7 Demand and supply phenomena

This chapter addresses the modelling of various demand and supply phenomena emerging on public transport networks: passenger information, congestion at stops and on board, service regularity. These phenomena affect route choice, either directly (information), or indirectly through travel costs (congestion); therefore they are to be made an integral part of transit assignment models, which shall then evolve from the basic frameworks presented in the previous Chapter 6.

The aim is then that of providing travel times and, in case of strategy models, also the hyperarc diversion probabilities, for a given arc flow pattern.

7.1 Strategies and information

Authors: Klaus Noekel, Guido Gentile, Michael Florian

This section is devoted to the modelling of the following phenomena in the context of transit assignment:

- strategic behaviour with respect to line vehicle arrivals at stops,
- information provision to passengers.

In the basic framework for frequency-based assignment (Section 6.2) passengers choose between complete alternative paths before starting their journey. Although shortest-path search seems to be a rational basis for this decision, a significant part of the generalized cost of each alternative is not known in advance and enters only as an expected value, i.e. the waiting times. Indeed, they derive from the random departure of lines from stops wrt to the passenger arrival. Then, actual waiting times encountered by the passenger as his journey unfolds may differ substantially from the expected values. Can the passenger reduce expected waiting times by postponing part of the route decision with the possibility of reacting strategically to random headways? This section explores route choice models in which passengers take decisions based on information they acquire during the trip, while on board or waiting at stops.

The convenience of passengers in adopting a strategic behaviour stems from the fact that it could be better to board a slower line that is arriving earlier than to wait for a faster line that will arrive later; here “slow” and “fast” do not refer to the commercial speed of the line but to the expected travel time to reach the destination once boarded the line, which may include further sub-strategies and other lines.

In the classical model of optimal strategies, passengers acquire information about the line served by the next arriving carrier at the stop, by simply looking at its signboard. That information is available only when the carrier is actually approaching the stop and becomes thus visible by the passenger. This is clearly not the situation of modern travel information systems, where passengers can know as soon as they reach the stop (or earlier when entering a station with several stops) a list of arrival times from an electronic panel, typically the next one for each line among all runs that already departed from the terminal. The internet revolution allows for an even higher degree of freedom, since passengers can access the same information above from home/work (computers), or also en-route through mobile device (smartphones).

In general, it turns out that the extent to which it is possible to reduce the expected generalized cost of the journey, by adapting during the trip the taken route to incoming information about line arrival times at stops, depends on additional assumptions about:

- the regularity of service,
- the passenger’s ability to observe service operation en-route, and
- the structure of the strategies considered by the passenger.

Each combination of assumptions about these aspects induces a different route choice model. In this section, some alternative sets of assumptions are reviewed, linked to the corresponding assignment model, and the results are compared in terms of the line shares and in terms of sensitivity against perturbations of input data.

7.1.1 Optimal strategies with exponential headways

Consider the transit network topology that was introduced in Section 6.2.2 for frequency-based models and the arc performances presented in Section 6.2.3. The cost associated with each pedestrian arc and each line segment is assumed constant. At each stop along the itinerary of every transit line the inter-arrival times of the vehicles (headway) are instead not constant, but their distribution is known; this induces random wait times.

Because several lines may serve the same stop, the passenger directed towards a given destination may choose to board the first arriving vehicle of a given line set, instead of waiting for one single line. Then, depending on which line arrives first, the journey will follow along different routes. This strategic approach implies a trade-off between lower wait time and higher expected costs to reach the destination once boarded

the line (which includes the in-vehicle time and possibly other transfers). The objective of the passenger is to minimize the expected total generalized cost, taking into account that the different time components of a transit journey (waiting, riding, walking) typically have different weights (comfort coefficients); in particular, waiting at stop is usually perceived as more onerous than riding on-board a vehicle, although this may change due to on-board overcrowding.

This model requires to compute the combined expected time for the arrival of the first vehicle for any subset of lines serving the same stop (with given headway distributions), as well as the probability that each line arrives first.

Using the terminology of Section 6.1.3, the stops are the diversion nodes, each distinct set of serving lines identifies a (waiting) hyperarc, the diversion probability of using each branch of the hyperarc is equal to the corresponding line probability, the conditional travel time of each branch is equal to the combined expected wait time.

7.1.1.1 Network topology

When formulating the optimal strategy model it is common practice to head the alighting arcs directly at the base node, and not at the stop node; while the return of the stop arc is eliminated. This modification of the network topology presented in Section 6.2.2 is depicted in Fugure 7.1 and is useful to have only one type of diversion arcs, i.e. the waiting arcs, exiting from the stop diversion node. Then, for each combination of waiting arcs $s^+ \subseteq A^{wait}$ exiting from each stop $s \in S$ an hyperarc $\breve{a}$ is introduced:

- the *waiting hyperarcs* $H^{wait} = \{\breve{a} \subseteq s^+ : \forall s \in S\}$.

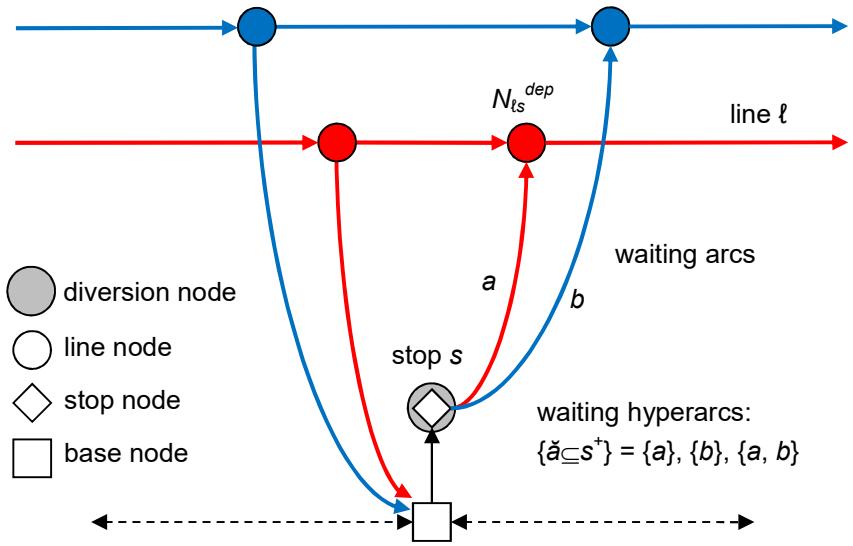

Figure 7.1. Topology of the transit network with boarding hyperarcs exiting from the stop diversion node.

The following exposition is based on the seminal work of Spiess and Florian (1989).

The arc performances presented in Section 6.2.3 provides a cost c_{ag} for each arc $a \in A$ and class $g \in G$. However, the cost of waiting arcs is given here by the non-temporal component only, because the wait cost is handled in a separate term, which depends also on the passenger destination. Moreover, each arc $a \in A$ is here characterized by a second attribute, i.e. the frequency f_a , which is (nearly) infinite with the exception of the waiting arcs, where it is equal to the frequency of the associated line (which may also be referred to as line a for short).

The solution of the (deterministic) route choice model for passengers of class $g \in G$ directed to each single destination $d \in D$ can be described as an acyclic subgraph $(N, \bar{A}_{dg} \in A)$, which is referred to as a *hypertree* by Nguyen and Pallottino (1988). This defines the topology of the optimal strategy from each origin, that is the most extended hyperpath on the hypertree having that origin and that destination. Because these solutions are independent, in the following the indices dg are omitted.

Recall that $a^- \in N$ and $a^+ \in N$ denote, respectively, the initial and final node of arc $a \in A$, while $i^+ \subseteq A$ and $i^- \subseteq A$ denote respectively, the forward and backward star of node $i \in N$.

For each node $i \in N$ define as $\bar{A}_i^+ = i^+ \cap \bar{A}$ the arcs exiting from i and belonging to the solution hypertree. If i is not a stop, then $\bar{A}_i^+$ is the (one) *successor arc* of the node towards the destination. If i is a stop diversion node, then $\bar{A}_i^+$ is the set of waiting arcs associated with those lines which the passenger will possibly board to reach the destination, i.e., the (one) *successor hyperarc* of stop i. These lines are conveniently referred to as the *attractive set*. Among those lines the passenger boards the vehicle that arrives first, and waits on average for their combined expected time of arrival.

7.1.1.2 Combined wait time and line shares

Consider the successor hyperarc $\bar{A}_i^+ = \breve{a} \subseteq i^+$ of stop $i \in S$. Let $t_{\breve{a}}$ be the expected wait time for the arrival of the first vehicle serving any of the waiting arc branches $a \in \breve{a}$, which is referred to as the *combined wait time*. Let $p_{a|\breve{a}}$ be the probability that arc $a \in \breve{a}$ corresponds to the first line served among the attractive set identified by $\breve{a}$. If line headways at stop i are independent and have an exponential distribution it is:

$$t_{\breve{a}} = \frac{1}{\sum_{b \in \breve{a}} f_b} \tag{7.1}$$

and

$$p_{a|\breve{a}} = \frac{f_a}{\sum_{b \in \breve{a}} f_b}, \quad \forall a \in \breve{a}\,. \tag{7.2}$$

The above formula can be obtained from the more general results of next Section 7.1.2.1 applied to the case of exponential headways.

The sum of the frequencies of all attractive lines is referred to as the *combined frequency* of the stop.

Interestingly, the above formula are also valid for the successor arc $a \in i^+$ of any other node $i \notin S$, where it is: $t_a = 0$, given that $f_a \to \infty$, and $p_{a|a} = 1$.

Equations (7.1) and (7.2) provide for each branch $a \in \breve{a}$ of hyperarc $\breve{a}$ the conditional travel time $t_{a|\breve{a}}$ (that are all equal to $t_{\breve{a}}$) and the diversion probability $p_{a|\breve{a}}$, respectively, which are the main variables of the strategy model based on hyperpaths presented in Section 6.1.3.

7.1.1.3 The greedy approach to compute the attractive lines

Consider the sub-problem of a class g user choosing among the lines i^+ available at stop $i \in S$ the attractive set $\bar{A}_i^+ = \breve{a}$ as part of a his/her trip towards destination d. The most difficult question in the computation of shortest hypertrees stems indeed from the search of the hyperarc $\breve{a}$ over the set i^+ which yields the minimum expected cost; after all, finding an optimal sub-set is a combinatorial problem.

Rearranging the Equations (6.30) based on (6.26), under the assumption that diversion arcs have only non-temporal costs, the following version of the Bellman equation provides the expected cost of diversion node $i \in N^{div}$:

$$w_i = Min\left(w_i(\breve{a}):\ \forall \breve{a} \subseteq i^+\right), \tag{7.3}$$

$$w_i(\breve{a}) = \gamma_i \cdot t_{\breve{a}} + \sum_{a \in \breve{a}} w_a \cdot p_{a|\breve{a}}\,. \tag{7.4}$$

The expected cost $w_i(\breve{a})$ to reach the destination from stop i as a function of the attractive set $\breve{a}$ is given by the sum of:

- the combined wait time $t_{\breve{a}}$, multiplied by the value of time γ_i (which is equal to the value of time γ_{ag} in Equation (6.67.f)),
- the remaining cost w_a to reach the destination once boarded each attractive line $a \in \breve{a}$, multiplied by the corresponding line share $p_{a|\breve{a}}$.

Assume that the remaining cost w_a to reach the destination once boarded each line $a \in i^+$ available at the stop

has been already determined; but recall that this is given by the cost of the waiting arc *a* plus the expected cost of its final node a^+ : $w_a = c_a + w_{a+}$. This sub-problem is in fact part of a more general recursive problem where the unknowns are the expected costs of all nodes (see Section 6.1.7).

Based on Equations (7.1) and (7.2), in the case of exponential headways the main function (7.4) of the optimal strategies Problem (7.3) becomes:

$$w_i(\breve{a}) = \frac{\gamma_i + \sum_{a\in\breve{a}} w_a \cdot f_a}{\sum_{a\in\breve{a}} f_a} . \tag{7.5}$$

Consider the case where another line, associated with arc $b\notin\breve{a}$, is added to the attractive set; based on (7.5) the new expected cost can be expressed through the following recursive formula:

$$w_i(\breve{a}\cup b) = \frac{\gamma_i + w_b \cdot f_b + \sum_{a\in\breve{a}} w_a \cdot f_a}{f_b + \sum_{a\in\breve{a}} f_a} = \frac{w_i(\breve{a})\cdot\left(\sum_{a\in\breve{a}} f_a\right) + w_b \cdot f_b}{f_b + \left(\sum_{a\in\breve{a}} f_a\right)} . \tag{7.6}$$

Because (7.6) is a weighted average with positive coefficients (the frequency of arc *b* and the cumulative frequency of stop *i*), then the expected cost at stop *i* can be improved if and only if a line whose remaining cost once boarded is lower than the current expected cost is added to the attractive set:

$$w_b < w_i(\breve{a}) \leftrightarrow w_i(\breve{a}\cup b) < w_i(\breve{a}) . \tag{7.7}$$

By exploiting the order of lines in terms of remaining costs, the complexity of finding the attractive set of lines can be dramatically reduced through the following *greedy algorithm*:

- starting from an empty set,
- add the lines in increasing order of remaining cost to reach the destination once boarded,
- stop when the remaining cost of the next line is higher than the current value of the expected cost.

The correctness of the greedy algorithm can be proved by contradiction. Assume the existence of a better attractive set which is not formed by the first best *n* lines whose remaining cost is lower that the resulting expected cost. This yields a value of expected cost through (7.5). Based on Equation (7.6), adding any line with a better remaining cost or subtracting any line with a lower remaining cost from this attractive set would improve the solution in terms of expected cost.

7.1.1.4 Model formulation as an optimization problem

Since the solution hypertree $\bar{A}$ is the unknown of the route choice problem (the optimal strategies), the model for a single destination and user class is formulated by using the following binary variables for each arc $a\in A$:

$$x_a = \begin{cases} 0, & \text{if } a \notin \bar{A} \\ 1, & \text{if } a \in \bar{A} \end{cases} . \tag{7.8}$$

The assignment model (for each destination and class) may now be stated as the following optimization problem to minimize the total cost suffered by passengers (i.e. both travel costs on arcs and wait costs at stops), subject to consistency constraints (i.e. to assign the demand along the solution hypertree):

$$Min\left(\sum_{a\in A} c_a \cdot q_a + \sum_{i\in S} \frac{\gamma_i \cdot q_i}{\sum_{a\in i^+} f_a \cdot x_a}\right) \tag{7.9.a}$$

subject to:

$$q_a = q_i \cdot \frac{f_a \cdot x_a}{\sum_{b\in i^+} f_b \cdot x_b}, \quad i = a^-, \quad \forall a \in A , \tag{7.9.b}$$

$$q_i = d_i + \sum_{a \in i^-} q_a, \quad \forall i \in N, \tag{7.9.c}$$

$$q_i \geq 0, \quad \forall i \in N, \tag{7.9.d}$$

$$x_a \in \{0,1\}, \quad \forall a \in A, \tag{7.9.e}$$

where q_a is the flow on arc a, q_i is the total flow at node i, d_i is the travel demand departing from node i, if any (these are flows of class g users directed toward destination d);. At first sight, Equations (7.9) constitute a mixed integer nonlinear optimization problem with unknowns q_i (real-valued) and x_a (integer-valued). Fortunately, however, this problem may be reduced to a simpler one by substituting the following flow conservation constraint for each node i and considering as unknown the arc flows $q_a \geq 0$:

$$\sum_{a \in i^+} q_a = q_i . \tag{7.10}$$

Indeed, by introducing new variables ω_i , which represent the total wait time at stop $i \in s$, defined as:

$$\omega_i = \frac{\gamma_i \cdot q_i}{\sum_{a \in i^+} f_a \cdot x_a} \tag{7.11}$$

one obtains the equivalent problem:

$$Min\left(\sum_{a \in A} c_a \cdot q_a + \sum_{i \in S} \omega_i \right) \tag{7.12.a}$$

subject to:

$$q_a = x_a \cdot f_a \cdot \omega_i, \quad \forall a \in i^+, \quad \forall i \in S, \tag{7.12.b}$$

$$\sum_{a \in i^+} q_a - \sum_{a \in i^-} q_a = d_i, \quad \forall i \in N, \tag{7.12.c}$$

$$q_a \geq 0, \quad a \in A, \tag{7.12.d}$$

$$x_a \in \{0,1\}, \quad \forall a \in A . \tag{7.12.e}$$

The objective function in Equation (7.12.a) is now linear and the 0-1 variables are only used in Equation (7.12.b) which are the only nonlinear constraints. These may be relaxed, yielding a linear program with real-valued unknowns q_a and ω_i :

$$Min\left(\sum_{a \in A} c_a \cdot q_a + \sum_{i \in S} \omega_i \right) \tag{7.13.a}$$

subject to:

$$q_a \leq f_a \cdot \omega_i, \quad \forall a \in i^+, \quad \forall i \in S, \tag{7.13.b}$$

$$\sum_{a \in i^+} q_a - \sum_{a \in i^-} q_a = d_i, \quad \forall i \in N, \tag{7.13.c}$$

$$q_a \geq 0, \quad a \in A . \tag{7.13.d}$$

It may be shown by using the extreme point properties of the solutions for a linear program, that Problem (7.13) is equivalent to Problem (7.12). The dual problem of this last linear program is:

$$Max\left(\sum_{i \in N} w_i \cdot d_i \right) \tag{7.14.a}$$

subject to:

$$\mu_a + c_a + w_j \geq w_i, \quad \forall a = (i,j) \in A, \tag{7.14.b}$$

$$\sum_{a \in i^+} f_a \cdot \mu_a = \gamma_i, \quad \forall i \in N, \tag{7.14.c}$$

$$\mu_a \geq 0, \quad a \in A, \tag{7.14.d}$$

where w_i and μ_a are the dual variables corresponding, respectively, to Equations (7.13.c) and Equation

(7.13.b).

Let ($\mathbf{q}^*$, $\boldsymbol{\omega}^*$) and ($\mathbf{w}^*$, $\boldsymbol{\mu}^*$) denote, respectively, the optimal solutions of the primal and dual linear programs. The weak complementary slackness conditions are:

$$\left(q_a^* - f_a \cdot \omega_i^*\right) \cdot \mu_a^* = 0, \quad i = a^-, \quad \forall a \in A \tag{7.15}$$

and

$$\left(\mu_a^* + c_a + w_j^* - w_i^*\right) \cdot q_a^* = 0, \quad \forall a = (i,j) \in A\,. \tag{7.16}$$

In both the primal and dual formulations, this transit assignment model has a close resemblance to the shortest path problem, and perfect correspondence is obtained when none of the arcs involves waiting; thus $f_a = \infty$ and $\omega_i = 0$.

7.1.1.5 Solution algorithm

The solution algorithm is composed of two parts. In a first pass, from the destination to all nodes (including origins), the successors (arc or hyperarc) and the expected cost (label) from each node to the destination are computed. In a second pass, from all nodes (including origins) to the destination, the demand is assigned to the arcs $a \in \bar{A}$ of the hypertree. The algorithm is stated below.

Table 7.1. Assignment algorithm on optimal strategies with exponential headways.

Part 1: Compute the optimal strategy

Step 1.1 (initialization):
$w_i \leftarrow \infty \ \forall i \in N$; $w_d \leftarrow 0$
$f_i \leftarrow 0 \ \forall i \in S$
$B \leftarrow A$; $\bar{A} \leftarrow \varnothing$

Step 1.2 (get the next arc to examine):
find $a \in B$ such that $c_a + w_{a+} \le c_b + w_{b+}$ for each $b \in B$; $B \leftarrow B\text{-}\{a\}$

Step 1.3 (do the Bellman check for arc $a = (i,j)$ and update the node labels):
if $w_i > c_a + w_j$ then:

if $f_a < \infty$ then $w_i \leftarrow \dfrac{w_i \cdot f_i + \left(c_a + w_j\right) \cdot f_a}{f_i + f_a}$; $f_i \leftarrow f_i + f_a$ otherwise $w_i \leftarrow c_a + w_j$ (7.17)

$\bar{A} \leftarrow \bar{A} + \{a\}$

Step 1.4 (loop until B is empty):
if $B = \varnothing$ then stop otherwise go to Step 1.2

Part 2: Assign the demand on the hypertree

Step 2.1 (preload the demand on the origins):
$q_i \leftarrow d_i \ \forall i \in N$
$q_a \leftarrow 0 \ \forall a \in A$

Step 2.2 (propagate the node flow to the successor arcs):

for each $a = (i,j) \in \bar{A}$ in decreasing order of $(c_a + w_j)$ do:

if $f_a < \infty$ then $q_a \leftarrow q_i \cdot \frac{f_a}{f_i}$ otherwise $q_a \leftarrow q_i$ (7.18)

$q_j \leftarrow q_j + q_a$

Special treatment is reserved to waiting arcs with finite frequency ($f_a < \infty$), otherwise the proposed algorithm is identical to a shortest tree method adopting the label setting approach by Dijkstra.

The auxiliary variables f_i contain the combined frequencies of all waiting arcs that exit from stop $i \in S$ and belong to the solution hypertree $\bar{A}$.

The convention $0 \cdot \infty = \gamma_i$ is used in the first update of expected cost for stop $i \in S$ when $f_i = 0$ and $w_i = \infty$.

Note that in Step 1.3 a line a whose remaining cost once boarded $w_a = c_a + w_{a+}$ is higher than the current expected cost w_i of stop i will not be included in the attractive set, while the lines are processed in order of remaining cost to reach the destination. Moreover, the label update of Step 1.3 is consistent with Equation (7.6). This is consistent with the greedy approach and thus ensures the success of the proposed algorithm to compute the shortest hypertree.

Finally, in Step 2.2 the flow propagation from stop i is consistent with the line shares of Equation (7.2).

Also, by using the primal and the dual formulations of the proposed transit assignment model one can prove that the proposed algorithm indeed finds the solution of Problem (7.9).

The algorithm is applied for each destination and class in turn.

7.1.1.6 Numerical example

In the following, the optimal strategy model with exponential line headways is calculated for the example network of Section 5.1.3. The assignment graph is the same of Section 6.2.5 and is also depicted in Figure 7.2 along with the arc costs and the demand flows.

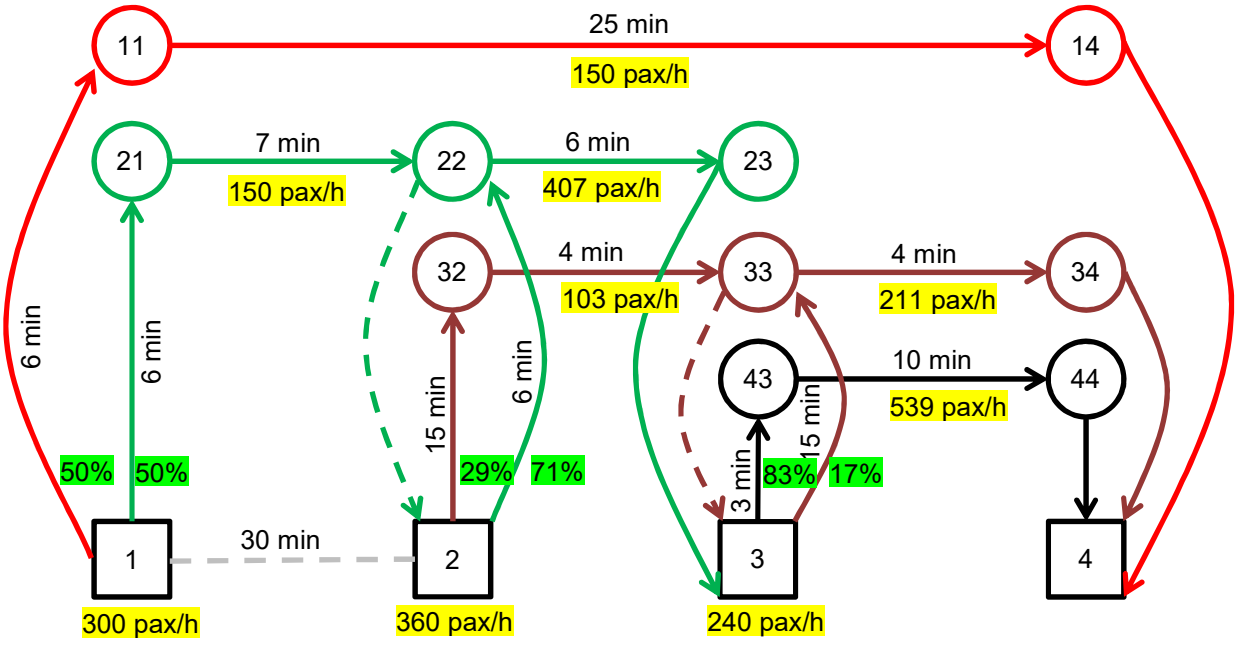

Figure 7.2. Input data and results of AoN assignment to optimal strategies applied to the example network.

The numerical computation presented in Table 7.2 results from a slight modification of the first pass of the algorithm described in Section 7.1.1.5: the next arc to visit is taken from the backward star of visited nodes

that are extracted in order of expected cost to destination, as in the shortest path algorithm of Section 6.2.5. The figures in brackets denote the Bellman updates of node costs and successors which are not convenient and/or are later replaced by a better solution.

Table 7.2. Shortest hypertree computation for destination node 4 following the optimal strategy algorithm.

node	expected cost	cumulative frequency	successor(s)	insertion order	extraction order
1	(49.07 = 30 + 19.07) (30.5 = 6 + 24.5) 27.75 = (30.5/6 + 25/6) / (1/6 + 1/6)	 (1/6) 1/3	(2) (21) 21,11	14	14
2	(23 = 15 + 8) 19.07 = (23/15 + 17.5/6) / (1/15 + 1/6) (57.75 = 30 + 27.75)	(1/15) 7/30	(32) 32,22 (1)	10	11
3	(19 = 15 + 4) 11.5 = (19/15 + 10/3) / (1/15 + 1/3)	(1/15) 2/5	(33) 33,43	8	8
4	0			1	1
11	25 = 25 + 0		14	5	12
14	0 = 0 + 0		4	2	2
21	24.5 = 7 + 17.5		22	13	13
22	17.5 = 6 + 11.5 (19.07 = 0 + 19.07)		23 (2)	12	10
23	11.5 = 0 + 11.5		3	11	9
32	8 = 4 + 4		33	9	6
33	4 = 4 + 0 (11.5 = 0 + 11.5)		34 (3)	6	5
34	0 = 0 + 0		4	3	3
43	10 = 10 + 0		44	7	7
44	0 = 0 + 0		4	4	4

Table 7.3. Attractive line shares for destination node 4.

stop	attractive set of lines	line	share
1	1, 2	1	1/2 = (1/6) / (1/6 + 1/6)
1	1, 2	2	1/2 = (1/6) / (1/6 + 1/6)
2	2, 3	2	5/7 = (1/6) / (1/6 + 1/15)
2	2, 3	3	2/7 = (1/15) / (1/6 + 1/15)
3	3, 4	3	1/6 = (1/15) / (1/15 + 1/3)
3	3, 4	4	5/6 = (1/3) / (1/15 + 1/3)

The shortest hypertree is identified recursively by the successor nodes. In particular, the shortest hyperpath from 1 to 4 is to board at stop 1 the first arriving between the red line and the green line; the former takes the passenger directly to the destination, while the latter requires to alight at stop 3; there, the passenger boards the first arriving between the maroon and the black line, both taking him/her to the destination. The dashed arcs of Figure 7.2 are not included in the solution hypertree. The arc flows can be easily determined by propagating the demand flows (depicted below the stops) along the solution hypertree, by applying the second pass of the algorithm and taking into account the line shares calculated in Table 7.3 (also depicted in green above the stops), thus obtaining the results depicted in yellow in Figure 6.8 for running arcs.

7.1.2 Regular headways and sequential observation

As explained in Section 6.2.1, exponentially distributed headways are just one extreme case in a spectrum. Indeed, exponential headways behave completely memory-less: if a passenger has already waited without success at the stop for a given period of time, s/he will have still to wait on average for the same time that

s/he expected to when s/he just reached the stop.

Other headway distributions correspond to higher service regularity. For example, the Erlang distribution offers a flexible representation for different degrees of regularity through its second parameter n, which is linked to the headway variation coefficient σ: $n = 1 / \sigma^2$. Its expected wait time provided by Equation (6.66) spans from the exponential case $1/f$ (for $n = 0$) to the case of constant headways $0.5/f$ (for $n \to \infty$).

In the following we refer to a given stop $i \in S$ and to the set L_s of lines serving it, each one associated with a waiting arc $a \in i^+$. Recall that the probability density function of the wait time for line $a \in i^+$ is related to the distribution of its headway (at that stop) through Equation (6.43). In the case of constant (deterministic) headways, the probability density function $\varphi_a{}^w(t)$ of waiting line a exactly for t and the probability $\bar{\Phi}_a^w(t)$ of waiting it for more than t are given, respectively, by (6.54) and (6.55); in the case of Erlang headways, these are given, respectively, by (6.52) (6.53).

The following exposition is based on the work of Gentile et al. (2002-2005).

7.1.2.1 Line shares and combined wait time

Consider the successor hyperarc $\bar{A}_i^+ = \breve{a} \subseteq i^+$ of stop $i \in S$. Assume that headways at stop i are independent and have a known distribution which may differ for each line serving the stop.

The probability $p_{a|\breve{a}}(t)$ that line a is boarded at time t is given by:

$$p_{a|\breve{a}}(t) = \varphi_a^w(t) \cdot \prod_{b \in \breve{a} - \{a\}} \bar{\Phi}_b^w(t) \quad , \forall a \in \breve{a} \,, \tag{7.19}$$

since the right hand side yields the probability that line a arrives at time t and all other attractive lines b have not yet arrived. Then, the line share is:

$$p_{a|\breve{a}} = \int_0^\infty p_{a|\breve{a}}(t) \cdot dt \quad , \forall a \in \breve{a} \,. \tag{7.20}$$

In the case of constant headways, this reduces to:

$$p_{a|\breve{a}} = f_a \cdot \int_0^{t_{\breve{a}}^{max}} \prod_{b \in \breve{a} - \{a\}} (1 - f_b \cdot t) \cdot dt \quad , \forall a \in \breve{a} \,, \tag{7.21}$$

where the maximum waiting time $t_{\breve{a}}^{max}$ is the minimum headway among the attractive lines:

$$t_{\breve{a}}^{max} = Min\left(\frac{1}{f_a} : \forall a \in \breve{a} \right). \tag{7.22}$$

The expected wait time $t_{\breve{a}}$ is given by:

$$t_{\breve{a}} = \int_0^\infty t \cdot \sum_{a \in \breve{a}} p_{a|\breve{a}}(t) \cdot dt \quad , \forall a \in \breve{a} \,. \tag{7.23}$$

where the integrand yields the wait time t multiplied by the probability that any line is boarded at t. Based on Equation (7.19) it is then:

$$t_{\breve{a}} = \int_0^\infty t \cdot \left(\prod_{a \in \breve{a}} \bar{\Phi}_a^w(t) \right) \cdot \left(\sum_{a \in \breve{a}} \frac{\varphi_a^w(t)}{\bar{\Phi}_a^w(t)} \right) \cdot dt \quad , \forall a \in \breve{a} \,. \tag{7.24}$$

The expected cost can be then retrieved from Equation (7.4).

As an alternative, the expected wait time can be obtained as:

$$t_{\breve{a}} = \int_0^\infty \prod_{a \in \breve{a}} \bar{\Phi}_a^w(t) \cdot dt \,, \tag{7.25}$$

where the integrand yields the probability that the wait time is higher than a given time t for all attractive lines a ; the proof of (7.25) is then similar to that of (6.48). In the case of constant headways, this reduces to:

$$t_{\breve{a}} = \int_0^{t_{\breve{a}}^{max}} \prod_{b \in \breve{a}} (1 - f_b \cdot t) \cdot dt \,. \tag{7.26}$$

The expected wait time $t_{a|\check{a}}$ conditional on boarding line a is given by:

$$t_{a|\check{a}} = \frac{\int_0^\infty t \cdot p_{a|\check{a}}(t) \cdot dt}{p_{a|\check{a}}} \quad ,\forall a \in \check{a} \,. \tag{7.27}$$

In the case of unbounded waiting time distributions, the computation can be addressed by cutting all tails at a suitable maximum headway h^{max} and scaling the original density of probability as follows:

$$\frac{\varphi_a^w(t)}{\left(1-\bar{\Phi}_a^w\left(h^{max}\right)\right)}\text{, for } t \le h^{max}\text{, and 0 otherwise.} \tag{7.28}$$

By observing Equation (6.65), which has general validity, some authors substitute the frequency of the line f_{ts} with $2 \cdot f_{ts} / (1-\sigma_{ts}^2)$ in the expression of the combined wait time and the line share (7.1) and (7.2), which are valid only for the case of exponential headways. As noted already in Section 6.2.1, this is an optimistic approximation that implies some coordination among different lines, which is unlikely to happen in practice.

7.1.2.2 Construction of the attractive set

Consider the sub-problem of a class *g* user choosing among the lines i^+ available at stop $i \in S$ the attractive set $\bar{A}_i^+ = \check{a}$ as part of a his/her trip towards destination *d*.

Assume that the remaining cost to reach the destination once boarded each line available at the stop has been already determined.

Equation (7.4) yielding the expected cost to reach the destination is still valid; but the greedy algorithm is not.

In general, it can be shown that if a line belongs to the solution attractive set, then also all lines with smaller remaining cost belong to it. This is also intuitive: why should the passenger let go a line which is better than another line s/he is willing to board?

Thus, the order of lines $a \in i^+$ in terms of remaining cost w_a still plays a role in the construction of the attractive set, which is formed by the first *x* lines. This is a general result.

However, it may happen that, different from the classical optimal strategies with exponential headways, the sequence of $w_i(\check{a}_x)$, where $\check{a}_x$ is the attractive set formed by the first *x* lines for $x = 1, \ldots, |i^+|$ in terms of w_a , shows more than one local minimum. In other words, it can happen that adding a line whose remaining cost is higher than the current expected cost may improve the solution. This also implies that in principle the solution attractive set could contain a line whose remaining cost is higher than the resulting expected cost.

Therefore, the algorithm should scan all such sets and evaluate for each of them the expected cost through Equation (7.4) to find the optimal solution:

$$\bar{A}_i^+ = \check{a}_{x^*} \quad , \quad x^* \leftarrow ArgMin\left(w_i(\check{a}_x) : x \in \left[1, \left|i^+\right|\right]\right). \tag{7.29}$$

7.1.2.3 Solution algorithm

The fact that best attractive set may contain a line whose remaining cost is higher than the resulting expected cost introduces possible cycles in the solution: the unlucky passenger that takes the bad line may alight as soon as possible (even at the same stop if the network topology allows it), go back to the stop and start waiting again.

There are many reasons why this should be avoided:

- in reality the situation that the passenger will find when getting back to the stop is strongly correlated with the unlucky one that s/e just left (instead for the model is a totally independent cast of dices);
- in theory a hyperpath does not contain cycles and thus our modelling framework cannot support them.

As mentioned in Section 6.1.7 the problem of cycles affects several models based on hyperpaths.

Fortunately, for the case at hand there is some evidence that the greedy approach is still valid if lines are ordered in terms of remaining cost to the destination including (instead of excluding) the waiting cost (for that

line only); this conjecture has not yet been proved nor rejected.

The original Optimal Strategies algorithm (Table 7.1) needs then to be adapted.

The filter $w_i > c_a + w_j$ of Step 1.3 on the remaining cost should not be applied at stops, where instead any new line a will be tested for inclusion and added to the attractive set if it improves the expected cost w_i . The tentative update is done using (7.21) and (7.25) in (7.4) instead of (7.17).

Moreover, we can opt to process arcs in order of:

- remaining cost to the destination including the waiting cost;
- remaining cost to the destination excluded the waiting cost,
- a predefined cost attribute (e.g. the distance on the graph from the stop to the destination).

Finally, in Equation (7.18) the ratio between line frequency and combined frequency, which yields the lane share in the case of exponential headways, is to be replaced with the probability obtained through (7.21).

The result is a heuristic which provides typically a good solution, but not necessarily an optimal strategy. Indeed, by reprocessing some arcs one may obtain a better hypertree.

Another approach which was found to work well in practice in Visum (2003), relaxes the requirement that all successor nodes of waiting arcs must have been processed before processing the stop, and substitutes estimation from upper and lower bounds where expected costs for successor nodes are not yet known.

7.1.3 Sequential observation and elapsed time

In Sections 7.1.1 and 7.1.2 it was assumed that passengers can only observe the next arriving vehicle at the current stop. This limited information constrains the possible decisions.

The simplest additional piece of information which can be obtained without any external support is the elapsed wait time. Whereas in the case of exponentially distributed headways this information is worthless, with growing regularity the passenger is able to revise his/her estimate of remaining wait times, and hence expected costs, while s/he is waiting. The effect becomes strongest with constant headways: if a line is served every 20 minutes, the estimated wait time at the beginning of waiting is 10 minutes; but after t minutes of waiting, the remaining wait time drops to (20-t)/2 minutes, until after at most 20 minutes the line must arrive with certainty.

Billi et al. (2003-2004) and PTV (2003) independently analyse the situation and generalize the notion of attractive line set, which is no longer constant, but varies as time is spent waiting at a stop.

Recall that any attractive set is formed by the first x lines in terms of remaining cost to reach the destination. Let then:

- $\breve{a}_{x(\tau)} \subseteq i^+$ be the attractive set of stop $i \in S$ that will be considered by the passenger at time $\tau \geq t$ of the wait after the elapsed wait time $t \geq 0$ and
- $w_i(t)$ be the expected cost after the elapsed wait time $t \geq 0$ resulting from the future application of the dynamic attractive set $\breve{a}_{x(\tau \geq t)}$.

Note that $w_i(t)$ is different from the expected cost $w_i(\breve{a}_{x(t)})$ calculated through Equation (7.4) for a constant attractive set $\breve{a}_{x(t)}$.

The key property of a "good" dynamic set $\breve{a}_{x(\tau)}$ is:

$$w_a \leq w_i(t) \Leftrightarrow a \in \breve{a}_{x(t)} \quad ,\forall t \geq 0 \quad ,\forall a \in i^+; \tag{7.30}$$

if after the elapsed wait time t arrives at stop i a line $a \in i^+$ whose remaining cost w_a to reach the destination is higher than the current expected cost $w_i(t)$, then the passenger has no convenience in boarding at and the line should not be included in the attractive set $\breve{a}_{x(t)}$; on the contrary, if its remaining cost w_a is lower than the current expected cost $w_i(t)$, then for the passenger it is convenient to board line a, which should then be included in the in the attractive set $\breve{a}_{x(t)}$.

Based on the above property of the dynamic attractive set derives an important property of the function $w_i(t)$, which can be proved to be monotone decreasing:

$$w_i(\tau) \leq w_i(t) \quad ,\forall \tau \geq t \quad ,\forall t \geq 0 \,. \tag{7.31}$$

The continuity of $w_i(t)$ is instead ensured by the continuity of the headway distribution functions.

As for (7.30), in the following we provide just an intuitive proof based on logical deduction. Our conjecture is that $w_i(t+dt) \le w_i(t)$ at any $t \ge 0$ for a small $dt > 0$. Without loss of generality, assume that the attractive set is constant during this small amount of time. The expected cost is composed by an expected waiting cost and an average (weighted by the line shares) of the remaining costs. The expected waiting cost decreases during dt because each line is more likely to arrive (the remaining wait time of a single line is a decreasing function of the elapsed time under mild assumptions). But the average remaining cost can increase if the line share of a costly line increases. Take this to the extreme case where the worst line is going to arrive at $t+dt$. Also in this case the expected cost decreases, because the remaining cost of that line is lower than the expected cost at t.

Based on (7.30) and (7.31), starting from $t = 0$, the expected cost $w_i(t)$ decreases with the elapsed time t and reaches progressively the remaining cost w_a of the initially attractive lines $a \in \breve{a}_{x(0)}$, which from that point shall exit the dynamic attractive set. Thus, each line $a \in i^+$ is attractive in a time interval $[0, \tau_a]$ for some $\tau_a \ge 0$ or never attractive at all (i.e. $\tau_a = 0$). In other words, while wait elapses, the lines drop out of the attractive set in decreasing order of their remaining cost.

Moreover, if the headway distribution of a line $a \in i^+$ is bounded, then the line will exit the dynamic attractive set at time τ_a not later than its maximum wait time h_a^{max} : $\tau_a \le h_a^{max}$.

The dynamic attractive set can be then defined as:

$$\breve{a}_{x(t)} = \left\{ a \in i^+ : w_a \le w_i(t) \right\} = \left\{ a \in i^+ : \tau_a \ge t \right\} . \tag{7.32}$$

7.1.3.1 Construction of the attractive set

The definition of the dynamic attractive set $\breve{a}_{x(t)}$ reduces to finding the times τ_a at which each line $a \in i^+$ drops out of the attractive set. Let $a1, a2, \dots, an$, with $n = |i+|$, be the lines in decreasing order of remaining cost, from the best to the worst; it is: $w_{a1} \le w_{a2} \le \dots \le w_{an}$; $\tau_{a1} \ge \tau_{a2} \ge \dots \ge \tau_{an}$. In the following the index a is dropped for the sake of simplicity. The expected cost $w_i(t)$ can then be denoted as $w_i(t, \{\tau_1, \tau_2, \dots, \tau_n\})$, thus showing explicitly its dependence on the dynamic set $\breve{a}_{x(\tau)}$.

The construction is done working backwards from the time $\tau_1 = h_1^{max}$ when a vehicle of the best line has arrived with certainty so that the expected cost is w_1 .

To find the time τ_2 when the second best line drops out of the attractive set, one needs to solve:

$$w_i\left(\tau_2, \{\tau_1, 0, \dots, 0\}\right) = w_2 . \tag{7.33}$$

For example, if the best line has constant headway this yields:

$$\tau_2 = h_1^{max} - 2 \cdot \frac{(w_2 - w_1)}{\gamma_i} . \tag{7.34}$$

The procedure follows recursively finding τ_k by solving:

$$w_i\left(\tau_k, \{\tau_1, \dots, \tau_{k-1}, 0, \dots, 0\}\right) = w_k . \tag{7.35}$$

There however two circumstances to take into account when constructing the attractive set in this way.

First, it can happen that the increasing expected cost (by proceeding backwards) does not reach w_k before time 0:

$$w_i\left(0, \{\tau_1, \dots, \tau_{k-1}, 0, \dots, 0\}\right) < w_k . \tag{7.36}$$

In this case, the procedure stops; the attractive set is constituted by the first k-1 lines: $\tau_h = 0 \ \forall h \ge k$. This index is recorded as $r^* = k$-1.

Second, it can happen that $\tau_k > h_k^{max}$, meaning that line k arrives with certainty before it drops out of the attractive set. In this case the procedure must be restarted from the time h_k^{max} and the expected cost w_k , considering that the wait ends at this time with a dynamic set constituted by the first k lines: $\tau_h = h_k^{max} \ \forall h \le k$. This index is recorded as $q^* = k$.

Thus, the attractive set spans only the relevant intervals of instant $[\tau_{k-1}, \tau_k]$ with k from q^* to r^*.

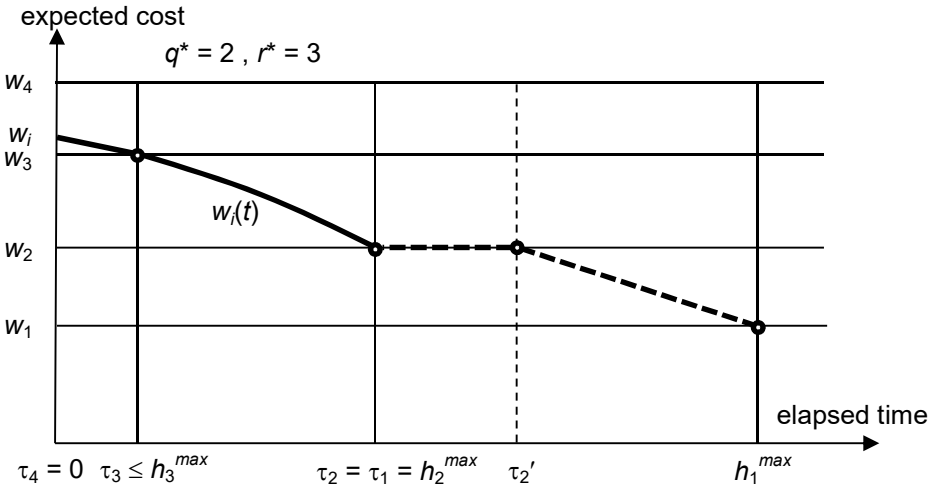

Figure 7.3. Expected cost as a function of the elapsed time for a dynamic attractive set. Note that the instant obtained as intersection of $w_i(t, \{h_1^{max}, 0, \dots, 0\})$ with w_2, denoted τ_2' is higher than h_2^{max}, and the procedure is then restarted from h_2^{max}.

7.1.3.2 Expected cost and line shares

The probability density function of the remaining wait time $\tau \geq t$ for line $a \in i^+$ after the elapsed time $t \geq 0$ is:

$$\varphi_a^w(\tau \mid t) = \frac{\varphi_a^w(\tau)}{\bar{\Phi}_a^w(t)}, \tag{7.37}$$

while the probability of waiting for more than $\tau \geq t$ after the elapsed time $t \geq 0$ is: is:

$$\bar{\Phi}_a^w(\tau \mid t) = \frac{\bar{\Phi}_a^w(\tau)}{\bar{\Phi}_a^w(t)}. \tag{7.38}$$

Assume that the passenger has waited without success until time $t \in [\tau_{k+1}, \tau_k]$ when the first k lines are attractive. The probability that a line h with $h \leq k$ is boarded at time τ is:

$$p_h(\tau \mid t) = \varphi_h^w(\tau \mid t) \cdot \prod_{1 \leq j \neq h \leq k} \bar{\Phi}_j^w(\tau \mid t) = \frac{\varphi_h^w(\tau)}{\bar{\Phi}_h^w(\tau)} \cdot \left(\prod_{j=1}^{k} \frac{\bar{\Phi}_j^w(\tau)}{\bar{\Phi}_j^w(t)} \right), \tag{7.39}$$

which yields the probability that line h arrives at time t and all other attractive lines j have not yet arrived; the latter product of the above equation yields the probability that the passenger will not board any of the attractive lines from time t to τ.

The expected cost at time t then given by:

$$w_i(t, \{\tau_1, \dots, \tau_k, 0, \dots, 0\}) = \int_t^{\tau_k} \left(\sum_{j=1}^{k} \left(\gamma_i \cdot (\tau - t) + w_j \right) \cdot \frac{\varphi_j^w(\tau)}{\bar{\Phi}_j^w(\tau)} \right) \cdot \left(\prod_{j=1}^{k} \frac{\bar{\Phi}_j^w(\tau)}{\bar{\Phi}_j^w(t)} \right) \cdot d\tau + w_k \cdot \left(\prod_{j=1}^{k} \frac{\bar{\Phi}_j^w(\tau_k)}{\bar{\Phi}_j^w(t)} \right). \tag{7.40}$$

The first term is the expected cost at time t if boarding occurs before time τ_k (on any attractive line at any instant $\tau \in [t, \tau_k]$), while the second term is the remaining cost of line k if boarding occurs later. The second term can be written in this compact form because the value of the expected cost at time τ_k is by construction equal to w_k, while the passenger will wait until τ_k only if no attractive line arrives in the meanwhile, which is yielded by the final product of the equation.

This formula is used as in (7.35) to obtain numerically the time τ_{k+1} during the backward computation of the integral from $t = \tau_k$ until $w_i(t)$ reaches w_{k+1} or $t = 0$. At the end of the recursive computation all τ_k with k from q^* to r^* are determined, the attractive set is $\bar{A}_i^+ = \breve{a} = \{q^*, \dots, r^*\}$ and the expected cost is $w_i = w_i(0)$.

Line probabilities can be computed afterwards, as follows:

$$p_{h|\breve{a}} = \sum_{k=Max(h,q^*)}^{r^*} \left(\prod_{j=k+1}^{r^*} \bar{\Phi}_j(\tau_j) \right) \cdot \int_{\tau_{k+1}}^{\tau_k} \frac{\varphi_h^w(\tau)}{\bar{\Phi}_h^w(\tau)} \cdot \left(\prod_{j=1}^{k} \bar{\Phi}_j^w(\tau) \right) \cdot d\tau \,. \tag{7.41}$$

Here the sum is taken over all intervals $[\tau_{k+1},\tau_k]$ in which line h may be boarded. The first product represents the probability that none of the lines k+1,…, r^* has arrived before being dropped, so that the passenger is still waiting at τ_{k+1}. The integral represents the probability of boarding line h during the interval $[\tau_{k+1},\tau_k]$, which is obtained like in (7.20).

For completeness, the combined wait time of the attractive set can be obtained from Equation (7.3):

$$t_{\breve{a}} = \frac{w_i - \sum_{a \in \breve{a}} w_a \cdot p_{a|\breve{a}}}{\gamma_i} \,. \tag{7.42}$$

The expected wait time $t_{a|\breve{a}}$ conditional on boarding line a can be obtained like in (7.27), once $p_{a|\breve{a}}(t)$ is defined based on (7.20) and (7.41).

7.1.3.3 Solution algorithm

Let us sum up the results again: a passenger can gain using a dynamic strategy. To this end, s/he defines a sequence of instant $\tau_1 \geq \tau_2 \geq \ldots \geq \tau_n$ at which the lines available at the stop are dropped from the attractive set in order of remaining cost to reach the destination. During each interval $[\tau_{k+1},\tau_k]$ the attractive set is constant and made-up by the first best k lines. So, if after an elapsed wait time $t \in [\tau_{k+1},\tau_k]$ one line $h \leq k$ arrives at the stop, the passenger boards it; other lines are ignored.

The assignment algorithm is the same described in 7.1, except that the formulas for the tentative label update and the line shares are replaced, respectively, by (7.40) and (7.41). The computational advantage wrt the case of fixed dynamic set is that, by construction, the expected cost is higher than all the remaining costs of the attractive lines: this implies the absence of cycles and thus the optimality of the algorithm.

7.1.4 Parallel observation

In the previous sections it is still assumed that passengers can only observe the next line to be served at a stop. With real-time passenger information this assumption becomes invalid and passengers can often see the next departure times for all lines serving a stop. Equivalently, for the case of fixed schedules, printed timetables may exist at transfer stops, and the passenger may inspect them when s/he reaches the transfer stop, although they were not considered when s/he planned the journey. In both cases, the actual remaining wait times for all lines serving the current stop become available at the beginning of the wait.

Gentile et al. (2002-2005) and VISUM (2003) independently propose a label-setting algorithm for computing expected costs from each stop to a given destination and assign flows consistently with the line shares resulting from a route choice based on optimal strategy.

Unlike the other cases, the passenger does not choose an attractive set $\bar{A}_i^+ = \breve{a}$ at stop i and then boards the first arriving vehicle of a line in $\breve{a}$. Stochasticity here plays a different role. A given passenger arriving at a stop observes all wait times t_a for each $a \in i^+$ simultaneously, and makes a deterministic choice based upon this information: he will simply choose the line a which minimizes the expected cost $\gamma_i \cdot t_a + w_a$. However, t_a represents here a random draw from the distribution of all possible departure times, whose distribution is linked to that of headways as shown in Section 6.2.1. Different draws may lead to different decisions and multiple paths, which form hyperpath.

7.1.4.1 Line shares and expected cost

As stated before, the shares are equal to the probability of the respective lines being optimal. More precisely, the following condition shall hold for line $a \in \breve{a}$ to be chosen:

$$\gamma_i \cdot t_a + w_a \leq \gamma_i \cdot t_b + w_b \quad , \forall b \in \breve{a} - \{a\} \,. \tag{7.43}$$

The probability $p_{\breve{a}|a}(t)$ that line a is taken at time t is given by:

$$p_{a|\check{a}}(t) = \varphi_a^w(t) \cdot \prod_{b \in \check{a}-\{a\}} \bar{\Phi}_b^w\left(t + \frac{w_a - w_b}{\gamma_i}\right) \quad ,\forall a \in \check{a}\,, \tag{7.44}$$

since the right hand side yields the probability that line a arrives at time $t_a = t$ and all other attractive lines b have a worst expected cost, i.e. $t_b \geq t_a + (w_a - w_b)/\gamma_i$. Then, the line share is given by (7.20).

Equation (7.44) resembles (7.19). The difference lies in the condition imposed on the lines $b \neq a$ when a service of line a arrives at time t; it is not sufficient that line b will not arrive before t, but b must be worse than a in terms of waiting cost plus remaining cost.

As headways are constant, the general formula reduces to:

$$p_{a|\check{a}} = f_a \cdot \int_0^{1/f_a} \prod_{b \in \check{a}-\{a\}} Mid\left(0, 1 - f_b \cdot \left(t + \frac{w_a - w_b}{\gamma_i}\right), 1\right) \cdot dt \quad ,\forall a \in \check{a}\,. \tag{7.45}$$

The expected wait time $t_{\check{a}}$ is given by:

$$t_{\check{a}} = \int_0^{\infty} t \cdot \sum_{a \in \check{a}} p_{a|\check{a}}(t) \cdot dt \quad ,\forall a \in \check{a}\,. \tag{7.46}$$

As headways are constant, the general formula reduces to:

$$t_{\check{a}} = \sum_{a \in \check{a}} f_a \cdot \int_0^{1/f_a} t \cdot \prod_{b \in \check{a}-\{a\}} Mid\left(0, 1 - f_b \cdot \left(t + \frac{w_a - w_b}{\gamma_i}\right), 1\right) \cdot dt \quad ,\forall a \in \check{a}\,. \tag{7.47}$$

The expected cost can be then retrieved from Equation (7.3). As an alternative, the expected cost can be obtained directly as:

$$w_i(\check{a}) = \gamma_i \cdot \int_0^{\infty} \prod_{b \in \check{a}} \bar{\Phi}_b^w\left(t - \frac{w_b}{\gamma_i}\right) \cdot dt\,, \tag{7.48}$$

where the integrand yields the probability that the total cost to destination is higher than a given value $\gamma_i \cdot t$ for all attractive lines b.

The expected wait time $t_{a|\check{a}}$ conditional on boarding line a is given by (7.27).

7.1.4.2 Construction of the attractive set

To determine the attractive set we simply have to evaluate the above Equations for $\check{a} = i^+$ an then find out which line have a positive diversion probability:

$$\check{a} = \left\{a \in i^+ : p_{a|i^+} > 0\right\}. \tag{7.49}$$

In case of unbounded headways, any line serving the stop has a positive probability to be attractive: there is always a small chance that the good lines are late and one has to board on a bad line. This applies also to lines that apparently take the passenger far away from the destinations.

From a computation point of view, this raises some issue in the shortest hypertree algorithm of Table 7.1. As explained in Section 7.1.2.3 the fact that the expected cost w_i may be smaller than the remaining cost w_a of some attractive line $a \in \check{a}$ would produce cycles in the solution, which is to be avoided.

7.1.5 Comparison among different waiting models

We briefly compare here the different models in terms of the numerical results they yield for the example network of Section 5.1.3. To simplify the presentation we only use the demand from Stop 1 to Stop 4 (300 pax/h); in this case, for each line, only one volume is relevant because all passenger board and alight the service at the same stop.

Table 7.4. Line volumes (pax/h) and travel times for different waiting models.

headway distribution	exponential	constant	constant	constant
information acquired	arriving line	arriving line	elapsed wait time	parallel observation

volume Line 1 - Red	150	150	188	216
volume Line 2 - Green	150	150	113	84
volume Line 3 - Maroon	0	0	0	0
volume Line 4 - Black	150	150	113	84
expected travel time 1→4	30min30sec	27min45sec	27min41sec	27min35sec

The base case is the classical Optimal Strategies with exponential headways and information only about the next line arriving at the stop.

Without additional information, regular operation (constant headways) does not provide enough additional cues to change route choice. Both attractive paths are executed with 50% probability each. The reduction of expected travel time from 30min30sec to 27min45sec is only due to the shorter expected wait time with constant headways.

The more information is available, the higher the share of the faster line, as passengers know when it is advantageous to pass up the slower line, although it departs first from Stop 1.

Interestingly, the effect on total expected travel time is minimal in this particular example, but the shift of volumes between lines is significant: the difference compared to the case without additional information is up to 50%.

Finally, note that the fastest route in terms of running time is to transfer at Stop 2 to Line 3. This option is never optimal due to the low frequency of the Maroon line.

A further effect of information provision arises, if passengers can walk to nearby stops. In that case it makes a difference whether dynamic information about waiting times at the distant stop are already available before walking there.

To illustrate this effect, consider a small extension of the example network. An additional Stop 5 can be reached by walking from Stop 2 (walking time = 2 min). From Stop 5 an additional service, Line 5 - Purple, runs to Stop 4, with the characteristics shown in Table 7.5.

Table 7.5. Characteristics of additional Line 5 - Purple.

line segment	length [km]	frequency [veh/h]	expected headway [min]	commercial speed [km/h]	vehicle capacity [pax]	running time [min]
(4,5)	10	15	4=60/15	40	80	15

In the extended network we compare the cases where passengers observe current waiting times only for their local stop, or also for the distant stops which are walkable from their current stop.

Table 7.6. Effect of information provision for distant stops.

headway distribution	constant	constant
information	parallel observation only for local stop	parallel observation also for distant stops
volume Line 1 - Red	196	118
volume Line 2 - Green	104	182
volume Line 3 - Maroon	0	85
volume Line 4 - Black	0	0
volume Line 5 - Purple	104	97
expected travel time 1→4	27min25sec	26min37sec

If no information is provided for the distant stop, passengers at Stop 2 need to take a deterministic decision based on expected remaining travel times whether to transfer at Stop 3 or at Stop 5. Table 7.6 shows that in this case passengers walk to Stop 5 and board Line 5, instead of transferring later to Line 4, as they do in the original example. The reason is that Line 5 is slow but frequent; this results in a gain of 10 sec.

If waiting time information for both options is already available at stop 2, passengers can decide depending

on the current situation. This results in a large time saving, more than in the previous cases, and diverts 78 passengers from the red Line to the options via Stop 2. Moreover, the possibility of adopting a strategic behavior at Stop 2 finally activates Line 3, which is fast but not frequent.

7.1.6 When to alight? Where to continue?

All models described above answer one question: which lines are boarded by passengers waiting at a stop? As explained in the previous sections, the answer depends on what a passenger can observe while waiting at a single stop.

But what if a passenger has a broader scope of options without currently being in person at a stop? This is the case when there are several origin stops for a trip or several stops from which to continue after a transfer. More elementary, the choice between remaining on board a line and making a transfer depends on what information becomes available to the passenger at which particular moment.

In the following, the question when to alight and where to continue is addressed, thinking of a passenger on board a line who is able to acquire some information on waiting times at next stops. Actually, current technology (passenger navigators for mobile devices) allows to acquire these information anywhere, which makes the route choice model further complicate.

7.1.6.1 No information

First consider the case where a passenger is on board a vehicle and has no information about wait times for onward connections. The decision whether or not to alight is deterministic because the passenger can estimate the cost of transferring at the next stop only on the basis of expected costs. Two questions arise:

- Is this decision really deterministic?
- If onward connections are available from several stops, should the choice set contain one option for each possible next boarding stop or just one for the transfer in general?

Based on the objective information available to the passenger the answer to question 1 must be “yes”. This implies that passengers on board a given service and travelling to the same destination will all alight at precisely the same (transfer) stop. While this rule is theoretically sound, it would limit severely the set of paths chosen for a given O-D pair.

In practice, passengers may use more paths due to a variety of reasons, including random taste variations and imperfect estimates of remaining travel time. This may cause difficulties calibrating a deterministic choice model to observed flows. One possible way to account for more realistic behaviour is to apply a discrete choice model (see Section 4.5) to the choice set containing two alternatives: remain-on-board and alight; the utility of alighting should include the expected cost of each possible next boarding stop, i.e. the satisfaction of all these alternatives. A stochastic route model based on sequential arc choices provides a suitable foundation for such an approach.

7.1.6.2 Information about local onward options while still on board

Assume now that full wait time information is available at the current stop, but it is displayed in a way (e.g., through count-down panels) that the passenger can access them while still on board. Even in the absence of an information system, passengers on board a line may observe other vehicles arriving and departing at the same stop. Such an observation does not require any technical device, but still improves the passenger estimate of wait times for the transfer options. In that situation, remaining on board and each alternative transfer to other lines from the current stop become simultaneous options within a single choice set (not sequential choices). The appropriate boarding model (one of 7.1.1-7.1.4) should then be applied to the entire choice set.

If the trip could alternatively continue from a different stop, for which no wait time information is locally available, then there is a prior choice (deterministic or stochastic) between transferring to such a distant stop and the local options.

7.1.6.3 Information about all onward options while still on board

Finally, suppose that even more information is available to the passenger on board: actual wait times are displayed not only for the current stop, but also for the next stops of the line including more distant stops reachable by a short walk. Assume that a smartphone application enables the passenger to simultaneously observe all lines with which s/he can continue the journey – regardless of his/her current position. Based on real-time data and short-term forecast about arrival times, the mobile device can suggest the best transfer stop and the best line to board there.

In such a situation the passenger will take a sequence of decisions (the real-time forecast may change during the trip) on a choice set which includes a wide set of lines serving stops reachable by a short walk. The choice model from Section 7.1.4 would be appropriate for this kind of situation.

A clear distinction shall be done between the alternatives among which the choice is made at each diversion node (basically all nodes are diversion here), that are all reachable lines, and the local alternatives physically connected with the node, that are all arcs of its forward star. In essence, the choice is modelled wrt the first set of alternatives (the lines) and then the results are aggregated wrt the second set of alternatives (the arcs) to apply the sequential route choice paradigm of Section 6.1.5.

7.1.7 Optimal strategies on diachronic graphs

Consider the space-time network introduced in Section 6.3. How is it possible to simulate optimal strategies in the framework of a schedule-based model?

To this aim we shall concentrate on the essence of the model presented in Section 7.1.2: when a vehicle of a line (in this case a run) departs from the stop, a passenger will board if, depending on his/her destination, it is convenient (i.e. less costly) to do so than keep waiting for other services. In that case, we say that the line is attractive for the passenger.

We are here assuming that the passenger is not informed of the exact timetable when making his/her route choice (at the origin) and becomes aware of the run departure times by observing the vehicles at the stop. The assignment on the diachronic graph will then reflect what happens in practice for a given schedule, which can be fixed, but unknown to the passenger, or a realization for a particular day.

Interestingly, no hyperpath representation is required by the proposed model (no diversion nor hyperarcs). Hence, there is no need to modify the stop topology with respect to that of Figure 6.11. The only thing we need in addition to the classical schedule-based model is the expected cost to reach the destination from the stop as it is perceived by uniformed passengers (which is different than the cost resulting in practice), because this allows to represent the binary en-route decision between boarding and keep waiting.

For this purpose we shall apply Equations (7.4) and (7.25). This requires to calculate the attractive set (which here is not represented as a hyperarc) and the determinants of the headway distribution for each line. The latter can be obtained through Equations (5.12) and (5.13) as in Section 6.3.3.

The attractive set (for a given class of users) can instead be built-up at stops in reverse chronological order and transmitted backward in time through waiting arcs of the diachronic graph, within the computation of the optimal strategies towards a given destination. The procedure differs from the computation of a shortest tree on the diachronic graph (see Section 6.3.6) in only one point: the remaining cost of the waiting arc is not given as usual by the sum of the arc cost plus the expected cost to destination of its final node, as in Equation (6.15), but it is provided by the combined cost of the attractive set for the arc.

More specifically, to each stop is associated a set of lines and for each line of the set the cost to reach the destination once boarded. Each time a stop node is visited to apply the Bellman relation during the optimal strategy procedure, the attractive set is updated in a different way depending on which type or arc provided the best local alternative:

- if the boarding arc prevails and the corresponding line is already present in the attractive set, than its cost to reach the destination once boarded is updated with the meaning cost of the arc;
- if the boarding arc prevails and the corresponding line is not present in the attractive set, then it is added with the remaining cost of the arc;
- if the combined cost of the attractive set resulting after the above updates is higher than the remaining cost of the boarding arc plus the cost of waiting for that line only, the attractive set is reinitialized with it consistently;

- if the waiting arc prevails, then no update occurs;
- if the stop arc prevails, heading to the pedestrian network, then the attractive set is reinitialized with an empty set.

The solution of this routing algorithm is a tree which can be used to propagate on the network the demand loads from the origins to the current destination.

7.1.8 Reference notes and concluding remarks

Early works on transit systems were mostly devoted to the analysis of service regularity and to the process of waiting at single stops served by several lines. They are mainly aimed at developing realistic bus headway distributions and consequent passenger wait time distributions, such as Power and Erlang, in the case of common lines, i.e. lines overlapping along part of their itinerary (e.g., Hasseltroem, 1981; Marguier, 1981; Gendreau, 1984; Marguier and Ceder, 1984; Jansson and Ridderstolpe, 1992; Bouzaıene-Ayari et al., 2001).

With reference to the case of independent exponential headways, Spiess and Florian (1989) introduced the notion of optimal strategies to describe the adaptive en-route behaviour of passengers at the stop who board the arriving carrier if it belongs to a given set of attractive lines, not necessarily common. Nguyen and Pallottino (1988) showed how strategies can be formalized through hyperpaths. These two seminal works provided the theoretical and algorithmic base for the development of assignment models on large transit networks ever since. The expected wait time at a stop is assumed equal to the inverse of combined frequencies for all attractive lines and the line shares are determined by multiplying the frequency and the expected wait time. This model can also be extended to the case of stochastic (logit) route choice (Nguyen et al., 1998).

As was pointed out also by these original contributions, the assumptions underlying this frequency-based model are inconsistent with statistical analysis of real-world data (e.g., Bellei and Gkoumas, 2010), since independent exponential headways are obtained only under highly irregular service conditions, which is a clearly undesirable for planning.

The more recently developed models and methods, not only have added rigor to the analysis and efficiency to the algorithms, but also have provided the possibility of reproducing several relevant transit phenomena: queuing of passengers at stops, discomfort on-board due to overcrowding, partially regular and correlated distributions of headways at stops, provision of information at stops, etc. Congestion and irregularity are treated in later sections. Here we concentrated on the role of information and its consequences on the behaviour of passengers at stops.

A sound formulation of the stop model which allows for more realistic headway distributions, ranging from deterministic to exponential, for example based on the distance from the first stop, as well as for different assumptions on the available information, including the provision of real-time estimation of vehicle arrivals using Variable Message Signals or Apps, is supported by the more recent work of Gentile et al. (2005); these aspects are essential for a good planning of transit systems (Shimamoto et al., 2005; Ren et al., 2009). If no real-time information is available, passengers may change their attractive set to minimize the expected cost, simply based on the time already spent waiting at the stop (Billi et al., 2004, Noekel and Wekeck, 2009).

The provision of information (Dziekan and Kottenhoff, 2007) has been analysed, not only in the framework of frequency-based models, but also in the framework of schedule-based models (Hickman and Wilson, 1995; Crisalli and Rosati, 2005). The interaction of many individuals receiving and transmitting real-time personalized information (crowd sourcing), for each Origin-Destination pair and desired departure or arrival time, is a new stream of research (Arentze, 2013; Nuzzolo et al., 2013).)

The algorithm presented in Section 7.1.7 for the computation of optimal strategies on diachronic graphs without hyperpaths is an original contribution of this book. The proposed approach inherits some similarity with the stochastic model of Cortés et al. (2013) for static transit networks, where the probability of boarding a line is a decreasing function of the difference between the remaining cost and expected cost to destination of keep waiting.

7.1.8.1 List of references

Arentze T.A. (2013) Adaptive, personalized travel information systems: A Bayesian method to learn users' personal preferences in multi-modal transport networks. IEEE Transactions on Intelligent Transportation Systems 14, 1957-1966.

Bellei G., Gkoumas K. (2010) Transit vehicles' headway distribution and service irregularity. Public Transport 2, 269-289.

Billi C., Gentile G., Nguyen S., Pallottino S. (2004) Rethinking the wait model at transit stops. Proceedings of of TRISTAN V, Guadeloupe, French West Indies. Also in Proceedings of MTIT 2003, Reggio Calabria, Italy.

Bouzaıene-Ayari B., Gendreau M., Nguyen S. (2001) Modelling bus stops in transit networks: A survey and new formulations. Transportation Science 35, 304-321.

Cortés C.E., Jara-Moroni P., Moreno E., Pineda C. (2013) Stochastic transit equilibrium. Transportation Research B 51, 29-44.

Crisalli U. and Rosati L. (2005) Transit services and user information: an application of schedule-based path choice and assignment models. Proceedings of European Transportation Forum 2005, Strasbourg.

Dziekan K., Kottenhoff K. (2007) Dynamic at-stop real-time information displays for public transport: effects on customers. Transportation Research A 41, 489-501.

Gendreau M. (1984) Une etude approfondie d'un modele d'equilibre pour l'affectation des passagers dans les reseaux de transport en commun. PhD thesis, Departement d'informatique et de recherche operationnelle, Universite de Montreal, Canada.

Gentile G., Nguyen S., Pallottino S. (2005) Route choice on transit networks with online information at stops. Transportation Science 39, 289-297. Also in Proceedings of the VI Congresso SIMAI 2002, Chia Laguna, Italy.

Hasseltroem D. (1981) Public transportation planning – A mathematical programming approach. PhD thesis, Department of Business Administration, University of Gothenburg, Sweden.

Hickman M.D., Wilson N.H.M. (1995) Passenger travel time and path choice implications of real-time transit information. Transportation Research C 3, 211-226.

Jansson K., Ridderstolpe B. (1992) A method for the route-choice problem in public transport systems. Transportation Science 26, 246-251.

Marguier, P.H.J. (1981) Optimal strategies in waiting for common bus lines. Master's thesis, Department of Civil Engineering, M.I.T., Cambridge, USA.

Marguier P.H.J., Ceder A. (1984) Passenger waiting strategies for overlapping bus route. Transportation Science 18, 207-230.

Nguyen S., Pallottino S. (1988) Equilibrium traffic assignment for large scale transit networks. European Journal of Operational Research 37, 176-186.

Nguyen S., Pallottino S., Gendreau M. (1998) Implicit enumeration of hyperpaths in a logit model for transit networks. Transportation Science 32, 54-64.

Nökel K., Wekeck S. (2009) Boarding and alighting in frequency-based transit assignment. Transportation Research Record 2111, 60-67.

Nuzzolo A., Crisalli U., Comi A., Rosati L. (2013) An advanced pre-trip planner with personalized information on transit networks with ATIS. Proceedings of 16th International IEEE Conference on Intelligent Transportation Systems, The Hague, 2146-2151.

PTV AG (2003) VISUM 9.0 Manual, available from PTV Group, Karlsruhe.

Ren H., Gao Z., Lam W.H.K., Long J. (2009) Assessing the benefits of integrated en-route transit information systems and time-varying transit pricing systems in a congested transit network. Transportation Planning and Technology 32, 215-237.

Shimamoto H., Kurauchi F., Iida Y. (2005) Evaluation on effect of arrival time information provision using transit assignment model. International Journal of ITS Research 3, 11-18.

Spiess H., Florian M. (1989) Optimal strategies: A new assignment model for transit networks, Transportation Research B 23, 83-102.

7.2 Discomfort: seating and crowding

Authors: Jan-Dirk Schmöcker, Guido Gentile

The previous section explained how route choice strategies depend on the information on vehicle arrivals available to passengers during their wait at stops as well as on the service regularity (headway distributions). However, route costs and hence assignment results can be further influenced by vehicle capacities in terms of discomfort and queuing, as described in this section and the following one.

In particular, this section presents equilibrium models with no strict capacity constraints where the congestion derives from discomfort. The aim is indeed to reproduce the following phenomena in the context of transit assignment:

- in-vehicle crowding;
- at stop crowding;
- seat capacity.

As explained in Section 5.1.2, discomfort is not only perceived as on-board crowding, but also as on-platform densities, as well as in specific pedestrian elements for circulation inside stations (e.g., stairs connecting to the platform).

The value of being able to sit while travelling is well documented in the behavioural literature. At higher densities, the more standing passengers are packed, the more likely they perceive this as uncomfortable and stressful. Hence, passengers will be willing to re-route on longer but less-congested routes.

Thus, discomfort for passengers on-board increases with in-vehicle loading, which can be measured by the saturation rate, i.e., number of passengers on board divided by the vehicle capacity. This is directly related to the seat availability and the density of standing passengers:

- for low/medium saturation rates, crowding discomfort is due to the lower probability of getting a seat (seat unavailability);
- for medium/high saturation rates, crowding discomfort is due to the closer physical distance with other passengers (privacy violation);
- for higher saturation rates, crowding discomfort is due to physical contact and pressure of other passengers (squeezing).

This section is structured into 4 main parts. In Section 7.2.1, we limit our attention to privacy violation and squeezing, which require just the specification of the functional form for the crowding coefficient. In Section 7.2.2, we address the essence of the seat availability modelling, that is how to describe the allocation mechanism taking into account the priority rules among different passenger flows, such as the chronological order of operations at stops, by amending the topology of the network model presented in sections 6.2.2 and 6.3.1. In Section 7.2.3, we describe the formulation of the equilibrium problem. Finally, in Section 7.2.1.1, we provide some numerical examples.

7.2.1 Overcrowding congestion

For the sake of simplicity, we refer here to the case of frequency-based assignment on static networks presented in Section 6.2, although the proposed formulation can be immediately extended to the case of schedule-based models on diachronic graphs presented in Section 6.3.

The task at hand is to specify the functional expression of the crowding discomfort coefficient $\gamma_{\ell sg}^{crowd}$ of segment $s \in S_\ell - S_\ell^+$ of line $\ell \in L$ for user class $g \in G$ introduced in Section 6.2.3. Possibly the most simple method to describe the discomfort caused by overcrowding is by introducing a multiplication factor to the running travel time for all passengers on-board, as in Equation (6.67.d). This can be done with a BPR-type function, similar to the cost functions considering the impact of road congestion to travel times:

$$\gamma_{\ell sg}^{crowd}\left(q_a\right) = 1 + \alpha_g^{crowd} \cdot \left(\frac{q_a}{\kappa_\ell^{veh} \cdot f_{\ell s}} \right)^{\beta_\ell^{crowd}} \quad , \quad \forall a = \left(N_{\ell s}^{dep}, N_{\ell s^{+\ell}}^{arr}\right) \in A^{run}, \forall g \in G \ . \tag{7.50}$$

where:

- q_a is the volume of passenger on the running arc a;
- $f_{\ell s}$ is the frequency of line ℓ at stop s, i.e., the flow of vehicles serving the line;
- $\kappa_\ell^{veh} \cdot f_{\ell s}$ is the capacity of line ℓ at stop s (the flow of vehicles multiplied by their individual capacity);
- $q_a / (\kappa_\ell^{veh} \cdot f_{\ell s})$ is the saturation rate (or occupancy) of vehicles on the line segment s;
- α_g^{crowd} and β_ℓ^{crowd} are the BPR coefficient and exponent for overcrowding congestion perceived by passengers of class g travelling on-board line ℓ (typical values are $\alpha_g^{crowd} = 1$ and $\beta_\ell^{crowd} = 2$).

The example in Section 4.5.4 provides more insights on the practical relevance of vehicle occupancy level in revealed passenger behavioural and willingness-to-pay.

The saturation rate can also be interpreted as the number of passengers $q_a / f_{\ell s}$ on board a single vehicle serving the line (the flow of passengers divided by the flow of vehicles) divided by its capacity κ_ℓ^{veh}.

Also the discomfort caused by overcrowding at the stop can be modelled by introducing a multiplication factor to the wait time, as in Equation (6.67.f). This can be done again with a BPR-type function which specifies the expression of the crowding discomfort coefficient γ_{sg}^{crowd} of stop $s \in S$ for user class $g \in G$ introduced in Section 6.2.3:

$$\gamma_{sg}^{crowd}(\mathbf{q}_A) = 1 + \alpha_g^{crowd} \cdot \left(\frac{\sum_{b \in s^+} q_b \cdot t_b}{\kappa_s^{stop}} \right)^{\beta_s^{crowd}} \quad , \quad \forall s = S, \forall g \in G \ , \tag{7.51}$$

where:

- the crowding discomfort depends on several arc flows, and thus in principle on the flow vector $\mathbf{q}_A$,
- the sum of the passenger flow q_b for each waiting arc b exiting from the stop s multiplied by the its expected time t_b yields the expected number of passengers waiting at the stop;
- κ_s^{stop} is the capacity stop of stop s;
- the ratio of the above two numbers yields the saturation rate of stop s;
- α_g^{crowd} and β_s^{crowd} are the BPR coefficient and exponent for overcrowding congestion perceived by passengers of class g waiting at stop s (typical values are $\alpha_g^{crowd} = 1$ and $\beta_s^{crowd} = 2$).

The formulas just introduced can be immediately extended to the case of schedule-based models based on diachronic graphs under the consideration that the arc loads represent in this case a number of passengers, which can directly be compared with the vehicle and stop capacity, respectively:

$$\gamma_{rsg}^{crowd}(q_a) = 1 + \alpha_g^{crowd} \cdot \left(\frac{q_a}{\kappa_\ell^{veh}} \right)^{\beta_\ell^{crowd}} \quad , \quad \boldsymbol{a} = \left(N_{rs}^{dep}, N_{rs^{+\ell}}^{arr} \right) \in A^{run} \ , \tag{7.52}$$

$$\gamma_{sgt}^{crowd}(q_a) = 1 + \alpha_g^{crowd} \cdot \left(\frac{q_a}{\kappa_s^{stop}} \right)^{\beta_s^{crowd}} \quad , \quad \forall \boldsymbol{a} = \left((s,t), (s,t+1) \right) \in A^{wait} \ . \tag{7.53}$$

7.2.1.1 Applications to the example network

The arc performance model of Equation (7.50) to reproduce crowding congestion is here applied jointly to the classical route choice model of Optimal Strategies presented in Section 7.1.1. The resulting equilibrium problem has been solved for the example network of Section 5.13 through the MSA, although better performing algorithms are available.

Given the dimension of the vehicles serving the lines (80 pax), the line capacities are way higher than the flows on the running arcs assigned to the shortest hyperpaths, as shown in Table 7.7. However, some congestion emerges and the discomfort on-board is slightly higher than the mere cost of travel time. Does the equilibrium mechanism actually change the flow pattern? Not necessarily.

Indeed, only if the cost on the uncongested shortest route (hyperpath, in this case) of a given O-D pair increases so much as to be higher than that of an alternative route we then observe some shift of flows.

Moreover, in the case of strategic behaviour, more paths are actually used by the same O-D pair (passengers board on the first arriving attractive lines), but their shares does not depend on cost which suffer congestion, but rather on frequencies which do not suffer congestion (at least in this basic model). As a consequence, the arc flows of Table 7.7 are exactly the same of those resulting in the numerical example of Section 7.1.1.6.

Table 7.7. Line volumes (pax/h) due to crowding congestion.

		Segment					
		1→2	2→3	3→4	1→4	2→1	production
line	[pax/h]	3.5 km	3 km	3 km	10 km	3.5 km	[pax*km/h]
1- Red	800				150		1500
2 - Green	800	150	407				1746
3 - Maroon	320		103	211			941
4 - Black	1600			539			1618
walk	INF	0				0	0

Let's now assume that the vehicles serving Line 2 and 3 are substituted by small vehicles with a limited on-board capacity of 8 pax. In this case the two lines get very congested and some passenger must divert to alternatives routes to ensure equilibrium. In particular, all users from Stop 1 will consider only Line 1. The results of the assignment are reported in Table 7.8

Due to the line share mechanism based on frequencies, the proportion of passengers boarding Line 2 and 3 at Stop 2 is unchanged wrt the previous case, although the costs for the two lines are different as the volume on-board (on Line 2 there are not anymore the 150 passengers that boarded at Stop 1).

We can then conclude that the transit assignment equilibrium based on Optimal Strategies is somehow more stable than that without strategic behaviour.

Note that, as expected, the capacity constraint is not satisfied by the equilibrium formulation with crowding congestion. Despite the presence of alternative routes (e.g. walking to Stop 1) based on the BPR model of Equation (7.50) the passengers departing from Stop 2 prefer to suffer a very high discomfort (there are around three times as much passengers on-board than the vehicle capacity). This is also due to the fact that there is no advantage in boarding Line 2 form Stop 1 instead that from Stop 2, since the discomfort on-board is suffered by all passengers; the seating capacity model presented in Section 7.2.2 would instead ensure priority for passengers already on-board).

Table 7.8. Line volumes (pax/h) due to crowding congestion with small vehicles for Lines 2 and 3.

		Segment					
		1→2	2→3	3→4	1→4	2→1	production
line	[pax/h]	3.5 km	3 km	3 km	10 km	3.5 km	[pax*km/h]
1- Red	800				300		2999
2 - Green	80	0	257				772
3 - Maroon	32		103	57			479
4 - Black	1600			543			1630
walk	INF	0				0	0

7.2.2 Seat availability

The main disadvantage of the approach proposed in the previous section to reproduce on-board discomfort is that the resulting model equally penalises all passengers on-board, independently of when they boarded the line vehicle. Therefore, it is not reflected that passengers who boarded earlier have a higher chance of

obtaining a seat and of not experiencing the (whole) disutility caused by over-crowding.

The differentiation of the discomfort experienced by sitting versus standing passengers is accomplished in this section by explicitly modelling the limited seat availability and the random process of passengers finding a seat. A main difficulty for this is though the representation of initially standing passengers who might be able to find a seat during their journey thanks to seated passengers that alight at stops, considering that the former have a priority over the newly boarding passengers. In general, this leads to a network model (with more nodes and arcs) and to an equilibrium model (with asymmetric cost functions) that is more complex than that introduced to represent standard overcrowding congestion.

In particular, it is here assumed that passengers who are already on board have priority over the newly boarding passengers in two ways:

- passengers arriving at a stop sitting are guaranteed a seat for the next line segment, so that they either alight or remain seated;
- passengers arriving at a stop standing who do not alight have priority over the passengers newly boarding, i.e., these passengers have a prior chance to occupy any seat that might become vacant thanks to alighting passengers.

This leads to a new network description based on hyperarcs (see Section 6.1.3) and to the introduction of "fail-to-sit" probabilities, as described in the following.

A different specialization of line nodes (see Figure 7.4) with respect to that proposed in Figure 6.6 is required, where each line layer is duplicated to represent the service for seating passengers and for standing passengers. Moreover, for each line, two additional nodes are introduced to represent placing:

- a *board placing node* to consolidate at each stop the waiting phase for both types of boarding passengers; who succeeds in getting a seat takes the *seat placing arc*, and who will have to stand takes the *stand placing arc*.
- a *stand placing node* that splits the dwelling arc to consolidate the flows of standing passengers who decide to remain on-board; who succeeds in getting a seat at the stop takes the *switch seating arc*, and who will have to stand takes the *keep standing arc.*

Therefore, in total, six nodes for each stop of line $\ell \in L$ are introduced:

- the *seating arrival node* $N_{\ell s}^{a\text{-}seat} \in N_\ell$, $\forall s \in S_\ell\text{-}S_\ell^-$;
- the *seating departure node* $N_{\ell s}^{d\text{-}seat} \in N_\ell$, $\forall s \in S_\ell\text{-}S_\ell^+$.
- the *standing arrival node* $N_{\ell s}^{a\text{-}stand} \in N_\ell$, $\forall s \in S_\ell\text{-}S_\ell^-$;
- the *standing departure node* $N_{\ell s}^{d\text{-}stand} \in N_\ell$, $\forall s \in S_\ell\text{-}S_\ell^+$;
- the *board placing node* $N_{\ell s}^{p\text{-}board} \in N_\ell$, $\forall s \in S_\ell\text{-}S_\ell^+$.
- the *stand placing node* $N_{\ell s}^{p\text{-}stand} \in N_\ell$, $\forall s \in S_\ell\text{-}S_\ell^+$.

The network is then built up by introducing the following types of arcs and hyperarcs:

- the *pedestrian arcs* $A^{walk} = E^{walk}$;
- the *stop arcs* $A^{stop} = \{(B_s^{stop}, s): \forall s \in S\} \cup \{(s, B_s^{stop}): \forall s \in S\}$;
- the *seat running arcs* $A^{r\text{-}seat} = \{(N_{\ell s}^{d\text{-}seat}, N_{\ell s[+\ell]}^{a\text{-}seat}): \forall s \in S_\ell\text{-}S_\ell^+, \forall \ell \in L\}$;
- the *seat placing arcs* $A^{p\text{-}seat} = \{(N_{\ell s}^{p\text{-}board}, N_{\ell s}^{d\text{-}seat}): \forall s \in S_\ell\text{-}S_\ell^+, \forall \ell \in L\}$;
- the *seat dwelling arcs* $A^{d\text{-}seat} = \{(N_{\ell s}^{a\text{-}seat}, N_{\ell s}^{d\text{-}seat}): \forall s \in S_\ell\text{-}S_\ell^-\text{-}S_\ell^+, \forall \ell \in L\}$;
- the *seat alighting arcs* $A^{a\text{-}seat} = \{(N_{\ell s}^{a\text{-}seat}, s): \forall s \in S_\ell\text{-}S_\ell^-, \forall \ell \in L\}$.
- the *stand running arcs* $A^{r\text{-}stand} = \{(N_{\ell s}^{d\text{-}stand}, N_{\ell s[+\ell]}^{a\text{-}stand}): \forall s \in S_\ell\text{-}S_\ell^+, \forall \ell \in L\}$;
- the *stand dwelling arcs* $A^{d\text{-}stand} = \{(N_{\ell s}^{a\text{-}stand}, N_{\ell s}^{p\text{-}stand}): \forall s \in S_\ell\text{-}S_\ell^-\text{-}S_\ell^+, \forall \ell \in L\}$;
- the *stand placing arcs* $A^{p\text{-}stand} = \{(N_{\ell s}^{p\text{-}board}, N_{\ell s}^{d\text{-}stand}): \forall s \in S_\ell\text{-}S_\ell^+, \forall \ell \in L\}$;
- the *stand alighting arcs* $A^{a\text{-}stand} = \{(N_{\ell s}^{a\text{-}stand}, s): \forall s \in S_\ell\text{-}S_\ell^-, \forall \ell \in L\}$;
- the *waiting arcs* $A^{wait} = \{(s, N_{\ell s}^{p\text{-}board}): \forall s \in S_\ell\text{-}S_\ell^+, \forall \ell \in L\}$;
- the *switch seating arcs* $A^{p\text{-}switch} = \{(N_{\ell s}^{p\text{-}stand}, N_{\ell s}^{d\text{-}seat}): \forall s \in S_\ell\text{-}S_\ell^+, \forall \ell \in L\}$;
- the *keep standing arcs* $A^{p\text{-}keep} = \{(N_{\ell s}^{p\text{-}stand}, N_{\ell s}^{d\text{-}stand}): \forall s \in S_\ell\text{-}S_\ell^+, \forall \ell \in L\}$;
- the *boarding hyperarcs* $H^{board} = \{\{(N_{\ell s}^{p\text{-}board}, N_{\ell s}^{d\text{-}seat}), (N_{\ell s}^{p\text{-}board}, N_{\ell s}^{d\text{-}stand})\}: \forall s \in S_\ell\text{-}S_\ell^+, \forall \ell \in L\}$;
- the *dwelling hyperarcs* $H^{dwell} = \{\{(N_{\ell s}^{p\text{-}stand}, N_{\ell s}^{d\text{-}seat}), (N_{\ell s}^{p\text{-}stand}, N_{\ell s}^{d\text{-}stand})\}: \forall s \in S_\ell\text{-}S_\ell^+, \forall \ell \in L\}$.

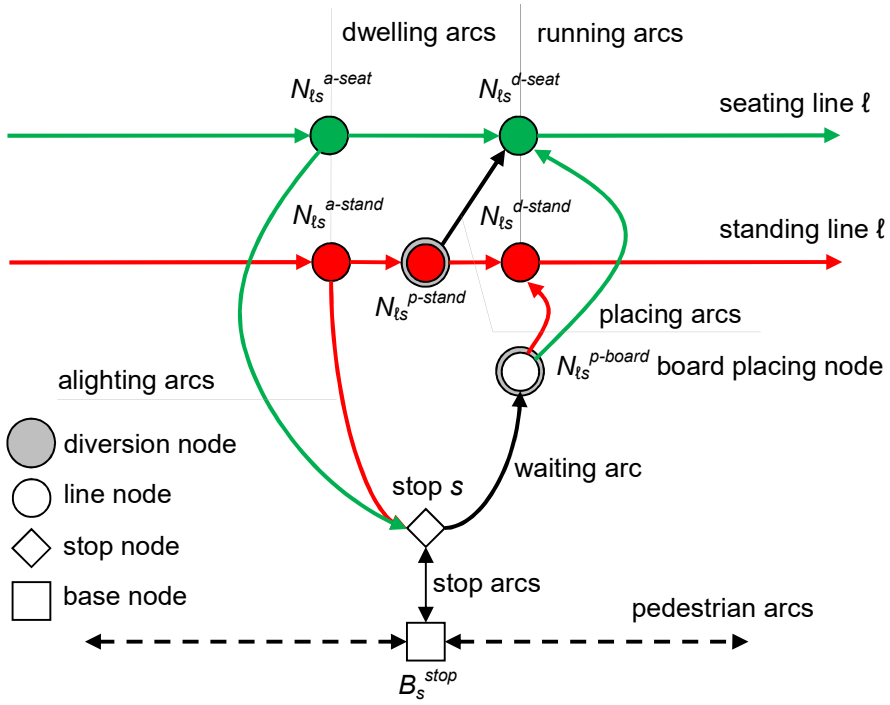

Figure 7.4. Network topology to represent seat availability and priority.

The diversion nodes are here the placing nodes: $N^{div} = N^{p\text{-}board} \cup N^{p\text{-}stand}$. Two different hyperarcs are introduced for each line stop to represent the probabilistic event of seating or standing:

- a boarding hyperarc, for newly boarding passengers;
- a dwelling hyperarc, for standing passengers who have priority over the newly boarding passengers in getting the seats left by alighting passengers.

With respect to the performance model presented in 6.2.3, the following changes:

- placing arcs ($A^{p\text{-}stand} \cup A^{p\text{-}seat} \cup A^{p\ switch} \cup A^{p\text{-}keep}$) and hyperarcs ($H^{board} \cup H^{dwell}$) are dummy (null cost);
- equations (6.67.c) apply to all dwelling arcs and branches;
- the crowding discomfort coefficient (7.50) applies only to the stand running arcs where the vehicle capacity κ_ℓ^{veh} is replaced with the standing capacity κ_ℓ^{stand}, while for seat running arcs it assumes a constant (lower) value $\gamma_{\ell g}^{seat}$.

Instead, a new model must be specified to provide the hyperarc diversion probabilities. Under the main assumption that all competing passengers, possibly belonging to different classes, have (on average) the same motivation in chasing any free seats, the *sit probability* is simply given by the ratio between supply and demand of seats; the probability is anyhow bounded between 0 and 1.

For the dwelling hyperarc, the supply is given by the seating capacity of the vehicle serving the line multiplied by the frequency at the stop (i.e., the flow of vehicles) reduced by the dwelling passengers that are already seated; the demand is given by the passengers that arrive at the stop standing on-board, reduced of the share of those who alight:

$$p_{a'/\breve{a}} = Mid\left(0, \frac{\kappa_\ell^{seat} \cdot f_{\ell s} - q_b}{q_d = q_{a'} + q_{a''}}, 1\right), \quad p_{a''/\breve{a}} = 1 - p_{a'/\breve{a}}$$
$$\forall \breve{a} = \{a', a''\} \in H^{dwell}$$
$$a' = \left(N_{\ell s}^{p-stand}, N_{\ell s}^{d-seat}\right) \in A^{p-switch}$$
$$a'' = \left(N_{\ell s}^{p-stand}, N_{\ell s}^{d-stand}\right) \in A^{p-keep}$$
$$d = \left(N_{\ell s}^{a-stand}, N_{\ell s}^{p-stand}\right) \in A^{d-stand}$$
$$b = \left(N_{\ell s}^{a-seat}, N_{\ell s}^{d-seat}\right) \in A^{d-seat} \tag{7.54}$$

For the placing hyperarc, the supply is given by the seating capacity of the vehicle serving the line multiplied by the frequency at the stop reduced by the dwelling passengers that are already seated and further reduced by the switching passengers; the demand is given by the boarding passengers:

$$p_{a'/\breve{a}} = Mid\left(0, \frac{\kappa_\ell^{seat} \cdot f_{\ell s} - q_b - q_e}{q_d = q_{a'} + q_{a''}}, 1\right), \quad p_{a''/\breve{a}} = 1 - p_{a'/\breve{a}}$$
$$\forall \breve{a} = \{a', a''\} \in H^{board}$$
$$a' = \left(N_{\ell s}^{p-board}, N_{\ell s}^{d-seat}\right) \in A^{p-seat}$$
$$a'' = \left(N_{\ell s}^{p-board}, N_{\ell s}^{d-stand}\right) \in A^{p-stand}$$
$$b = \left(N_{\ell s}^{a-seat}, N_{\ell s}^{d-seat}\right) \in A^{d-seat}$$
$$e = \left(N_{\ell s}^{p-stand}, N_{\ell s}^{d-seat}\right) \in A^{p-switch}$$
$$d = \left(s, N_{\ell s}^{p-board}\right) \in A^{wait} \tag{7.55}$$

As mentioned already, for both types of hyperarcs it is assumed: $t_{\breve{a}} = 0$. These equations allow to apply the sequential model presented in Section 6.1.5.

The presence of fail-to-sit probabilities provided by Equations (7.54) and (7.55) ensures that the seating capacity of the vehicle is never exceeded.

For what concerns route choice, equation (6.28) ensures that the expected cost for reaching the destination when boarding a given line results from the average of seating and standing weighted with the sit and fail to sit probability, respectively. In turn, the cost of standing includes the possibility of seating at next stops.

Noteworthy, this model implies that the alighting decision is not predetermined anymore: passengers who have obtained a seat might prefer to transfer later, whereas standing passengers are more likely to transfer earlier. The fact that the diversion probability of the seat alighting arc is different than that of the stand alighting arc can be indeed well reflected by the proposed network structure, as the expected costs to reach a destination of the seat line nodes are typically lower than those of the corresponding stand line nodes.

However, in the proposed model it is not possible to check if a seat becomes available for certain after alighting of other passengers, and then decide whether to alight at the current stop, as this may be not possible or too stressful for a passenger. If this feature is instead desirable, it requires some modification of the network.

There are two major differences between the hyperarcs just introduced for modelling seating and those for modelling attractive line sets (introduced in Section 7.1):

- for seating, there is no choice to be made – the probabilities are determined by a physical random event, in fact there is just one exiting hyperarc;
- the resulting diversion probabilities (no choice probabilities) depend (asymmetrically) on passenger flows – while the attractive line set depends solely on given headways and remaining costs (which may depend indirectly on flows), here the assignment model is necessarily congested, leading to an equilibrium problem.

The scheme of Figure 6.3 can be applied considering the sequential model based on hyperarcs of Section 6.1.5. In particular, the fail-to-sit probabilities are computed by the performance model as an additional cost function, as they depend on the arc flows. Consistency is found only at equilibrium.

The extension to schedule-based models of the proposed approach is straightforward and requires just to apply the duplication of the line sub-network as in Figure 7.4 to each single run of the diachronic graph introduced in Section 6.3.1.

The example in Section 4.5.4 provides more insights on the practical relevance of seat availability in

revealed passenger behavioural and willingness-to-pay.

Although, crowding and seating have been presented separately, the two concepts can easily be considered simultaneously in the same model. This is as simple as including the BPR-type discomfort coefficient of Equation (7.50) in the expanded seat-availability network of Figure 7.4. However, travellers that are seated perceive crowding very differently to those standing; for the sake of simplicity, we can assume that for the two kinds of running arcs (seating and standing) there are two different line discomfort coefficient $\gamma_{\ell g}^{line}$, denoted $\gamma_{\ell g}^{seat}$ and $\gamma_{\ell g}^{stand}$, respectively, and that the crowding discomfort coefficient $\gamma_{\ell sg}^{crowd}$ affected by congestion applies only to the latter.

7.2.3 Static equilibrium models with discomfort cost functions

Discomfort congestion due to on-board overcrowding yielded by equation (7.50) is separable, because the cost of the running arc depends on the flows of the same arc only. The resulting equilibrium problem is then a rather simple extension of classical traffic assignment models on road networks. This will hence lead to iterative methods that relocate passengers away from crowded line (or run) segments until an equilibrium solution is reached, such as the fixed-point algorithm presented in Section 6.1.8.

In the case of deterministic behaviour where passenger choose a route with minimum cost, if only one class of users is considered, the transit assignment problem can be formulated as an equivalent minimization program with unknown flows q_{ad} of users traveling on arc $a \in A$ to destination $d \in D$, whose objective function is the well-known sum of arc cost integrals (Beckmann, 1956):

$$Min\left(\sum_{a \in A} \int_0^{q_a} c_a(q) \cdot dq \right), \tag{7.56.a}$$

subject to the consistency (node flow conservation) and non-negativity constraints:

$$\sum_{a \in i^+} q_{ad} - \sum_{a \in i^-} q_{ad} = \begin{cases} 0, & \forall i \in N / O / \{d\} \\ d_{id}, & \forall i \in O / \{d\} \\ -\sum_{o \in O} q_{od}, & i = d \end{cases}, \quad \forall d \in D, \tag{7.56.b}$$

$$q_{ad} \geq 0, \quad \forall a \in A, \quad \forall d \in D, \tag{7.56.c}$$

$$q_{ad} = \sum_{d \in D} q_{ad}, \quad \forall a \in A. \tag{7.56.d}$$

This leads to a convex optimisation problem in terms of arc flows that may be efficiently solved with several iterative methods, ranging from Frank Wolfe to Gradient Projection (Bertsekas, 1999). Most of such equilibrium algorithms involve the following cyclic sequence of steps:

0. start from a feasible flow pattern which satisfies non-negativity and consistency constraints;
1. calculate the new performance pattern through the arc cost functions at the current flow iterate;
2. determine the search direction, which implies to apply the route choice model based on the new costs and to carry out the consequent flow propagation of travel demand;
3. find a step in the search direction such that the new iterate of flows possibly leads to an improvement of the objective function;
4. check the distance to equilibrium (for example through the relative gap); if it does not meet the stop criteria then go back to step 1.

In gradient projection algorithms (including bush-based methods, such as LUCE (Gentile, 2014) and Algorithm B (Dial, 2006) the route choice probabilities (path-based or arc-based) obtained in step 2 are not a direct blind application of the route choice model but rather try to incorporate the consequences on the equilibrium of such choices.

As further well known from the road assignment case, multiple equilibria may though be possible in terms of route flows (and arc flows, if multiple classes are considered), while the uniqueness of equilibrium is ensured only in terms of arc volumes. However, uniqueness requires (as a sufficient condition) the strict monotonicity of the arc cost function, while here the only cost actually depending on flows is that of running arcs through

the crowding discomfort coefficient. Therefore, uniqueness does not hold true for pedestrian arcs that are not affected by congestion.

Another possible approach (directly derived from road traffic assignment) to the formulation of equilibrium problems with overcrowding discomfort on transit networks is the interpretation of the Lagrangian multipliers of a mathematical program with explicit capacity constraints as the additional cost on running arcs due to congestion ($\gamma_{\ell sg}^{crowd} - 1$). If the arc flow is below the line capacity, then the crowding discomfort coefficient is one; if the flow equals the capacity constraint, then the additional cost of discomfort can be positive and the crowding coefficient can be higher than one:

$$\begin{cases} \gamma_{\ell sg}^{crowd} = 1, & \text{if } q_a < \kappa_{\ell}^{veh} \cdot f_{\ell s} \\ \gamma_{\ell sg}^{crowd} \geq 1, & \text{if } q_a = \kappa_{\ell}^{veh} \cdot f_{\ell s} \end{cases} \quad , \quad \forall a = \left(N_{\ell s}^{dep}, N_{\ell s^+ \ell}^{arr} \right) \in A^{run}, \forall g \in G \ . \tag{7.57}$$

Lam et al. (1999) address the transit assignment problem with strict capacity constraints for stochastic (logit) route choice. The resulting model is basically an extension of Bell (1995) so solve equilibrium problems on road networks.

Other methods to incorporate capacity constraints will be presented in the next Section 7.3.

To address the combination of overcrowding congestion with the route choice model based on optimal strategies discussed in Section 7.1.1, Problem (7.13) is suitably extended (Spiess and Florian, 1989). The objective function (7.56) of the equivalent minimization program is changed to:

$$\sum_{a \in A} \int_0^{q_a} c_a(q) \cdot dq + \sum_{d \in N} \sum_{i \in S} \omega_{id} \ , \tag{7.58}$$

where the additional unknowns ω_{id} represent the total wait time at stop $i \in S$ of passengers traveling towards destination $d \in D$; moreover, the following constraints involving the frequency f_a of each line associated with a waiting arc $a \in i^+$ are to be considered:

$$q_{ad} \leq f_a \cdot \omega_{id}, \quad \forall a \in i^+, \quad \forall i \in S, \quad \forall d \in D \ . \tag{7.58.a}$$

In this case, the above step 2 requires to integrate the hyperpath-based algorithm presented in Table 7.1 for the uncongested case, leading to efficient methods (Wu et al. 1994).

Discomfort congestion due to overcrowding at stops yielded by equation (7.51) is non-separable, because the cost of each waiting arc depends on the flows of all waiting arcs at the stop; moreover, the Jacobian of the arc performance function is not symmetric. The resulting equilibrium problem cannot then be formulated as a non-linear optimization program like those presented in this section; to this end we can use a variational inequality problem or a fixed-point problem, as in Bellei et al. (2000). The same is true for the seat availability model based on hyperpaths presented in Section 7.2.2, where the fail to sit probabilities (7.54) and (7.55) depend on several arc flows at the stop.

Specifying flow-dependent arc costs and diversion probabilities is not just applied within static frequency-based models, but also within schedule-based models on space-time networks. The extension of both discomfort congestion models for overcrowding and seat availability to the latter framework is rather straightforward and does not merit particular considerations. The same is true for the equilibrium models presented in this section: there is no substantial difference from a mathematical point of view between a static frequency-based assignment and a schedule-based assignment on space-time networks.

7.2.4 Reference notes and concluding remarks

7.2.4.1 Crowding congestion

Congestion functions for the representation of discomfort due to overcrowding on-board and at stops were proposed by several authors. References to the resulting equilibrium models have been provided already in Section 7.2.3.

Practitioners in Tokyo (Morichi et al., 2001; Kato et al., 2010), where crowding discomfort is a severe problem, rely on a disaggregate assignment model based on discrete choice theory where choice sets of paths are created a priori and passengers are then split between routes based on probit or logit probabilities

in the context of a stochastic user equilibrium. In their application, a congestion model analogous to (7.50) is used with a fixed exponent $\beta_{\ell g}^{crowd} = 2$. They found that the parameter $\alpha_{\ell g}^{crowd}$ associated with crowding congestion is significant in all choice models and that its evaluation depends mainly on the trip purpose.

7.2.4.2 Seating congestion

Tian et al. (2007) describe a schedule-based model that considers passenger congestion effects including seat availability. They formulate an equilibrium model for a many-to-one network applicable for the morning commute into the city centre of large metropolitan areas. Reducing the model to a many-to-one network has the advantage that it avoids the problem of standing passengers being able to find a seat during the journey due to alighting passengers. Using a schedule-based model allows to represent explicitly the optimal departure time. The paper illustrates that at equilibrium some long distance commuters will travel before and some will travel after the peak, while the spread in optimal departure times increases the longer the travel distance, as the travel costs of standing gain in importance compared to the early or late arrival penalties.

Sumalee et al. (2009) have developed a stochastic assignment model on transit networks that explicitly considers the effect of seat availability on route choice as well as departure time choice. They consider priorities of on-board passengers over newly boarding passengers and further assume that passengers who are travelling for a longer distance and passengers who have stood for a longer time have a higher motivation in chasing any free seats. This assumption introduces a further complexity in the model as 'the past' has to be considered in modelling travellers' behaviour at each decision point, while the probability of getting a seat is not simply given by the ratio of supply and demand. Indeed, this kind of seat allocation is solved by a simulation approach.

Leurent (2012) and Schmöcker et al. (2011) have suggested two frequency-based approaches to consider seating capacity and standing discomfort. Compared to Sumalee et al. (2009) both models are simpler in that they do not consider individual passengers' desire to sit depending on their journey length and standing time, which avoids the introduction of a simulation approach. The representation of priorities among passenger flows and the reflection that seated passengers do not suffer from crowding effects is the main focus for both models. The main idea is the introduction of "fail-to-sit" probabilities.

In Schmöcker et al. (2011) this is achieved through the introduction of second layer of nodes and arcs for each line representing the seated service. The seat availability model presented in Section 7.2.2 founds its roots in this work.

In Leurent (2012) this is achieved through the introduction of line legs that represent each combination of boarding and alighting nodes for standing and seating, which are used in route choice and flow propagation. Leurent and Liu (2009) applied the latter approach to the Paris network and provided further evidence that considering seat availability can indeed have a significant effect on line loadings (with changes by up to 30%) and on the overall passenger cost.

7.2.4.3 List of reference

Beckmann M., McGuire C., Winston C. (1956) Studies in the economics of transportation. New Haven, Connecticut, Yale University Press.

Bell M.G.H. (1995) Stochastic user equilibrium assignment in networks with queues. Transportation Research B 29, 125-137.

Bellei G., Gentile G., Papola N. (2000) Transit assignment with variable frequencies and congestion effects. Proceedings of the 8th Meeting of the EURO Working Group on Transportation, Rome, Italy.

Bertsekas D.P. (1999) Nonlinear programming. Second Edition. Athena Scientific, Belmont, Massachusetts.

Dial R. (2006) A path-based user-equilibrium traffic assignment algorithm that obviates path storage and enumeration. Transportation Research B 40, 917-936.

Gentile G. (2014) Local User Cost Equilibrium: a bush-based algorithm for traffic assignment. Transportmetrica A 10, 15-54.

Kato H., Kaneko Y., Inoue M. (2010) Comparative analysis of transit assignment: evidence from urban railway system in the Tokyo Metropolitan Area. Transportation 37, 775-799.

Lam W.H.K., Gao Z.Y., Chan K.S. Yang N. (1999) A stochastic user equilibrium assignment model for congested transit networks. Transportation Research B 33, 351-368.

Leurent F. (2012) On seat capacity in traffic assignment to a transit network. Journal of Advanced Transportation 46, 112-138.

Leurent F., Liu K. (2009) On seat congestion, passenger comfort and route choice in urban transit: A network equilibrium assignment model with application to Paris. Proceedings of the 88th Annual Transportation Research Board Meeting, Washington DC.

Morichi S., Iwakura S., Morishige S., Itoh M., Hayasaki S. (2001) Tokyo metropolitan rail network long-range plan for the 21st century. Proceedings of the 80th Annual Meeting of Transportation Research Board, Washington DC.

Schmöcker J.-D., Fonzone A., Shimamoto H., Kurauchi F., Bell M.G.H. (2011) Frequency-based transit assignment considering seat capacities. Transportation Research B 45, 392-408.

Spiess H., Florian M. (1989) Optimal strategies: a new assignment model for transit networks. Transportation Research B, 23, 83-102.

Sumalee A., Tan Z.J., Lam W.H.K. (2009) Dynamic stochastic transit assignment with explicit seat allocation model. Transportation Research B 43, 895-912.

Tian Q., Huang H.-J., Yang H. (2007) Commuting equilibria on a mass transit system with capacity constraints. Proceedings of the 17th International Symposium on Transportation and Traffic Theory (ISTTT), ed.s R. Allsop, M.G.H. Bell, and B.G. Heydecker, Elsevier, London, 261-384.

7.3 Passenger queuing

Authors: Guido Gentile, Valentina Trozzi

The major limitation of the models that represent vehicle capacity as a discomfort due to overcrowding (in contrast to that caused by seat unavailability) lays in the fact that it is not possible to reproduce the priority of on-board passengers with respect to those waiting at the stop. Thus, all passengers suffer the same cost, as if everybody alighted the vehicle at each stop and attempted re-boarding it. In reality, when overcrowding is very heavy and the crush capacity is reached on-board, no further passenger is able to get on the vehicle. Then, an oversaturation queue of passengers waiting at the stop is formed. But clearly this phenomenon does not affect the passengers that are already on-board.

This type of severe transit congestion due to vehicle capacity affects many transport systems, like busways and railways both in developing and developed countries, mostly in metropolitan contexts, and is lately receiving increasing attention by modellers and operators.

This chapter presents equilibrium models with capacity constraints and is devoted to the modelling of the following phenomena in the context of transit assignment:

- oversaturation queues of passengers waiting at stops,
- mingling and fail-to-board probabilities vs. FIFO and service bottlenecks.

7.3.1 Queuing congestion

Passenger queuing occurs when a vehicle departing from a stop $s \in S_\ell - S_\ell^+$ has not enough remaining capacity to accommodate on-board all travellers that are waiting for that line $\ell \in L$ (possibly among other lines of the attractive set). More specifically, a residual queue remains unserved at the stop when the flow of passengers wishing to board (arc b) is higher than the capacity of the line (given by the capacity of the vehicle κ_ℓ^{veh} multiplied by the frequency of the line at that stop $f_{\ell s}$) reduced by the flow of dwelling passengers (arc d) that are already on-board:

$$q_b > \kappa_\ell^{veh} \cdot f_{\ell s} - q_d \quad d = \left(N_{\ell s}^{arr}, N_{\ell s}^{dep}\right) \in A^{dwell},\ b = \left(s, N_{\ell s}^{dep}\right) \in A^{wait} \ . \tag{7.59}$$

If the above condition occurs, some passengers are not able to board the arriving carrier serving the line and will have to wait for a next departure. The additional wait time due to the lack of space on-board increases, on average, not only with the number of passengers wishing to board, but also with the number of dwelling passengers that are already on-board. The latter clearly have a priority on the former with respect to the occupation of the available vehicle space and are not affected by passengers attempting to board (unless discomfort is considered). Queuing congestion is thus patently non-separable.

It is important to distinguish two different queuing phenomena that occur at stops:

- the queue formed by passengers that are waiting for the next arrival of a line and will be actually able to step on-board (under-saturation queue), which is an unavoidable phenomenon that depends on the nature of the service and its discontinuous availability in time;
- the queue where some waiting passengers will not be able to board the next arriving carrier (over-saturation queue), which characterizes a critical functioning state of the system; in this case some passengers may have to wait for several carrier arrivals before being able to board.

In this section the focus is on over-saturation queues, as the under-saturation queues have been analysed indirectly in Section 6.2.1 through the modelling of wait times.

In general, the mechanism of passenger queuing is determined by the stop layout and behavioural attitudes. There are two main possible assumptions for passenger (over-saturation) queuing:

- *mingling*, and
- *FIFO*.

For stations with long platforms, it is generally assumed that travellers mingle, which implies that no priority rule is satisfied. Thus, in cases of oversaturation, a passenger who reaches the stop just before the carrier

arrives may be lucky and board the approaching vehicle, while those who arrived earlier may be unlucky and forced to continue waiting. A common modelling assumption is that all passengers waiting along the platform have the same chance of boarding the next approaching vehicle. A similar situation occurs at bus stops if the social culture of passengers is such that no priority is recognized to travellers arrived earlier at the stop.

On the other hand, in some countries for some transit systems (including buses) it happens that First In First Out (FIFO) queues arise at stops, with boarding priority for passengers arrived earlier. Polite queuing is experimented more and more around the world when congestion at stops becomes a recurrent fact, as this passenger behaviour ensures a reduction of waiting time variance (but the same expected value).

Moreover, for stations or stops with very crowded platforms, the mingling mechanism is not anymore valid, if extremely severe congestion occurs. In this case, a large queue of passengers forms and may even spillback on the access ways to the platform including stairs. Therefore, the queueing should be divided into two parts, first FIFO and then mingling.

Furthermore, we can distinguish two queuing mechanism depending on the stop layout:

- the stop is designed (with barriers) to have physically separate queues for each line;
- passengers arriving at the stop join a single mixed queue regardless of their attractive line(s).

The first instance is very common in coach and train terminals. In this case, should congestion occur and no real-time information be available, passengers cannot behave strategically because they must join one specific queue as soon as they reach the stop. It may then be difficult to change queue in order to take advantage of events occurring while they are waiting (e.g., if another attractive line arrives first). Consequently, the stop shall be modelled as a group of separate stops, each of which is served by one line only.

The second type of stop layout is more common in urban public transport networks. In this case, if congestion occurs, users arriving at the stop join the unique queue (regardless of their choice set) and (try to) board the first line of their attractive set that becomes (actually) available. In the case of mingling, each passenger waiting at the stop has the same probability to succeed in boarding an arriving vehicle that is attractive to him/her. In the case of FIFO queue, passengers who do not board the arriving vehicle at the stop because it is not attractive to them can be overtaken by other passengers, and the priority rule is valid only among the passengers actually interested in the departing line. So, if a passenger in the queue does not board, then the next one will if the service is in his/her attractive set; this process starts with the first passengers and is repeated until there is available capacity on-board.

In general, two main modelling approaches are possible to represent crush capacity:

- *soft capacity constraints*, and
- *strict capacity constraints*.

In the first case, the vehicle capacity can be exceeded by the number of on-board passengers. Congestion affects the cost pattern inducing additional impedance on waiting arcs through a suitable arc cost function. Then, the route choice model will indirectly tend to lower the on-board flow exceeding the line capacity. However, relevant capacity violations can result at equilibrium when no alternative route is available.

In the second case, the vehicle capacity will never be exceeded by the number of on-board passengers. Strict capacity constraints can be satisfied in several ways:

- introducing a discontinuity in the arc cost function of waiting arcs with a vertical asymptote (or simply, a very steep impendence) when on-board flows approach the line capacity, which can be done also in the context of a static assignment model but requires (to ensure the existence of a solution) the presence of an alternative (possibly uncongested) path (e.g., on the pedestrian network);
- removing the flow in excess from the boarding arc (if the model is static) and may be injecting it in the following temporal layer of a quasi-dynamic assignment model, or in the waiting arc for next runs in a schedule-based assignment model;
- explicitly reproducing the queuing phenomenon in the context of a within-day dynamic assignment model.

In the following, two methods are proposed to represent mingling queues, which can be developed in the

framework of frequency-based models on static networks or schedule-based models on space-time networks (*effective frequency*, by De Cea and Fernandez, 1993; and *fail-to-board probability*, by Kurauchi et al., 2003). One last method is proposed to represent FIFO queues, which requires the within-day dynamic simulation of macroscopic flows (*bottleneck model* with variable exit capacity, by Meschini et al., 2007).

7.3.2 Effective frequency

The fundamental idea behind the method of effective frequency is that, for a passenger who is waiting a given line at a stop, the probability to succeed in boarding its next approaching vehicle (which is the same for all mingling travellers without considering any boarding priority) decreases on average with the level of on-board congestion. The latter is expressed by the saturation rate of the next line segment (running arc *a*), where the waiting flow (arc *b*) and the dwelling flow (arc *d*) merge.

Therefore, rather than the nominal frequency $f_{\ell s}$, it is assumed that passengers will consider at stop $s \in S_\ell - S_\ell^+$ an *effective frequency* $f_{\ell s}^{eff}$ that is lower than the nominal one, and reduced by the following BPR term:

$$f_{\ell s}^{eff}(q_a) = \frac{f_{\ell s}}{1 + \alpha_\ell^{queue} \cdot \left(\frac{q_a}{\kappa_\ell^{veh} \cdot f_{\ell s}} \right)^{\beta_\ell^{queue}}}, \quad a = \left(N_{\ell s}^{dep}, N_{\ell s^+}^{arr} \right) \in A^{run} , \tag{7.60}$$

where:

- $q_a / (\kappa_\ell^{veh} \cdot f_{\ell s})$ is the saturation rate of the vehicle in the next line segment;
- α_ℓ^{queue} and β_ℓ^{queue} are the BPR coefficient and exponent for the queuing congestion (typical values are $\alpha_\ell^{queue} = 1$ and $\beta_\ell^{queue} = 4$).

The expected wait time at the stop (and also the split of passenger among attractive lines in the strategy models presented in Section 7.1) is, hence, calculated by applying the same equations that are valid in the uncongested case, whereas the nominal frequency is substituted with the effective frequency. When waiting for a single line, based on Equation (6.65), the wait time at stop $s \in S_\ell - S_\ell^+$ is then given by:

$$t_{\ell s}^{wait}(q_a) = \frac{0.5}{f_{\ell s}^{eff}(q_a)} \cdot (1 + \sigma_{\ell s}) . \tag{7.61}$$

The effective frequency method has been the first (computationally tractable) way to incorporate capacity constraints in a transit assignment model. However, it leads to the overloading of some services.

After all, representing this congestion phenomena in a static framework is somewhat disappointing, since queuing is intrinsically dynamic. In order to partly overcome this fault, an alternative formulation of the method can be considered by incorporating the following congestion function obtained from queuing models:

$$f_{\ell s}^{eff}(q_b, q_d) = f_{\ell s} \cdot \left(1 - \left(\frac{q_b}{Max\left(q_b \, , \, \kappa_\ell^{veh} \cdot f_{\ell s} - q_d \right)} \right)^{\chi_\ell^{queue}} \right) , \quad \begin{matrix} b = \left(s, N_{\ell s}^{dep} \right) \in A^{wait} \\ d = \left(N_{\ell s}^{arr}, N_{\ell s}^{dep} \right) \in A^{dwell} \end{matrix} . \tag{7.62}$$

where:

- χ_ℓ^{queue} is the exponent for the ratio between (demand) the waiting flow and (supply) the remaining capacity (typical values is $\chi_\ell^{queue} = 4$); the ratio is bounded to one and the possible 0/0 reads 1.

In this case, the congestion level is expressed as the ratio between the flow of passengers willing to board (arc *b*) and the remaining on-board capacity, given by the line capacity minus the dwelling flow (arc *d*). When the saturation rate approaches one, the effective frequency becomes null and the wait time infinite. Consequently, a strict capacity constraint can be enforced, with line loads never exceeding the available on-board space.

Nevertheless, using such formulation introduces a discontinuity in the arc cost function, and affects the mathematical properties that ensure the existence of equilibrium, as well as the convergence of solution algorithms; especially so, if the overall capacity of the transit network is insufficient to transport the whole demand. This issue can be partly tackled by introducing a suitable pedestrian network composed of arcs with

infinite capacity and finite (but relatively high) cost, so that a walking path is always available between every O-D pair.

In general, the method of effective frequency may result in travel times that are unrealistically high. Indeed, as in any static assignment model, we are not able to reproduce the accumulation capacity of the network: in reality, exceeding flows are temporarily stored into queues that build-up and vanish during the peak, while passengers can progress towards their destination after a finite delay.

A practical way of representing queues in the context of static assignment is obtained by coupling optimal strategies (see Section 7.1) and effective frequencies into a user equilibrium model (see Section 6.1.8). As congestion increases, more (and hence slower) lines are included in the attractive set. Moreover, if all lines are congested, some passengers would rather walk than continue to wait. This leads to a *stability condition*: passengers waiting at a stop would consider an attractive set that is never completely saturated and therefore each of them would be able to board the first arriving vehicle for at least one of the attractive lines.

7.3.2.1 Applications to the example network

Here we ideally continue the numerical tests of Section 7.2.1.1, by substituting the crowding congestion with the queueing congestion.

The arc performance model of Equation (7.60) is here applied jointly to the classical route choice model of Optimal Strategies presented in Section 7.1.1. The equilibrium problem has been solved for the example network through the MSA, and the resulting flows are reported in Table 7.9, assuming small vehicles for Line 2 and 3.

Differently from the results of Table 7.8 where the crowding congestion is reproduced, in the case of (non-separable) queueing congestion a relevant number of passengers departing from Stop 2 prefer to walk to Stop 1 and then to board Line 2, even if the same Line 2 is available directly at Stop 2. This is because here the passenger already on-board have priority over those boarding at the stop; the former who boarded Line 2 at Stop 1 do not suffer any congestion at Stop 2 unlike the latter.

Despite the effective frequency model is intended to reproduce vehicle capacities, as we can see from Table 7.9 these are represented as soft constrains, in the sense that they can be (and are in our case) not (at all) satisfied.

Table 7.9. Line volumes (pax/h) for queuing congestion with effective frequency.

		Segment					
		1→2	2→3	3→4	1→4	2→1	production
line	[pax/h]	3.5 km	3 km	3 km	10 km	3.5 km	[pax*km/h]
1- Red	800				308		3080
2 - Green	80	102	266				1155
3 - Maroon	32		86	96			544
4 - Black	1600			496			1489
walk	INF	0				110	384

To overcome this drawback, we finally present the results of a similar equilibrium model with queue congestion where Equation (7.60) is substituted with Equation (7.62). The latter has a strict capacity constraint, but to ensure the existence of a solution requires the presence of some alternative non-congested route, e.g. a pedestrian network. The results reported in Table 7.10 show how the capacity constraints are now actually satisfied.

Table 7.10. Line volumes (pax/h) for queuing congestion with strict capacity constraint.

		Segment					
		1→2	2→3	3→4	1→4	2→1	production

line	[pax/h]	3.5 km	3 km	3 km	10 km	3.5 km	[pax*km/h]
1- Red	800				549		5479
2 - Green	80	80	80				514
3 - Maroon	32		32	32			191
4 - Black	1600			319			960
walk	INF	0				329	1140

Yet, this is achieved through very high costs for boarding passengers, which may be unrealistic, and also requires many MSA iteration to reach equilibrium. For example, from the results of Table 7.11, we can see that the expected cost to reach the destination for the passengers departing from Stop 2 is three times that of the uncongested case. Thus, a fully satisfactory representation of capacities can only be achieved in a dynamic context, where passenger queues are explicitly simulated, as will be shown in the following sections.

Table 7.11. Expected costs of different congestion models.

cost [min] from origin [stop]	optimal strategies	crowding large vehicles	crowding small vehicles	queuing small vehicles	strict capacity small vehicles
1	27.75	29.71	34.51	29.82	31.87
2	19.07	21.68	53.35	59.80	61.87
3	11.50	12.74	14.15	12.85	13.00

7.3.3 Fail-to-board probability

The method presented in this section is meant to reproduce strict capacity constraints by developing one step further the same idea underlying the alternative formulation of effective frequencies, given by Equation (7.62). When mingles queues occur at the stop, the probability to succeed in boarding the next approaching vehicle for a passenger who waits a given line is assumed equal to the ratio between supply and demand or, more specifically, the remaining capacity available on-board, given by the line capacity minus the dwelling flow (arc *d*), and the flow of waiting passengers who wish to board (arc *b*).

This implies that, in case of oversaturation, at stop $s \in S_\ell - S_\ell^+$ some travellers will fail to board line $\ell \in L$. Here, the aim is to represent this phenomenon explicitly (on flows) and not implicitly (on costs) through its effects on the perceived frequency. Like for the fail-to-sit probability (see Section 7.2.2), the result is conveniently achieved by means of a network model based on hyperarcs through the specification of diversion probabilities.

Few changes to the schemes of Figure 6.6 are required in the stop topology to reproduce the fail-to-board probability (see Figure 7.5). In order to represent this event in topological form, a *service node* $N_{\ell s}^{serv}$ is introduced to split the waiting arc in two (like in the seat availability model); its second part is then called *boarding arc*. Furthermore a *failure arc* is added to transfer back to the stop node the passengers who do not succeed in boarding the next vehicle serving the line and shall start waiting again. The following types of arcs and hyperarcs are then introduced or modified:

- the *waiting arcs* $A^{wait} = \{(s, N_{\ell s}^{serv}): \forall s \in S_\ell - S_\ell^+, \forall \ell \in L\}$;
- the *boarding arcs* $A^{board} = \{(N_{\ell s}^{serv}, N_{\ell s}^{dep}): \forall s \in S_\ell - S_\ell^+, \forall \ell \in L\}$;
- the *failure arcs* $A^{fail} = \{(N_{\ell s}^{serv}, s): \forall s \in S_\ell - S_\ell^+, \forall \ell \in L\}$;
- the *service hyperarcs* $H^{serv} = \{\{(N_{\ell s}^{serv}, N_{\ell s}^{dep}), (N_{\ell s}^{serv}, s)\}: \forall s \in S_\ell - S_\ell^+ , \forall \ell \in L\}$.

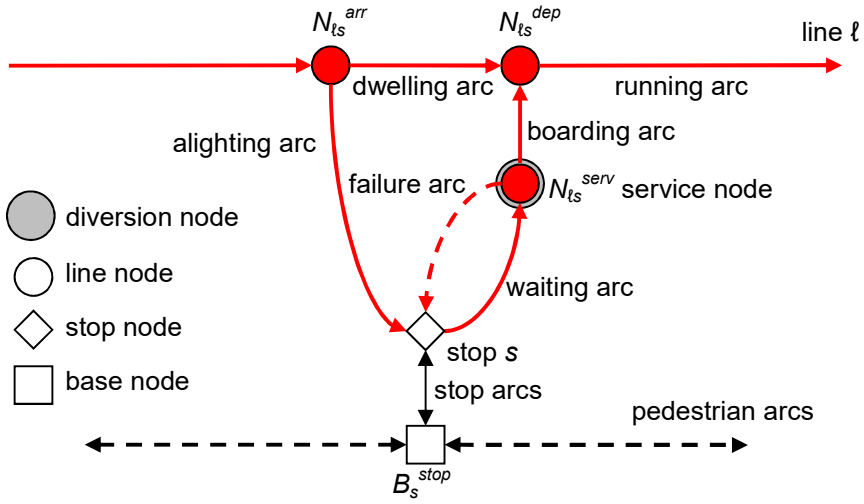

Figure 7.5. The stop topology to reproduce the fail-to-board probability.

Note that different names have been adopted for the splitting node and the resulting arc in the fail-to-board probability model with respect to the seat availability model, so that the two models can be combined without confusion. The possible transfer arcs are connected to the service node.

The diversion nodes are here only the service nodes: $N^{div} = N_{\ell s}^{serv}$. Service hyperarcs are introduced for each line stop to represent the probabilistic event of succeeding and getting on-board vs. failing and keep waiting.

Under the main assumption that all competing passengers, possibly belonging to different classes, have (on average) the same motivation in getting on-board, the *boarding probability* is simply given by the ratio between supply and demand of on-board places; the probability is anyhow bounded between 0 and 1:

$$\begin{aligned} p_{a/\breve{a}} &= Mid\left(0, \frac{\kappa_\ell^{veh} \cdot f_{\ell s} - q_d}{q_b}, 1\right), \\ p_b^{fail} &= p_{e/\breve{a}} = 1 - p_{a/\breve{a}} \end{aligned} \qquad \begin{aligned} &\forall \breve{a} = \{a, e\} \in H^{serv} \\ &a = \left(N_{\ell s}^{serv}, N_{\ell s}^{dep}\right) \in A^{board} \\ &e = \left(N_{\ell s}^{serv}, s\right) \in A^{fail} \\ &b = \left(s, N_{\ell s}^{serv}\right) \in A^{wait} \\ &d = \left(N_{\ell s}^{arr}, N_{\ell s}^{dep}\right) \in A^{dwell} \end{aligned} \quad . \tag{7.63}$$

Like for seat availability, or this type of hyperarcs it is assumed: $t_{\breve{a}} = 0$.

The fail-to-board probability p_b^{fail} is clearly the complement to 1 of the boarding probability. This schema allows for different models, from static or quasi-dynamic, to dynamic macroscopic or microscopic models.

In static models, the failing arc is not actually coded in the network, because it is not possible to cast passengers back to the stop at a later time and its presence would create absorbing cycles, which are difficult to handle, while the capacity constraint would be violated. Thus, the flow exiting from the waiting arc and entering the boarding arc (that in this case is the only branch of the boarding hyperarc) is scaled by the boarding probability, while the rest is eliminated from the network.

In dynamic models, including schedule-based models with space-time network, passengers who fail to board are transferred back to the stop node.

Two approaches are available to represent the cost of failure:

- if the failure arc is not coded, then a non-temporal cost component is to be introduced on the waiting arc to represent the risk of fail-to-board; the passengers who failed to board are eliminated from the model (or swapped to the next temporal layer);
- if the failure arc is explicitly coded, then the risk of failure is represented by the hyperarc diversion, which will possibly take the passengers back to the stop where a new wait begins. The expected cost to reach the destination from the service node will, then, be given as the weighted average between the cost of the departure node plus the boarding arc and the cost of the stop node plus zero (the

failure arc is dummy). By construction, the cost of the service node is lower than the cost of the stop node (because the wait for one vehicle arrival has already been paid, although fail-to-board can occur) and the resulting increment due to the weighted average represents the cost of failure. No passenger is eliminated from the network.

In the first case, with respect to the performance model presented in 6.2.3 few things change:

- on the boarding arc, the travel time is null and the boarding fee is paid;
- on the waiting arc, the travel time and comfort are expressed by Equation (6.67.f), where the non-temporal cost is given by the risk of failure:

$$t_a = 0 \quad , \quad \gamma_{ag} = \gamma_g^{vot} \quad , \quad c_{ag}^{nt} = c_{\ell s}^{bfee} \cdot \gamma_g^{mfee} \quad , \forall a = \left(N_{\ell s}^{serv}, N_{\ell s}^{dep}\right) \in A^{board} \ , \tag{7.63.a}$$

$$t_a = t_{\ell s}^{wait}\left(\mathbf{q}_A\right) \quad , \quad \gamma_{ag} = \gamma_g^{vot} \cdot \gamma_g^{wait} \cdot \gamma_{sg}^{stop} \cdot \gamma_{sg}^{crowd}\left(\mathbf{q}_A\right) \quad , \quad c_{ag}^{nt} = p_a^{fail} \cdot c_{ag}^{fail} \quad , \begin{matrix} \forall a = \left(s, N_{\ell s}^{serv}\right) \in A^{wait} \\ \forall a = \left(N_{\ell' s}^{arr}, N_{\ell s}^{serv}\right) \in A^{trans} \end{matrix} \ . \tag{7.63.b}$$

All waiting passengers suffer from a cost due to the risk of fail-to-board, which is additional to the temporal cost of waiting for the arrival of the boarded service. This expected cost of failing is obtained multiplying the fail-to-board probability p_a^{fail} by the additional cost in the case of failure c_{ag}^{fail}. The latter is given by the *risk-averseness coefficient* γ_g^{risk} of class $g \in G$ users towards (abnormal) delays (since failing to board is perceived as a malfunctioning of the system), multiplied by the value of time γ_{ag} of the waiting arc, multiplied by the average additional wait time conditional on failing. In turn, this additional wait time is given by the expected headway (the inverse of the frequency), multiplied by the number of arriving carriers a waiting passenger will fail to board on average before boarding, which is equal to one over the probability of not failing. Then, we have:

$$c_{ag}^{fail} = \gamma_g^{risk} \cdot \gamma_{ag} \cdot \frac{1}{f_{\ell s}} \cdot \frac{1}{\left(1 - p_a^{fail}\right)} \ . \tag{7.64}$$

The term at the denominator $f_{ts} \cdot (1\text{-}p_a^{fail})$ can also be seen as a sort of effective frequency and its inverse as a sort of effective expected headway; this coincides with the average additional time that the passenger has to wait if s/he fails to board the first arriving vehicle.

The above failing cost tends to infinity as the fail to board probability goes to one. The amount of passengers who will accept the risk of failing is a result of the equilibrium mechanism.

This schema is also suitable for quasi-dynamic models. When propagating flows, temporal layers are processed in chronological order, and passengers who fail to board are transferred back to the stop node, in the *next* temporal layer, when they will have to wait again (note than the route choice is calculated based on the arc costs of the *current* layer).

The cost expression (7.64) might be too severe as nobody will accept risking if the fail-to-board probability is close to one. However, queuing is a dynamic phenomenon that is related to a temporary lack of capacity. In reality, passengers might know from experience that congestion at stops will eventually decrease after the peak. Instead, equation (7.64) evaluates the failure costs as if congestion lasts forever.

The second case completely overcomes this fault, but requires dynamic assignment models. In the case of fully dynamic models, the cost expression Equation (7.64) is not necessary, since the failure arc takes with a given probability the passenger back to the stop node at a later time. At this time the cost of the stop node intrinsically includes the additional delays due to queuing and is higher than the cost of the departure node.

Like in the case of seat availability, the diversion probabilities (and not 'choice' probabilities) are determined by a physical random event and depend (asymmetrically) on passenger flows. Differently from the case of multiple attractive lines here the assignment model is necessarily congested, leading to an equilibrium problem. The scheme of Figure 6.3 can be applied considering the sequential model based on hyperarcs of Section 6.1.5.

The example in Section 4.5.4 provides more insights on the practical relevance of fail-to-board probability in revealed passenger behavioural and willingness-to-pay.

The idea of fail-to-board probability can be applied also in the case of schedule-based services modelled

through a space time network. In this case the service node N_{rs}^{serv} may be introduced to split the boarding arc just to isolate the diversion node; indeed, there whould be no need of such a node, because the waiting phase and the boarding phase have already dedicated separate arcs; the failure arc is headed at the next node in time of the same stop. With respect to the network model presented in Section 6.3.1 the following arc and hyperarcs are introduced or modified:

- the *service arcs* $A^{serv} = \{((s, t^{-}(\theta_{rs} - t_{\ell}^{board})), N_{rs}^{serv}): \forall s \in S_{\ell}\text{-}S_{\ell}^{+}, \forall r \in R_{\ell}, \forall \ell \in L\}$;
- the *failure arcs* $A^{fail} = \{(N_{rs}^{serv}, (s, t^{-}(\theta_{rs} - t_{\ell}^{board}) +1)): \forall s \in S_{\ell}\text{-}S_{\ell}^{+}, \forall r \in R_{\ell}, \forall \ell \in L\}$;
- the *boarding arcs* $A^{board} = \{(N_{rs}^{serv}, N_{rs}^{dep}): \forall s \in S_{\ell}\text{-}S_{\ell}^{+}, \forall r \in R_{\ell}, \forall \ell \in L\}$;
- the *service hyperarcs* $H^{serv} = \{\{(N_{rs}^{serv}, N_{rs}^{dep}), (N_{rs}^{serv}, (s, t^{-}(\theta_{rs} - t_{\ell}^{board}) +1))\}: \forall s \in S_{\ell}\text{-}S_{\ell}^{+}, \forall r \in R_{\ell}, \forall \ell \in L\}$.

Equation (7.63) can be immediately extended to the case of schedule-based models based on diachronic graphs under the consideration that the arc loads represent in this case a number of passengers, which can directly be compared with the vehicle capacity:

$$p_a^{fail} = 1 - Mid\left(0\,,\ \frac{\kappa_{\ell}^{veh} - q_d}{q_a}\,,\ 1\right) \quad , \quad \begin{matrix} a = \left(\left(s, t^{-}\left(\theta_{rs} - t_{\ell}^{board}\right)\right), N_{rs}^{serv}\right) \in A^{serv} \\ d = \left(N_{rs}^{arr}, N_{rs}^{dep}\right) \in A^{dwell} \end{matrix} \quad , \tag{7.65}$$

7.3.4 Bottleneck model with variable exit capacity

We address here the case where the stop layout and the travellers' behaviour are such that passengers have to join a FIFO queue and respect the boarding priority of those who arrived before them.

The FIFO queuing process at a stop can be seen as a gate system, and it works similarly to the access of a cableway.

think what happens to access a cableway. As soon as passengers reach the stop, they join the queue before the gate and start waiting. But only passengers after the gate will be actually able to board the next arriving carrier. Thus, in general, an ideal gate separates the two phases: passengers before the gate are queuing (over-saturation delay due to congestion), while passengers after the gate are waiting for the next arrival (under-saturation delay due to the discontinuity of the service).

However, this scheme does not apply when passengers waiting at the stop go to the different destinations and thus may have different attractive sets partially overlapping. In this case, if a line that arrives at the stop is not attractive for a passenger in the queue s/he can be overtaken by the next one, if the service is in his/her attractive set, until there is available capacity on-board. The result is a sort of mixed queue for all lines serving the stop.

For modelling convenience, we imagine the presence of separate queues for each line, and those queues are joined with certain probabilities by passengers that include the corresponding lines in their attractive sets. The under-saturation delay due to the discontinuity of the service is spent before (and not after) joining the queue on hyperarcs whose line shares spit passengers among the above queues.

Few changes to the schemes of Figure 6.6 or Figure 7.1 are then required in the stop topology to reproduce FIFO queuing (see Figure 7.6). To separate the over-saturation queue form the under-saturation queue, a *queue node* $N_{\ell s}^{que}$ is introduced to split the waiting arc in two; its second part is then called *queueing arc*. The following types of arcs are then introduced or modified:

- the *waiting arcs* $A^{wait} = \{(s, N_{\ell s}^{que}): \forall s \in S_{\ell}\text{-}S_{\ell}^{+}, \forall \ell \in L\}$;
- the *queuing arcs* $A^{que} = \{(N_{\ell s}^{que}, N_{\ell s}^{dep}): \forall s \in S_{\ell}\text{-}S_{\ell}^{+}, \forall \ell \in L\}$.

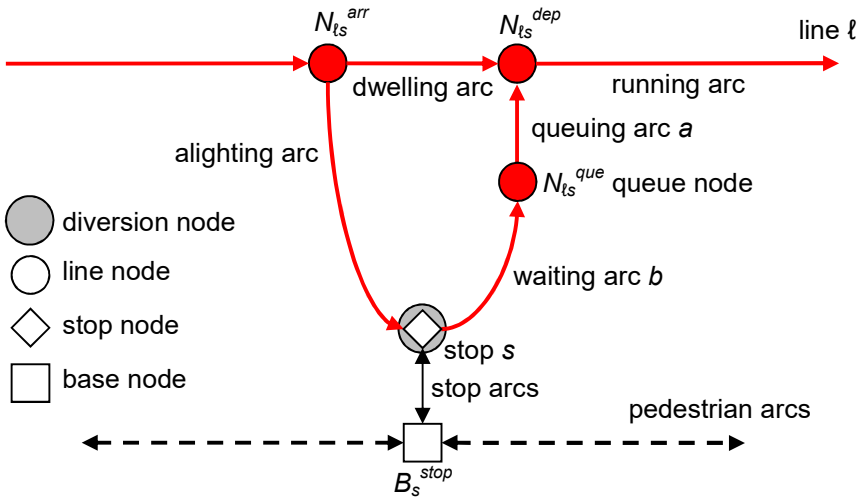

Figure 7.6. The stop topology to reproduce FIFO queuing.

Static assignment is not a proper modelling framework to reproduce queuing phenomena. In the following, a dynamic macroscopic model for frequency-based assignment is then introduced adopting the framework presented in Section 6.4. It is then assumed that all variables are (in general) continuous functions of the day time (also called temporal profiles), and transit services are conceived as a continuous flow of supply with 'instantaneous capacity' (which is expressed in terms of passengers per hour instead of passengers per vehicle). This allows to reproduce the effect of time-discrete services through the temporal profile of the average wait times.

Let's consider then the supply side of the equilibrium problem, where the aim is to provide for given arc flows the exit times and the comfort coefficients of each arc, as well as the diversion probabilities of each hyperarc, which are all used in the route choice model.

When it is assumed that passengers follow a FIFO protocol, the exit time (profile) from the queuing arc of a specific line for a given entry flow (profile) can be calculated by means of the bottleneck model proposed in Meschini et al. (2007) that explicitly reproduces the formation and dispersion of passenger queues. The main assumption is that the capacity of the bottleneck, given by flow of line vehicles (the frequency) multiplied by the vehicle capacity and reduced by the flow of dwelling passengers (the remaining capacity), is continuous in time but not constant. Indeed, the flow of passengers using the line is not constant in time due to demand modulation; moreover, the presence of time-varying dwelling delays due to boarding and alighting passenger flows induces frequency modulation in time, as already in Section 6.4.5.

The mathematical formulation of the model works on cumulative flows and considers the cumulative number of passengers joining the queue of a line and the cumulative remaining on-board capacity as an input, and the cumulative number of passengers leaving the queue and boarding the line as an output. If the remaining on-board capacity does not suffice to accommodate the flow of passengers who are ready to board after the under-saturation wait at the stop, a queue builds up which will dissipate only if/when the remaining on-board capacity is greater than the inflow of arriving passengers from the waiting arc. Let:

- $\kappa_a(\tau)$ be the instantaneous remaining capacity at time τ, which is available at the end of the queueing arc a for passengers wishing to board line $\ell \in L$ at stop $s \in S_\ell - S_\ell^+$.

This is equal to the capacity of one vehicle κ_ℓ^{veh} multiplied by the flow of vehicles departing from the stop, i.e., the departure line frequency $f_{\ell s}^{dep}(\tau)$, reduced by the flow of passengers exiting from the dwelling arc d, the latter being equal to the entry flow at the earlier time $\tau - t_{\ell s}^{dwell}$ under the assumption of constant dwell time:

$$\kappa_a(\tau) = \kappa_\ell^{veh} \cdot f_{\ell s}^{dep}(\tau) - q_d^{out}(\tau) \quad , \quad a = \left(N_{\ell s}^{que}, N_{\ell s}^{dep}\right) \in A^{que} . \tag{7.66}$$

Let's recall that the cumulative inflow $q_a^{cin}(\tau)$ and outflow $q_a^{cout}(\tau)$ of the queuing arc a at time τ are given by the integral of the instantaneous inflow and outflow, respectively; analogously, let's define $\kappa_a^{cum}(\tau)$ as the cumulative remaining capacity:

$$q_a^{cin}(\tau) = \int_0^{\tau} q_a^{in}(\vartheta) \cdot d\vartheta \quad , \quad q_a^{cout}(\tau) = \int_0^{\tau} q_a^{out}(\vartheta) \cdot d\vartheta \quad , \quad \kappa_a^{cum}(\tau) = \int_{\theta}^{\tau} \kappa_a(\vartheta) \cdot d\vartheta \tag{7.67}$$

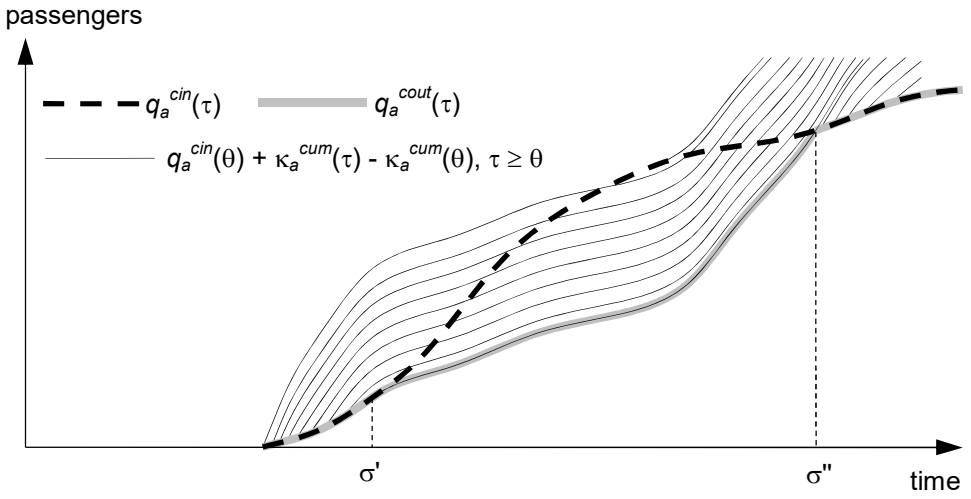

Figure 7.7. Bottleneck with time-varying capacity. The cumulative outflow is the lower envelop of the profiles family for each θ , with $\tau \geq \theta$, obtained from the vertical translation of the cumulative remaining capacity that goes through point (θ, $q_a^{cin}(\theta)$). No queue is present when $q_a^{cin}(\tau)$ prevails. Here, the queue arises at time σ' and vanishes at time σ''.

Based on the Newell-Luck minimum principle (stating that among all possible flow state the more restrictive one holds), the cumulative outflow $q_a^{cout}(\tau)$ of the queuing arc *a* at time τ is the lower envelope of all possible temporal profiles that would result if the queue would start at any previous time $\theta \leq \tau$; in this case the outflow would be given by the inflow until the queue begins at time θ and by the time varying capacity $\kappa_a(\tau)$ from θ until τ (see Figure 7.7):

$$q_a^{cout}(\tau) = Min\left(q_a^{cin}(\theta) + \kappa_a^{cum}(\tau) - \kappa_a^{cum}(\theta), \forall \theta \leq \tau\right). \tag{7.68}$$

The exit time $\theta_a(\tau)$ of the queuing arc *a* for a passenger who enters it at time τ can be obtained as in (6.77) on the basis of the cumulative inflows and outflows assuming that the FIFO rule (no overtaking) holds:

$$q_a^{cout}\left(\theta_a(\tau)\right) = q_a^{cin}(\tau). \tag{7.69}$$

In the context of commuting trips, passengers know by previous experience:

- the (average) number of carriers $n_a(\tau)$ they must let go (because other passengers who arrived earlier at the stop have priority) before being able to board each line $\ell = L_a$, if queuing starts at a given time τ.

This is equal to the number of vehicle passing from τ to $\theta_a(\tau)$:

$$n_a(\tau) = \int_{\tau}^{\theta_a(\tau)} f_{\ell s}^{dep}(\vartheta) \cdot d\vartheta \quad , \quad a = \left(N_{\ell s}^{que}, N_{\ell s}^{dep}\right) \in A^{que}. \tag{7.70}$$

If there is no over-saturated queuing then $n_a(\tau) = 0$.

Correspondingly, the average frequency $f_a(\tau)$ perceived by passengers while queuing is given by the ratio between the number of vehicles $n_a(\tau)$ passing from τ to $\theta_a(\tau)$ and the duration of this time interval:

$$f_a(\tau) = \frac{n_a(\tau)}{\theta_a(\tau) - \tau}. \tag{7.71}$$

Let's now calculate the exit time of the waiting arc $b \in A^{wait}$, which shall take into account for the service being not continuous in time.

In presence of over-saturation queues, waiting is related not to the arrival of just one line vehicle with a known headway distribution, but to the consecutive arrivals of $n_a(\tau)+1$ vehicles. Indeed, because the service is discontinuous, one vehicle is waited anyhow by all passengers, even if no oversaturation queue occurs.

Under the assumption that the headway of line L_a , which is experienced by a passenger who started queuing at time τ , is exponentially distributed with a constant frequency equal to $f_a(\tau)$ during the whole time spent waiting, then the wait time before $n_a(\tau)+1$ carrier arrival occur is a stochastic variable having a Gamma probability density function (which is the continuous version of the Erlang distribution introduced in Section 6.2.1):

$$\varphi_b^w(\tau,t)=\begin{cases}\dfrac{f_a(\tau)^{\left(n_a(\tau)+1\right)}\cdot Exp\left(-f_a(\tau)\cdot t\right)\cdot t^{n_a(\tau)}}{\Gamma\left(n_a(\tau)\right)}, & \text{if } t\geq 0\,.\\ 0\,, \quad \text{otherwise}\end{cases} \tag{7.72}$$

This corresponds to the worst situation in terms of headway irregularity. The other extreme case is constant (deterministic) headways, that corresponds to the best possible regularity:

$$\varphi_b^w(\tau,t)=\begin{cases}f_a(\tau), & \text{if } 0\leq t-\left(\theta_a(\tau)-\tau\right)+\leq\dfrac{1}{f_a(\tau)}\\ 0, \quad \text{otherwise}\end{cases}\,. \tag{7.73}$$

In case of a singleton attractive set with one line only, the expected wait time $E(\varphi_b{}^w(\tau,t))$ for the first $n_a(\tau)+1$ arrivals can be approximated as in (6.65) by a convex linear combination of the two extreme cases (exponential headways and deterministic headways) through the square of the variation coefficient $\sigma_a{}^2$, which is an input of the model.

To the waiting arc it is associated only the additional wait time (due to discontinuous service) for a passenger who starts queueing in τ, while a (possibly significant) part of the waiting time is already accounted for in the queuing time $\theta_a(\tau)$-τ with $a = (b^+)^+$. The entry time of the waiting arc b is then equal to:

$$\theta_b^{-1}(\tau)=\theta_a(\tau)-E\left(\varphi_b^w(\tau,t)\right)\,. \tag{7.74}$$

From the exit time by inversion of the temporal profile it is possible to obtain the entry time.

When a set of attractive lines is considered by the passenger who will board the first available vehicle, using distributions like (7.72) and (7.73) in the equations of Section 7.1, the combined wait times $t_{b|\check{b}}(\tau)$ conditional to take line $b\in\check{b}$ and the line shares (diversion probabilities) $p_{b|\check{b}}(\tau)$ can be obtained for each hyperarc $\check{b}\in s^+$ of stop $s\in S$. Again, in the conditional exit times the queuing times have to be deducted:

$$\theta_{b|\check{b}}^{-1}(\tau)=\theta_a(\tau)-t_{b|\check{b}}(\tau)\,. \tag{7.75}$$

Note that the parameters of the distributions (as well as the remaining costs needed by some strategy models) for the entire wait time are evaluated at time τ when possible queuing stars (waiting starts earlier) and refer only to the period of time while the passenger is queuing. Clearly, this assumption is made for modelling convenience. In the case of strategies this introduces a further approximation, which is minor if the change in time of headway distribution parameters due to congestion is slow with respect to waiting times:

- the diversion probabilities applied to the passengers entering the waiting branch $b\in\check{b}$ at time τ are calculated wrt the headway distributions perceived by passengers who at time τ are exiting this branch and start queuing;
- the conditional exit times of passengers entering each waiting branch $b\in\check{b}$ at time τ refer to (slightly) different headway distributions.

Note that the calculation of conditional exit times is necessary to implement the dynamic hyperarc model proposed in Section 6.4.4. A possible approximation which may simplify the model is: $t_{b|\check{b}}(\tau) = t_{\check{b}}(\tau)$.

7.3.5 Impulse flows and run capacity constraint

An alternative approach to the representation of schedule-based supply can be achieved through a standard graph, like the static transit network of Figure 6.6, by adapting the macroscopic model for dynamic transit

assignment of Sections 6.4 to the presence of runs and their capacity constraints.

In essence, a time-discrete flow model is considered when referring to running and dwelling arcs, where all the passengers on board of a run are assumed to cross any section along the line at the same instant, thus forming a dense point-packet or impulse flow. On the contrary, a time-continuous flow model is considered when referring to the pedestrian and stop arcs. Waiting and alighting arcs concentrate continuous flows into discrete flows and spread discrete flows into continuous flows, respectively. Thus, we will have run loads for running arcs and dwell arcs, as well as for waiting arcs (the boarding impulse outflow) and alighting arcs (the impulse inflow), but also a temporal profile for waiting arc inflows and alighting arc outflows.

This requires the definition of proper temporal profiles of the exit time for waiting arcs and alighting arcs, to compress and decompress the passenger flows.

Consider first the waiting arc $b \in A^{wait}$.

For each run $r \in R_\ell$ of line $\ell = L_b$ the following capacity constraint is to be satisfied:

$$q_{br} \le \kappa_\ell^{veh} - q_{dr} \quad d = \left(N_{rs}^{arr}, N_{rs}^{dep}\right) \in A^{dwell},\ b = \left(s, N_{rs}^{dep}\right) \in A^{wait} \ , \tag{7.76}$$

where q_{br} and q_{dr} are the loads of passengers boarding run r at stop s and of those dwelling that are already on-board. By definition the exit time of all passengers that board run r from stop s coincides with the departure time θ_{rs}. To take into account the effects of this capacity constraint, we shall determine the:

- time $\rho_{br} \le \theta_{rs} - t_\ell^{board}$ when the last passenger that achieves boarding run r (or would achieve to do so, in case of null inflows) arrives at the stop and enters the waiting arc (t_ℓ^{board} is a safe departure margin).

Then we have:

$$\theta_b(\tau) = \begin{cases} \theta_{rs} & r : \rho_{b\,r-1} < \tau \le \rho_{br} \\ \infty & \tau > \rho_{b\,R_\ell^+} \end{cases} . \tag{7.77}$$

The instants ρ_{br} can be determined recursively following the run order from $r = 1$ to $r = R_\ell^+$ (for the sake of simplicity, runs are referred here through their integer order in the sequence and R_ℓ).

To this end let's initialize $\rho_{b0} = -\infty$. The passengers willing to take line ℓ that arrive at the stop later than $\rho_{b\,r-1}$ shall board the successive run r until their number overcomes the residual capacity $\kappa_\ell^{veh} - q_{dr}$, which happens at a specific instant denoted σ_{br}:

$$\kappa_\ell^{veh} - q_{dr} = q_b^{cin}\left(\rho_{b\,r-1}\right) - q_b^{cin}\left(\sigma_{br}\right) \ , \tag{7.78}$$

or their arrival at the stop is too late to board run r, which happens at time $\theta_{rs} - t_\ell^{board}$. Then we have:

$$\rho_{br} = Min\left(\sigma_{br}, \theta_{rs} - t_\ell^{board}\right) \ . \tag{7.79}$$

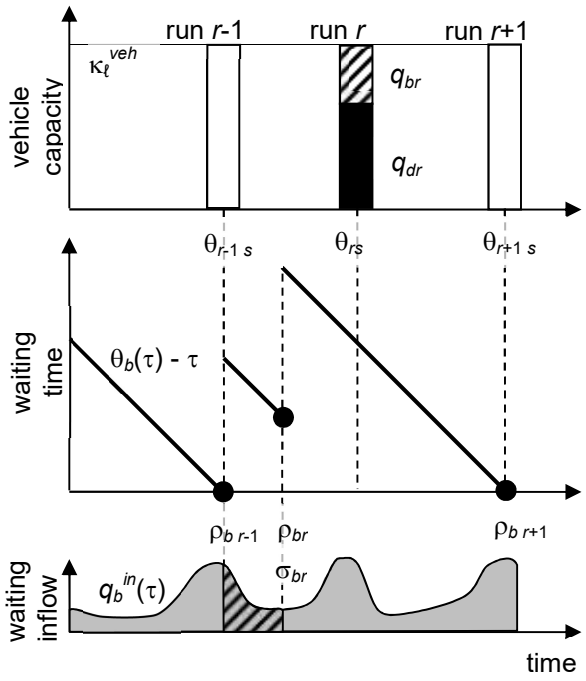

Figure 7.8. Saw-tooted waiting time for given residual capacities and inflows; here $t_\ell^{board} = 0$.

The proposed waiting model reproduces the priority of passengers arrived earlier at the stop. It is then consistent with FIFO queuing, unlike the fail-to-board model for schedule-based systems presented in Section 7.3.3, which is instead consistent mingling queuing. The main difference is that in the latter model new passengers arriving at the stop will influence the waiting time of those who arrived earlier.

Note that the model yields discontinuities in the travel time pattern, although this is coherent with the real phenomenon. Indeed, the waiting time profile has the saw-tooth shape depicted in Figure 7.8, where each run r will be taken by the passengers that entered the waiting arc during the time interval $(\rho_{b\,r-1}, \rho_{br}]$. Then, the boarding load can be calculated as the integral of the waiting inflow in this interval:

$$q_{br} = q_b^{cin}\left(\rho_{br}\right) - q_b^{cin}\left(\rho_{b\,r-1}\right) . \tag{7.80}$$

Consider now the alighting arc $a \in A^{alight}$.

Strictly speaking, the exit time of a associated with run r depends on the position of the passenger in the alighting load q_{ar} . Therefore, from the first to the last user in this load the exit time (after the arrival time τ_{rs} of the run and the additional alighting time t_ℓ^{alight}) varies linearly from 0 to the ratio q_{ar} / κ_ℓ^{alight} between the number of alighting passengers and the alighting capacity of vehicle doors.

For what concerns route choice, we can assume a risk adverse behaviour, such that all the alighting passengers will perceive the same travel time:

$$\theta_{ar} = \tau_{rs} + \frac{q_{ar}}{\kappa_\ell^{alight}} + t_\ell^{alight} . \tag{7.81}$$

On the other hand, when propagating the alighting passengers on the pedestrian network, we will spread them uniformly:

$$q_a^{out}\left(\tau\right) = \kappa_\ell^{alight} \quad 0 \le \tau - \tau_{rs} - t_\ell^{alight} \le \frac{q_{ar}}{\kappa_\ell^{alight}} . \tag{7.82}$$

The model proposed in this section for schedule-based services can also be used to extend the dynamic macroscopic model for frequency-based services presented in Section 7.3.4 to networks with mixed services. Indeed, the former can be seen as a particular instance of the latter under the assumption that no waiting is considered but only queuing, while the departure frequencies at stops are given by an impulse flow of vehicles representing each single run rather than by smooth temporal profiles. In this framework it is also possible to represent the propagation of such a discontinuous frequency from the first stop based on the

model of Section 6.4.5, with the possibility of representing also its modulation from stop to stop due dynamic phenomena, including dwelling congestion (see Section 7.4). The drawback of this approach is the dense temporal discretization that is needed to clearly distinguish the individual runs in the resulting temporal profiles.

7.3.6 Reference notes and concluding remarks

7.3.6.1 Mingling queuing

As shown in this section, the representation of mingling passengers queues at stops can be developed in the framework of frequency-based models on static networks and schedule-based models on space-time networks using two different approaches: effective frequency (De Cea and Fernandez, 1993) and fail-to-board probability (Kurauchi et al., 2003).

In the context of frequency-based models, static assignment with optimal strategies can be improved by considering effective frequency with strict capacity constraints (Wu et al., 1994; Cominetti and Correa, 2001; Cepeda et al., 2006).

Bell and Schmocker (2004) apply instad the approach of fail-to-board probabilities to quasi-dynamic model; Schmöcker et al. (2008) extend this approach to strategy-based route choice.

Schedule-based models with mingling queues have been developed by several authors.

Carraresi et al. (1996) consider a multicommodity flow model with strict capacity constraints.

Tian et al. (2007) introduced in-vehicle congestion through a bulk-queue model and analyzed the theoretical properties of the equilibrium flows.

Hamdouch and Lawphongpanich (2008) have explored the possibility of considering hyperpaths on space-time networks where the strategic behaviour of waiting passengers derives from the uncertainty of boarding the arriving vehicle due to capacity constraints. The extension of fail-to-board probabilities to schedule-based models presented in Section 7.3.3 founds its roots in this work.

Nuzzolo et al. (2012) applied the effective frequency approach to stochastic assignment models on the diacronich graph.

7.3.6.2 FIFO queuing

An early attempt to model FIFO queues was made by Bouzaïene-Ayari (1998) by using a bulk queue model, but the complexity of the formulations practically prevents the analysis of network equilibrium on large networks.

Poon et al. (2004) use a time-increment simulation to load passenger demand onto the network and the available capacity of each vehicle is updated dynamically. After each simulation run, the passenger arrival and departure profiles at all stations are recorded and these are used to predict dynamic queuing delays. From such delays, minimum paths are updated and used for the next simulation run. The user equilibrium assignment problem is solved iteratively by the method of successive averages. A similar approach is adopted in Teklu (2008), within a day-to-day assignment model, and in Leurent et al. (2012) within a genral framework for meso-simulation of transit networks.

Indeed, space-time networks are not suitable for FIFO modelling because passenger flows on arcs are mingled by construction. The more complex dynamic models based on macroscopic flows can instead serve for the purpose. In particular, the model presented in Section 7.3.5 has been proposed in Papola et al. (2007).

By contrast, in the frequency-based realm, the definition of a supply model for dynamic assignment is not equally simple because different runs of the same service are not distinguished and thus it is not immediately possible to evaluate the capacity available on a certain line/stop at a certain time of the analysis period. Indeed, the majority of available models with capacity constraints, are developed in a static setting only.

Meschini et al. (2007), whose bottleneck model has been presented in Section 7.3.4, is among the very few dynamic models for frequency-based transit assignment with FIFO queues. It makes use of a macroscopic

representation of vehicle and passenger flows as (upper semi) continuous functions of time (temporal profiles). Transit services are then considered as a continuous flow of vehicles with an instantaneous capacity. The model allows however to represent the average effect of time-discrete services on wait times. It should be noticed that this continuous availability of the transit vehicles, though questionable from a phenomenal point of view, is consistent with the basic assumption of the frequency-based modelling framework, where passengers conceive all the runs of the same line as a unitary supply facility. Trozzi et al. (2013a, 2013b) extended this approach to strategies and information.

7.3.6.3 List of references

Bell M.G.H., Schmocker J.-D. (2004) A solution to the congested transit assignment problem. In Scheduled-Based Dynamic Transit Modeling: Theory and Applications, ed.s N.H.M. Wilson and A. Nuzzolo, Springer, New York, 263-280.

Bouzaiene-Ayari B., Gendreau M., Nguyen S. (1998) Passenger assignment in congested transit networks: A historical perspective. Equilibrium and advanced transportation modelling, ed.s. P. Marcotte and S. Nguyen, Kluwer Academic Publishers, 304-321.

Carraresi P., Malucelli F., Pallottino S. (1996) Regional mass transit assignment with resource constraints. Transportation Research B 30, 81-89.

Cepeda M., Cominetti R., Florian M. (2006) A frequency-based assignment model for congested transit networks with strict capacity constraints: characterization and computation of equilibria. Transportation Research B 40, 437-459.

Cominetti R., Correa J. (2001) Common lines and passenger assignment in congested transit networks. Transportation Science 35, 250-267.

De Cea J., Fernandez J.E. (1993) Transit assignment for congested public transport systems: An equilibrium model. Transportation Science 27,133-147.

Hamdouch Y., Lawphongpanich S. (2008) Schedule-based transit assignment model with travel strategies and capacity constraints. Transportation Research B 42, 663-684.

Kurauchi F., Bell M.G.H., Schmöcker J.-D (2003) Capacity constrained transit assignment with common lines. Journal of Mathematical Modelling and Algorithms 2-4, 309-327.

Leurent F., Chandakas E., Poulhès A. (2012) A passenger traffic assignment model with capacity constraints for transit networks. Procedia - Social and Behavioral Sciences 54, Proceedings of EWGT2012, 772-784.

Meschini L., Gentile G., Papola N. (2007) A frequency based transit model for dynamic traffic assignment to multimodal networks. Proceedings of the 17th International Symposium on Transportation and Traffic Theory (ISTTT), ed.s R. Allsop, M.G.H. Bell, and B.G. Heydecker, Elsevier, London, 407-436.

Nuzzolo A., Crisalli U., Rosati L. (2012) A schedule-based assignment model with explicit capacity constraints for congested transit networks. Transportation Research C 20, 16-33.

Papola N., Filippi F., Gentile G., Meschini L. (2007) Schedule-based transit assignment: a new dynamic equilibrium model with vehicle capacity constraints. In Schedule-based modeling of transportation networks. Theory and applications, ed.s N.H.M. Wilson and A.Nuzzolo, Springer, 145-171.

Poon M.H., Wong S.C., Tong C.O. (2004) A dynamic schedule-based model for congested transit networks. Transportation Research B 38, 343-368.

Teklu F. (2008) A stochastic process approach for frequency-based transit assignment with strict capacity constraints. Networks and Spatial Economics 8, 225-40.

Trozzi V., Gentile G., Bell M.G.H., Kaparias I. (2013a) Dynamic User Equilibrium in public transport networks with passenger congestion and hyperpaths. Transportation Research B 57, 266-285.

Trozzi V., Gentile G., Bell M.G.H., Kaparias I. (2013b) Effects of countdown displays in public transport route choice under severe overcrowding. Networks and Spatial Economics, published on-line.

Schmoeker J.-D., Bell M.G.H., Kurauchi F. (2008) A quasi-dynamic capacity constrained frequency-based transit assignment model. Transportation Research B 42, 925-945.

Wu J.H., Florian M., Marcotte P. (1994) Transit equilibrium assignment: a model and solution algorithms. Transportation Science 28, 193-203.

7.4 Service perturbations

Authors: Ektoras Chandakas, Moshen Babaei, Oded Cats, Pieter Vansteenwegen, Guido Gentile

This section addresses the problem of reproducing service perturbations due to non-recurring, unpredictable events as well as to systematic, predictable events that affect the regular operation of public transport. The lack of service regularity (irregularity) is intended here as any deviation of the actual run arrivals at the stops from the planned schedule. The focus is then on the average effects of minor service perturbations on route choices and network performances occurring on a daily basis, rather than on the real-time management of major service disruptions. In particular, this section is devoted to the modelling of the following phenomena in the context of transit assignment:

- service irregularity due to supply and demand uncertainties;
- propagation of perturbations along the line;
- pairing and bunching of vehicles;
- correlation of headway distributions among lines at various stops;
- dwell time dependence on boarding and alighting flows;
- impact of dwell times on the service frequency;
- time varying frequency along the line;
- lines operated with a fixed number of vehicles;
- stop berthing capacities as a constraint to frequencies;
- reliability and robustness of transit networks also wrt coincidences.

The desired output of these assignment models is an expected flow and cost pattern, that shall take into account, not only the service perturbations caused by random passenger loads and events on the transit network, but also the possible countermeasures in terms of behavioural strategies by users and control strategies by operators.

The management of public transport operations is a demanding task due to a multitude and variety of factors that impact service performances.

On the one hand, exceptional events (such as extreme weather conditions, infrastructure malfunctions and demand peaks) can occasionally lead to service perturbations of great intensity; these events should be simulated through specific scenarios with local modifications of the demand and supply model. On the other hand, smaller events (such as accidents, run cancellations and demand fluctuations) occur more frequently (usually on a daily basis) and may lead to minor perturbations. But these operation dis-functionalities imply widespread reductions in service capacity and speed, causing a systematic increase in the travel time of individual passengers. In general, both types of events influence the appeal of public transportation and attract the attention of city authorities.

The characteristics of the transit network make it an open system sensible to the external environment. Thus, service perturbations that affect public transport are both endogenous and exogenous.

Service perturbations may have a relevant impact on passenger flows, which are the ultimate output of transit assignment models, and vice versa, through two different mechanisms. On one (demand) side, route choice is influenced by the service regularity. On the other (supply) side, variations of vehicle and passenger arrival flows at stops induce variations in boarding and alighting loads, with recursion on dwell times. The relevance of both aspects is confirmed by theoretical and empirical studies.

In the following, first, the causes of the supply and demand randomness affecting the transit system are identified along with their impact on the service operation. Then, the models that allow coping with these phenomena are described.

7.4.1 Supply and demand uncertainties

A broad range of factors, both internal and external to public transport operation, can introduce unreliability in a transit system. The external factors are mainly related to uncertainties on the actual state of the road network required for the realization of transit supply. Public transport operators are faced with problems such as: work zones, road incidents, adverse weather conditions; but they also have to cope with the unavoidable

effects caused by the mixed use of roads, such as: congestion on shared transit lanes and presence of traffic signals. The internal factors are mainly related to uncertainties on the actual provision of the physical resources required to deliver transit services, either influenced by economical aspects such as, lack of crew and vehicles, or by poor production technologies such as deficient monitoring and information. Both produce the malfunctioning of system operations and hence can be revealed in indicators, such as irregular (not on-time) dispatching. These exogenous factors can implicitly or explicitly lead to stochastic values for the supply side characteristics of public transport, such as vehicle running time, vehicle departure time and headways.

However, additional uncertainties can be due to endogenous variations in boarding and alighting flows and dwell times. These phenomena, in addition to the passenger demand fluctuations, are also sensitive to vehicle characteristics and the type of technologies used in the system operation. For example, the method of fare collection can affect boarding times and their variability, while the application of a holding police can reduce the variance of headways.

The impact of a specific source of uncertainty on the reliability of different public transport modes may be significantly different. For example, in a rail transit line running on a fixed guideway, since the effect of congestion due to mixed traffic is limited, running times between stations can be treated as deterministic (non-random) parameters. On the other hand, such an assumption will not hold true in non-exclusive road lanes with day-to-day travel time and flow fluctuations.

Whatever the source of uncertainty, the service unreliability can be characterized in terms either of the deviation of vehicle arrivals from the schedule arrival times (for schedule based models), or of the variations of the vehicle headways (for frequency-based models). Figure 7.9 shows how the supply-side uncertainty and the demand-side uncertainty can both lead to service irregularity in a line-based analysis.

Poor on-time dispatching (i.e., departure from the terminal) caused by a malfunctioning in system operations associated with human errors or technical failures may result in running and dwell time uncertainty, due to the within-day dynamic nature of travel times, especially so if the transit routes share road space with other traffic. The variability in running times and dwell times may in turn cause an uncertainty on the number of vehicles actually available at the terminal to perform the next service runs, thus further affecting on-time dispatching.

Poor on-time dispatching by extent has an impact on vehicle headways for a given fleet size, thus causing uncertainty in arrival times at stops. The vehicle arrival times are also stochastic in nature, since running times and the dwell times are time-dependent and inherently random (e.g., delayed vehicle's door shutting and departure from the stop, road congestion and traffic lights).

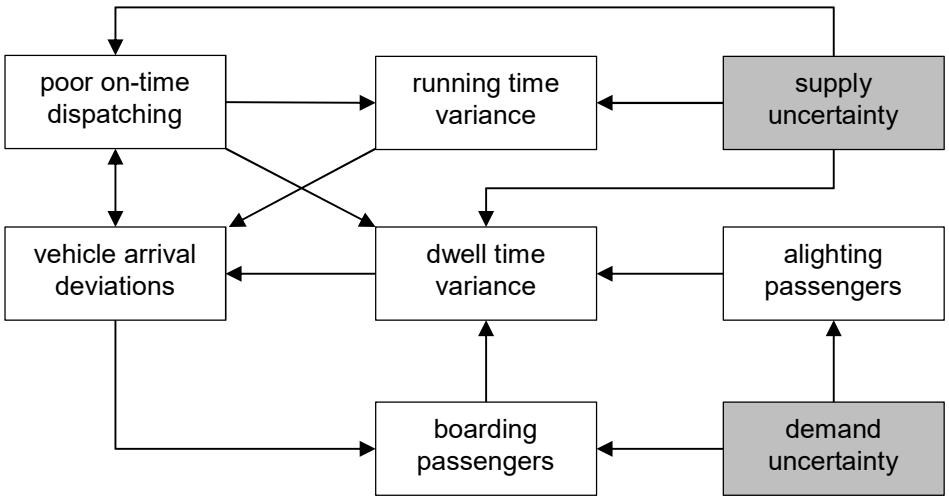

Figure 7.9. The flow diagram of the interactions between service irregularity and the uncertainties on both demand and supply-sides.

This interplay would increase in complexity by considering the demand-side uncertainties, as described in the following.

In addition to the supply-side uncertainty, the demand variability also can lead to service perturbations. The deviations of the vehicle arrival times may lead to the variability on the stock of passengers that have arrived at the stop and wish to board. Hence, the dwell time at a stop, as a function of both the boarding and alighting (and, in highly congested systems, on-board) flows of passengers, will inherit some variance. The alighting passengers have necessarily boarded the vehicle at preceding stops; which means that the dwell time at a given stop can be expressed as a function of the number of passengers boarding at that stop and at preceding stops.

For high frequency lines, since an excess in dwell time at a given stop (or in the next running time) generally leads to an increase in the number of passenger waiting at the following stop, then there is a relation between the dwell time at downstream stops, leading to the "bus bunching" phenomenon, as discussed later in detail. These relations can increase in complexity if the other intervening parameters, such as the vehicle capacity constraint, are included. Instead, in case of low frequency lines (e.g., headway of 15 minutes or more) passengers will arrive at the stop only a few minutes before the scheduled time and not continuously (like in the former case); thus the number of waiting passengers will not increase in case of a delay.

Any variability in the variables is thus to be considered here in the context of the service operation, rather than from the point of view of the passenger. For example, suppose a situation where traffic congestion is the only source of running time uncertainty; if congestion varies only on a day-to-day basis, then it can be assumed fixed within the analysis of a specific day. Thus, for a single day the running time variability cannot be accounted for as a cause of headway variation (or service irregularity) and the headway should be considered as constant. On the contrary, if the running times (traffic congestion) are assumed to vary within the analysis period (e.g., one or two-hour period), this will certainly lead to service irregularity (albeit without using control strategies or flexible fleet size).

Irrespective of the uncertainty sources, it is convenient to reflect the service irregularity on the basis of a probability distribution function $\varphi^h(h)$ for the inter-arrival times of successive vehicles at a particular transit stop, say headway h, or only on the basis of its statistical determinants, i.e., the frequency $f = 1/E(h)$ and the variation coefficient (square) $\sigma^2 = Var(h) / E(h)^2$, that are related to the mean and the standard deviation of the headway.

7.4.2 Distribution of boarding passengers and dwell times

This section investigates the interaction between demand uncertainty and service perturbations and more precisely how the headway irregularity can cause additional variation in the number of boarding passengers.

We can generally assume, if the headway is not too large, that the passenger arrival rate at a particular stop follows a Poisson distribution (typical of rare events) independent of the vehicle departure process. This clearly implies that passengers do not synchronize their arrival at the boarding stop to the line time-table, i.e., we are considering a frequency-based setting rather than a schedule-based setting, as the former is more appropriate in the case of irregular services.

Let $q = q_{\ell s}$ be the average rate of passenger arrivals (events) at stop $s \in S_\ell\text{-}S_\ell^+$ of line $\ell \in L$ (this is also equal to the flow on the corresponding waiting arc) and $h = h_{\ell s}$ be the constant (fully regular, for the moment) headway. The number n of boarding passengers accumulated during the headway will be a Poisson random variable with equal mean and variance (by definition, for Poisson variables):

$$E(n) = Var(n) = q \cdot h \,. \tag{7.83}$$

Let $\kappa = \kappa_\ell^{board}$ be the vehicle boarding capacity introduced in Section 5.1.2.5. It is assumed that the vehicle capacity constraint does not influence boarding, that $t_0 = t_\ell^{do}$ is the *door operation time*, and that the dwell time $t = t_{\ell s}^{dwell}$ at the stop is linearly dependent on the number of boarding passengers (for example, the alighting passengers can use other large doors):

$$t = t_0 + \frac{n}{\kappa} \,. \tag{7.84}$$

Subscripts and superscripts are here removed for the sake of simplicity.

Then, the expected value and the variance of the dwell time are, respectively:

$$E(t) = t_0 + \frac{E(n)}{\kappa} = t_0 + \frac{q \cdot h}{\kappa} \,, \tag{7.85}$$

$$Var(t) = \frac{Var(n)}{\kappa^2} = \frac{q \cdot h}{\kappa^2} = \frac{E(t) - t_0}{\kappa} . \quad (7.86)$$

The above variance measures only the effect of demand uncertainty on the dwell time distribution (and hence on service perturbations) for the case of constant headway. The relationships becomes more complex if the other intervening parameters are taken into account.

For example, suppose headway is also a random variable denoted by h, with a mean of $E(h)$ and a variance of $Var(h)$. The mean and the variance of the number of boarding passengers n can be calculated by conditioning on the headway (law of total variance), respectively, as follows:

$$E(n) = E_h(E(n|h)) = E(q \cdot h) = q \cdot E(h) , \quad (7.87)$$

$$Var(n) = E_h(Var(n|h)) + Var_h(E(n|h)) = E(q \cdot h) + Var(q \cdot h) = q \cdot E(h) + q^2 \cdot Var(h). \quad (7.88)$$

Compared to the case of constant (non-random) headway of Equation (7.83), the above variance of the number of boarding passengers has increased by $q^2 \cdot Var(h)$. One may refer to this as the effect of service perturbations on endogenous demand uncertainty. In fact, there is no change in the expected number of boarding passengers between Equations (7.83) and (7.87), while the variance increased as noted above. In a similar vein the statistical determinants of the dwell time can be calculated using (7.87) and (7.88) in the first Equations of (7.85) and (7.86):

$$E(t) = t_0 + \frac{q \cdot E(h)}{\kappa} , \quad (7.89)$$

$$Var(t) = \frac{q \cdot E(h) + q^2 \cdot Var(h)}{\kappa^2} . \quad (7.90)$$

Based on (6.64), we can also rewrite the dwelling variance (7.90) in terms of the service variation coefficient σ and, alternatively, of the line frequency $f = 1/E(h)$ or the expected dwell time $E(t)$:

$$Var(t) = \frac{q}{f \cdot \kappa^2} + \left(\frac{q}{f \cdot \kappa}\right)^2 \cdot \sigma^2 = \frac{E(t) - t_0}{\kappa} + \left(E(t) - t_0\right)^2 \cdot \sigma^2 . \quad (7.91)$$

Compared to the case of constant (non-random) headway of Equation (7.86), the above variance of the number of boarding passengers has increased by the second term on the right hand side of (7.91).

Clearly, the rate of passengers q attracted to a transit stop (called 'demand' here and assumed as a constant input) can itself be dependent on the inherent variability of the system characteristics in the context of an assignment model.

7.4.3 Emergence of headway irregularity and vehicle bunching

The previous sections outlined several inherent sources of uncertainty that affect transit operations. These sources include: dispatching time from the origin terminal, traffic congestion, delays at intersections, driver behaviour, travel demand and dwell time at stops. These stochastic factors are connected through the relation between the headway of successive vehicles, the number of waiting passengers and the dwell times, as well as the propagation of delays through the stop chain of the line itinerary. These interrelations result with a positive feedback loop that may cause the amplification of random variations.

This section illustrates the phenomenon where a vehicle running late picks up more passengers and hence is further delayed, while the succeeding vehicle progressively catches up; this process is called *pairing*, or *bunching*. In the following, we will formalise the mechanism underlying the formation of vehicle bunching.

Let us consider the case of a service, line $\ell \in L$, that has a relatively short planned headway of $h_\ell = 1/f_\ell$ (e.g., $h_\ell \leq$ 15min) and by extent assume a spontaneous (Poissonian) arrival of the passengers at a constant rate $q_s = q_{\ell s}$ at each stop $s \in S_\ell$.

For the sake of simplicity, stops and runs are referred here through their integer order in the sequence S_ℓ and R_ℓ , respectively.

The departure time θ_{rs} of run $r \in R_\ell$ from stop $s \in S_\ell$ is decomposed into the summation of riding times t_{ri}^{run} and dwell times t_{ri}^{dwell} of previous stops $i \leq s$ as follows.

$$\theta_{rs} = \sum_{i=1}^{s-1} t_{ri}^{run} + t_{r\,i+1}^{dwell} ; \quad (7.92)$$

The (departure) headway $h_{rs} = \theta_{rs} - \theta_{r\text{-}1\ s}$ at stop s between run r and the preceding run r-1 can be obtained, based on (7.92), as a function of the headway at a certain upstream stop $j < s$ as follows:

$$h_{rs} = \theta_{rs} - \theta_{r-1s} = \sum_{i=1}^{s-1} t_{ri}^{run} + t_{r\,i+1}^{dwell} - t_{r-1\,i}^{run} - t_{r-1\,i+1}^{dwell} = h_{rj} + \sum_{i=j}^{s-1} t_{ri}^{run} + t_{r\,i+1}^{dwell} - t_{r-1\,i}^{run} - t_{r-1\,i+1}^{dwell} \tag{7.93}$$

Let's introduce a new variable representing the relative difference between the actual headway and the fixed planned headway, called *headway deviation*:

$$\alpha_{rs} = \frac{h_{rs}}{h_\ell} - 1 . \tag{7.94}$$

As mentioned earlier, both supply and demand are subject to stochastic discrepancies. For example, an exogenous factor could lead to irregular dispatching from the first stop and result with an headway at the first stop that is different from the planned one. Moreover, traffic conditions, driver behaviour or irregular passenger activity at stops may yield running times between stops and/or dwell times at stops that are either shorter or longer than usual. These hence result with an actual headway at a some stop j that is longer or shorter than the planned headway, i.e., $\alpha_{rj} \neq 0$.

The following demonstrates how these initial exogenous discrepancies would then be further reinforced by the endogenous interactions between supply and demand. Let us consider the following conditions:

- expected dwell time of run r at stop s can be approximated, like in (7.84), as a linear function of the number of boarding passengers $E(n_{rs})$, so that $t_{rs}^{dwell} = t_0 + E(n_{rs}) / \kappa$;
- passengers' arrival at each stop s follows a Poisson process, so that the expected number of boarding passengers that waits at stop s for run r is: $E(n_{rs}) = q_s \cdot h_{rs}$;
- the preceding vehicle run followed the planned headway so that $\alpha_{r\text{-}1\ i} = 0$, $\forall i = j, \ldots, s$;
- running times between stops are assumed to be independent of headways and constant among runs;
- the passengers rate q_s at stop s is constant in time.

Under these conditions, the deviation of the headway at stop s from the planned headway can be expressed as a function of the headway deviation at an upstream stop j :

$$\alpha_{rs} = \frac{\alpha_{rj}}{\prod_{i=j+1}^{s} \left(1 - \frac{q_i}{\kappa}\right)} \tag{7.95}$$

We now prove the above expression.

Applying Equation (7.93) to two consecutive stops i and i+1 under the assumption of constant running time among runs we get:

$$h_{ri+1} = h_{ri} + t_{r\,i+1}^{dwell} - t_{r-1\,i+1}^{dwell} \tag{7.96}$$

Dividing each side by the planned headway h_ℓ and using $t_{rs}^{dwell} = t_0 + q_s \cdot h_{rs} / \kappa$ we get:

$$\frac{h_{r\,i+1}}{h_\ell} = \frac{h_{ri}}{h_\ell} + \frac{q_{i+1}}{\kappa} \cdot \frac{h_{r\,i+1}}{h_\ell} - \frac{q_{i+1}}{\kappa} \cdot \frac{h_{r-1\,i+1}}{h_\ell} \tag{7.97}$$

Finally, subtract -1 to both sided and recall that $h_{r\text{-}1\ i+1} = h_\ell$; using (7.94), after rearranging we get:

$$\alpha_{r\,i+1} = \frac{\alpha_{ri}}{\left(1 - \frac{q_{i+1}}{\kappa}\right)} . \tag{7.98}$$

Applying the above formula for $i = j$+1, …, s , by induction we get Equation (7.95):

$$\alpha_{rj+1} = \frac{\alpha_{rj}}{\left(1-\frac{q_{j+1}}{\kappa}\right)}$$

$$\alpha_{rj+2} = \frac{\alpha_{rj+1}}{\left(1-\frac{q_{j+2}}{\kappa}\right)} = \frac{\alpha_{rj}}{\left(1-\frac{q_{j+1}}{\kappa}\right)\cdot\left(1-\frac{q_{j+2}}{\kappa}\right)} \quad . \tag{7.99}$$

$$\alpha_{rj+3} = \frac{\alpha_{rj+2}}{\left(1-\frac{q_{j+3}}{\kappa}\right)} = \frac{\alpha_{rj}}{\left(1-\frac{q_{j+1}}{\kappa}\right)\cdot\left(1-\frac{q_{j+2}}{\kappa}\right)\cdot\left(1-\frac{q_{j+3}}{\kappa}\right)}$$

Note that the product in Equation (7.95) is smaller than 1 and positive like each one of its elements. Thus, it provides an amplification effect which increases with the number of stops and with the demand flow rate at each stop. If $\alpha_{rj} = 0$ then also $\alpha_{rs} = 0$ and the system is in equilibrium. In other words, the headway remains at the same level of downstream without amplification effects. If however $\alpha_{rj} > 0$ then $h_{rs} > h_{rj}$, while if $\alpha_{rj} < 0$ then $h_{rs} < h_{rj}$. Therefore, along a line an amplification of the variation can be observed which leads to the bunching effect.

Equation (7.95) also implies that the amplification rate is independent of the planned headway, and depends rather on the average dwell times.

Note that in reality the amplification rate is even higher. Indeed, the dwell time depends in practice also on the alighting loads, which are a fixed portion of the passenger on-board for each stop; but a late vehicle also accumulates more passengers on-board.

The effects of headway irregularity on transit line performance is twofold. Not only the variance of the dwell times is affected by the variability of headways, as shown in this Section, but also (and more important) expected wait times increase with it, as shown in Section 6.2.1. Reproducing service irregularity in frequency-based models for transit assignment is then primarily attained by properly defining the variation coefficients along the stops of each line.

ITS can greatly help in resolving headway irregularity issues and increase the reliability of services. For example, AVL data can be used to identify schedule discrepancies and vehicle bunching, which allows to suggest interventions for adjusting the planned time-table. A holding policy can also be implemented to control headways in real-time; when a vehicle is catching up the previous run, then the driver is invited to slow down along a run segment between stops or waiting a few more seconds before departing from a stop.

Although these technologies are readily available from the market, there is some resistance in drivers labour unions in implementing fleet control. Although the potential benefits are enormous, still a minority of transport operators exploit these crucial tools to the full extent.

7.4.4 Dwelling congestion

As shown in the previous sections, vehicle dwelling at stops is a phenomenon that can have a different impact on the service operation and on its quality perceived by passengers, depending on the transport system. In particular, metro and busses, that have more stops and frequent service, are more affected than trains and coaches.

By dwell time, we define the period a vehicle is immobilized at a station to allow passenger alighting and boarding. Independently to the transport system, vehicle dwelling is composed of a series of processes:

- doors opening after the vehicle is safely positioned at stop;
- passenger flow time (alighting and boarding);
- doors remain open without passenger flow;
- doors closing, safety control and vehicle departing.

The first and last processes are independent of the vehicle loads; they are linked to the door operation and can be regrouped into the door manoeuvre time. However, the intermediate processes are related to the passenger loads: boarding and alighting flows, as well as the vehicle on-board load and the stock of

travellers on the platform. Therefore, we can define the dwell time as a function of the passenger flow vector, which depends on the exchange capacity and the interface between vehicle and platform. Consequently the dwell times produce a connection between passenger volumes and service operations.

In the following we refer to the network model of Figure 6.6, in the context of a frequency-based assignment.

The dwell time of a vehicle is related to the flows of passenger alighting and boarding it at the stop. The capacity of doors gives the service rate of passengers that can get in and out the vehicle. Thus, the dwell time at stop $s \in S_\ell - S_\ell^- - S_\ell^+$ of line $\ell \in L$ can be assumed to depend on the ratio between the number of passengers alighting (arc *a*) and boarding (arc *b*) the vehicle (that are given by the corresponding flows divided by the line frequency) and the corresponding door capacity (or flow rates), as follows:

$$t_{\ell s}^{dwell}(q_a, q_b) = t_\ell^{do} + Max\left(t_\ell^{ab}, \frac{q_a}{\kappa_\ell^{alight} \cdot f_{\ell s}} + \frac{q_b}{\kappa_\ell^{board} \cdot f_{\ell s}}\right) \quad , a = \left(N_{\ell s}^{arr}, s\right) \in A^{alight}, b = \left(s, N_{\ell s}^{dep}\right) \in A^{board}, \tag{7.100}$$

where t_ℓ^{do} is the door operation time (which includes margins for safety control) and t_ℓ^{ab} is the minimum dwell time for alighting and boarding.

If doors for boarding and alighting are separate, we can instead assume the following expression:

$$t_{\ell s}^{dwell}(q_a, q_b) = t_\ell^{do} + Max\left(t_\ell^{ab}, \frac{q_a}{\kappa_\ell^{alight} \cdot f_{\ell s}}, \frac{q_b}{\kappa_\ell^{board} \cdot f_{\ell s}}\right), \tag{7.101}$$

where only the time for the most congested operation between boarding and alighting is considered and clearly the capacities are reduced accordingly.

The dwelling congestion is clearly non-separable.

The Transit Capacity and Quality of Service Manual (TCQSM, TRB, 2003) suggests a range of values for the parameters of Equations (7.100)-(7.101), depending on vehicle and service operating characteristics, i.e. on the transport system. For busses, the capacities take values in the range of 2.5-4.2 sec/pax for boarding and 2.1-3.3 sec/pax for alighting; for the rail and metro, values in the range of 1.38-3.97 sec/pass for boarding and 1.11-4.21 sec/pax have been observed.

In the first place, the door capacity has been considered to be fixed. However, this capacity can possibly be reduced by the effects of on-board and on platform overcrowding, since the difficulty of moving inside the carrier and exchanging loads between the vehicle and the stop increases with the loads of passenger.

In this case, we can multiply the dwell time (of arc *d*), or equivalently reduce the door capacities, by the following two BPR term:

$$1 + \alpha_\ell^{dwell} \cdot \left(\frac{q_d}{\kappa_\ell^{veh} \cdot f_{\ell s}}\right)^{\beta_\ell^{dwell}} + \alpha_\ell^{dwell} \cdot \left(\frac{\sum_{b \in s^+} q_b \cdot t_b}{\kappa_s^{stop}}\right)^{\beta_\ell^{dwell}} \quad , d = \left(N_{\ell s}^{arr}, N_{\ell s}^{dep}\right) \in A^{dwell} \tag{7.102}$$

where:

- $q_d / (\kappa_\ell^{veh} \cdot f_{\ell s})$ is the saturation rate of the dwelling vehicle (separable);
- the sum of the passenger flow q_b for each waiting arc *b* exiting from the stop *s* multiplied by the its expected time t_b yields the expected number of passengers waiting at the stop, κ_s^{stop} is the capacity of stop *s* and the ratio of the above two numbers yields the saturation rate of stop *s* (non-separable), like in Equation (7.51);
- α_ℓ^{dwell} and β_ℓ^{dwell} are the BPR coefficient and exponent for dwelling congestion (typical values are $\alpha_\ell^{dwell} = 1$ and $\beta_\ell^{dwell} = 2$).

Finally, consider that not all congestion phenomena can be well represented in a schedule-based model; indeed, service delays are inconsistent with the idea of a fixed timetable. In particular, dwelling congestion is the main internal effect influencing travel times and can thus not be represented in that framework.

Instead, simulation-based models for transit assignment, where the movement and interaction among individual vehicles and travellers is represented explicitly, allow to track each run of the line taking also into account the perturbations to the service timetable (e.g. due to dwelling congestion), given the dispatching from the first stop. In that framework it is also possible to simulate rules on how dispatching is modified if a vehicle to make the scheduled run is not available due to delays of other runs. This level of representation

enables the explicit modelling of passenger flows at stops as well as their impact on dwell times and service reliability. The stochastic and dynamic interaction between supply and demand can emulate the evolution of the headway variability along the line, which may results with the bunching phenomenon.

7.4.5 Impacts of dwell times on the service frequency

The mechanisms described in the previous sections capture the impact of the demand and supply variability on dwell times. Nevertheless, few approaches exist for handling the effects of passenger traffic and travel time variability on operation frequencies.

Three main frequency adaptation mechanisms are described in the following. In the first case, if the number of vehicles operating a line is fixed, an excess of dwell time and running time may condition the rate of vehicles passing at stops. In the second case, the maximum service provided is related to the berthing capacity of the station. In the third case, frequency in a dynamic setting can actually vary in time and along the line due to the within-day variability of running times and of dwell times (in particular), which are affected by time-varying flows; this issue has been already addressed in Section 6.4.5.

7.4.5.1 Fixed number of vehicles for each line

The service operation links the fleet size and the journey time of a line $\ell \in L$ to its service frequency.

Let's assume for the sake of simplicity that line ℓ is circular, i.e., the run time from the last stop takes the vehicle back to the first stop, while the terminal times are represented as dwell times.

The journey time t_ℓ^{cycle} of the vehicle to make a complete cyclic trip through all line stops S_ℓ and get back to the first stop, including the possible terminal times, is dependent on the traffic conditions and on the dwell time $t_{\ell s}^{dwell}$ of each stop $s \in S_\ell$. Based on Equations (7.100)-(7.101) the dwell time depends on the boarding and alighting flows as well as on the line frequency: $t_{\ell s}^{dwell}(q_a, q_b, f_\ell)$. Then the (cycle) journey time depends on the passenger flow vector $\mathbf{q}$ and on the line frequency f_ℓ :

$$t_\ell^{cycle}\left(\mathbf{q}, f_\ell\right) = \sum_{s \in S_\ell} t_{\ell s}^{run} + \sum_{s \in S_\ell} t_{\ell s}^{dwell}\left(q_a, q_b, f_\ell\right). \tag{7.103}$$

If the number N_ℓ of vehicles operating transit line ℓ is constant, the service frequency is equal to this number divided by the cycle journey time of the vehicles:

$$f_\ell = \frac{N_\ell}{t_\ell^{cycle}\left(\mathbf{q}, f_\ell\right)}. \tag{7.104}$$

The line frequency shall then be obtained by solving the above nonlinear equation for f_ℓ , which can be addressed as a fixed-point problem.

7.4.5.2 Limited stationing capacities of platforms

In the planning horizon, the fleet size is practically adjustable, while the scarce resource pertains to the node capacities of the support infrastructure (e.g. the stations of the rail network). Here, the platform berthing capacity is addressed as a scarce resource.

At any stop $s \in S$ a passing vehicle of line ℓ blocks the platform for a certain period, given by the dwell time plus an operating margin t_ℓ^{om} (mainly introduced for safety reasons). As already stated, the dwell time $t_{\ell s}^{dwell}(q_a, q_b, f_\ell)$ depends on the boarding and alighting flows as well as on the line frequency; this can be assumed null if the line does not serve the stop and passes without stopping. Given the set of lines L_s passing from the stop, the perceptual occupation α_s^{occ} of the platform is given by the sum for all such lines of the dwell time plus the operating margin multiplied by the line frequency (assuming that times and frequency are expressed in consistent units, e.g. h and 1/h):

$$\alpha_s^{occ}\left(\mathbf{q}, \mathbf{f}\right) = \sum_{\ell \in L_s} f_\ell \cdot \left(t_{\ell s}^{om} + t_{\ell s}^{dwell}\left(q_a, q_b, f_\ell\right)\right). \tag{7.105}$$

Clearly, this perceptual occupation cannot be greater than one; therefore, if the platform does not suffice to accommodate all lines, then the conflicting frequencies shall be reduced proportionally, starting from a given

desired value f_ℓ^{des}, to satisfy the capacity constraint:

$$f_\ell = Min\left(1, \frac{1}{\alpha_s^{occ}(\mathbf{q},\mathbf{f})} : \forall s \in S\right) \cdot f_\ell^{des} . \quad (7.106)$$

This problem result is an equilibrium among lots of stops and lines. The line frequency shall then be obtained by solving the above nonlinear system of equations for **f** , which can be addressed as a fixed-point problem.

Note that this model explains only part of the node performance (that connected to platform capacity), as it does not considers other relevant aspects of service operation in rail stations, such as the management of track conflicts.

7.4.6 Reliability and robustness

Reliability and robustness are key performance indicators of public transport services. These qualities are deemed crucial by travellers and directly affects their mode choice, hence supporting modal shift from car to transit. However, disruptions and breakdowns can never be completely avoided in public transportation services. Moreover, congestion on road and rail networks is continuously increasing. All these cause in many cases relevant delays.

Planners should try to minimize the negative impact of these unavoidable delays on both the service quality for the passengers and the service costs for the operators. Furthermore, transport authorities started now to realise that the most important effect of congestion is not that average travel times increase, but that travel times become highly unreliable, i.e. robustness and reliability are at least as important as efficiency. That insight should now be translated more and more into research and practice in order to redesign transport systems and make these more attractive to passengers.

7.4.6.1 Definitions of reliability

Reliability can be defined as the ability of an item to perform a required function, under given environmental and operational conditions and for a stated period of time. In statistical terms, reliability is the probability that a system, possibly consisting of many components, will function correctly.

On this basis, three indicators have been defined to evaluate the reliability of transportation systems:

- Connectivity reliability considers the probability that a pair of nodes in a network remains connected. A special case of this index is the terminal reliability that is concerned with the existence of at least one path between each origin-destination (O-D) pair.
- Capacity reliability refers to the probability that the network capacity can accommodate a certain travel demand at a required level of service.
- Travel time reliability is defined as the probability that a trip between a given O-D pair can be completed successfully within a specified time interval.

The first two indicators can be referred to the supply-side while the third can be basically referred to the demand-side of transportation. They have been mostly defined to assess road network performance under uncertainty. But transit systems have specific attributes that differentiate their assessment from private transport systems:

1. Vehicles depart from stops with scheduled headways, leading to wait times for the passengers.
2. Capacity of vehicles (or the seat capacity) is limited, and therefore some passengers may fail to board the first arriving vehicle (or may fail to get a seat) at the stop.
3. Passengers may have to transfer to another line(s) to complete a single trip,
4. Passengers have to walk to transit stops.

Items 1 and 2 have motivated researchers to define a number of reliability indicators that are different from the abovementioned three general indicators. The adherence to the scheduled arrival or departure of vehicles affects the arrival pattern of passengers at stops and hence affects their wait time probability distribution. From passengers' viewpoint, the following two questions related to wait times may be arisen with

respect to the service reliability:

- How much is the service punctual?
- Does the service arrive at the stop regularly?

For high frequency services, e.g. with headways shorter than 10-15 min, where passengers tend to arrive at stops randomly instead of coordinating with vehicle arrivals even if the timetable is published, the headway regularity would be more important, and, therefore, an indicator accounting for headway variability may be more appropriate for assessing the service reliability. On the other hand, in case of less frequent services, the degree of punctuality may better represent the service reliability. The key difference between these two concepts can be illustrated by the following example: if a transit service is systematically two minutes late the punctuality is poor while the regularity is perfect.

Several probability-based indicators can be defined to assess the punctuality or regularity of a service, e.g.:

- the percentage of services arriving on-time;
- the percentage of services arriving more than 5 min late;
- the average of percentage deviations from the mean headway.

Furthermore, besides using data from observations for a post assessment, indicators can be calculated from simulation results in order to predict the service reliability in advance. This implies introducing random variables, or at least standard deviations, at the level of service operation, with particular reference to headways. In such context other indicators of reliability can be defined where the perspective of the passenger is also taking into account, such as the probability that:

- journey travel times are less than a given threshold;
- line headways at stops are larger than a given threshold;
- passenger wait times are less than a given threshold;
- each passenger can board the first arriving vehicle at stops;
- each passenger can get a seat when boarding at stops.

7.4.6.2 Definitions of robustness

In general, the robustness of a service is how well it performs in practice, under realistic and thus uncertain circumstances. This is more than requiring that the system will work in practice (like reliability).

However, when designing a public transport service, a more specific definition of robustness is required. Actually, many different definitions of robustness exist; the classical one focuses on minimizing the effects of disruptions and delays.

A first definition of robustness is about schedule adherence after disruptions. This is closely related to defining robust a service where the propagation of delays is prevented as much as possible.

A second definition requires that the schedule should remain free of conflicts (a conflict occurs when two vehicles request to use the same platform at the same time), even in the worst-case scenario. In order to accommodate delays, a lot of buffers will be needed in the timetable.

A third definition tries to bridge the typical gap between timetabling and dispatching. While scheduling, the recovery strategies should be taken into account explicitly since robustness is achieved if the timetable does not cause conflicts during the execution.

A fourth definition concentrates on minimizing missed transfers. The idea is that all connections should be guaranteed as long as the delays are limited to a certain amount.

Unfortunately, all these definitions ignore the efficiency of the system and the first three also ignore the passenger perspective. When only the reliability is cared about, then inserting a high number of long buffer times in the timetable would make a system robust. Nevertheless, passengers and operators would obviously not be satisfied, since travel times and production resources increase.

Recently, the passenger perspective became more important when robustness is discussed. A simple way to achieve that is by using passenger loads as weights, when the delays of vehicles are evaluated. A more sophisticated way to consider the passenger's perspective is to minimize the total travel time of all users 'in

practice', i.e. considering the missed transfers and not just the planned travel times.

Focussing on the actual travel times automatically leads to a trade-off between the classical interpretation of robustness (reliability) and the efficiency of the system. As a result, the included buffer times will not be too short, because then a lot of conflicts will occur and passengers will miss many transfers, but the buffer times will also not be too long, since this would directly yield too longer travel times. Therefore, this definition of passenger robustness is comprehensive and embraces most of the classical ones.

7.4.6.3 Strategies to obtain robustness

Here, a number of strategies to obtain robustness are discussed.

In order to obtain a reliable system and to guarantee an attractive service passenger robustness, as defined above, should be put forward in every stage of network design.

This starts with the design of the infrastructure. For buses, this leads to separate lanes in congested areas and getting priority at traffic lights. For rail based transit, this involves providing sufficient capacity and alternative tracks when something goes wrong.

Also during the planning of line itineraries, robustness should be considered. This is certainly not common practice yet. Nevertheless, decisions made in this stage can significantly influence the robustness of the system. For instance, when the length of the lines is decided: obviously, the service on longer lines has a higher chance of being perturbed and these disturbances have a larger influence. Furthermore, the number of passengers that will require a transfer is decided in this stage; shorter lines imply more transfers (which is bad for efficiency), but may produce less missed transfers (which is good for reliability).

For the planning of timetables, many different approaches have been developed in order to obtain passenger robustness. All these methods intend to optimize the size and the position of buffer times in the schedule. In this way, the propagation of delays should be minimized and vehicles should get some time to recover from delays. In this context, it is actually better to make an explicit distinction between buffer times and time supplements. Time supplements are added to nominal running and dwelling times in the timetable in order to give vehicles more possibilities to arrive/depart on time. Therefore, supplements are directly included into the (planned) travel times of passengers. Buffer times are instead scheduled between two vehicles using the same part of the network (e.g., a track or road, a platform or stop, etc.) and are not included into the travel time of passengers, except for transfer and waiting times, but affect the capacity of the infrastructure. Both supplements and buffers will also be limited by the objective to minimize the passenger travel times.

Naturally, also when making (local) dispatching decisions passenger robustness should be strived for. This becomes especially for synchronization: when a high number of passengers needs to transfer from vehicle 1 to vehicle 2, a dispatcher will decide if and how long the departure of vehicle 2 will be delayed when vehicle 1 is expected to arrive late. Sufficiently delaying vehicle 2 may guarantee that no passenger will miss his transfer. At the same time, passengers already in vehicle 2 will be delayed, and vehicle 2 might also delay other vehicles later or generate conflicts.

Obviously, avoiding disruptions by appropriate maintenance strategies is also important when striving for robustness.

7.4.7 Reference notes and concluding remarks

7.4.7.1 Boarding passengers and travel times

The variability of boarding passengers independent of the vehicle departure process has been studied by Holroyd and Scraggs (1966), van Oort and van Nes (2009). It is generally assumed that, if the headway is not sufficiently large, the passenger arrival rate at a particular stop follows a Poisson distribution.

Numerous studies have been conducted to investigate the effect of sub-hourly variations in running times on the headway variation, e.g. Osuna and Newell (1972) and Adebisi (1986). Similar studies asses the importance of dwell times on the transit service, e.g. Vuchic (2006) and Lai et al. (2011). All these analysis do not introduce explicitly a direct connection to the passenger flows.

Lin and Wilson (1992) consider dwell times critical for determining the system performance and the quality of

service. They identify three direct effects: the dwell time directly affects the vehicle cycle time; at the stop level a dwelling vehicle occupies the platform obstructing the following vehicles; and the dwell time is believed to be a major factor for headway variability and vehicle bunching. These effects are also thoroughly discussed in the Transit Capacity and Quality of Service Manual (TRB, 2003).

The bouncing model presented in Section 7.4.3 is an original contribution of this book.

7.4.7.2 The dwelling process

Whereas Equation (7.100) is used widely by many practitioners, various types of dwell time functions can be found in the literature. Each research is focused on the influence of a particular phenomenon on the capacities and dwell time. We here list some principal findings.

Firstly, it is generally agreed that a positive correlation exists between the number of boarding and alighting passengers and the dwell time, which gives rise to an equilibrium probelms (Bellei et al., 2000; Babazadeh and Aashtiani, 2005). Second, the dwell time determinants are influenced by various sources of variability, either positively or negatively. They are negatively affected by congestion factors, such as the platform crowding and by the in-vehicle load (Fritz, 1983; Aashtiani, 2002). They are further influenced by physical factors, such as the vertical gap between the vehicle and the platform and the door width (Fernandez et al., 2011). Particularly, the boarding flow rate depends on the operation characteristics, such as front door boarding for buses, and the type of fare control mechanism (TRB, 2003; Fernandez et al., 2010). Third, according to Harris (2005), the boarding and alighting capacities are not constant throughout the same boarding/alighting group, but they vary according to the passenger's position in that. In fact, the fastest alighting rates are detected on the early exiting passengers, while the fastest boarding rates on passengers in the middle of the group. Finally, Szplett and Wirashinghe (1984) show that the distribution of passengers on a platform is not uniform, but depends on the position of entry and exit points, and the dwell time is subject to the flow of the door with the maximum utilization.

The main approach for calculating vehicle's dwell times is by making a statistical analysis of an appropriate dataset in order to determine a suitable function that fits the records, while establishing a set of significant attributes. The data collection methods continuously evolve, but we can distinguish the field observation surveys (Lin and Wilson, 1992; TRB, 2003), the automatic passenger counters (Rajbhandari et al., 2003), the field experiments (Harris, 2005) and the laboratory experiments (Fernandez et al., 2010).

An alternative approach for the estimation of the dwell time is the use of pedestrian micro-simulators to model alighting and boarding passengers, such as the cellular automata (Zhang et al., 2008). By defining the behaviour of passengers against obstacles and attractions at an individual level, the simulation allows to reproduce a range of complex phenomena that emerge at a macroscopic level on the platform, during the vehicle's dwelling. This way, the effect of numerous infrastructure set-up and rolling stock compositions can be tested.

7.4.7.3 Reliability and robustness indicators

Ceder (2007) classified different indicators associated with reliability problems from different viewpoints (i.e., planning indicators, operational indicators and maintenance indicators).

Classical reliability indicators have been introduced by Iida and Wakabayashi (1989) and Asakura (1996).

The distinction between punctuality and regularity and their determinants is discussed in Okrent (1974), Bowman and Turnquist (1981), Carey (1999), van Oort and van Nes (2009).

Several authors have made efforts to link reliability indicators to the result of assignment and simulation models with the aim of taking into account the passenger perspective e.g., Yin et al. (2004), Chen et al. (2009), Babaei et al. (2013).

Various definitions of robustness and comparisons between them, as well as some examples of how to obtain robustness in practice, are discussed in, among others, Goverde (2005), Kroon et al. (2008), Schobel et al. (2009), Cicerone et al. (2009), Van Oort (2011) and Dewilde et al. (2011; 2014).

The key objective remains how to improve these indicators through correct management of transit services (Abkowitz M., Tozzi J., 1987). Monitoring and management can be greatly enhanced today thanks to AVL data (El-Geneidy et al., 2011).

7.4.7.4 Traffic assignment with supply variability

A number of transit assignment models have been developed to account for the uncertainty of vehicle arrivals.

Introducing headway variations in frequency-based assignment models is a key element in reproducing service irregularity, as it was repeatedly shown so far. One limitation of all models presented in this book lays in the fact that service headway distributions at stops are assumed independent among different lines, which of course is not often true in reality. Shimamoto et al. (2010) developed an assignment model that takes into account the correlation between vehicle arrivals of different lines. Wait times and flows are sampled from a normal distribution with a correlation matrix that is a function of the number of boarding and alighting passengers.

Yang and Lam (2006) introduced a reliability-based assignment model to congested transit network to simulate unreliable services. Szeto et al. (2011) formulate as a Nonlinear Complementarity Problem a risk-aversive transit assignment in which in-vehicle travel time, waiting time and capacity are considered as stochastic variables; both their means and variances are incorporated into the formulation.

In schedule-based models for transit assignment the representation of individual trips enables to account for the temporal distribution of reliability problems. Initially, schedule-based models were developed based on the assumption that vehicles run with perfect punctuality and hence considered arrival and departures times to be deterministic. Service irregularity can be modelled either implicitly by adding a random term to the perceived utility function (e.g., Nielsen, 2004) or explicitly by simulating vehicle runs and dwell time as interdependent random variables. The latter was used in stochastic schedule-based model developed by Nuzzolo et al. (2001). Huang and Peng (2002) developed a path choice models for transit systems that include various stochastic processes, such as the departure time, the travel time and the probability to make a successful transfer.

The simulation-based approach to transit assignment (Cats, 2011) can support the modelling of various sources of service uncertainty – traffic conditions, dispatching regime from the terminals, dwell time at stops and their relationship with passengers' flows. The explicit modelling of these processes within a dynamic simulation of transit operations will contribute to a more realistic reproduction of supply uncertainty, compared with introducing independent stochastic processes referring to separate system elements. Emulating the dynamics of these sources enables to analyse their impacts and potential methods to prevent them. In particular, the bunching problem arises from the interaction between supply and demand variability and could be therefore captured by simulating individual vehicles and travellers and how they move throughout the network. This allows mimicking the way in which system reliability evolves over time and escalates along the route. Furthermore, the impact of service perturbations could be embedded into the dynamic route choice model.

7.4.7.5 Variable service frequencies

Few approaches exist for handling the effects of passenger and vehicle traffic on operation frequencies.

If the vehicle fleet is fixed, the vehicle cycle time determines the frequency of the transit services, which become a variable of the equilibrium model. In Bellei et al. (2000) the assumption that the frequency of the transit line is fixed is relaxed under the consideration that the number of passengers boarding and alighting will influence the dwell time. In this line of research Lam et al. (2002) propose a stochastic model for frequency-based assignment. In addition, Meschini et al. (2007) propose a dynamic assignment model with dynamic propagation of the line frequency.

Harris (2005) and Harris and Anderson (2007) consider the dwell time at stations and the occupation of the station platform as the critical factor for determining the performance and the capacity of high duty guided lines (metro and commuter rail), where signaling take a relevant role (Lai et al., 2011). This effect is treated by the restrained frequency model, which is introduced by Leurent et al. (2011).

7.4.7.6 List of references

Abkowitz M., Tozzi J. (1987) Research contributions to managing transit service reliability. Journal of Advanced Transportation 21, 47-65.

Adebisi O. (1986) A mathematical model for headway variance of fixed-route buses. Transportation Research B 20, 59-70.

Aashtiani H., Iravani H. (2002) Applications of dwell time function in transit assignment model. Transportation Research Record 1887, Paper 02-3498.

Asakura Y. (1996) Reliability measures of an origin and destination pair in a deteriorated road network with variable flows. In Transportation Networks: Recent Methodological Advances, ed. M.G.H. Bell, Pergamon Press, Oxford, 273-288.

Babaei M., Schmocker J-D., Shariat-Mohaymany A. (2014) The impact of irregular headways on seat availability. Transportmetrica A 10, 483-501.

Babazadeh A., Aashtiani Z.H. (2005) Algorithm for equilibrium transit assignment problem. Transportation Research Record, 1923, 227-235.

Bellei G., Gentile G., Papola N. (2000) Transit assignment with variable frequencies and congestion effects. Proceedings of the 8th Meeting of the EURO Working Group on Transportation, Rome, Italy.

Bowman L.A., Turnquist M.A. (1981) Service frequency, schedule reliability and passenger wait times at transit stops. Transportation Research A 15, 465-471.

Carey M. (1999) Ax ante heuristic measures of schedule reliability. Transportation Research B 33, 473-494.

Cats O. (2011) Dynamic modeling of transit operations and passenger decisions. PhD Thesis, KTH Royal Institute of Technology, Stockholm, Sweden.

Ceder A (2007) Public Transit Planning and operation: Theory, modeling and practice. Butterworth-Heinemann, Oxford, UK.

Chen X., Yu L., Zhang Y., Guo J. (2009) Analyzing urban bus reliability at the stop, route and network levels. Transportation Research A 43, 722-734.

Cicerone S., D'Angelo G., Di Stefano G., Frigioni D., Navarra A., Schachtebeck M., Schöbel A. (2009) Recoverable robustness in shunting and timetabling. In Robust and Online Large-Scale Optimization: Models and Techniques for Transportation Systems, ed.s R.K. Ahuja, R.H. Möhring and C.D. Zaroliagis, Lecture Notes in Computer Science, 28-60.

Dewilde T., Sels P., Cattrysse D., Vansteenwegen P. (2011) Defining robustness of a railway timetable. Proceedings of 4th International Seminar on Railway Operations Modelling and Analysis, Rome, Italy, 1-20.

Dewilde T., Sels P., Cattrysse D., Vansteenwegen P. (2014) Improving the robustness in railway station areas. European Journal of Operational Research 235, 276-286.

El-Geneidy A.M., Horning J., Krizek K.J. (2011) Analyzing transit service reliability using detailed data from automatic vehicular locator systems. Journal of Advanced Transportation 45, 66-79.

Fernandez R. (2011) Experimental study of bus boarding and alighting times. Proceeding of ETC 2011, Glasgow, UK.

Fernandez R., Zegers P., Weber G., Tyler N. (2010) Influence of platform height, door width and fare collection on bus dwell time: Laboratory evidence from Santiago de Chile. Proceedings of TRB Annual Meeting, Washington DC.

Fritz M. (1983) Effect of crowding on light rail passenger boarding times. Transportation Research Record 908, 43-50.

Goverde R.M.P. (2005) Punctuality of railway operations and timetable stability analysis. Ph.D. thesis, Delft University of Technology, Delft, The Netherlands.

Holroyd E.M., Scraggs D.A. (1966) Waiting times for buses in Central London. Traffic Engineering and Control 8, 158-160.

Huang R., Peng Z.R. (2002) Schedule-based path-finding algorithms for transit trip-planning systems. Transportation Research Record 1783, 142-148.

Harris N.G. (2005) Train boarding and alighting rates at high passenger loads. Journal of Advanced Transportation 40, 249-263.

Harris N.G., Anderson R.J. (2007) An international comparison of urban rail boarding and alighting rates. Proceedings of the Institution of Mechanical Engineers. Journal of Rail and Rapid Transit 221, 521-526.

Kroon L., Maroti G., Retel Helmrich M., Vromans M.J.C.M., Dekker R. (2008) Stochastic improvement of cyclic railway timetables. Transportation Research B 42, 553-570.

Iida Y., Wakabayashi H. (1990) An approximation method of terminal reliability of road network using partial minimal path and cut set. Proceedings of the 5th WCTR, 367-380.

Lai Y.C., Wang S.H., Jong J.C. (2011) Development of analytical capacity models for commuter rail operations with advanced signaling systems. Proceedings of the 90th Annual Meeting of Transportation Research Board, Washington DC.

Lam W.H.K., Zhou J., Sheng Z.-H. (2002) A capacity restraint transit assignment with elastic line frequency, Transportation Research B, 36, 919-938.

Lin T.-M., Wilson N.H.M (1992) Dwell time relationships for light rail systems. Transportation Research Record 1361, 287-295.

Leurent F., Chandakas E., Poulhes A. (2011) User and service equilibrium in a structural model of traffic assignment to a transit network. Procedia - Social and Behavioral Sciences 20, Proceedings of EWGT2011, 495-505.

Meschini L., Gentile G., Papola N. (2007) A frequency based transit model for dynamic traffic assignment to multimodal networks. Proceedings of the 17th International Symposium on Transportation and Traffic Theory (ISTTT), ed.s R. Allsop, M.G.H. Bell, and B.G. Heydecker, Elsevier, London, 407-436.

Nielsen O.A. (2004) A large-scale stochastic multi-class schedule-based transit model with random coefficients. In Schedule-based Dynamic Transit Modeling: theory and applications, ed.s N.H.M Wilson and A. Nuzzolo, Kluwer Academic Publisher, 53-78.

Nuzzolo A., Russo F., Crisalli U. (2001) A doubly dynamic schedule-based assignment model for transit networks. Transportation Science 35, 268-285.

Okrent M.M. (1974) Effects of transit service characteristics on passenger waiting time. Master Thesis, Northwestern University, Department of Civil Engineering, Evanston, Illinois.

Osuna E.E., Newell G.F. (1972) Control strategies for an idealized public transportation system, Transportation Science 6, 52-72.

Rajbhandari R., Chien S., Daniel (2003) Estimation of bus dwell time with automatic passenger counter information. Transportation Research Record 1841, 120-127.

Shimamoto H., Kurauchi F., Schmöcker J.-D. (2010) Transit assignment model incorporating the bus bunching effect. Proceeding of 12th World Congress on Transport Research, Lisbon, Portugal.

Schobel A., Kratz A. (2009) A bicriteria approach for robust timetabling. In Robust and Online Large-Scale Optimization: Models and Techniques for Transportation Systems, ed.s R.K. Ahuja, R.H. Möhring and C.D. Zaroliagis, Lecture Notes in Computer Science, 119-144.

Szplett D., Wirasinghe S.C. (1984) An investigation of passenger interchange and train standing time at LRT Stations: Alighting, boarding and platform distribution of passengers. Journal of Advanced Transportation 18, 1-12.

Szeto W.Y., Solayappan M., Jiang Y. (2011) Reliability-based transit assignment for congested stochastic transit networks. Computer-Aided Civil and Infrastructure Engineering 26, 311-26.

TRB (2003) Transit Capacity and Quality of Service Manual. On-line report prepared for the Transit Cooperative Research Program.

Van Oort N. (2011) Service reliability and urban public transport design. Ph.D. thesis. Netherlands TRAIL Research School, Delft, The Netherlands.

Van Oort N., Van Nes R. (2009) Regularity analysis for optimizing urban transit network design. Public Transport 1, 155-168.

Vuchic V.R. (2006) Urban transit: operations, planning and economics. Wiley, New York, USA.

Yang L., Lam W.H.K. (2006) Probit-type reliability-based transit network assignment. Transportation Research Record 1977, 154-163.

Yin Y., Lam W.H.K., Miller M.A. (2004) A simulation-based reliability assessment approach for congested

transit network. Journal of Advanced Transportation 38, 27-44.

Zhang Q., Han B., Li D. (2008) Modeling and simulation of passenger alighting and boarding movement in Beijing metro stations. Transportation Research C 16, 635-649.

7.5 Fares

Authors: David Watling, Guido Gentile, Klaus Noeke, Michael Florian

It is important to first appreciate the sheer complexity of dealing with public transport fares, if aiming to represent all important facets. There exist many different ways of paying for public transport, e.g.:

- walk-on single/return fares,
- advance fares, perhaps determined by some yield-management approach (such as in airlines),
- daily or monthly passes,
- family or group discount tickets,
- multi-trip tickets,
- smart-cards with a maximum daily fare,
- etc. …

In addition, some fares give different levels of flexibility in terms of services that can be used. In a real network there will likely be a mix of people paying fares in different ways. As well as different mechanisms for paying, there will be different fare levels for different types of traveller, e.g. concessionary fares for elderly or disabled people, or for young people and students. In addition, in some cities there may be a mix of kinds of service, including different qualities of service, possibly operated by different companies, and these may be priced differently (e.g. express/air-conditioned buses versus regular buses). There will also exist different abilities/willingness to pay for a given fare, as might be reflected in different values of time.

7.5.1 The question of whether fares need to be included

Unlike travel time, waiting time, discomfort, failure-to-board, etc., it is more difficult to associate some of the fare structures described above with a particular trip. For example, even if we knew that someone made *n* trips using a certain pass, do they really associate 1/*n* times the cost of the pass with each trip when making choices?

For modelling the demand for public transport we may wish to explicitly consider how demand varies according to these different types of ticket and segmentations of the population, or at the other extreme to aggregate all the possibilities into an average fare per passenger journey, as two possible treatments of this problem.

On the other hand, given our focus on modelling route choice, it will be the case in many situations that we can justify neglecting fares, since the fare paid will be invariant to the route chosen, especially if we are considering networks with a single fare structure for all transport modes, or a network with a single dominant public transport mode. Even if in some cases the fare on a particular origin-destination movement may vary with the route chosen, if this happens relatively rarely then we might justify neglecting fares as an approximation. This pragmatic situation is the one commonly adopted in practice, and is summarised well by the guidance from the UK Department for Transport (DfT, 2007):

‘ Fares need not be included in the assignment, provided that they do not influence route choice; matrices of fares can be added to the generalised cost after the assignment and before passing cost matrices to a demand model or appraisal package. Where fares can influence route choice then it is essential to include them in the assignment. It is accepted that the complexity of some fare systems may prevent them from being represented exactly in the assignment model, but the model representation needs to be *acceptable*. Acceptability can be gauged from whether the assignment model validates or not. ‘

Therefore, a key first question is whether fares need to be included at all in the assignment stage, since in many cases the routes chosen will not be sensitive to the fare levels.

Particular cases in which fares may need to be included are where there are multiple types of public transport modes with different fare levels, or where there are a significant number of cases in which a single origin-destination trip may include combinations of different kinds of transport modes. In these cases, it is difficult to make the separation between the demand for each type of transport mode and the route choice for each mode, since the choice of mode type and route are inter-related. Having said this, there are many other complexities in dealing with combined modes which mean that even in such cases an explicit consideration of fares is often not a high priority. However, it is not so rare to find real-life examples in which it seems more difficult to justify the approximation of neglecting fares. Such a case is in which high and low quality modes

may offer competing options on the same corridor, the high quality mode typically being faster, more comfortable but also with a higher fare and perhaps less frequent.

The remarks given above are general ones in that they are not specific to a particular modelling approach; in particular, they apply equally to frequency-based and schedule-based approaches. In the next section, then, we consider the particular considerations for each type of approach.

7.5.2 Transit route choice including fares

So far in Part 3 we assumed that route choice models of any kind work on a graph in which arcs are labelled with generalised cost. We assumed generalised cost to be given by the monetization of travel times and discomfort, plus monetary costs, i.e.: fares.

This poses a challenge, because according to Section 5.2.1.6 a wide range of fare schemes exist in practice. Only some of them are additive in the sense that the total fare for a complete trip can be found by summing a line segment attribute over all segments of the trip. If this is the case, then the line segment attribute can be incorporated into the arc generalised cost and will take effect in route choice.

Many fare schemes are not additive, however, including simple schemes like distance-based and zone-based fares with degressive fare amounts. Here the fare amount is an attribute of the complete path and cannot be broken down to arc level.

For the schedule-based models of Section 6.3 this does not pose a problem, because the evaluation proceeds in three distinct steps:

1. Search paths
2. Calculate generalised cost for each path
3. Split demand between paths

The three steps are carried out sequentially, and conceptually we can assume that the calculation is done separately for each OD pair. Step 1 returns complete paths from O to D. Therefore, step 2 can apply any fare model, however complex. With degressive fare tables, we know exactly how long the complete trip is, or how many zones are traversed. At the end of step 2 we have a choice set with complete paths and exact generalised costs per path, and we can apply a choice model in step 3.

Section 6.2 and Section 7.1 described that frequency-based route choice models are evaluated differently, because the choice set does not consist of individual paths, but of hyperpaths or strategies. A single pass over the network backwards from a destination towards all origins combines all three steps (search, cost calculation, split), and for all OD pairs with a given destination. During the pass node labels are computed which represent expected cost from an intermediate node to the destination, and this conflicts with fares which are only defined at complete path level.

7.5.2.1 Application to the example network

Consider the example network from Section 5.13 emended by the inclusion of Line 5 – Purple and Stop 5, as in Section 7.1.5. We define a degressive, distance-based fare scheme as follows. A single ticket applies to the complete trip. Each line segment has a distance of 1. The fare for a total distance of 1 costs 2 units, any longer distance costs 3 units.

Recall the Optimal Strategies algorithm or its generalizations. The calculation proceeds backwards from the destination to all origins and sets labels at all intermediate nodes. These labels represent the expected cost from the node to the destination, the assumption being that this cost is the same for all passengers waiting for a service at this node, regardless of their origin. But is it?

We focus on the node label updates for Stop 2, while computing route choice towards destination Stop 4. These passengers have a choice of travelling via Stop 3 (various possible hyperpaths, distance 2, fare 3) or via Stop 5 (distance 1, fare 2). The monetary component favours the path via Stop 5, and will influence route shares. Now consider a passenger from Stop 1 who gets off the Line 2 at Stop 2 and evaluates transfer options. Any route from 1 to 4 via 2 will have a distance of at least 2, so the fare will always be 3. In this case the monetary component is neutral at Stop 2. We should therefore expect route shares to be different

between passengers originating or transferring at Stop 2.

We apply the case with complete information from Section 7.1.5. The second column of Table 7.12 repeats the last column of Table 7.6, which ignore fares.

The third column shows how route choice changes, if travellers from stop 1 consider fare. A value of time of 20 units/h is assumed. The red line now attracts a much higher proportion of travellers because it costs only 2 units, compared to 3 units for all other paths. Travellers who choose the green line still get off at Stop 2 and split between the maroon and purple lines according to the same shares as before. These results were produced setting fictitious arc costs so that they sum up to exact fares for origin Stop 1.

We now run the algorithm with fictitious arc costs set up to sum to exact fares from Stop 2. The fourth column shows route choice for travellers originating from Stop 2. They also choose between the maroon and purple lines, but the shares reflect the fact that for them the purple line is cheaper.

Table 7.12. Line shares (%) with and without the effect of fares.

fares	ignored	exact solution	exact solution
origin of travelers	Stop 1	Stop 1	Stop 2
volume Line 1 - Red	40	81	0
volume Line 2 - Green	60	19	0
volume Line 3 - Maroon	28	9	40
volume Line 4 - Black	0	0	0
volume Line 5 - Purple	32	10	60

7.5.2.2 The relevance of approximations

The experiment demonstrates that non-additive fares indeed lead to route shares at intermediate nodes which differ by origin stop. Unfortunately this implies that exact route choice for all origins cannot be computed in a single application of the algorithm, because the arc costs differ by origin. If an exact solution is essential, the algorithm needs to be run separately for each origin. This, of course, increases run time by a factor equal to the number of origins and may not be feasible.

Practical alternatives use approximations. The simplest approximation seeks to assign costs to arcs which reflect the effect of fares on average. More complex approximations are possible. Example: if in the real systems separate tickets have to be purchased for each leg of the trip, each according to a possibly non-additive fare scheme, then it may be feasible within practical runtime / memory constraints to duplicate the working graph by boarding stop within each leg, labelling the arcs with costs corresponding to each possible boarding stop. An exact solution can then be computed in a single application of the algorithm, albeit on an expanded working graph, as described in the next section.

7.5.3 Representation of complex fares via journey levels

It is well known that arc-based models can handle additive fares quite easily, while non-additive fares are difficult to simulate. In the latter case, a specific monetary cost should be associated in principle with each relevant path of the transit network, which requires their explicit enumeration. For example, the fee paid for a trip may depend on the sequence of lines or transport systems taken by the passenger. Limitations on the number of allowed transfers and constraints such as must use rules are also nontrivial.

In particular, integrated fare schemes cannot be easily reproduced through the arc cost model presented in Section 6.2.3, where the data structure considers only fees for boarding a line and for running on a section between two consecutive stops, without taking into account if the passenger has already paid for other transit services during the same journey.

Instead, even if the metropolitan transit network is operated by several independent transport companies, the passenger is often able to surf more freely the available services, without paying for each used line and/or section. This is because an integrated fare system with several forms of discount is organized or coordinated by a mobility agency. For example passengers may pay full fare at initial boarding but reduced or no fare on transfer boarding of the same transport system.

However, as shown in the example of Figure 7.10, it is often impossible to reduce a complex fare structure to some linear form which could be reproduced by means an arc based model. Therefore, to avoid excessive model distortions a greater effort shall be made to explicitly simulate the rules that determine the actual fees of trips from origin to destination.

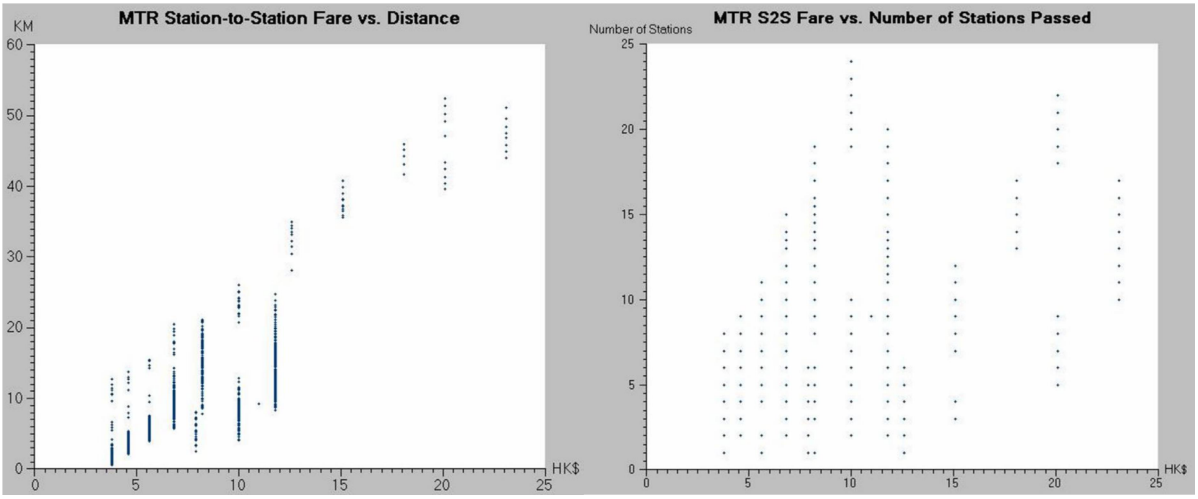

Figure 7.10. From station to station matrix of Hong Kong MTR shows non-additive fares, that are non-linear wrt to distance nor wrt to the number of station passed (these plots were provided by Michael Florian, INRO).

Sequential route choice models have no memory, since computations are done backward from destination to origin(s); it is possible to consider what will happen from the current node to the destination by introducing additional node labels (for example to know the number of transfers), but not what happened to reach that node. This information can though be kept and utilized, for example to apply proper fares, in route choice algorithms by means of journey levels which add memory to arc-based models, as illustrated in the following.

The concept of *journey level* is here introduced as an innovative paradigm to model the monetary costs paid by users resulting from a variety of transit fare schemes, allowing to simulate rebates on trips which include multiple transport systems (e.g. bus plus metro) as well as must use rules and limitations on the number of transfers.

A journey level reflects the information accumulated along a trip in terms of which transport systems, and possibly in which order, have been used by the passenger so far. This requires the construction of a more complex assignment network. In practice, to represent a relevant state of the journey a portion of the transit network is duplicated into a parallel layer. Each journey level includes a subset of lines and all walking arcs, possibly with the exceptions of connectors to origins and destinations. The alighting arcs are headed directly at the base node corresponding to the stop. Each stop serves only one transport system. Each journey level is then connected to other subsequent levels through inter-level stop arcs between the base node of the previous level and the stop node of the next level, on which integrated fares and discounts (negative fares) can be applied. The assignment network results then a bush (i.e., an acyclic graph) of journey levels, starting with origin nodes and ending with destination nodes, so that the route choice model may allow a limited set of feasible sequences of levels. In principle, each journey level is characterized by distinct arc attributes, including any of the generalized costs associated with walking, waiting, boarding and riding. How the journey layers are formed and to which other layers are connected depends on the fare scheme to reproduce; the examples that follow will help to clarify how the proposed approach is applied in practice.

In the first example, different perceived costs are modelled for initial boarding vs. transfers, because transfer boarding are penalized more by passengers.

- Level 0. The pedestrian network including centroids and connectors.
- Level 1. The whole transit network, excluding origin connectors, but including destination connectors and all stop arcs.
- Level 0 → Level 1. All stop arcs, with a discount for initial boarding.

In the second example, passengers must use at least one train line on a transit network with bus lines.

- Level 0. The pedestrian network including origin connectors, but excluding destination connectors; the bus lines.
- Level 1. The whole transit network (with both bus and train lines), excluding origin connectors, but including destination connectors and all stop arcs.
- Level 0 → Level 1. All stop arcs heading to train lines.

In the third example, transfers within the same transport system are free of charge.

- Level 0. The pedestrian network including connectors.
- Level 1. The bus lines and the pedestrian network, excluding origin connectors, but including destination connectors and stop arcs heading to bus lines with free of charge transfer (between busses).
- Level 0 → Level 1. All stop arcs heading to bus lines, with the one time bus fare (say 4€).
- Level 2. The train lines and the pedestrian network, excluding origin connectors, but including destination connectors and stop arcs heading to train lines with free of charge transfer (between trains).
- Level 0 → Level 2. All stop arcs heading to train lines, with the one time train fare (say 8€).
- Level 3. The whole transit network (with both bus and train lines), excluding origin connectors, but including destination connectors and all stop arcs with free of charge transfer.
- Level 1 → Level 3. All stop arcs heading to train lines, with the one time train fare.
- Level 2 → Level 3. All stop arcs heading to bus lines, with the one time bus fare.

In the fourth example, transfers within the same transport system are free of charge and there is a discount for taking both transport system. With respect to the third example, at the interchange between levels 1→3 and 2→3 the discount (say 2€) shall be applied to the one time fare (see Figure 7.11).

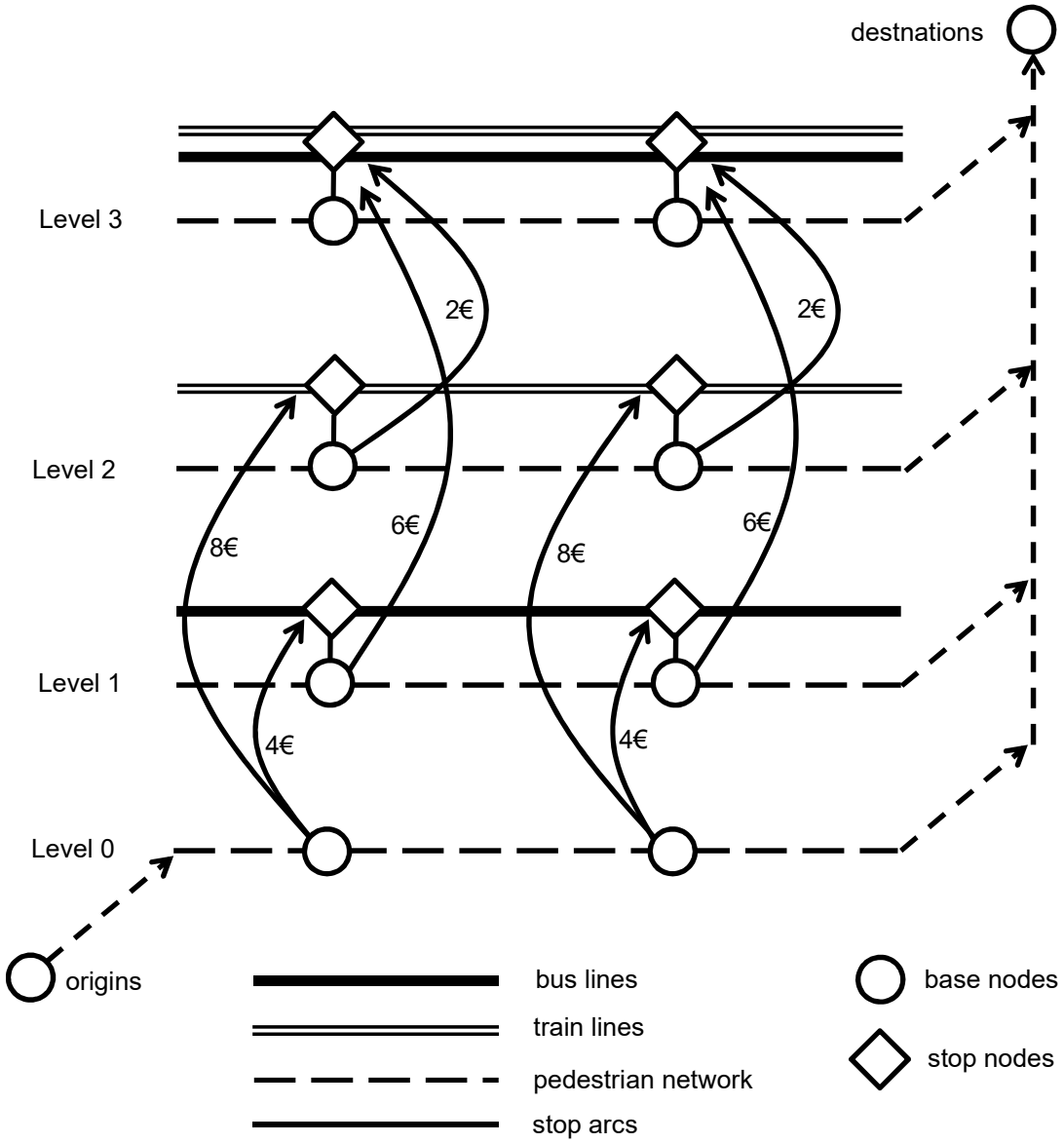

Figure 7.11. Simulation of fare discount for getting both trains and busses through journey levels.

Note that the journey level approach can be seen as an extension of the multimodal network approach presented in Section 5.1.1.2, based on the duplication of transport system subnetworks. In general, however, the network augmentation is not for free from a modelling point of view; indeed, any congestion phenomena will involve the sum of the flow across the several arc replica, and this makes the arc cost function non-separable, with negative implications on the possibility of proving the equilibrium uniqueness.

From an algorithm point of view, the arcs that lead from one level to another need not to be coded explicitly as a proper multi-label scheme can be implemented; this shows relevant computational advantages wrt the network augmentation.

Slight changes in the availability of connections among journey levels and centroids can determine relevant modifications to the fare system that is reproduced. Thus, this great flexibility requires a highly conscientious modeller.

The journey level approach permits to handle a variety of fare schemes that depend on the sequence of transport systems taken during the trip and the fare rules that apply to discounts between them. But, this approach requires more computation time and there are other complex fares that cannot be addressed this way.

7.5.4 Reference notes and concluding remarks

Different approaches on if and how to model transit fares in assignment models are proposed by several authors (e.g., Whelan and Johnson, 2004; Owen and Philips, 1987; Nielsen, 2000; Horn, 2003; Garcia & Marin, 2005; Hamdouch et al, 2007).

The methodology presented here based on network layers for the simulation of more complex fare schemes is a quite recent contribution by Constantin and Florian (2015). A similar network construction for nonlinear highway tolls is also used in Lo and Chen (2000) and in Morosan and Florian M. (2015). The same approach has been applied to multi-modal journeys by Lo et al. (2003, 2004).

7.5.4.1 List of references

Constantin I., Florian D. (2015) Integrated fare modelling with strategy-based transit assignment. In Proceedings of CASPT15, Rotterdam.

DfT (2007) Model structures and traveller responses for public transport schemes. Transport Analysis Guidance 3.11.1. UK Department for Transport, London, UK.

Garcia R., Marin A. (2005) Network equilibrium with combined modes: models and solution algorithms. Transportation Research B 39, 223-254.

Hamdouch Y., Florian M., Hearn D.W., Lawphongpanich S. (2007) Congestion pricing for multi-modal transportation systems. Transportation Research B 41, 275-291.

Horn M.E.T. (2003) An extended model and procedural framework for planning multi-modal passenger journeys. Transportation Research B 37, 641-660.

Lo H.K., Chen A. (2000) Traffic equilibrium problem with route-specific costs: formulation and algorithms. Transportation Research B: Methodological 34, 493-513.

Lo H.K., Yip C.W., Wan K.H. (2003) Modeling transfer and non-linear fare structure in multi-modal network. Transportation Research B 37, 149-170.

Lo H.K., Yip C.W., Wan Q.K. (2004) Modeling competitive multi-modal transit services: a nested logit approach. Transportation Research C 12, 251-272.

Morosan C.D., Florian M. (2015) A network model for capped link-based tolls. EURO Journal on Transportation and Logistics 4, 223-236.

Nielsen O.A. (2000) A stochastic transit assignment model considering differences in passenger utility functions. Transportation Research B 34, 377-402.

Owen A.D., Phillips G. D. A. (1987) The characteristics of railway passenger demand: An econometric investigation. Journal of Transport Economics and Policy 21, 231-253.

Whelan G., Johnson D. (2004) Modelling the impact of alternative fare structures on train overcrowding. International Journal of Transport Management 2, 51–58.