Applications of Transportation Supply Models

Introduction

Network models and related algorithms are powerful tools for modeling transportation systems. A network model is a simplified mathematical description of the physical 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. 1.

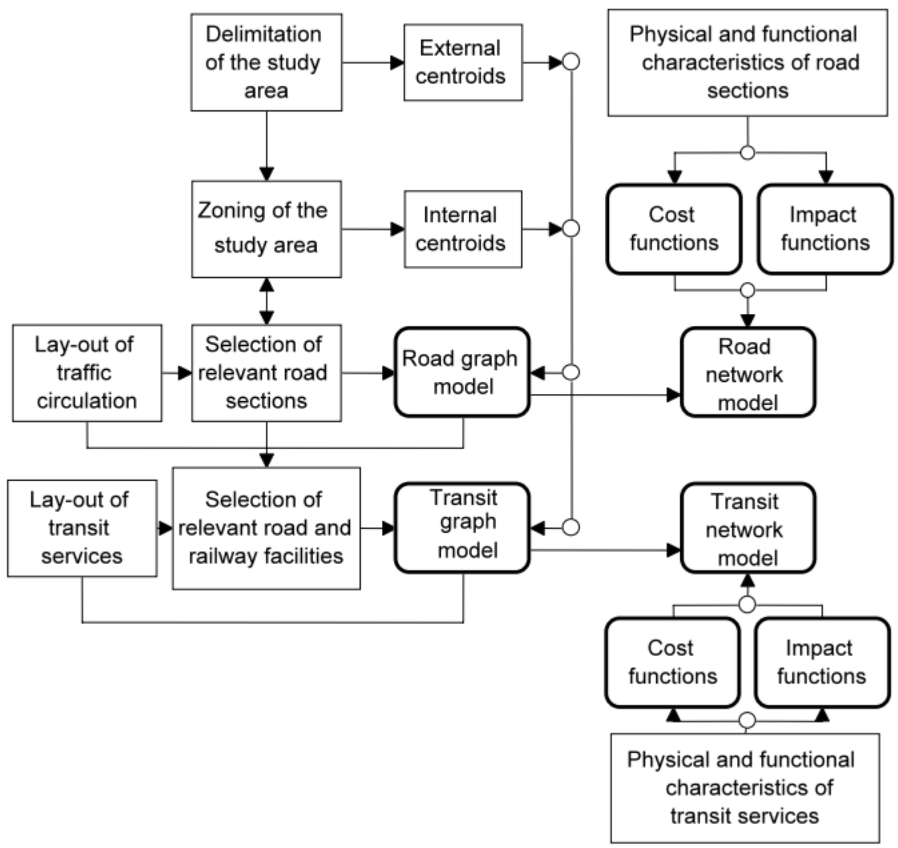

Fig. 1 Functional phases for the construction of an urban bimodal network model.

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. 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 a road) and scheduled services (such as trains or buses).

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. The centroids can approximate the internal points within a traffic zone. In general, the zone centroid is a fictitious node, that is, a node that 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 with 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 for 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. 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. 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, 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.

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.

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 shows the graph representing part of the urban road network.

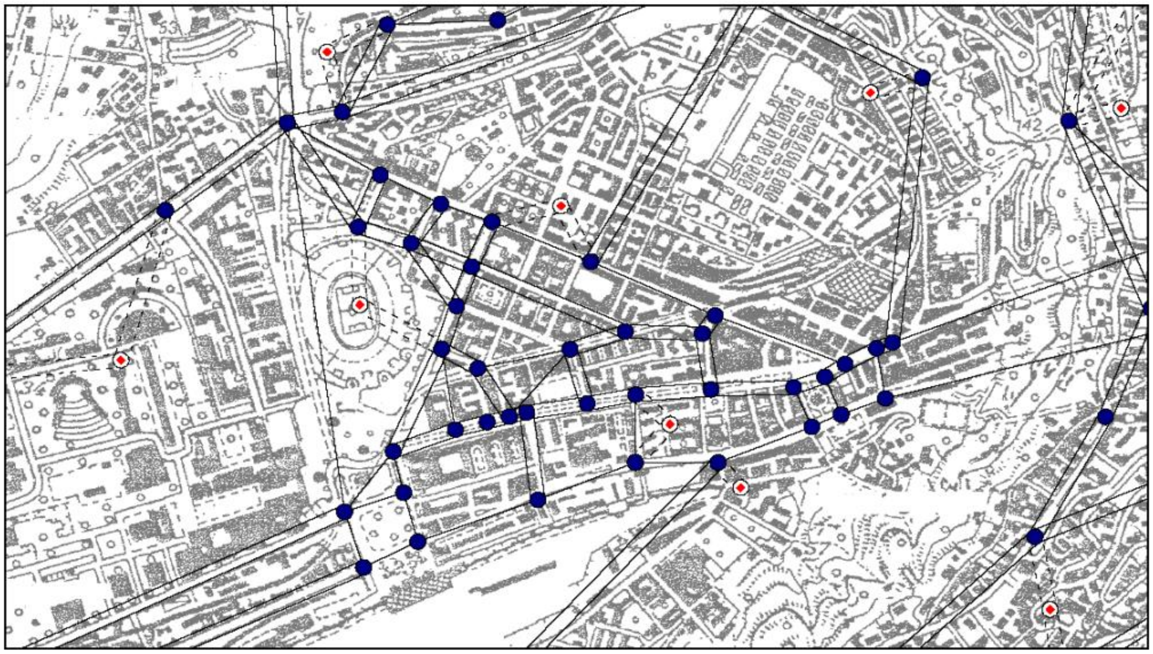

Fig. 2 Example of a graph representing part of an urban road system.

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. 3).

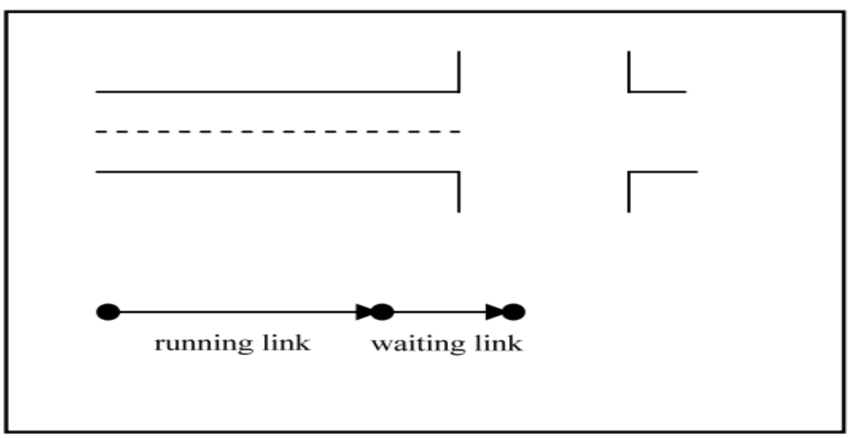

Fig. 3 Representation of a road intersection with running and waiting for links.

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 4 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 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. 5). 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.

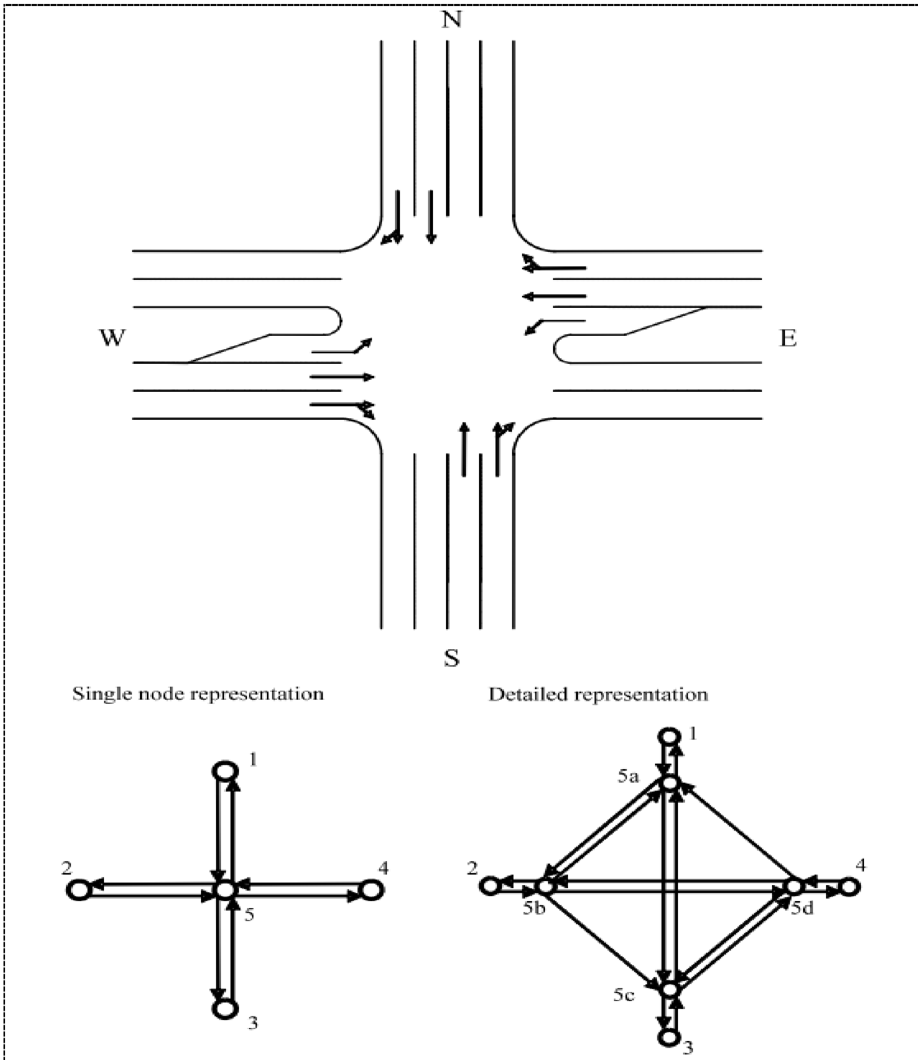

Fig. 4 Graphs for a road intersection.

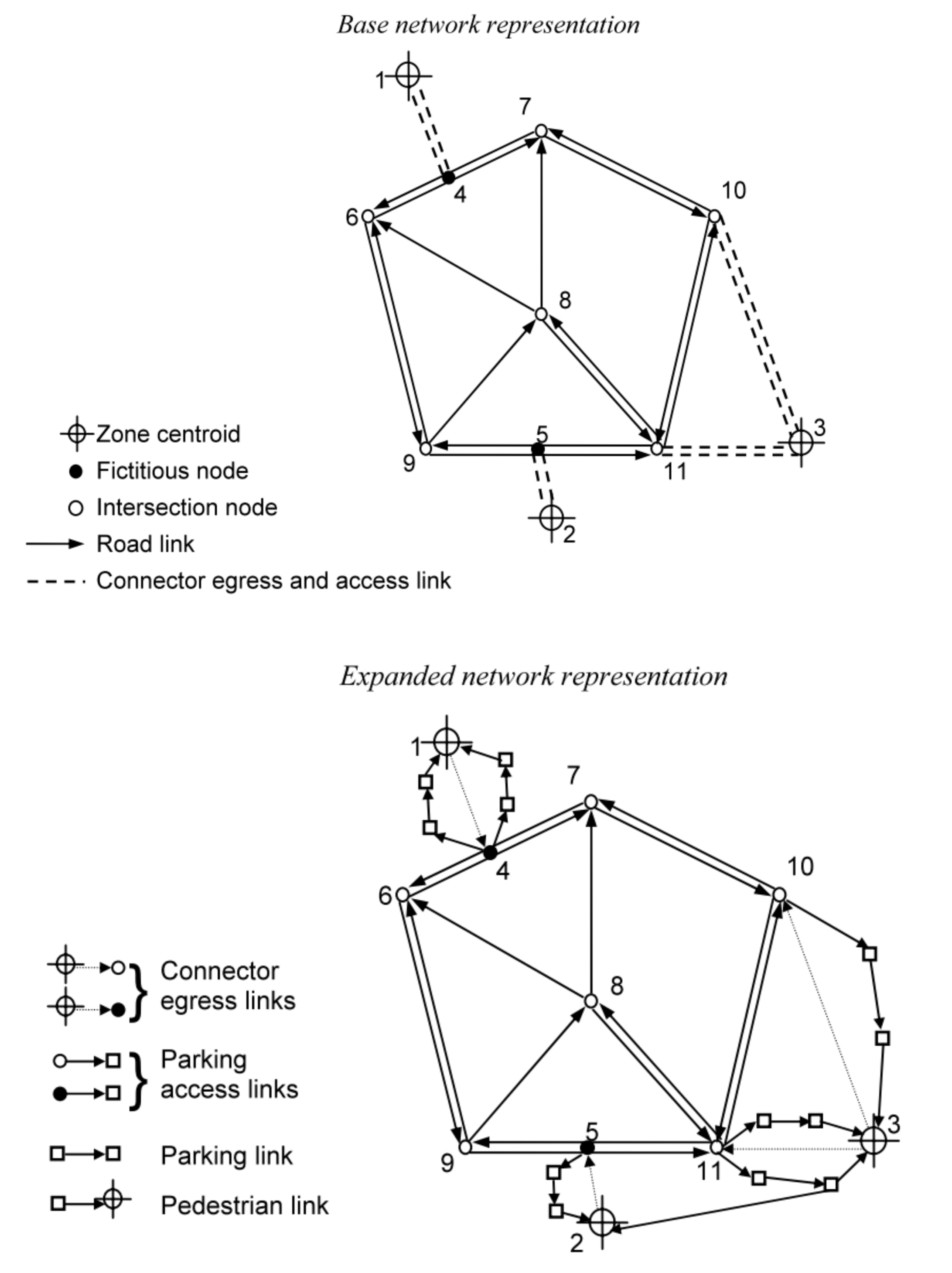

Fig. 5 Explicit representation of parking supply.

Link Performance and Cost Functions

The generalized transportation cost of a road link is usually made up of 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:

$$c_a(f) = \beta_1 tr_a(f) + \beta_2 tw_a(f) + \beta_3 mc_a(f) \qquad 1$$

Where:

$tr_a(f)$ is the function relating the running time on the link a to the flow vector.
$tw_a(f)$ is the function relating the waiting time on the link a to the flow vector.
$mc_a(f)$ is the function relating the monetary cost on the link a to the flow vector.
The dependence on physical and functional variables b_a, and parameters γ, has been omitted for simplicity's sake. It has been assumed that homogenization coefficients may differ for the different time components. Furthermore, not all of the components are present for each link; for example, 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.

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:

$$mc_a = mc_{toll} + mc_{fuel}(f)$$

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. 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 lecture. 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 intro period 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:

$$tr_a = L_a / v_a(f_a) \quad (2)$$

Where:

tr_a is the running time on the 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 the 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 extra-urban 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.

$$tr_a(f_a) = \frac{L_a}{v_{oa}} + (\frac{L_a}{v_{ca}} - \frac{L_a}{v_{oa}})(\frac{f_a}{Q_a})^4 \quad (3)$$

Where:

L_a is the length of link a.

v_{oa} is the free-flow average speed.

v_{ca} is the average speed with flow equal to the 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 link a, N_a, and lane capacity, Q_{ua}.

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_{oa}, 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. Parameters γ_1 and γ_2 are typically estimated on empirical data.

Figure 5 shows a diagram 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. 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.

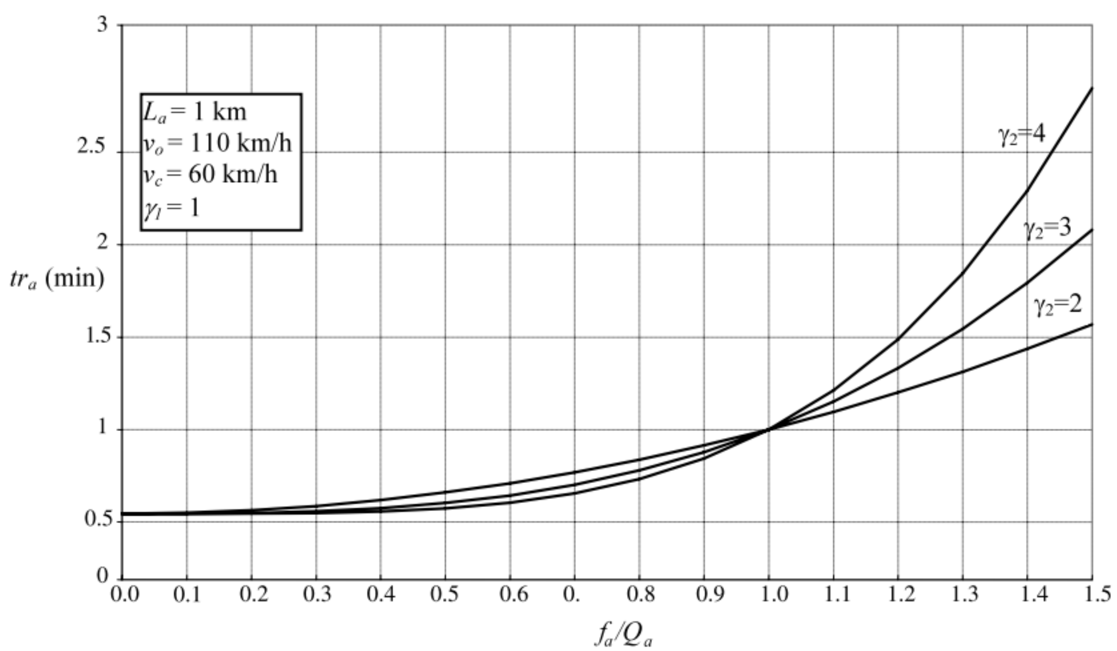

Fig. 5 Motorway travel time function for different values of some parameters.

b) Extra urban Road Links Users traveling on an extra-urban 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 (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.

$$tr_a(f_a, f_{a^*}) = \frac{L_a}{v_{0a}} + \gamma_a\left(\frac{L_a}{v_{ca}} - \frac{L_a}{v_{0a}}\right)\left(\frac{f_a + f_{a^*}}{Q_{aa^*}}\right)^{\gamma_2} \quad 4$$

Where, apart from the symbols introduced previously, the link in the opposite direction is denoted by a^* and the 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:

$$v_a = 29.9 + 3.6Lu_a - 0.6P_a - 13.9T_a - 10.8D_a - 6.4S_a + 4.7Pv_a - 1.0E - 04\frac{\left(\frac{f_a}{Lu_a}\right)^2}{1+T_a+D_a+S_a} \quad 5$$

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].
S_a is the percentage of the length of an 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 the link a in Equiv. vehicles/hour.

The travel time on link a may thus be calculated by multiplying the time obtainable from (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):

$$tr_a = \frac{L_a}{v_a}.c(L_a) = \frac{L_a}{v_a}.\frac{1}{1-\exp(-0.47-0.48E-2.L_a)} \qquad 6$$

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 interrupting flow (delays at intersections are thus included):

$$\begin{cases} tr_a = \left({L_a}/{v_{0a}}\right)\left(1 + \frac{\gamma f_a}{Q_a - f_a}\right) & for\ f_a \le \delta Q_a \\ tr_a = tangent\ approximation & for\ f_a > \delta Q_a \end{cases} \qquad 7$$

With $\delta < 1\ and\ Q_a = link\ capacity$. Also, see Fig. 6.

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.

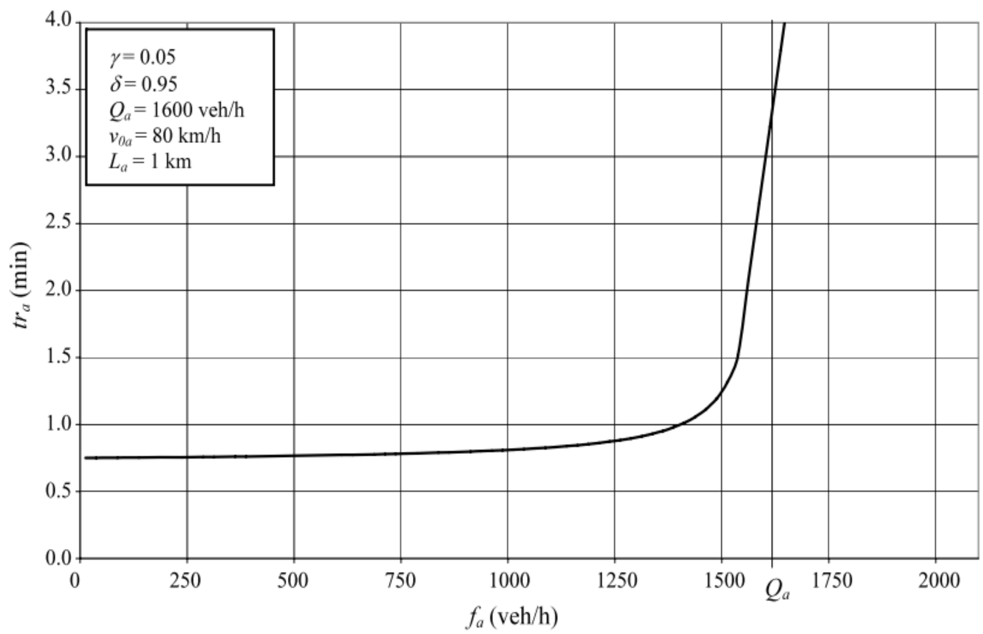

Fig. 6 Hyperbolic travel time cost function.

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:

$$tw_a^u(f_a) = T_s + (T_s^2 + \sigma_s^2).\frac{f_a}{2}.\frac{1}{1-f_a/Q_a} \qquad 8$$

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 (8) is derived from the assumption of a queuing system M/G/1 (∞, FIFO) with Poisson arrivals and general service time.

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 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 (8) is known as "stochastic delay."
Moreover, the average delay computed with (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 (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, $t_w(\alpha Q_a)$ with an angular coefficient equal to the derivative of (8) computed at this point:

$$tw_a(f_a) = tw_a(\alpha Q_a) + K(f_a - \alpha Q_a) \quad 9$$

With

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

Figure 7 shows the relationships (8) and (9) for some values of the parameters.

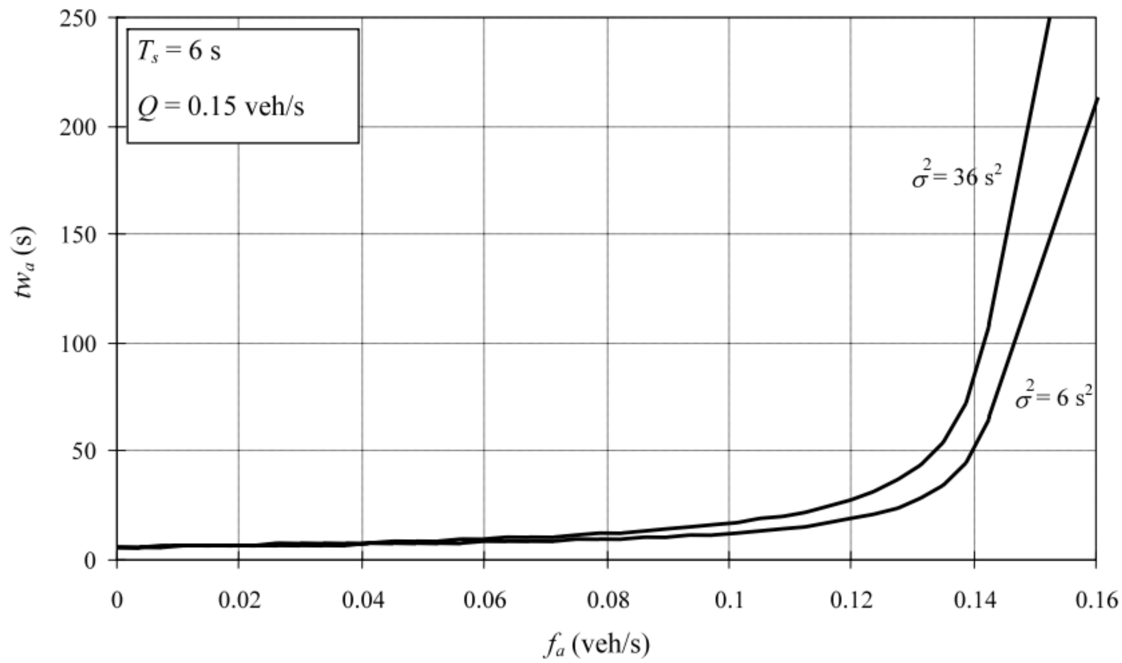

Fig. 7 Waiting time functions (8) and (9) at toll-barrier links.

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. The deterministic average (oversaturation) delay tw_a^d is then equal to:

$$tw_a^d = T_s + (\frac{f_a}{Q_q} - 1)\frac{T}{2} \qquad 10$$

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

Note that in this case the assumption of intro period 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, remove this assumption.

The average delay tw_a can be calculated by combining the stochastic undersaturation average delay tw_a^u expressed by (8) with the deterministic average oversaturation delay tw_a^d, expressed by (10). The combined delay function is such that the deterministic delay function is its oblique asymptote (see Fig. 8). The following equation results.

$$tw_a(f_a) = T_s + (T_s^2 + \sigma^2)\frac{f_a}{2} + \frac{T}{4}\left\{\frac{f_a}{Q_a} - 1 + \left[(\frac{f_a}{Q} - 1)^2 + \frac{4(\frac{f_a}{Q_a})}{Q_a T}\right]^{1/2}\right\} \qquad 11$$

b) Signal-Controlled Intersection Links Queuing and delay phenomena at signalized intersections can be obtained from the queuing theory results. 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 intersections), without lanes reserved for right or left turns.

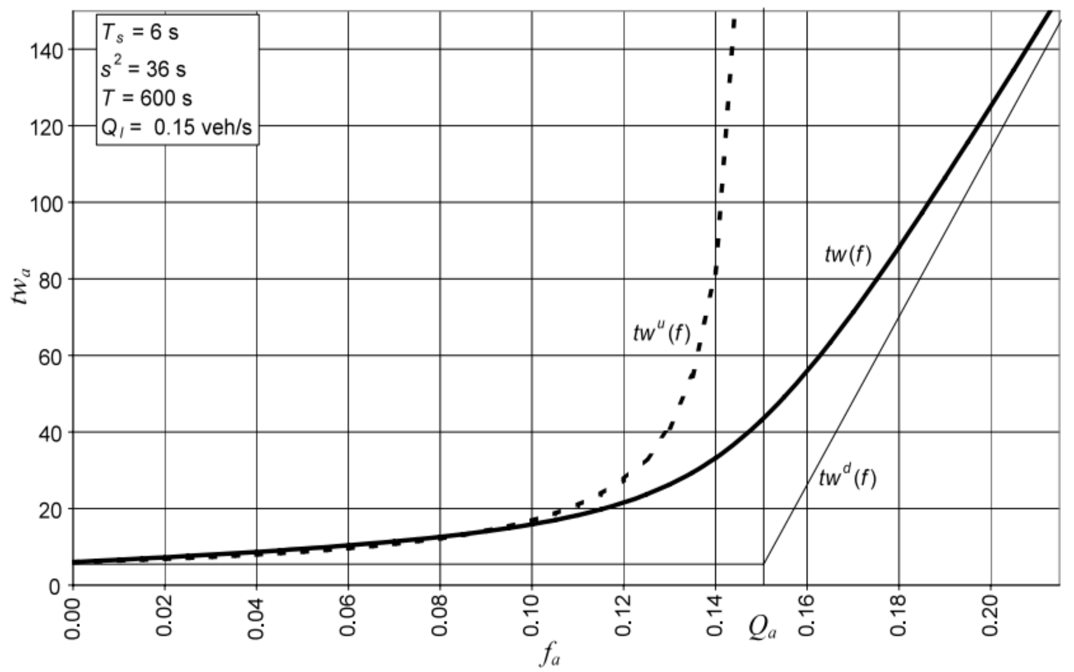

Fig. 8 Under- and oversaturation waiting time functions for toll barrier links.

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. 9 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.

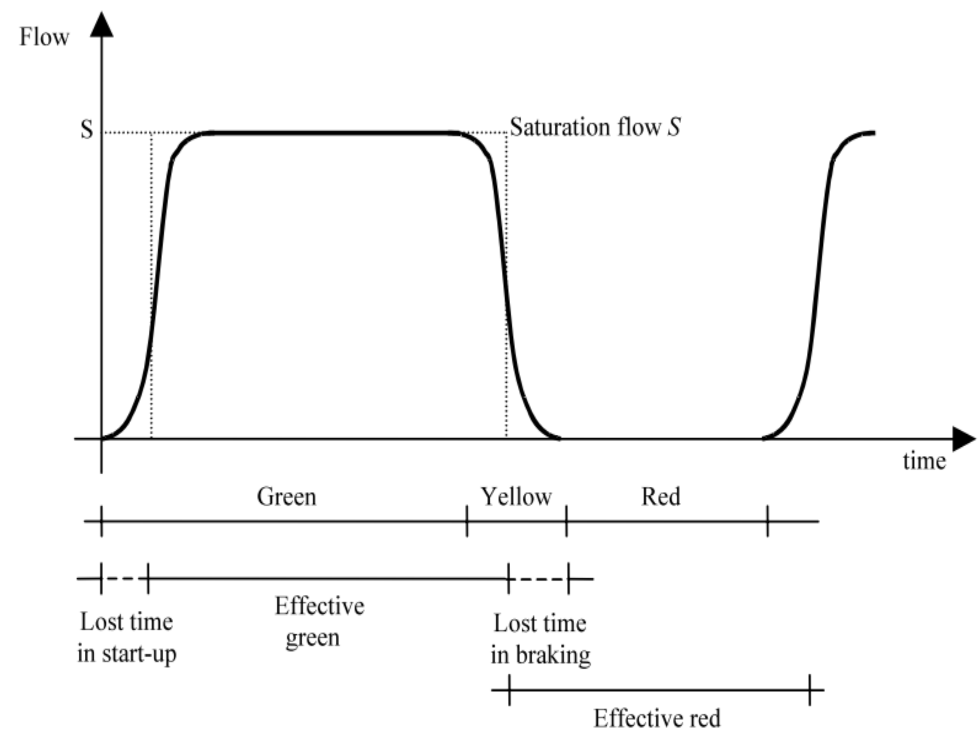

Fig. 9 Discharge flow from signal-controlled intersection in relation to cycle phases.

Below, to simplify the notation, we omit the index of link a. Moreover, to facilitate application of the results, 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.
μ = G/T_c be the effective green/cycle ratio for the approach.

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

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

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

$$S.G = \mu.S.T_c$$

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 (μ = 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:

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

Thus, the approach can be defined as undersaturated if:

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

That is:

$$\bar{u} < \mu.S \qquad 12$$

On the other hand the approach is defined oversaturated if:

$$\bar{u} \geq \mu.S \qquad 13$$

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 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 describe the operation of the intersection is used.

Figure 10 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,

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

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.

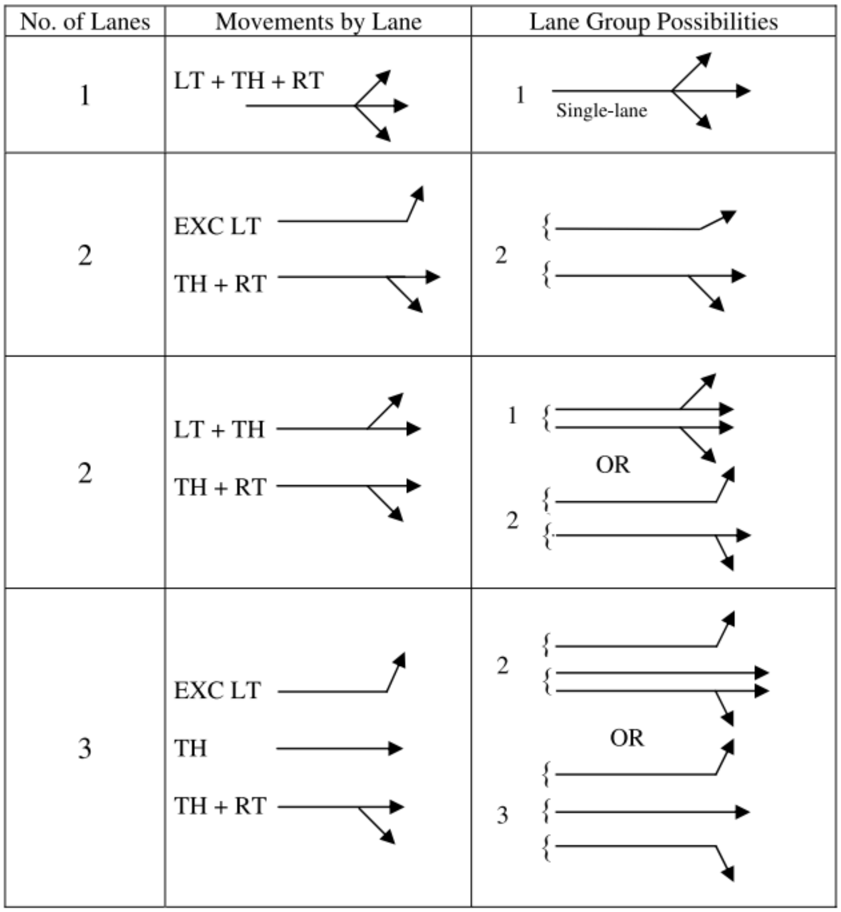

Fig. 10 Typical lane groups for the HCM method for calculating saturation flow.

The first six adjustment factors not connected with the type of turning maneuvers are reported in Fig. 11. Once the approach capacity Q_1 = μS is known, we may calculate the queue length and mean waiting time tw_a , using models derived from different approaches.

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_{HV}

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. 11 Adjustment factors in the HCM method for saturation flow.

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 the consequent assumptions on users' behavior. The approach followed in this lecture 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 intro period 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 extra-urban 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 dealing with intro period dynamic systems.

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. 12. 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).

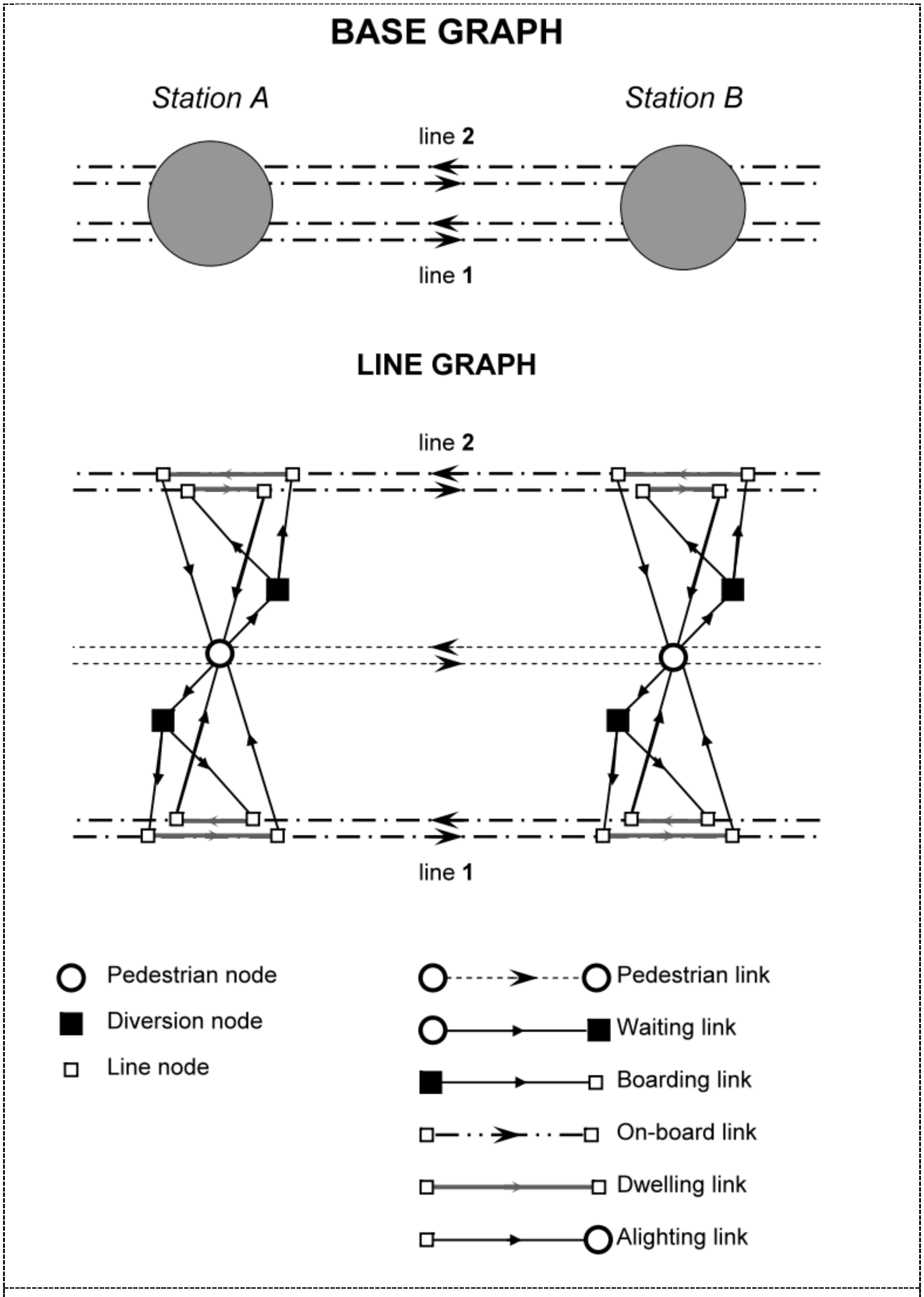

Fig. 12 Line-based graph for urban transit systems.

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_{a1}, 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 with 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 onboard 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.

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

$$Tb_a = \frac{L_a}{v_a(b_a, \gamma_a)} \qquad 14$$

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.), the infrastructure (slope, the radius of bends, etc.), and the circulation regulations on the physical section and the type of service represented. Relationships of this type can be deduced from mechanics to which specialized texts should be referred. 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.

<u>The *waiting time*</u> is the average time that users spend between their arrival at the stop/station and the arrival of the line (or lines) on the board. Waiting time is usually expressed as a function of the line frequency $\emptyset_{ln}$, 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 rate (consistent with the within-day stationarity assumption), the average waiting time is:

$$Tw_{ln} = \frac{\theta}{\emptyset_{ln}} \qquad 15$$

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. 13.

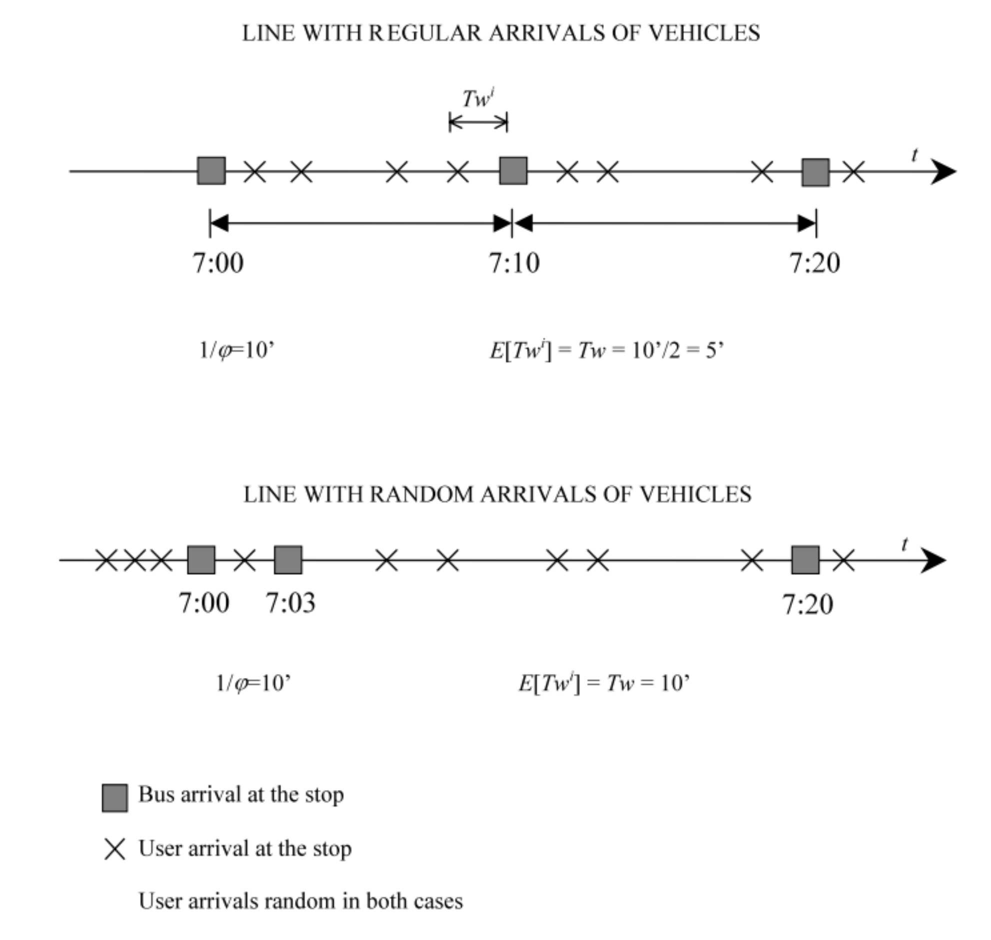

Fig. 13 Arrivals and waiting times at a bus stop.

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 (15) by using the cumulated frequency Φ_m of the set of attractive lines:

$$Tw_{ln} = \frac{\theta}{\Phi_m} \quad with\ \Phi_m = \sum_{ln \in Ln_m} \varphi_{ln} \qquad 16$$

Expression (16) 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 (16) is often used also for intermediate values of θ. These expressions of average waiting times are applied to path choice models for transit systems.

Access/egress times are also usually modeled through very simple performance functions analogous to expression (14):

$$Ta_{ln} = \frac{L_{ln}}{v_{al}(b_{ln}, \gamma_{ln})}$$

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

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 relating the dwelling time at a stop Td_{ln} to the user flows boarding and alighting the vehicles of each line:

$$Td_{ln} = \gamma_1 + \gamma_2 (\frac{f_{al/(a)} + f_{br(a)}}{Q_D}) \gamma_3 \quad (17)$$

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.
γ_1, γ_2, γ_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:

$$Tw_{ln} = \frac{\theta}{Q_{\ln(.)}} (\frac{f_{b(.)} + f_{w(.)}}{Q_{ln}}) \quad (18)$$

Where $Q_{\ln(.)}$ 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.
- the sum of the user flow staying on board $f_{b(.)}$ and the user flows willing to board, $f_{w(.)}$.
- the line capacity Q_{ln}. This formula is valid only for $f_{b(.)} + f_{w(.)} > Q_{ln}$.

Note that both performance functions (17) and (18) 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 :

$$dc_a = \gamma_3 f_a + \gamma_4 (\frac{f_a}{Q_a})^{\gamma_5} \quad (19)$$

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