection is computed from an “ideal” saturation flow rate, usually 1900 equivalent passenger cars per hour of green time per lane (pcphgpl), adjusted for a variety of prevailing conditions that are not ideal. The method can be summarized by the following expression,

$$S = S_0 \cdot N \cdot F_w \cdot F_{HV} \cdot F_g \cdot F_p \cdot F_{bb} \cdot F_a \cdot F_{RT} \cdot F_{LT}$$

where

S	is the saturation flow rate for the specific lane group, expressed as a total for all lanes in the lane group under prevailing conditions, in vphg
S_0	is the ideal saturation flow rate per lane, usually 1900 pcphgpl
N	is the number of lanes in the lane group
F_w	is the adjustment factor for lane width (12 ft or 3.66 m lanes are standard)
F_{HV}	is the adjustment factor for heavy vehicles in the traffic flow
F_g	is the adjustment factor for approach grade
F_p	is the adjustment factor for the existence of a parking lane adjacent to the lane group and the parking activity in that lane
F_{bb}	is the adjustment factor for the blocking effect of local buses that stop within the intersection area
F_a	is the adjustment factor for the area type
F_{RT}	is the adjustment factor for right turns in the lane group
F_{LT}	is the adjustment factor for left turns in the lane group

The first six adjustment factors not connected with the type of turning maneuvers are reported in Fig. 2.26.

Once the approach capacity $Q_l = \mu S$ is known, we may calculate the queue length and mean waiting time tw_a , using models derived from different approaches.

Application of Queuing Models From (2.4.12) and (2.4.13) it is clear that the results discussed in Sect. 2.2.2 hold for a queuing system representing a signalized intersection approach. In this context, the server's capacity Q coincides with the actual capacity of access: $Q = \mu \cdot S$. The latter is the weighted mean between the zero value of the "red" period and that equal to S for the "green" period, with $\mu = G/T_c$.

In the case in which access occurs in undersaturation conditions, the queue length may be calculated using (2.2.18) in which capacity assumes alternatively a value of zero, in intervals of length R (intervals of effective red), and a value of S , in intervals of length G (intervals of effective green) (see Fig. 2.27). As the system is undersaturated, at the end of each interval of effective green the queue is zero: $n_u(I \cdot T_c) = 0 \forall i$, where i stands for the progressive number of cycles. Thus, for each interval of effective red we have $n(\tau_0) = 0$ with $\tau_0 = I \cdot T_c$ and, setting $Q = 0$ in (2.2.18), the queue length is equal to:

$$n_u^R(\tau) = \bar{u}(\tau - I \cdot T_c) \quad I \cdot T_c \leq \tau \leq I \cdot T_c + R \quad (2.4.14)$$

The queue length reaches a maximum value at the end of the red-time, equal to:

$$n_u^R(I \cdot T_c + R) = \bar{u}R = \bar{u}(1 - \mu)T_c$$

Thus, at the beginning of the interval of effective green we have $n(\tau_0) = \bar{u}(1 - \mu)T_c$ with $\tau_0 = I \cdot T_c + R$, and the queue length in a certain instant τ of the interval is given by (2.2.18) with $Q = S$:

$$n_u^G(\tau) = \max\{0, \bar{u}(1 - \mu)T_c - (S - \bar{u})(\tau - I \cdot T_c - R)\}$$

ADJUSTMENT FACTOR FOR AVERAGE LANE WIDTH F_w									
Average lane width, W (FT)	8	9	10	11	12	13	14	15	16
F_w	0.867	0.900	0.933	0.967	1.000	1.033	0.067	1.100	1.133
ADJUSTMENT FACTOR FOR HEAVY VEHICLES F_{HW}									
Percentage of heavy vehicles (%)	0	2	4	6	8	10	15	20	
F_{HW}	1.000	0.980	0.962	0.943	0.926	0.909	0.870	0.833	
Percentage of heavy vehicles (%)	25	30	35	40	45	50	75	100	
F_{HW}	0.800	0.769	0.741	0.714	0.690	0.667	0.571	0.500	
ADJUSTMENT FACTOR FOR APPROACH GRADE F_g									
Grade (%)	-6	-4	-2	0	+2	+4	+6	+8	≥ 10
F_g	1.030	1.020	1.010	1.000	0.990	0.980	0.970	0.960	0.950
ADJUSTMENT FACTOR FOR PARKING F_p									
F_p	No. of parking maneuvers per hour								
No. of lanes in lane group	No parking	0	10	20	30	≥ 40			
1	1.000	0.900	0.850	0.800	0.750	0.700			
2	1.000	0.950	0.925	0.900	0.875	0.850			
3 or more	1.000	0.967	0.950	0.933	0.917	0.900			
ADJUSTMENT FACTOR FOR BUS BLOCKAGE F_{bb}									
F_{bb}	No. of buses stopping per hour								
No. of lanes in lane group	0	10	20	30	≥ 40				
1	1.000	0.960	0.920	0.880	0.840				
2	1.000	0.980	0.960	0.940	0.920				
3 or more	1.000	0.987	0.973	0.960	0.947				
ADJUSTMENT FACTOR FOR AREA TYPE F_a									
Type of area									F_a
CBD (Center Business District)									0.900
All other areas									1.000

Fig. 2.26 Adjustment factors in the HCM method for saturation flow

$$I \cdot T_c + R \leq \tau \leq I \cdot T_c + R + G \tag{2.4.15}$$

The time period (within the green) in which the queue is exhausted is (see (2.4.15)):

$$\Delta\tau_0 = \frac{\bar{u}(1 - \mu)T_c}{(S - \bar{u})}$$

The queue in undersaturation conditions therefore shows a periodic time trend, with zero values at the end of effective green time (i.e., at the beginning of the red interval) and maximum values at the end of the effective red interval (see Fig. 2.27).

However, if the system is in oversaturation conditions ($\bar{u} \geq \mu \cdot S$), the total queue length is obtained by summing the queue length in undersaturation to the queue length in oversaturation (see Fig. 2.28). The queue length in undersaturation, $n_u(\tau)$, is obtained once again by (2.4.14) and (2.4.15), for an arrivals rate equal to capacity

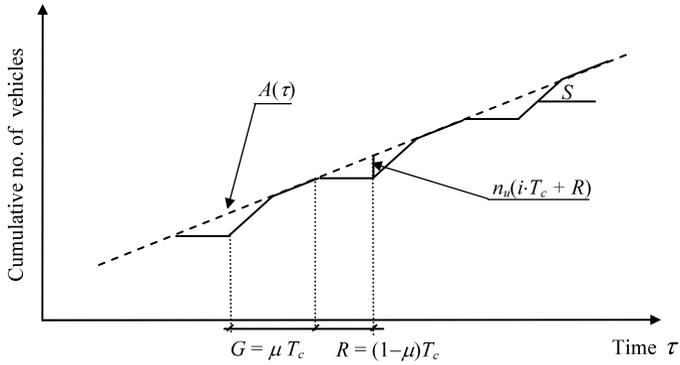


Fig. 2.27 Deterministic queuing model for signalized intersections, undersaturated conditions

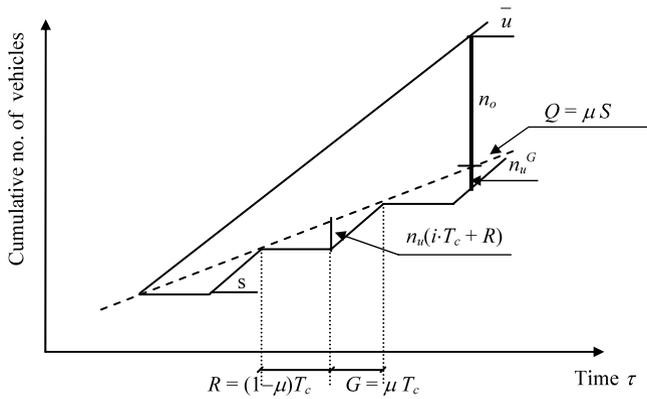


Fig. 2.28 Deterministic queuing model for signalized intersections, oversaturated conditions

$(\bar{u} = \mu \cdot S)$:

$$n_u^R(\tau) = \mu \cdot S(\tau - I \cdot T_c) \quad I \cdot T_c \leq \tau \leq I \cdot T_c + R \quad (2.4.16)$$

$$n_u^G(\tau) = \mu \cdot S(1 - \mu)T_c - S(1 - \mu)(\tau - I \cdot T_c - R) \quad I \cdot T_c + R \leq \tau \leq I \cdot T_c + R + G \quad (2.4.17)$$

The *oversaturated queue length* can be computed with the queue obtained from (2.2.18) with $Q = \mu \cdot S$, $\tau_0 = 0$ and $n(\tau_0) = 0$ (see Fig. 2.28):

$$n_0(\tau) = (\bar{u} - \mu \cdot S)\tau \quad (2.4.18)$$

The expressions of queue length allow us to determine the deterministic delay at intersections, as described below.

For undersaturated conditions $\bar{u} < \mu S$, the average individual delay tw_{US} can easily be obtained from the evolution over time of the queue length, as described

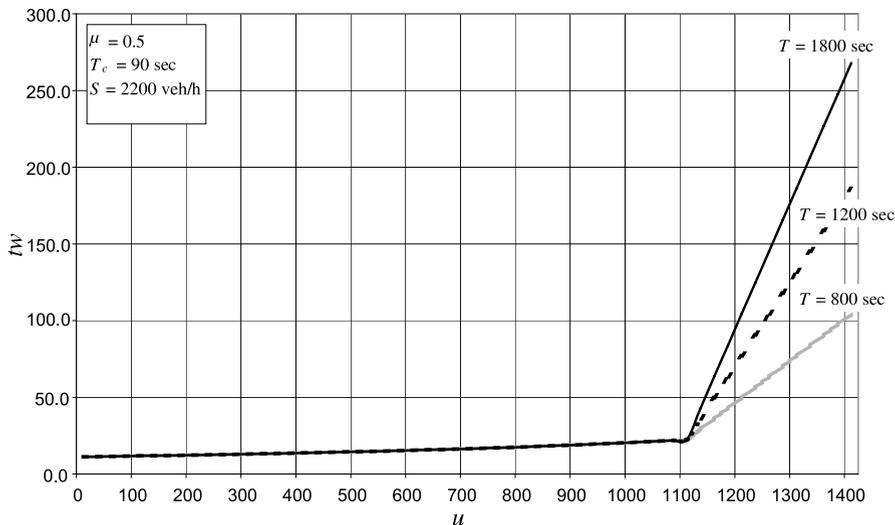


Fig. 2.29 Deterministic delay function at a signalized intersection

by (2.4.14) and (2.4.15):

$$tw_{US} = \frac{T_c[1 - \mu]^2}{2[1 - \bar{u}/S]} \tag{2.4.19}$$

In oversaturated conditions, $\bar{u} > \mu S$, for the deterministic case, the queue length, and respective delay, would tend theoretically to infinity. In practice, however, oversaturation lasts only for a finite period of time T , and the average delay tw_{OS} can be calculated from the evolution over time of queue length as described by (2.4.16) through (2.4.18):

$$tw_{OS} = \frac{T_c[1 - \mu]}{2} + \frac{T}{2}[(\bar{u}/\mu S) - 1] \tag{2.4.20}$$

Note that the first term is the value of (2.4.19) for $\bar{u} = \mu \cdot S$. The delay for the arrival flows can be computed through (2.4.19) for $\bar{u} < \mu \cdot S$, and through (2.4.20) for $\bar{u} \geq \mu \cdot S$, as depicted in Fig. 2.29. Note that the diagram depicted in Fig. 2.29 shows an increase in average delay also for flows below the capacity. This is due to the increase in the undersaturated delay expressed by (2.4.19).

Stochastic delay models are based on the results of queuing theory. More precisely, a signalized intersection is considered to be a $M/G/1 (\infty, FIFO)$ system. Therefore, the average delay is (see Sect. 2.2.2.3):

$$tw_q^{st}(u) = \frac{(\bar{u}/\mu S)^2}{2\bar{u}(1 - \bar{u}/\mu S)} \tag{2.4.21}$$

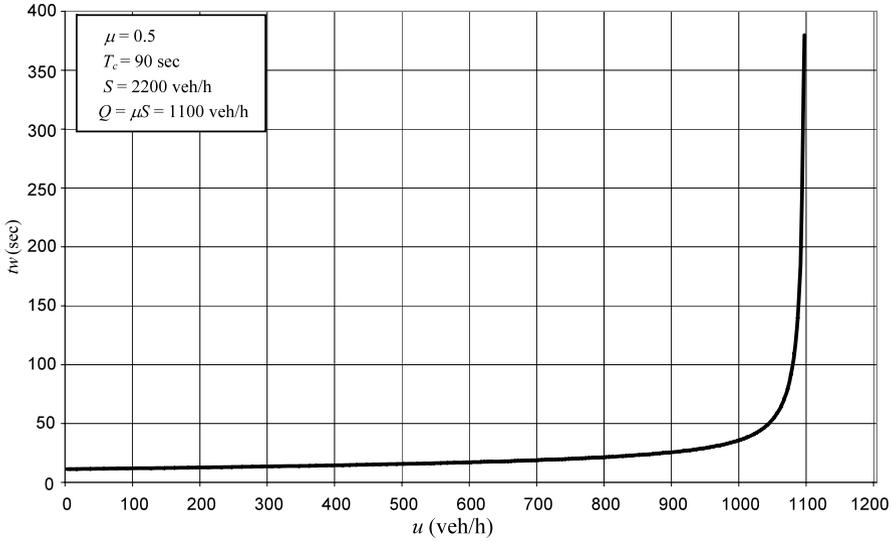


Fig. 2.30 The Webster delay model

Overall Delay Models The total (mean individual) delay equals the sum of the deterministic and the stochastic terms (introduced in the previous section), and sometimes, of terms calibrated through experimental observations.

One of the best known expressions is *Webster’s three-term* formula, proposed for an isolated intersection under the assumption of random (Poisson) arrivals and undersaturation conditions ($f_a/Q_a < 1$) (see Fig. 2.30):

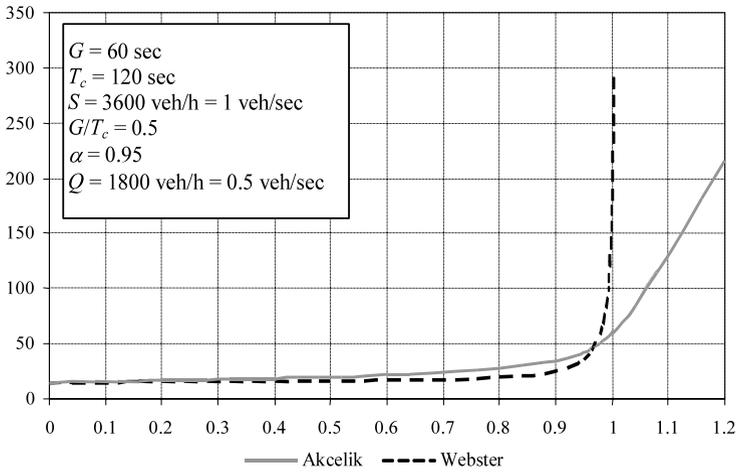
$$\begin{aligned}
 tw_a(f_a) = & \frac{T_c(1 - \mu)^2}{2(1 - f_a/S_a)} + \frac{(f_a/Q_a)^2}{2f_a(1 - f_a/Q_a)} \\
 & - 0.65(Q_a/f_a^2)^{1/3}(f_a/Q_a)^{2+\mu}
 \end{aligned}
 \tag{2.4.22a}$$

where

- T_c is the cycle length
- μ is the effective green to cycle length ratio for the lane group represented by link a
- Q_a is the capacity of the lane group represented by link a

The first term expresses the deterministic delay (see (2.4.19)), the second is the stochastic delay due to the randomness of the arrivals (see (2.4.21)), and the third term is an adjustment term obtained by simulation results. This term amounts to about 10% of the sum of the other two, hence its established use in practical applications of Webster’s two-term formula:

$$tw_a(f_a) = 0.9 \left[\frac{T_c(1 - \mu)^2}{2(1 - f_a/S_a)} + \frac{(f_a/Q_a)^2}{2f_a(1 - f_a/Q_a)} \right]
 \tag{2.4.22b}$$



f	f/Q	Akcelik	Webster
0.00	0.00	15.00	15.00
0.10	0.20	16.67	16.87
0.20	0.40	18.75	19.26
0.25	0.50	20.00	20.77
0.30	0.60	21.93	22.61
0.40	0.80	27.95	28.45
0.50	1.00	60.00	
0.60	1.20	216.75	

Fig. 2.31 Waiting time functions at a signalized intersection

The delay given by (2.4.22) tends to infinity for an arrivals flow f_a , which tends to capacity $Q = \mu \cdot S$ (see Fig. 2.30). Thus Webster’s formula cannot be used to simulate delays at oversaturated signalized intersections. To overcome this limit, it is possible to apply the two heuristic methods described for (2.4.8).

The first method applies (2.4.22) for values of f_a up to a percentage α of the capacity whereas for higher values a linear approximation of the function is used, thereby ensuring the continuity of the function and its first derivative:

$$tw_a(f_a) = tw_a(\alpha Q_a) + \left. \frac{d}{df} tw_a(f) \right|_{f_a=\alpha Q_a} \cdot (f_a - \alpha Q_a) \quad f_a \geq \alpha Q_a \quad (2.4.23)$$

The second method computes the oversaturation delay combined with the stochastic delay, deforming the stochastic delay function so that it has an oblique asymptote defined by the deterministic delay. Based on these considerations, Akcelik’s formula

was proposed:

$$\begin{aligned}
 tw_a(f_a) &= \frac{0.5T_c(1 - \mu_a)^2}{1 - \mu_a X_a} & X_a \leq 0.50 \\
 tw_a(f_a) &= \frac{0.5T_c(1 - \mu_a)^2}{1 - \mu_a X_a} + 900 \cdot T \cdot \left\{ X_a - 1 \right. \\
 &\quad \left. + \left[(X_a - 1)^2 + \frac{8(X_a - 0.5)}{\mu_a S_a T} \right]^{1/2} \right\} & 0.50 \leq X_a \leq 1 \quad (2.4.24) \\
 tw_a(f_a) &= 0.5T_c(1 - \mu_a) + 900 \cdot T \cdot \left\{ X_a - 1 \right. \\
 &\quad \left. + \left[(X_a - 1)^2 + \frac{8(X_a - 0.5)}{\mu_a S_a T} \right]^{1/2} \right\} & X_a > 1
 \end{aligned}$$

where $X_a = f_a/Q_a$ is the flow/capacity ratio, the times tw_a and T_c are expressed in seconds, S_a in pcph, and T is the duration of the oversaturation period in hours. Equation (2.4.24) is compared with the Webster formula in Fig. 2.31 for a value of $T = 0.5$ h.

Note that application of the previous formulae for calculating saturation flows, capacities, and average waiting times (delays) in the case of multiple lane groups requires an “exploded” representation of the intersection with several links corresponding to the relevant lane groups and their maneuvers (see Fig. 2.18). For example, in the case of an exclusive right-turn lane a single link can represent such a movement and the associated delay. Sometimes, to simplify the representation, fewer links than lane groups are used; in this case the total capacity of all lane groups is associated with the single link and the resulting delay is associated with the whole flow.

From a mathematical point of view the delay functions discussed so far are separable only if the traffic-signal regulation (assumed known) is such as to exclude interference between maneuvers represented by different links. For example, this is the case for the three-phase regulation scheme of a T-shaped intersection shown in Fig. 2.32. However, if the phases allow conflict points, for example, left turns from the opposite direction with through flows during the same phase, nonseparable cost functions may be necessary, which take account of the reciprocal reduction in saturation flow for maneuvers in conflict, such as for the two-phase scheme for the X intersection in Fig. 2.33.

In general, if a single node represents the entire intersection, the effects of individual maneuvers and lane groups are impossible to distinguish and separable functions are adopted, with a single value of saturation flow, reduced to account for the interfering turns.

In the case of control systems at signalized intersections, the control parameters (cycle length T_c , ratio μ of green time to cycle length) depend on flows arriving at the access roads which converge at the intersection. In this case the delay functions are different and definitely nonseparable.

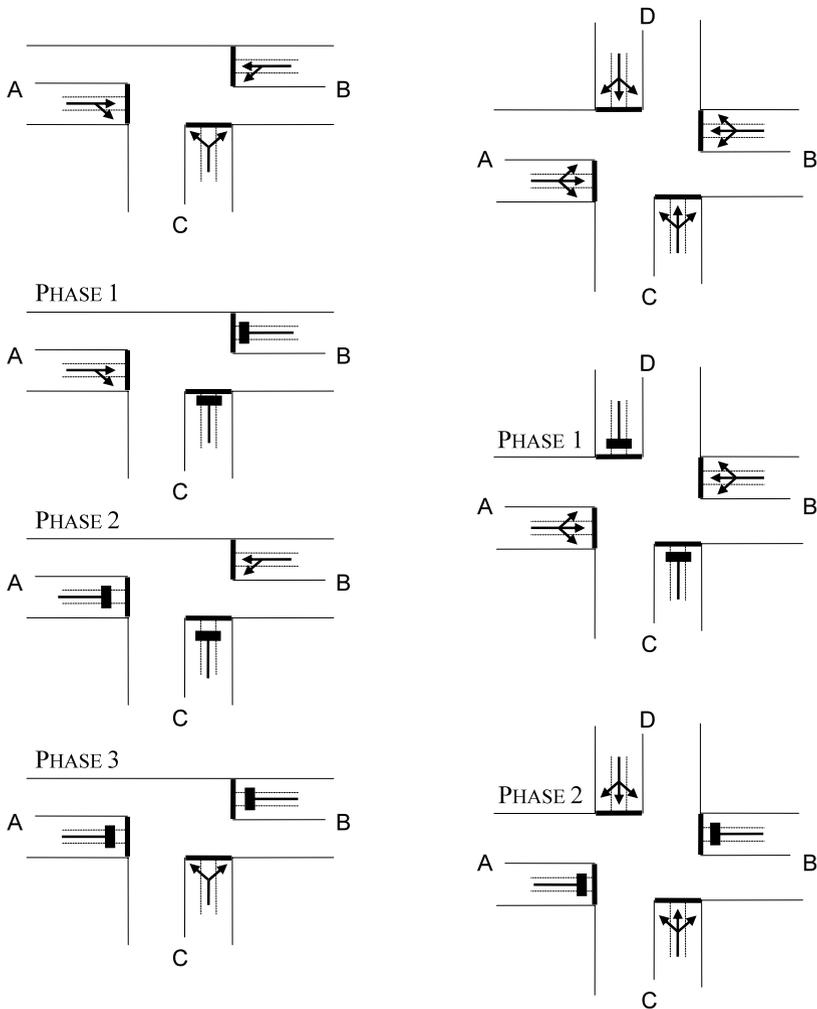
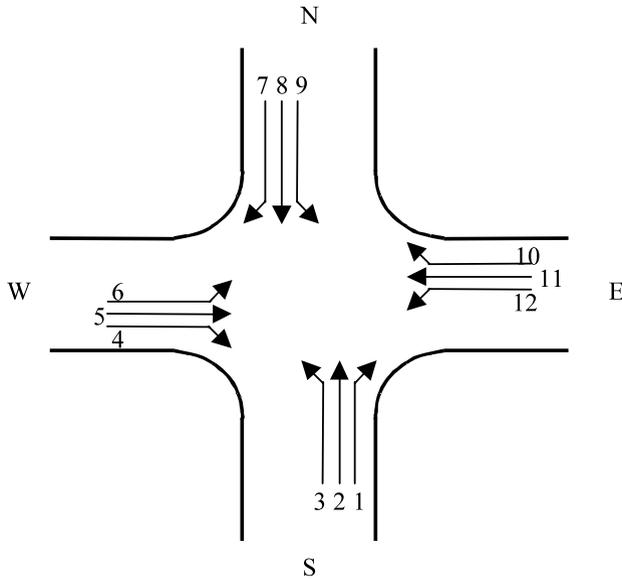


Fig. 2.32 Examples of traffic light phases for 3- and 4-arm intersections

Finally, in the case of networks of interacting intersections (i.e., so close as to affect one another), further regulation parameters must be introduced; hence, calculation of the delay cannot be performed with the formulae presented, but requires more detailed flow simulation models along the road sections joining a pair of adjacent intersections.

(c) *Priority Intersections* To complete the survey of the delay functions, *priority intersections* (i.e., intersections regulated by give-way rules rather than traffic lights) need to be considered. Empirical functions are often used to express average delays; these functions are nonseparable in that right-of-way rules cause delays due to con-



Maneuver		Flows influencing the delay
Direction South-North, right-hand turn	1	1,2,3,5,9
Direction South-North, crossing	2	1,2,3,5,6,9,10,11,12
Direction South-North, left-hand turn	3	1,2,3,5,6,7,8,11,12

Fig. 2.33 Flow conflicts for computing delays at a priority intersection

flicts between flows. As an example, the delay corresponding to the maneuvers at a 4-arm intersection can be calculated by means of the following HCM function.

$$tw_a(f) = \exp(-0.2664 + 0.3967 \ln(f_{conf}) + 3.959A(\ln(f_{conf}) - 6.92)) \quad (2.4.25)$$

where

$tw_a(f)$ is the waiting time expressed in seconds

f_{conf} is the total conflicting flow, which varies according to the maneuver as shown in Fig. 2.33

$A = 1$ if $f_{conf} > 1062$ vehicles/h, 0 otherwise

(d) *Parking Links* Monetary cost (fares) and search time are the most important performance attributes connected to links representing parking in a given area. In general, these attributes differ for links representing different parking types (facilities). The more sophisticated models of search time take into account the congestion effect through the ratio between the average occupancy of the parking facilities of type p , represented by link a , and the parking capacity Q_l .

The average search time can be calculated through a model assuming that available parking spaces of type p are uniformly distributed along a circuit, possibly

mixed with parking spaces of different types (e.g., free and priced parking). If occupancy of a given parking type at the beginning and end of the reference period is inferior to capacity, the following expression can be obtained.

$$ts_a(f_a) = \frac{L_p}{v_s} \frac{1}{occ_2(f_a) - occ_1} \cdot \frac{Q_{tot} \cdot (Q_p + 1)}{Q_p} \cdot \ln\left(\frac{1 + Q_p - occ_1}{1 + Q_p - occ_2(f_a)}\right) - \frac{(Q_{tot} - Q_p)}{Q_p} \quad (2.4.26)$$

where

- $ts_a(f_a)$ is the search time in minutes
- f_a is the flow on parking link a
- L_p is the average length of a parking space
- v_s is the average search speed for a free parking space
- occ_1 is the parking occupancy at the beginning of the reference period
- occ_2 is the parking occupancy at the end of the reference period, depending on flow assigned to the parking link and the turnover rate
- Q_p is the parking capacity of type p corresponding to link a
- Q_{tot} is the total capacity of all parking types mixed with type p in the zone

If one or both occ are above capacity, similar but formally more complicated formulas can be obtained. These expressions are not reported here.

2.4.2 Supply Models for Scheduled Service Transportation Systems

Discontinuous and nonsimultaneous transportation services can be accessed only at given points and are available only at given instants. Typical examples are scheduled services (buses, trains, airplanes, etc.), which can be used only between terminals (bus stops, stations, airports, etc.) and are available only at certain instants (departure times). Scheduled services can be represented by different supply models according to their characteristics and to the consequent assumptions on users' behavior (see Sect. 4.3.3.2). The approach followed in this chapter is based upon the modeling of *service lines*, that is, a set of scheduled runs with equal characteristics. This approach is consistent with the assumption of intraperiod stationarity and with path choice behavior, typical of high frequency and irregular urban transit systems.

If service frequency is low and/or it is assumed that the users choose specific runs, it is necessary to represent the service with a different graph known as a *run graph* or *diachronic graph*. This is usually the case with extraurban transportation services (airplanes, trains, etc.), which have low service frequencies and are largely punctual. In this case, however, the assumption of within-day stationarity does not hold. Indeed, the supply characteristics are often nonuniform within the reference period (arrival and departure times of single runs may be nonuniformly spaced). Furthermore, in order to simulate the traveler's behavior *desired departure* or *arrival times* should be introduced. For these reasons run-based supply models are described in Chap. 7 dealing with intraperiod dynamic systems.

2.4.2.1 Line-based Graph Models

If the scheduled services have high frequencies (e.g., one run every 5–15 min) and low regularity, it is usually assumed that the users do not choose an individual run, but rather a service line or a group of lines. A *service line* is a set of runs sharing the same terminals, the same intermediate stops, and the same performance characteristics, as in the case of an urban bus or underground lines. In this case a *line graph* is typically used. In this graph, nodes correspond to stops, more precisely to the relevant events occurring at the stops. *Access nodes* represent the arrival of the user at the stop, the *stop node*, or *diversion node*, represents the boarding of a vehicle, and the *line nodes* represent the arrival and departure of vehicles of a given line at a given stop. The links represent activities or phases of a trip: access trips between access nodes (*access links*), waiting at the stop (*waiting links*), boarding and alighting from the vehicles of a line (*boarding* and *alighting links*), the trip from one stop to another of the same line (*line links*), and vehicle dwelling at the stop (*dwelling links*).

Essentially, each stop is represented by a subgraph such as that shown in Fig. 2.34. The graph representing an entire public transportation system can be built by combining the *line graph* and the *access graph* through the stop subgraphs. Access links may represent different access modes depending on the system modeled. In urban areas, they may represent pedestrian connections or, sometimes, undifferentiated “access modes” including local transit lines to the main network of bus and rail services. The line graph is completed by adding nodes and links allowing entry/exit from the centroids to the stops; in the urban context this usually occurs through pedestrian nodes and links or through road links connected to park-and-ride facilities (nodes).

2.4.2.2 Link Performance and Cost Functions

The typical performance attributes used in line-based supply models are travel time components related to different trip phases and monetary costs. Travel times can be decomposed into on-board travel times T_b , dwelling times at stops T_d , waiting times T_w , boarding times T_{br} , alighting times T_{al} , and access/egress times T_a , which may correspond to walking or driving time for urban transit networks. In general, a single time component is associated to each link and the coefficients β , homogenizing travel times into costs (disutilities) are different. In fact, several empirical studies have shown that waiting and walking times have coefficients two to three times larger than that of on-board time for urban transit systems.

Performance functions used in many applications do not take congestion into account, at least with respect to flows of transit users, as it is assumed that services are designed with some extra capacity with respect to maximum user flows.

On-board travel time of a transit link can be obtained through a very simple expression:

$$Tb_a = \frac{L_a}{v_a(\mathbf{b}_a, \boldsymbol{\gamma}_a)} \quad (2.4.27)$$

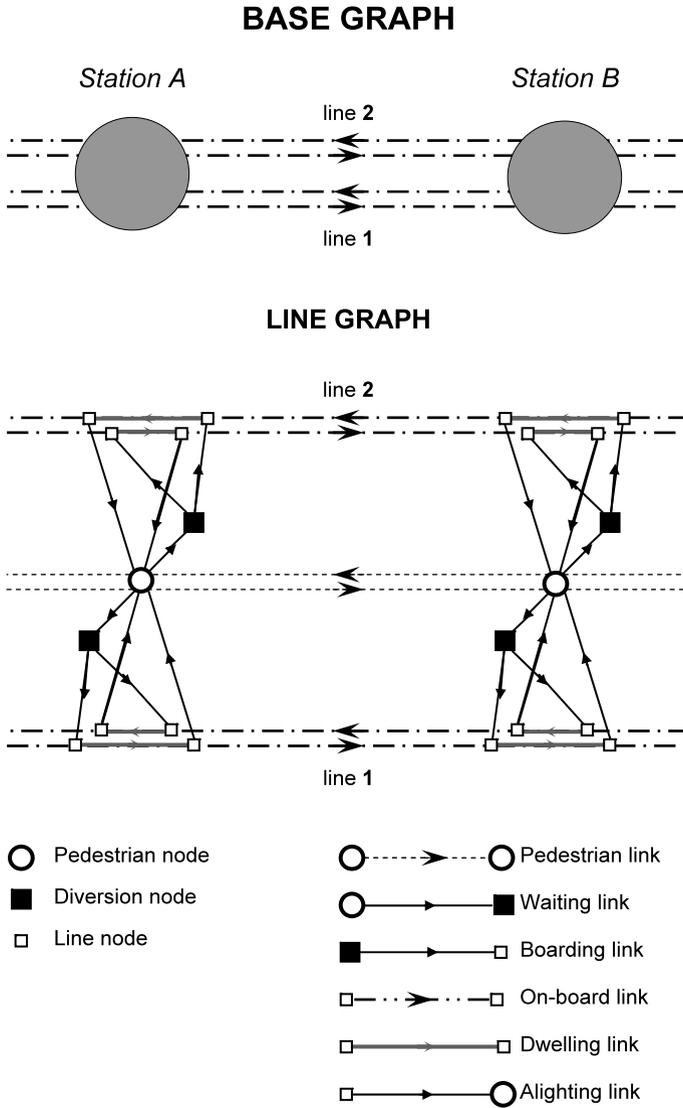


Fig. 2.34 Line-based graph for urban transit systems

where vector b_a includes the relevant characteristics of the transit system represented by link a , and vector γ_a comprises a set of parameters. The average speed is strongly dependent on the type of right-of-way. For exclusive right-of-way systems, such as trains, the average speed v_a can be expressed as a function of the characteristics of the vehicles (weight, power, etc.), of the infrastructure (slope, radius of bends, etc.), of the circulation regulations on the physical section and the type of service represented. Relationships of this type can be deduced from mechanics for

which specialized texts should be referred to. For partial right-of-way systems, such as surface buses, the average speed depends on the level of protection (e.g., reserved bus lane) and the vehicle flows on the links corresponding to interfering movements. Performance functions of this type typically derive from descriptive models.

The *waiting time* is the average time that users spend between their arrival at the stop/station and the arrival of the line (or lines) they board. Waiting time is usually expressed as a function of the line *frequency* φ_{ln} , that is, the average number of runs of line ln in the reference period. When only one line is available the average waiting time Tw_{ln} will depend on the regularity of vehicle arrivals and the pattern of users' arrivals at the stop. It can be shown that, under the assumption that users arrive at the stop according to a Poisson process with a constant arrival rate⁶ (consistent with the within-day stationarity assumption), the average waiting time is:

$$Tw_{ln} = \frac{\theta}{\varphi_{ln}} \quad (2.4.28)$$

where θ is equal to 0.5 if the line is perfectly regular (i.e., the headways between successive vehicle arrivals are constant), and it is equal to 1 if the line is "completely irregular" (i.e., the headways between successive arrivals are distributed according to a negative exponential random variable); see Fig. 2.35.

In the case of several "*attractive lines*," that is, when the user waits at a diversion node m for the first vehicle among those belonging to a set of lines Ln_m , the average waiting time can again be calculated with expression (2.4.28) by using the *cumulated frequency* Φ_m of the set of attractive lines:

$$Tw_{ln} = \frac{\theta}{\Phi_m} \quad \text{with} \quad \Phi_m = \sum_{ln \in Ln_m} \varphi_{ln} \quad (2.4.29)$$

Expression (2.4.29) holds in principle when vehicle arrivals of all lines are completely irregular. In this case cumulated headways can still be modeled as a negative exponential random variable, with a parameter equal to the inverse of the sum of line frequencies. In practice, however, expression (2.4.29) is often used also for intermediate values of θ .

These expressions of average waiting times are revisited in Sect. 4.3.3.2 on path choice models for transit systems.

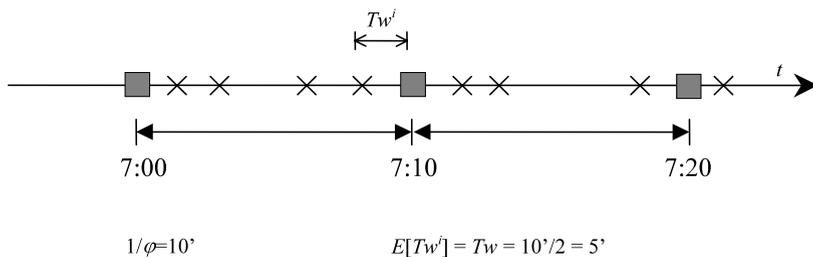
Access/egress times are also usually modeled through very simple performance functions analogous to expression (2.4.27):

$$Ta_{ln} = \frac{L_{ln}}{v_{al}(\mathbf{b}_{ln}, \gamma_{ln})}$$

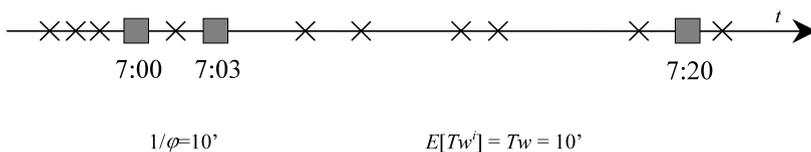
where v_{al} represents the average speed of the access/egress mode. Also in the case of pedestrian systems, it is possible to introduce congestion phenomena and correlate

⁶To be precise, it is assumed that users' arrival is a Poisson process; that is, the intervals between two successive arrivals are distributed according to a negative exponential variable.

LINE WITH REGULAR ARRIVALS OF VEHICLES



LINE WITH RANDOM ARRIVALS OF VEHICLES



■ Bus arrival at the stop

× User arrival at the stop

User arrivals random in both cases

Fig. 2.35 Arrivals and waiting times at a bus stop

the generalized transportation cost with the pedestrian density on each section by using empirical expressions described in the literature.

More detailed performance models introduce congestion effects with respect to user flows both on travel times and on comfort performance attributes. An example of the first type of function is that relating the *dwelling time* at a stop Td_{ln} to the user flows boarding and alighting the vehicles of each line:

$$Td_{ln} = \gamma_1 + \gamma_2 \left(\frac{f_{al(a)} + f_{br(a)}}{Q_D} \right) \gamma_3 \tag{2.4.30}$$

where

$f_{al(a)}$ is the user flow on the alighting link

$f_{br(a)}$ is the user flow on the boarding link

Q_D is the door capacity of the vehicle

$\gamma_1, \gamma_2, \gamma_3$ are parameters of the function

Another example is the function relating the average waiting time to the flow of users staying on board and those waiting to board a single line. This function takes into account the “refusal” probability, that is, the probability that some users may not be able to get on the first arriving run of a given line because it is too crowded and have to wait longer for a subsequent one. In the case of a single attractive line l the waiting time function can be formally expressed as

$$T_{wln} = \frac{\theta}{\varphi_{ln}(\cdot)} \left(\frac{f_{b(\cdot)} + f_{w(\cdot)}}{Q_{ln}} \right) \quad (2.4.31)$$

where $\varphi_{ln}(\cdot)$ is the actual available frequency of line ln , that is, the average number of runs of the line for which there are available places. It depends on the ratio between the demand for places – sum of the user flow staying on board $f_{b(\cdot)}$ and the user flow willing to board, $f_{w(\cdot)}$ – and the line capacity Q_{ln} . This formula is valid only for $f_{b(\cdot)} + f_{w(\cdot)} > Q_{ln}$.

Note that both performance functions (2.4.30) and (2.4.31) are nonseparable, in that they depend on flows on links other than the one to which they refer.

Discomfort functions relate the average riding discomfort on a given line section represented by link a , dc_a , to the ratio between the flow on the link (average number of users on board) and the available line capacity Q_a :

$$dc_a = \gamma_3 f_a + \gamma_4 \left(\frac{f_a}{Q_a} \right)^{\gamma_5} \quad (2.4.32)$$

where, as usual, γ_3 , γ_4 , and γ_5 are positive parameters, usually with γ_5 larger than one expressing more-than-linear effect of crowding.

Reference Notes

The application of network theory to the modeling of transportation supply systems can be found in most texts dealing with mathematical models of transportation systems, such as Potts and Oliver (1972), Newell (1980), Sheffi (1985), Cascetta (1998), Ferrari (1996), and Ortuzar and Willumsen (2001). All of these, however, deal primarily or exclusively with road networks. The presentation of a general transportation supply model and its decomposition into submodels as described in Fig. 2.14 is original.

Performance models and the traffic flow theory are dealt with in several books and scientific papers. The former include Pignataro (1973), the ITE manual (1982), May (1990), McShane and Roess (1990), the Highway Capacity Manual (2000), and the relevant entries in the encyclopaedia edited by Papageorgiou (1991). Among the latter, the pioneering work of Webster (1958), later expanded in Webster and Cobbe (1966) and those of Catling (1977), Kimber et al. (1977), Kimber and Hollis (1978), Robertson (1979), and Akcelik (1988) on waiting times at signalized intersections. In-depth examinations of some aspects of traffic flow theory can be found in Daganzo (1997).

For a theoretical analysis of queuing theory, reference can be made to Newell (1971) and Kleinrock (1975).

The work of Drake et al. (1967) reviews the main speed–flow–density relationships, and gives an example of their calibration. The linear model was proposed by Greenshields (1934). References to nonstationary traffic flow models are in part reported in the bibliographical note to Chap. 7.

A review of the road network cost functions can be found in Branston (1976), Hurdle (1984), and Lupi (1996). The study of Cartenì and Punzo (2007) contains experimental speed–flow relationships for urban roadways, reported in the text (2.4.5) and updates the work by Festa and Nuzzolo (1989). The cost function for parking links (2.4.26) was proposed by Bifulco (1993).

Supply models for scheduled services have traditionally received less attention in the scientific community. The line representations of scheduled systems are described in Ferrari (1996) and in Nuzzolo and Russo (1997).

Several authors, such as Seddon and Day (1974), Jolliffe and Hutchinson (1975), Montella and Cascetta (1978), and Cascetta and Montella (1979), have studied the relationships between waiting times and service regularity in urban transit systems. Congested performance models discussed in Sect. 2.4.2 have been proposed by Nuzzolo and Russo (1993), and other models for waiting time at congested bus stops are quoted in Bouzaiene-Ayari et al. (1998). Mechanics of motion is treated in detail in several classical books. For an updated bibliographical note see Cantarella (2001).



<http://www.springer.com/978-0-387-75856-5>

Transportation Systems Analysis

Models and Applications

Cascetta, E.

2009, XVIII, 742 p. 100 illus., Hardcover

ISBN: 978-0-387-75856-5