Springer Tracts on Transportation and Traffic

Guido Gentile
Klaus Noekel *Editors*

Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems

COST Action TU1004 (TransITS)

cost
EUROPEAN COOPERATION
IN SCIENCE AND TECHNOLOGY

Springer

Springer Tracts on Transportation and Traffic

Volume 10

Series editor

Roger P. Roess, New York University Polytechnic School of Engineering, New York, USA
e-mail: rpr246@nyu.edu

About this Series

The book series "Springer Tracts on Transportation and Traffic" (STTT) publishes current and historical insights and new developments in the fields of Transportation and Traffic research. The intent is to cover all the technical contents, applications, and multidisciplinary aspects of Transportation and Traffic, as well as the methodologies behind them. The objective of the book series is to publish monographs, handbooks, selected contributions from specialized conferences and workshops, and textbooks, rapidly and informally but with a high quality. The STTT book series is intended to cover both the state-of-the-art and recent developments, hence leading to deeper insight and understanding in Transportation and Traffic Engineering. The series provides valuable references for researchers, engineering practitioners, graduate students and communicates new findings to a large interdisciplinary audience.

More information about this series at http://www.springer.com/series/11059

Guido Gentile · Klaus Noekel
Editors

Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems

COST Action TU1004 (TransITS)

Springer

Editors
Guido Gentile
Sapienza University
Rome
Italy

Klaus Noekel
PTV Group
Karlsruhe
Germany

ISSN 2194-8119 ISSN 2194-8127 (electronic)
Springer Tracts on Transportation and Traffic
ISBN 978-3-319-25080-9 ISBN 978-3-319-25082-3 (eBook)
DOI 10.1007/978-3-319-25082-3

Library of Congress Control Number: 2015953814

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Contents

Contributors

Ingmar Andreasson Logistik Centrum Göteborg AB, V Frölunda, Sweden

Moshen Babaei Civil Engineering Department, Faculty of Engineering, Bu-Ali Sina University, Hamadan, Iran

Maria Bordagaray University of Cantabria, Santander, Spain

Oded Cats Department of Transport and Planning, Delft University of Technology, GA, Delft, The Netherlands; Department of Transport Science, Royal Institute of Technology (KTH), Stockholm, Sweden

Ektoras Chandakas Laboratory on City, Mobility and Transportation, Ecole des Ponts ParisTech, University Paris-East, Paris, France; Transamo, Transdev Group, Paris, France

Francesco Corman Delft University of Technology, CD, Delft, The Netherlands

Umberto Crisalli Department of Enterprise Engineering, University of Rome Tor Vergata, Rome, Italy

Luigi dell'Olio University of Cantabria, Santander, Spain

Valentina Fini Dipartimento di Ingegneria Civile, Edile e Ambientale, Università di Roma La Sapienza, Rome, Italy

Michael Florian CIRRELT, University of Montreal, Montréal, QC, Canada

Achille Fonzone Transportation Research Institute, Edinburgh Napier University, Edinburgh, UK

Markus Friedrich University of Stuttgart, Stuttgart, Germany

Guido Gentile Dipartimento di Ingegneria Civile, Edile e Ambientale, Università di Roma La Sapienza, Rome, Italy

Selini Hadjimitriou University of Modena and Reggio Emilia, Reggio Emilia, Italy

Younes Hamdouch United Arab Emirates University, Al Ain, United Arab Emirates

Sara Hernandez Transport Research Centre (TRANSyT-UPM), Universidad Politécnica de Madrid, ETSI Caminos, Canales y Puertos, Madrid, Spain

Irina Jackiva Transport and Telecommunication Institute, Rīga, Latvia

Kjell Jansson Stockholm, Sweden

Ioannis Kaparias City University London, London, UK

Karl Kottenhoff Department of Transport Science, KTH, Stockholm, Sweden

Anders Langeland University of Stavanger, Stavanger, Norway

Odd Larsen Molde University College, Molde, Norway

Fabien Leurent Laboratory on City, Mobility and Transportation, Ecole des Ponts ParisTech, University Paris-East, Paris, France

Andrés García Martínez Transport Research Centre (TRANSyT-UPM), Universidad Politécnica de Madrid, ETSI Caminos, Canales y Puertos, Madrid, Spain

Andrés Monzón Transport Research Centre (TRANSyT-UPM), Universidad Politécnica de Madrid, ETSI Caminos, Canales y Puertos, Madrid, Spain

Efthia Nathanail University of Thessaly, Pedion Areos, Volos, Greece

Klaus Noekel PTV AG, Karlsruhe, Germany

Agostino Nuzzolo Department of Enterprise Engineering, University of Rome Tor Vergata, Rome, Italy

Emilio Picasso University of Buenos Aires, Buenos Aires, Argentina

Maria Nadia Postorino Università Mediterranea di Reggio Calabria, Reggio Calabria, Italy

Sebastián Raveau Department of Transport Engineering and Logistics, Pontificia Universidad Católica de Chile, Macul Santiago, Chile

Alicia Rodriguez Universidad Carlos III de Madrid, Madrid, Spain

Rosaldo Rossetti Faculdade de Engenharia, Universidade do Porto, Porto, Portugal

Jens Schade TU Dresden, Dresden, Germany

Jan-Dirk Schmöcker Department of Urban Management, Kyoto University, Kyoto, Japan

Valentina Trozzi Strategy and Service Development, Transport for London, London, UK

Pieter Vansteenwegen KU Leuven, Leuven, Belgium

Francesco Viti University of Luxembourg, Luxembourg, Luxembourg

David Watling Leeds University, Leeds, UK

Notation

The notation utilized in this book copes with the following rules and assumptions.

A variable is a quantitative characteristic of an object/element; it is composed by the following:

- a variable identifier (only one letter, except for indicators), that specifies the nature of the characteristic (e.g., cost, flow, time, probability)
- some subscripts, that specify the referred object/element (e.g., node, origin, destination, mode, class, arc, line, run) which typically belongs to a discrete set; if the subscript is an integer index (e.g., 1, 2, …, n), it implicitly refers to the object/element at a certain position of an ordered list or vector
- a (possible) superscript, that specifies the sub-types of a same variable identifier (e.g., waiting time, in-vehicle time); the use of few letters ensures a self-explanatory notation

The following typographic rules also hold:

- the same symbol can be used as a variable, subscript or superscript, possibly in different contexts
- scalars, subscripts and superscripts are denoted in lower case, italic, not bold
- indicators (denoted possibly with one or more letters) are in upper case, italic, not bold
- set are denoted in upper case, italic, not bold
- vectors and matrices are denoted in lower case, not italic, bold
- square brackets can be used instead of subscripts to avoid nested subscripts
- parameters and coefficients are typically denoted with Greek letters

Notation Map

Subscripts

i, j	Node, vertex, generic indices
a, b	Arc, edge
$\breve{a}$	Hyperarc
o	Origin
d	Destination
z	Zone
m	Mode of transport, nest of alternatives
y	Turn
k, h	Route, path, hyperpath, tree, bush, generic choice alternative
ℓ	Line
s	Stop
r	Run
t, e	Time index, instant, interval referring to the initial instant
x	Space index, point, segment referring to the initial point
v	Vehicle
u	Person, user, individual, passenger, demand component
g	User class, person group, demand segment
f	Dictionary of parameters, arc type, function
c	Attribute
n	Iteration, year, outcome

Variables

q	Flow, volume, number of passengers, number of vehicles, cumulative flow
d	Demand
c	Cost, disutility
h	Headway, time interval
t	Travel time, elapsed time
τ, θ	Clock time; use τ for entry time and θ for exit time
κ	Capacity
f	Frequency
s	Speed
k	Density

l	Length, distance, travelled space
ξ	Space progressive
x	Generic input
y	Generic output
f	Generic function
p	Probability
u	Random utility
v	Systematic utility
w	Satisfaction, expected cost, minimum cost
ε	Random error term
a	Attribute, characteristic

Sets

$\mathfrak{R}$, $\mathfrak{R}_+$, $\mathfrak{R}_{++}$, $\mathfrak{R}_-$, $\mathfrak{R}_{--}$	Real numbers, non-negative, positive, non-positive, negative
$\mathfrak{I}$, $\mathfrak{I}_+$, $\mathfrak{I}_{++}$, $\mathfrak{I}_-$, $\mathfrak{I}_{--}$	Integer numbers, non-negative, positive, non-positive, negative
X	Generic set
N	Nodes
$A \subseteq N \times N$	Arcs
$O \subseteq N$	Origin (node)s
$D \subseteq N$	Destination (node)s
Z	Zones
M	Modes, nests
B	Vertices
E	Edges
H	Hyperarcs
Y	Turns
K	Routes, paths, hyperpaths, trees, bushes, choice set
L	Lines
S	Stops
R	Runs
$T = \{0, 1, \ldots, t, \ldots, \eta\}$	Instants of the time discretization, intervals
$\mathrm{T} = \{\tau_t : t \in T\}$	Clock times
$\aleph = \{0,1,\ldots, x, \ldots, \nu\}$	Points of the space discretization, segments
$\Xi = \{\xi_x : x \in \aleph\}$	Progressives
U	Users, individuals, demand components
V	Vehicles
G	Classes of users
F	Dictionaries, arc types, functions
C	Attributes, characteristics

Associations

$F_a \in F$	Parameters of edge $a \in E$
$F_g \in F$	Parameters of class $g \in G$
$F_s \in F$	Parameters of stop $s \in S$
$F_\ell \in F$	Parameters of line $\ell \in L$
$F_{\ell s} \in F$	parameters of stop $s \in S_\ell$ of line $\ell \in L$
$O_z \subseteq O$	Origin(s) of zone $z \in Z$
$D_z \in D$	Destination of zone $z \in Z$
$O_u \in O$	Origin of demand component $u \in U$
$D_u \in D$	Destination of demand component $u \in U$
$G_u \in G$	Class of demand component $u \in U$
$O_k \in O$	Origin of route $k \in K$
$D_k \in D$	Destination of route $k \in K$
$M_k \in M$	Mode of route $k \in K$
$L_a \in L$	Line of arc $a \in A$ (if any)
$L_r \in L$	Line of run $r \in R$

Operators

$\|X\|$	Cardinality of the generic set X, number of its elements
$Bool(\cdot)$	Boolean function; applies to a Boolean expression x: $Bool(x) = 1$, if x = TRUE, $Bool(x) = 0$, if x = FALSE
$Pr(\cdot)$	Probability; applies to an event or condition among random variables
$E(\cdot)$	Expected value; applies to a random variable
$Var(\cdot)$	Variance; applies to a random variable
$SD(\cdot)$	Standard deviation; applies to a random variable
$Cov(\cdot, \cdot)$	Covariance, applies to two random variables
$Max(\cdot)$, $Min(\cdot)$	Maximum, minimum; apply to a set or real numbers
$Sin(\cdot)$, $Cos(\cdot)$, $Tan(\cdot)$	Sine, cosine, tangent; apply to a real number
$Log(\cdot)$, $Exp(\cdot)$	Logarithm, exponential; apply to a real number
$\cup, -, \times$	Union, subtraction, product between two sets
$\leftarrow$	Assign value on the right of the arrow to variable on the left
$\Rightarrow, \Leftarrow, \Leftrightarrow$	Implications
$\varphi(\cdot)$	Probability density function
$\Phi(\cdot)$	(Cumulative) distribution function
$\bar{\Phi}(\cdot) = 1 - \Phi(\cdot)$	Complementary distribution function

Topology

$Z^{int} \subseteq Z$	Internal zones
$Z^{ext} \subseteq Z$	External zones
$(B, E \subseteq B \times B)$	Base network
$(N, A \subseteq N \times N)$	Assignment graph
$(a \in A) =$	Generic arc
$(i \in N, j \in N) = ij$	
$A_m \subseteq A$	Arcs of mode $m \in M$
$a^- \in N$	Tail (initial node) of arc $a \in A$
$a^+ \in N$	Head (final node) of arc $a \in A$; in case of hyperarcs this is a set
$i^+ = \{a \in A: a^- = i\}$	Forward star of node $i \in N$; this is a set of arcs, not of nodes
$i^- = \{a \in A: a^+ = i\}$	Backward star of node $i \in N$; this is a set of arcs, not of nodes
$k = (N_k, A_k)$	Acyclic sub-graph (path, hyperpath, bush, tree)
$N_k \subseteq N$	Nodes of acyclic sub-graph k
$A_k \subseteq (A \cap N_k \times N_k)$	Arcs of acyclic sub-graph k
$i_k^+ = i^+ \cap A_k$	Successor arcs of $i \in N_k$
$i_k^- = i^- \cap A_k$	Predecessor arcs of $i \in N_k$
$k^- = \{i \in N_k : i_k^- = \varnothing\}$	Origin nodes of acyclic sub-graph k
$k^+ = \{i \in N_k : i_k^+ = \varnothing\}$	Destination nodes of acyclic sub-graph k
$od \in O \times D$	Generic OD pair
K_{odm}	Routes connecting on the network origin $o \in O$ to destination $d \in D$ on mode $m \in M$
$K_m = \cup_{od \in O \times D} K_{odm}$	Routes of mode $m \in M$
$K \in K_m \Rightarrow A_k \subseteq A_m$	Generic route on mode $m \in M$
$K = \cup_{m \in M} K_m$	Set of all routes defined on the multimodal network
Δ_{ak}	Number of times that a user travelling on route $k \in K$ passes through arc $a \in A$
$\Delta_{\breve{a}k}$	Probability of using hyperarc $\breve{a} \in H$ when traveling on route $k \in K$
$N^{div} \subseteq N$	Diversion nodes
$A^{div} = \{i^+ : i \in N^{div}\}$	Diversion arcs
$\breve{a} \subseteq i^+: i \in N^{div}$	A hyperarc is a set of arcs (its branches) exiting from a same diversion node
$H \subseteq \{\breve{a} \subseteq i^+: i \in N^{div}\}$	Set of hyperarcs
$(y \in Y) = (a \in A, b \in A)$ $= ab$	Generic turn
$y^- \in A$	Tail (initial arc) of turn $y \in Y$
$y^+ \in A$	Head (final arc) of turn $y \in Y$
$y^\circ \in N$	Centre (via node) of turn $y \in Y$
$Y_i \subseteq i^- \times i^+$	Turns of node $i \in N$
$t^+(\tau) \in T \cup \eta + 1$	The next time index of instant $\tau \geq \tau_0$

$t^{-}(\tau) \in T \cup \eta + 1$	The previous time index of instant $\tau \geq \tau_0$
$\tau_{\eta+1} = \infty$	Additional instant
h_t	Duration of interval $t \in T$

Symbols

Transit service

$S_\ell \subseteq S$	Stops sequence of line $\ell \in L$; an ordered set with no repetitions
$S_\ell^- \in S_\ell$	First stop of line $\ell \in L$
$S_\ell^+ \in S_\ell$	Last stop of line $\ell \in L$
$s_\ell^- \in S_\ell$	Previous stop of stop $s \in S_\ell - S_\ell^-$ of line $\ell \in L$
$s_\ell^+ \in S_\ell$	Successive stop of stop $s \in S_\ell - S_\ell^+$ of line $\ell \in L$
$R_\ell \subseteq R$	Runs sequence of line $\ell \in L$; an ordered set with no repetitions
$R_\ell^- \in R_\ell$	First run of line $\ell \in L$
$R_\ell^+ \in R_\ell$	Last run of line $\ell \in L$
$r^{-\ell} \in R_\ell$	Previous run of run $r \in R_\ell - R_\ell^-$ of line $\ell \in L$
$r^{+\ell} \in R_\ell$	Successive run of run $r \in R_\ell - R_\ell^+$ of line $\ell \in L$
τ_{rs}	Arrival time of run $r \in R_\ell$ at stop $s \in S_\ell - S_\ell^-$
θ_{rs}	Departure time of run $r \in R_\ell$ at stop $s \in S_\ell - S_\ell^+$
$\theta_r = \theta_{rs},\ s = S_\ell^-$	Scheduled departure of run $r \in R_\ell$
t_{rs}^{run}	Running time of run $r \in R_\ell$ on the line segment s from stop $s \in S_\ell - S_\ell^+$ to s_ℓ^+
t_{rs}^{dwell}	Dwelling time of run $r \in R_\ell$ at stop $s \in S_\ell - S_\ell^- - S_\ell^+$
$t_{\ell st}^{run}$	Running time of line segment s from stop $s \in S_\ell - S_\ell^+$ to $s_\ell{}^+$ during interval $t \in T$
$t_{\ell st}^{dwell}$	Dwelling time at stop $s \in S_\ell - S_\ell^- - S_\ell^+$ during interval $t \in T$
$B_{\ell s} \subseteq E$	Edge sequence of line segment $s \in S_\ell - S_\ell^+$; an ordered set with no repetitions
$l_{\ell s}$	Length of the line segment $s \in S_\ell - S_\ell^+$
l_a	Length of edge $a \in E$
s_a^{walk}	Walking speed of edge $a \in E$
s_{at}	Commercial speed of edge $a \in B$ during interval $t \in T$
t_ℓ^{stop}	Stop time of line $\ell \in L$
t_ℓ^{alight}	Alighting time of line $\ell \in L$
t_ℓ^{board}	Boarding time of line $\ell \in L$
t_ℓ^{do}	Door operation time of line $\ell \in L$

t_{ℓ}^{ab}	Minimum dwell time for alighting and boarding of line $\ell \in L$
$h_{\ell st}$	Headway of line $\ell \in L$ at stop $s \in S_{\ell} - S_{\ell}^{+}$ during interval $t \in T$
$\varphi_{\ell st}^{h}(h)$	Headway distribution of line $\ell \in L$ at stop $s \in S_{\ell} - S_{\ell}^{+}$ during interval $t \in T$
$h_{\ell s}^{max}$	Maximum headway of the distribution of line $\ell \in L$ at stop $s \in S_{\ell} - S_{\ell}^{+}$
h_{rs}^{dep}	Departure headway of run $r \in R_{\ell} - R_{\ell}^{-}$ from stop $s \in S_{\ell} - S_{\ell}^{+}$
h_{rs}^{arr}	Arrival headway of run $r \in R_{\ell} - R_{\ell}^{-}$ to stop $s \in S_{\ell} - S_{\ell}^{-}$
α_{rs}	Headway deviation of run $r \in R_{\ell} - R_{\ell}^{-}$ to stop $s \in S_{\ell} - S_{\ell}^{-}$
$f_{\ell st}$	Frequency of line $\ell \in L$ at stop $s \in S_{\ell} - S_{\ell}^{+}$ during interval $t \in T$
$f_{\ell s}^{dep}(\tau)$	Departure frequency of line $\ell \in L$ at stop $s \in S_{\ell} - S_{\ell}^{+}$ at time τ
$f_{\ell s}^{arr}(\tau)$	Arrival frequency of line $\ell \in L$ at stop $s \in S_{\ell} - S_{\ell}^{+}$ at time τ
f_a	Frequency of line $L_a \in L$ associated with arc $a \in A^{wait}$, otherwise ∞
$\sigma_{\ell st}$	Irregularity or variation coefficient of line $\ell \in L$ at stop $s \in S_{\ell} - S_{\ell}^{+}$ during $t \in T$
$n_{\ell s}$	Parameter of the Erlang headway distribution of line $\ell \in L$ at stop $s \in S_{\ell} - S_{\ell}^{+}$
$\varphi_{\ell st}^{w}(t)$	Waiting time distribution of line $\ell \in L$ at stop $s \in S_{\ell} - S_{\ell}^{+}$ at instant $t \in T$
$t_{\ell st}^{wait}$	Expected waiting time of line $\ell \in L$ at stop $s \in S_{\ell} - S_{\ell}^{+}$ at instant $t \in T$
κ_{ℓ}^{veh}	Vehicle capacity of line $\ell \in L$
κ_{ℓ}^{seat}	Seating capacity of line $\ell \in L$
κ_{ℓ}^{stand}	Standing capacity of line $\ell \in L$
κ_{ℓ}^{crush}	Crush capacity of line $\ell \in L$
κ_{ℓ}^{door}	Door capacity of line $\ell \in L$
κ_{ℓ}^{board}	Boarding capacity of line $\ell \in L$
κ_{ℓ}^{alight}	Alighting capacity of line $\ell \in L$
κ_{s}^{stop}	Stop capacity of stop $s \in S$
γ_{sg}^{stop}	Discomfort coefficient of stop $s \in S$ for class $g \in G$
$\gamma_{\ell g}^{line}$	Discomfort coefficient of line $\ell \in L$ for class $g \in G$
C^{cs}	Set of comfort attributes for each stop
C^{cl}	Set of comfort attributes for each line
a_{sc}	Value of comfort attribute $c \in C^{cs}$ for stop $s \in S$

$a_{\ell c}$	Value of comfort attribute $c \in C^{cl}$ line $\ell \in L$
β_{cg}^{stop}	Utility coefficient of stop comfort attribute $c \in C^{cs}$ for stop $s \in S$ and class $g \in G$
β_{cg}^{line}	Utility coefficient of line comfort attribute $c \in C^{cl}$ for line $\ell \in L$ and class $g \in G$
c_{kg}^{fare}	Fare of route $k \in K$ and class $g \in G$
$c_{\ell s}^{kfee}$	Kilometric fee of line segment $s \in S_\ell - S_\ell^+$
$c_{\ell s}^{bfee}$	Boarding fee of stop $s \in S_\ell - S_\ell^+$
α_ℓ^{dwell}	BPR coefficient for dwelling congestion of line $\ell \in L$
β_ℓ^{dwell}	BPR exponent for dwelling congestion of line $\ell \in L$
α_ℓ^{queue}	BPR coefficient for queuing congestion of line $\ell \in L$
β_ℓ^{queue}	BPR exponent for queuing congestion of line $\ell \in L$
χ_ℓ^{queue}	Exponent for the saturation of the remaining capacity of line $\ell \in L$
β_ℓ^{crowd}	BPR exponent for crowding congestion of line $\ell \in L$
β_s^{crowd}	BPR exponent for crowding congestion of stop $s \in S$

Public Transport Network

$E^{walk} \subseteq E$	Set of walkable edges
$B_z^{orig} \subseteq B$	Origin vertex associated with zone $z \in Z$
$B_z^{dest} \subseteq B$	Destination vertex associated with zone $z \in Z$
$B_s^{stop} \subseteq B$	Vertex associated with stop $s \in S$
N^{base}	Set of base nodes
N^{stop}	Set of stop nodes
N_ℓ	Set of line nodes of line $\ell \in L$
$N_{\ell s}^{arr}$	Arrival node of line $\ell \in L$ at stop $s \in S_\ell - S_\ell^-$
$N_{\ell s}^{dep}$	D node of line $\ell \in L$ from stop $s \in S_\ell - S_\ell^+$
N_r	Set of run nodes of run $r \in R$
N_{rs}^{arr}	Arrival node of run $r \in R$ from stop $s \in S_\ell - S_\ell^+$
N_{rs}^{dep}	Departure node of run $r \in R$ from stop $s \in S_\ell - S_\ell^+$
$N_{\ell s}^{a\text{-}seat}$	Seating arrival node of line $\ell \in L$ at stop $s \in S_\ell - S_\ell^-$
$N_{\ell s}^{d\text{-}seat}$	Seating departure node of line $\ell \in L$ at stop $s \in S_\ell - S_\ell^-$

$N_{\ell s}^{a\text{-}stand}$	Standing arrival node of line $\ell \in L$ at stop $s \in S_\ell - S_\ell^-$
$N_{\ell s}^{d\text{-}stand}$	Standing departure node of line $\ell \in L$ at stop $s \in S_\ell - S_\ell^-$
$N_{\ell s}^{p\text{-}board}$	Board placing node of line $\ell \in L$ at stop $s \in S_\ell - S_\ell^-$
$N_{\ell s}^{p\text{-}stand}$	Stand placing node of line $\ell \in L$ at stop $s \in S_\ell - S_\ell^-$
$N_{\ell s}^{serv}$	Service node of line $\ell \in L$ at stop $s \in S_\ell - S_\ell^+$
$N_{\ell s}^{que}$	Queue node of line $\ell \in L$ at stop $s \in S_\ell - S_\ell^+$
A^{walk}	Set of pedestrian arcs
A^{stop}	Set of stop arcs
A^{run}	Set of running arcs
A^{dwell}	Set of dwelling arcs
A^{wait}	Set of waiting arcs
A^{board}	Set of boarding arcs
A^{alight}	Set of alighting arcs
A^{trans}	Set of transfer arcs
A^{dest}	Set of destination arcs
$A^{r\text{-}seat}$	Set of seat running arcs
$A^{p\text{-}seat}$	Set of seat placing arcs
$A^{d\text{-}seat}$	Set of seat dwelling arcs
$A^{a\text{-}seat}$	Set of seat alighting arcs
$A^{r\text{-}stand}$	Set of stand running arcs
$A^{d\text{-}stand}$	Set of stand dwelling arcs
$A^{p\text{-}stand}$	Set of stand placing arcs
$A^{a\text{-}stand}$	Set of stand alighting arcs
$A^{p\text{-}switch}$	Set of switch seating arcs
$A^{p\text{-}keep}$	Set of keep standing arcs
A^{fail}	Set of failing arcs
A^{serv}	Set of service arcs
A^{que}	Set of queuing arcs
H^{wait}	Set of waiting hyperarcs
H^{board}	Set of boarding hyperarcs
H^{dwell}	Set of dwelling hyperarcs
H^{serv}	Set of service hyperarcs

Class Parameters

γ_g^{vot}	Value of time for users of class $g \in G$
γ_g^{walk}	Walking discomfort coefficient for users of class $g \in G$
γ_g^{wait}	Waiting discomfort coefficient for users of class $g \in G$

c_g^{tran}	Transfer cost for users of class $g \in G$
γ_g^{risk}	Risk-averseness coefficient for users of class $g \in G$
γ_g^{mfee}	Fee multiplier for users of class $g \in G$
γ_g^{del}	Delay coefficient for users of class $g \in G$
γ_g^{adv}	Anticipation coefficient for users of class $g \in G$
t_g^{ant}	Maximum anticipation wrt desired departure time for users of class $g \in G$
t_g^{del}	Maximum delay wrt desired departure time for users of class $g \in G$
α_g^{crowd}	Overcrowding congestion coefficient for users of class $g \in G$
α_g^{learn}	Cost learning exponential filter for users of class $g \in G$
α_g^{choup}	Choice updating exponential filter for users of class $g \in G$
α_g^{cred}	Credibility exponential filter for users of class $g \in G$
t_g^{wmax}	Maximum walking time on the pedestrian network for users of class $g \in G$
t_g^{stop}	Reference stop time for users of class $g \in G$
γ_g^{loss}	Discomfort coefficient for lost opportunities for users of class $g \in G$

Performances

c_{ag}	Generalized cost of arc $a \in A$ for users of class $g \in G$
c_{ag}^{nt}	Non-temporal cost of arc $a \in A$ for users of class $g \in G$
t_a^0	Free-flow travel time of arc $a \in A$
t_a	Travel time of arc $a \in A$
γ_{ag}	Value of time on arc $a \in A$ for users of class $g \in G$
c_{kg}	Generalized cost of route $k \in K$ for users of class $g \in G$
c_{kg}^{na}	Non-additive cost of route $k \in K$ for users of class $g \in G$
$\tilde{c}_{kg}^{n}$	Forecasted cost of path $k \in K$ for users of class $g \in G$ in day n
$p_{a\|\breve{a}(dgm)}$	Diversion probability of using branch $a \in \breve{a}$ of hyperarc $\breve{a} \in H$ (for users of class $g \in G$ that travel on mode $m \in M$ to destination $d \in D$)

$t_{a\|\breve{a}}$ (dgm)	Conditional travel time for using branch $a \in \breve{a}$ of hyperarc $\breve{a} \in H$ (for users of class $g \in G$ that travel on mode $m \in M$ to destination $d \in D$)
$t_{\breve{a}\ (dgm)}$	Combined travel time of hyperarc $\breve{a} \in H$ (for users of class $g \in G$ that travel on mode $m \in M$ to destination $d \in D$)
γ_{ig}	Value of time on arcs exiting the diversion node $i \in N^{div}$ for class $g \in G$ users
$c_{a\|\breve{a}\ g}$	Conditional cost of using branch $a \in \breve{a}$ of hyperarc $\breve{a} \in H$ for class $g \in G$ users
$c_{\breve{a}g}$	Combined cost of hyperarc $\breve{a} \in H$ for users of class $g \in G$
$c_{\breve{a}dgm}$	Combined cost of hyperarc $\breve{a} \in H$ for users of class $g \in G$ that travel on mode $m \in M$ to destination $d \in D$
γ_{sgt}^{crowd}	Crowding discomfort coefficient of class $g \in G$ user at stop $s \in S$ at instant $t \in T$
$\gamma_{\ell sg}^{crowd}$	Crowding discomfort coefficient of segment $s \in S$ of line $\ell \in L$ for class $g \in G$
γ_{rsg}^{crowd}	Crowding discomfort coefficient of segment $s \in S$ of run $r \in R_\ell$ for class $g \in G$
$f_{\ell s}^{eff}$	Effective frequency of line $\ell \in L$ at stop $s \in S_\ell - S_\ell^+$
$p_{\ell s}^{fail}$	Fail-to-board probability of line $\ell \in L$ at stop $s \in S_\ell - S_\ell^+$
$c_{\ell s}^{fail}$	Additional cost in the case of fail-to-board line $\ell \in L$ at stop $s \in S_\ell - S_\ell^+$
$\theta_a(\tau)$	Exit time from arc $a \in A$ of a passenger who enters it at time τ
θ_{at}	Exit time from arc $a \in A$ of a passenger who enters it at instant $t \in T$
m_{aet}	Share of users that enter arc $a \in A$ during interval $e \in T$ and exit it during $t \in T$
$\kappa_a(\tau)$	Instantaneous remaining capacity at time τ available at the end of arc $a \in A^{que}$
$\kappa_a^{cum}(\tau)$	Cumulative remaining capacity of arc $a \in A^{que}$ at time τ
$n_a(\tau)$	Number of carriers passengers must let go before being able to board each line $\ell = L_a$, if queuing on arc $a \in A^{que}$ starts at a given time τ
ρ_{br}	Time when the last passenger that achieves boarding run r arrives at the stop and enters the waiting arc $b \in A^{wait}$

Route Generation

c_k^{imp}	Search impedance of path $k \in K$
t_k^{tot}	Total travel time of path $k \in K$
n_k^{trans}	Number of transfers of path $k \in K$
c_k^{fare}	Fare of path $k \in K$
β^{time}, β^{trans}, β^{fare}	Attribute multipliers
α^{imp}	Relative tolerance wrt the least impedance path
χ^{imp}	Additive absolute tolerance wrt the least impedance path
χ^{trans}	Maximum number of transfers

Attributes and Information

$\hat{a}_{kcu}^n$	Value of attribute $c \in C$ on path $k \in K_u$ forecasted by traveller $u \in U$ for day n
$\alpha_u^{PK\ n}$	Weight of Prior Knowledge in day n for traveller $u \in U$
$\alpha_u^{TE\ n}$	Weight of Travel Experience in day n for traveller $u \in U$
α_u^{RTI} ^{n}	Weight of Real-Time Information in day n for traveller $u \in U$
$\tilde{a}_{kcu}^1$	Value of attribute $c \in C$ on path $k \in K_u$ prior known by traveller $u \in U$ for day 1
$\tilde{a}_{kcu}^n$	Value of attribute $c \in C$ on path $k \in K_u$ anticipated by traveller $u \in U$ for day n
$\tilde{a}_{kcu}^{RTI}$	Value of attribute $c \in C$ on path $k \in K_u$ got from RTI by traveller $u \in U$ for day n

Stop and Run Choices

t_{is}^{walk}	Walking time on the shortest path from i to B_s^{stop} on the pedestrian network
v_s^{idgt}	Systematic utility of stop $s \in S$ for passengers of class $g \in G$ directed toward destination $d \in D$ that at instant $t \in T$ are in vertex $i \in B$
p_s^{idgt}	Probability of choosing stop $s \in S$ for passengers of class $g \in G$ directed toward destination $d \in D$ that at instant $t \in T$ are in vertex $i \in B$

$v_{r\|r'}^{sdgt}$	Systematic utility of run r conditional on arrival of run r' for users of class $g \in G$ directed to destination $d \in D$ who reached stop $s \in S$ at instant $t \in T$
$p_{r'\|r'}^{sdgt}$	Probability of run r conditional on arrival of run r' for users of class $g \in G$ directed to destination $d \in D$ who reached stop $s \in S$ at instant $t \in T$
p_r^{sdgt}	Unconditional probability of run r for users of class $g \in G$ directed towards destination $d \in D$ who reached stop $s \in S$ at instant $t \in T$

Flows and Route Choice

$d_{odmg(t)}$	Flow of class $g \in G$ users that travel on mode $m \in M$ (and depart during time interval $t \in T$) from origin $o \in O$ directed to destination $d \in D$
$\mathbf{d}_{ODmg(t)}$	O-D matrix of class $g \in G$ users that travel on mode $m \in M$ (and depart during time interval $t \in T$)
q_{kg}	Flow of class $g \in G$ users that travel on route $k \in K$
$\tilde{q}_{kg}^{n}$	Flow of class $g \in G$ on route $k \in K$ given by the Route Choice Model in day n
q_{ag}	Flow of class $g \in G$ users travelling on arc $a \in A$
q_{ag}^{nlm}	Flow of class $g \in G$ users on arc $a \in A$ given by the Network Loading Map
q_a	Volume of arc $a \in A$
ω_{ag}	Equivalency coefficient on arc $a \in A$ for class $g \in G$ users
q_a^0	Base volume of arc $a \in A$
p_{kg}	Probability that class $g \in G$ users take (choose) route $k \in K$
p_{admg}	Probability that class $g \in G$ users that travel on mode $m \in M$ directed to destination $d \in D$ take (choose) arc $a \in A$ conditional on being at its tail
w_{idmg}	Expected cost to reach destination $d \in D$ from node $i \in N$ perceived by users of class $g \in G$ that travel on mode $m \in M$
q_{idmg}	Flow traversing *node* $i \in N$ of class $g \in G$ users that travel on mode $m \in M$ directed to destination $d \in D$ users
w_{admg}	Remaining cost to reach destination $d \in D$ from node $a^- \in N$ for class $g \in G$ users that travel on mode $m \in M$ and take (choose) arc $a \in A$

$p_{\breve{a}dmg}$	Probability that class $g \in G$ users that travel on mode $m \in M$ directed to destination $d \in D$ take (choose) hyperarc $\breve{a} \in H$ conditional on being at its tail
$w_{\breve{a}dmg}$	Remaining cost to reach destination $d \in D$ from node $\breve{a}^- \in N$ for class $g \in G$ users that travel on mode $m \in M$ and take (choose) hyperarc $\breve{a} \in H$
$q_{ag}^{out}(\tau)$	Outflow of class $g \in G$ users from arc $a \in A$ at time τ
$q_{ag}^{int}(\tau)$	Inflow of class $g \in G$ users to arc $a \in A$ at time τ
$q_{ag}^{cout}(\tau)$	Cumulative outflow of class $g \in G$ users from arc $a \in A$ at time τ
$q_{ag}^{cint}(\tau)$	Cumulative inflow of class $g \in G$ users to arc $a \in A$ at time τ

Optimal Strategies

$\bar{A}_{(dg)} \in A$	Set of arcs in the solution hypertree provided by the (deterministic) route choice model (for passengers of class $g \in G$ directed to destination $d \in D$)
$\bar{A}_i^+ = i^+ \cap \bar{A}$	Arcs exiting from node $i \in N$ and belonging to the solution hypertree
$w_i(\breve{a})$	Expected cost to reach the destination from stop $i \in N$ as a function of the attractive set $\breve{a} \subseteq i^+$
x_a	Binary variable denoting if arc $a \in A$ belongs to the solution hypertree $\bar{A}$
ω_i	Total wait time at stop $i \in s$ (for a given class and destination)
μ_a	Dual variable of arc $a \in A$ in the optimal strategies solution
$p_{\breve{a}\|a}(t)$	Probability that the line $L_a \in L$ of arc $a \in \breve{a}$ is boarded after a wait time t
$t_{\breve{a}}^{max}$	Minimum headway among the attractive lines $\breve{a}$
$\breve{a}_x$	Attractive set formed by the first x lines at the stop in terms of remaining costs
$\breve{a}_{x(\tau)}$	Attractive set that will be considered by the passenger at time $\tau \geq t$ of the wait after the elapsed wait time $t \geq 0$
$w_i(t)$	Expected cost after the elapsed wait time $t \geq 0$ resulting from the future application of the dynamic attractive set $\breve{a}_{x(\tau \geq t)}$

τ_k	Elapsed wait time after which the kth line exits from the attractive set
$p_k(\tau\|t)$	Probability that the kth line is boarded at time τ after an elapsed wait time t

Demand Models

d_{og}^{gen}	Flow of class $g \in G$ users produced (or generated) from origin $o \in O$
d_{dg}^{att}	Flow of class $g \in G$ users attracted from destination $d \in D$
d_{odg}	Flow of class $g \in G$ users that travel from origin $o \in O$ to destination $d \in D$
w_{odg}	Satisfaction of class $g \in G$ to travel from origin $o \in O$ to destination $d \in D$
w_{odmg}	Satisfaction of class $g \in G$ to travel on mode $m \in M$ from origin $o \in O$ to $d \in D$
a_{zc}^{zone}	Value of landuse attribute $c \in C^{zone}$ in zone $z \in Z$
C^{zone}	Set of zone landuse attributes
β_{cg}^{gen}	Coefficient of generation attribute $c \in C^{zone}$ for class $g \in G$
β_{cg}^{att}	Coefficient of attraction attribute $c \in C^{zone}$ for class $g \in G$
a_{mgc}^{mod}	Value of modal split attribute $c \in C^{mod}$ for mode $m \in M$ and class $g \in G$
C^{mod}	Set of modal split attributes
β_{cg}^{mod}	Coefficient of modal split attribute $c \in C^{mod}$ for class $g \in G$
β_{g}^{mod}	Coefficient of generalized cost for mode choice of class $g \in G$
β_{g}^{route}	Coefficient of generalized cost for route choice of class $g \in G$

Estimation of Assignment

$\mathbf{d}$, $\mathbf{d}^0$, $\mathbf{d}^{LB}$, $\mathbf{d}^{UB}$	Vectors of demand parameters: calibrated, initial, lower bound, upper bound
$\boldsymbol{\delta}$, $\boldsymbol{\delta}^0$, $\boldsymbol{\delta}^{LB}$, $\boldsymbol{\delta}^{UB}$	Vectors of supply parameters: calibrated, initial, lower bound, upper bound
$\mathbf{q}$, $\mathbf{q}^m$	Vectors of arc flows: resulting from the assignment model, measured on the field

$q(\mathbf{d},\boldsymbol{\delta})$, $t(\mathbf{d},\boldsymbol{\delta})$	Functionals of the assignment model in terms of: traffic flows and travel times
z_1, z_2, z_3, z_4	Distance functions
ψ_1, ψ_2, ψ_3, ψ_4	Weights of the distance functions in the objective function
$\mathbf{M}$, $m_{a\ od}$	Assignment matrix; its elements are the fractions of demand d_{od} using each arc a

Prospect Theory

N	Set of outcomes
p_n	Objective probability of the nth outcome
v_{kn}	Utility of alternative $k \in K$ in the nth outcome
v_k^0	Reference point of alternative $k \in K$
Δv_{kn}	Gain or loss in the nth outcome wrt the reference point of alternative $k \in K$
$\pi(p)$	Pi function
$h(\Delta v)$	Value function
w_{kn}^+, w_{kn}^-	Cumulative probability of positive/negative the nth outcome of alternative $k \in K$
γ, δ	Parameters of the pi function
λ, α, β	Parameters of the value function

Random Utility

K_u	Choice set of alternatives (e.g., paths) considered by traveller $u \in U$
u_{uk}	Perceived utility by traveller $u \in U$ for alternative $k \in K_u$
v_{uk}	Systematic utility for traveller $u \in U$ associated to alternative $k \in K_u$
ε_{uk}	Random error term for traveller $u \in U$ associated to alternative $k \in K_u$
p_{uk}	Probability that traveller $u \in U$ chooses alternative $k \in K_u$
w_u	Satisfaction (or Expected Maximum Utility) of traveller $u \in U$
a_{ukc}	Value of the attribute $c \in C$ associated to alternative $k \in K_u$
β_c	Coefficient of attribute $c \in C$
θ	Scale parameter of the MNL model

$\delta_m = \theta_m/\theta_0$	Ratio of scale parameters of nest $m \in M$ in the NL and CNL model
α_c	Transformation parameter of attribute $c \in C$ in the Box-Cox model
CF_k	Commonality factor of alternative $k \in K$
SC_{kh}	Similarity coefficient of two alternatives $k \in K$ and $h \in K$
α_{mk}	Degree of inclusion of alternative $k \in K$ in nest $m \in M$
π_{uk}	Observed probability that traveller $u \in U$ chooses alternative $k \in K_u$
$L(\boldsymbol{\beta})$	Likelihood function
$LL(\boldsymbol{\beta})$	Log-likelihood function
σ_c	Standard deviation of parameter $c \in C$
t_c	t-test of parameter $c \in C$
LR	Likelihood ratio
ρ^2	Rho-square
MRS_{cj}	Marginal rates of substitution between attributes $c \in C$ and $j \in C$
ELA_{kch}	Elasticity of alternative $k \in K$ wrt an attribute $c \in C$ of alternative $h \in K$

Indicators

$K_s \subseteq K$	Set of routes which include boarding at stop $s \in S$
$K_\ell \subseteq K$	Set of routes which include running on line $\ell \in L$
$A_\ell^{run} \subseteq A^{run}$	Set of running arcs of line $\ell \in L$
q_{ktg}	Flow on route $k \in K$ of class $g \in G$ users departing during interval $t \in T$
VOL_{kt}	Volume of route $k \in K_{od}$ during interval $t \in T$
VOL_k	Volume of route $k \in K_{od}$
PBS_s	Passengers Boarding (at) Stop $s \in S$
PRL_ℓ	Passengers Riding Line $\ell \in L$
PKR_k	Passenger-Kilometres of Route $k \in K$
PKL_ℓ	Passenger-Kilometres of Line $\ell \in L$
SKL_ℓ	Service-Kilometres of Line $\ell \in L$
ALL_ℓ	Average Loading of Line $\ell \in L$
DIS_k	Distance of route $k \in K$
ATD_{od}	Average Travelled Distance for origin-destination pair $od \in O \times D$
t_{kt}	Travel time of route $k \in K$ for passengers departing at instant $t \in T$

ATT_{odt}	Average Travel Time for origin–destination pair $od \in O \times D$ and interval $t \in T$
c_{ktg}	Generalized cost of route $k \in K$ for passengers of class $g \in G$ departing at $t \in T$
AGC_{odgt}	Average Generalized Cost for pair $od \in O \times D$ class $g \in G$ and interval $t \in T$

Cost Benefit Analysis

NPV	Net Present Value
B_t	Benefits in year t
C_t	Costs in year t
r	Discount rate
PVB	Present Value of future Benefits
PVC	Present Value of future Costs
BCR	Benefit-Cost Ratio
IRR	Internal Rate of Return

Introduction

This book is the main output of the COST Action TU1004: Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems, which was called TransITS for short.

Cost Actions are projects funded by the EU to support a network of researchers. The focus is in this case on transit assignment and ITS technologies. In very productive four years of cooperation, more than 100 researchers from all continents have been involved in TransITS with various roles.

The specific purpose of this book is to provide to a wide community of possible readers, ranging from policy makers to practitioners, from master and Ph.D. students to researchers, a comprehensive source to understand how transit assignment can support the appraisal of investments in ITS technologies for operators and passengers. This is a unique contribution that is really missing in the literature.

The book is not a simple collection of papers. It is really a joint work of multiple hands with a very precise structure, where about 40 different authors (see the List of Contributors for more details) accepted to give a specific contribution with a uniform style and notation (which is not trivial). Of course, a huge coordination and sometimes rewriting effort by the Editorial Board (Guido Gentile, Fabien Leurent, Klaus Noekel, Francesco Viti) was necessary to achieve the desired result. Each chapter and section of the book then went through the careful reading of cross-reviewers. This ambitious project was concluded successfully, and we are all very proud of this book.

The result is a relevant piece of work, with around 700 dense pages articulated in 4 Parts and 10 Chapters (see the Table of contents for more details).

Part I introduces the use of ITS in public transport. The motivations for improving public transport in the context of sustainable cities and regions together with the importance of ITS technologies are first discussed (Chap. 1). The forms of public transport are then illustrated not only presenting the different types of systems and vehicles but also including organization and product issues (Chap. 2). Finally, the state of the art of ITS solutions for public transport is presented with several examples and applications, together with the conceptual link to the variables

of the transit assignment models that are used for the appraisal of such innovative systems (Chap. 3).

Part II introduces the idea of planning and modelling public transport. Modelling is presented not as a means in itself, but as a tool for rational decision-making about investment into transport. Passenger route choice models, the core of this book, form just a part of a hierarchy of travel demand models. An overview of this hierarchy is given and the common mathematical framework explained. Particular emphasis is given to random utility models (Chap. 4). The stage is set for Part 3 by defining a standard terminology for the input and output data of such models. The final section describes how ITS produces a large part of these data and how they can be used to build and validate models (Chap. 5).

Part III presents the theory of transit assignment and discusses the basic modelling frameworks: schedule-based, frequency-based, simulation-based. Particular emphasis is given to the representation of strategic behaviour through hyperpaths and to the dynamic aspects of transit simulation including within-day and day-to-day assignment models (Chap. 6). The focus is more on the demand and supply requirements rather than on equilibrium algorithms. The proposed models contain several advancements and original contributions which are designed to capture in the models the effects of real-world phenomena such as passenger information, vehicle capacity and operational stability (Chap. 7).

Part IV examines how transit modes interact with other modes in reality, both competing and complementing, and how these effects can be reflected in multimodal networks (Chap. 8). A review of which models presented in the book have found their way from theory into practice, in the form of both commercially available software and of prototypical academic implementations is presented; their use is also illustrated with the help of two case studies (Chap. 9). Finally, open challenges are described and directions for future research are proposed (Chap. 10).

We want to thank the Cost Office in Bruxelles (Thierry Goger, Carmencita Malimban, Mickael Pero, Andrea Tortajada) for continuously assisting our work with precise indications on managing the Action and the EU Commission for granting a suitable budget, in particular for this book.

We want to thank Prof. Mike Bell for conceiving and starting our Cost Action together with a small group of researchers that became larger and larger during the project. His guidance and the support from other senior researchers in the field (Michael Florian and Ingmar Andreasson, among the others) was crucial to give the right perspective to this work.

We want to thank SISTeMA—PTV Group (Lorenzo Meschini, Claudio Petrocelli) for being the grant holder of our Cost Action and taking care of the administration.

We want to thank the Scientific Secretary of our Cost Action, Valentina Trozzi, Ph.D.; without her patient and meticulous organization work, corroborated by top-level technical skills, we would all have been lost.

We want to thank the kind consideration of our publisher Springer (Oliver Jackson) and the precious work of our professional proof readers, who are not merely native speakers but really skilled researchers in transport assignment models

(Luana Chetcuti, and Nishanthi Venkatesan, among the others); they did a great job in finalizing the book to a printable version.

We want to thank all the people who were involved in the Action and actively participated in our meetings, conferences and training schools; many of their ideas and experiences are now part of this book.

We want to thank our institutions for giving us the time to participate in the Action.

We finally want to thank our families for the support and the cheering received; a large part of this book was written at nights and weekends.

Guido Gentile
Klaus Noekel

This book is based upon work from COST Action TU1004, supported by COST (European Cooperation in Science and Technology).

COST (European Cooperation in Science and Technology) is a pan-European intergovernmental framework. Its mission is to enable break-through scientific and technological developments leading to new concepts and products and thereby contribute to strengthening Europe's research and innovation capacities. It allows researchers, engineers and scholars to jointly develop their own ideas and take new initiatives across all fields of science and technology, while promoting multi- and interdisciplinary approaches. COST aims at fostering a better integration of less research intensive countries to the knowledge hubs of the European Research Area. The COST Association, an International not-for-profit Association under Belgian Law, integrates all management, governing and administrative functions necessary for the operation of the framework. The COST Association has currently 36 Member Countries (www.cost.eu).

“COST is supported by the EU Framework Programme Horizon 2020”

Part I
Public Transport in the Era of ITS - Francesco Viti

Chapter 1
Public Transport in the Era of ITS: The Role of Public Transport in Sustainable Cities and Regions

Xavier Roselló, Anders Langeland and Francesco Viti

Transportation is one of the most pervasive activities in any society or economy, since it enables the movement of people and goods from where they are to where they wish or they are planned to be. Among other transportation services, public transport systems contribute to a large share of the movement of people and therefore have an extremely important socio-economic role. Having a paramount importance for the quality and the stability of any socio-economic system, public transport planning, design and operations contribute to the equilibrium and the sustainable evolution of any region. Its importance for the founding, shaping and growth of urban agglomerations has been widely recognised, and the planning and design of transport services have had a major role in determining the locations of cities, their size, form and structure.

This chapter describes the central role of public transport for sustainable development, in comparison with alternative transport options and in particular the private transport modes.

Providing an unbiased and universal definition of sustainable mobility, or more generally of sustainable cities and regions, is not a trivial task. In the eyes of the city and its citizens, the term is undoubtedly a synonym of liveability, which encompasses a broad range of performance measures, both quantitative (e.g., the energy consumed/wasted, the pollution produced, the accidents, injuries and/or deaths caused) and qualitative (e.g., the risks involved, the stress caused). It is perhaps easier to generally accept the direct relation between the benefits that some characteristics of the transportation systems provide to sustainability. These are, for example, the increased accessibility thanks to faster or more reliable transport services, the reduced

F. Viti (✉)
University of Luxembourg, 6 rue Coudenhove-Kalergi, 1359 Luxembourg, Luxembourg
e-mail: francesco.viti@uni.lu

A. Langeland
University of Stavanger, Ullandhaug, 4036 Stavanger, Norway
e-mail: anders.langeland@uis.no

X. Roselló
Autoritat del Transport Metropolità, Muntaner, 321, 08021 Barcelona, Spain
e-mail: xavier.rosello.m@gmail.com

G. Gentile and K. Noekel (eds.), *Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems*, Springer
Tracts on Transportation and Traffic 10, DOI 10.1007/978-3-319-25082-3_1

negative externalities to the environment and the other citizens thanks to greener and more energy efficient vehicles, or the extra opportunities and indirect impacts to the society and the economy that new public transport services can offer (e.g., requalification of city centres, new employments, social justice and fairness). These benefits make public transport investments a very important political instrument and a fundamental driver for economic competitiveness of a region.

Sustainable cities and regions set specific transport policies aiming at achieving minimum modal split targets favouring the use of non-motorised modes of travelling (walking and cycling) as well as collective transport alternatives (public and private shared transport systems). These policies rely on the accuracy of the models described in this book, which provide an estimation of the demand attracted by public transport, the passenger flows moved by it and the performance of the level of service, which in turn is needed to estimate the attracted demand.

Key aspect in this chapter is the tight and mutual relation between public transport structure and organisation and (sustainable) urban development, through the essential concepts of mobility and accessibility. It is described therefore how public transport planning and design is determined and can be a determinant of cities and regions expansion, and how it can help at guaranteeing a sustainable development, i.e., how expansion can be controlled, in order to limit the essential problem of urban sprawl, is limited, and how it can efficiently be designed and managed in order to achieve modal split targets aimed at keeping energy consumption and environmental externalities within the international standards.

Together with the following two chapters, this chapter helps the reader to understand the complex interactions between public transport service design and organisation, including the adopted engineering and telecommunication technologies, and the mobility within, and sustainability of, our cities.

1.1 Accessibility and Social Exclusion

Transport and mobility are indispensable elements of socio-economic development in any country, and between countries. The free movement of people and goods is one of the foundations underpinning the creation of the European Union. Exercising this right is one of the main reasons for creating this common area of exchange and community. In fact, the European Commission's own White Paper on transport, *European transport policy for 2010: time to decide,* recognises that mobility is a right and even a conquest resulting from the community transport policy implemented since 1992. Ultimately, free movement is one of the key expressions of the concept of freedom. The right to mobility is therefore a right recognised in Europe, but this does not mean the right to a specific mobility model. Actually, a right to sustainable, universal mobility is recognised, and these two adjectives appear in the EU's *White Paper on Transport.*

Historically, the term of *transport and mobility* is often confused, whereas the difference between them is significant: to refer to journeys made by people from

origin to destination, the word traditionally used was "transportation". However, when the non-motorised modes such as walking and cycling, often referred to also as soft or active, became object of study and legislation, another word was used to refer to them: "mobility", because they could no longer fit in the traditional "transportation", as conventionally conceived. With time, semantic slip continued and "mobility" was used finally, and this is its current sense, to refer to the set of journeys made by a person or a group of people using any transport mode. Instead, the word "transport" is usually reserved to the infrastructure and vehicle system that supports and enables mobility in mechanical modes.

Mobility is an essential element in any urban and regional development, as it represents the link between activities both in space and time, and contributes to determine activity patterns and trip chains. Mobility has an important social and economical impact, given the traditional view of cities as markets and social interaction centres; improved mobility due to faster and more comfortable transportation means has favoured the spatial development and interaction of different activity centres. Mobility can also be defined as a measure of the ability to move efficiently from one activity location to another. This measure is directly influenced by the layout of the available transportation network and the level of service it offers.

The aim of mobility is therefore access, not movement. The aim of citizen's mobility is to enable access to a place or service, not travelling per se. Moreover, there is a wide range of individual conditions and access priorities, which makes accessibility a relative concept, not easily measured with simple, quantitative parameters such as travel times or distances.

Considering the large variety of objectives in people's mobility is a basic premise in designing an inclusive mobility policy that incorporates all options and does not marginalise or restrict any citizen. As far as possible, we need to offer as many options as possible to use the city and territory and to guarantee the access to the desired activities within a satisfactory level of accessibility. However, a public policy of mobility cannot take this condition for granted: on the contrary, the right circumstances need to be established to ensure that the universal right to access can be exercised under reasonable conditions of efficiency and cost, and this also involves a free choice of the mode of transport.

In many regions, the urban landscape is changing, and monocentric cities are transformed into polycentric urban areas, which in turn have an impact on transportation characteristics such as trip lengths and mode choices. These features indirectly affect social characteristics such as urban densities and the distribution/dispersion of activities. Mobility may on the other hand be a determining factor in *social exclusion. Social exclusion* can occur to individuals or entire communities of people, which are systematically blocked from rights, opportunities and resources (e.g., housing, employment, health care, civic engagement, democratic participation and due process) that are normally available to members of society and which are key to social integration, for instance, if lack of access to public transport or a vehicle prevents a person from getting to a job, training course, job centre or doctor's surgery and entertainment venues. Car ownership and access to a car vary widely. For instance, in Stavanger, Norway, only 7 % of the

population is without access to a car (source: Travel survey 2012), while in the metropolitan region of Barcelona, more than half of the population does not have regular access to a car (source: the 2013 Daily Mobility Survey in the metropolitan region of Barcelona), either because they do not have a driving licence or a car, or because they cannot access the car(s) that belongs to the household. Hence, it is difficult to see how the right to universal, sustainable mobility can be guaranteed with a model based on the predominance of private vehicle use over that of other modes of transport, as is the current situation.

The planning of mobility must therefore bear in mind the fact that guaranteeing access entails a strong social component. The current model depicts that people who cannot use a car constantly have fewer opportunities to access jobs, centres of study, leisure or other facilities. For some social groups, limited mobility is often in addition to other inequalities for reasons of wealth, age or gender. That is why choosing one mobility model or another can either encourage equality or increase inequality. It is therefore entirely reasonable to define as fair a mobility model as possible. The current model is neither immutable nor an inherent right.

The right to mobility infringes on other rights. The right to health and the right to a good quality environment are also universally recognised rights in the Declaration of Human Rights, among others. The current modal structure of mobility in many European cities is partly responsible for these rights being violated. Levels of air pollution and accidents are two easily recognisable consequences that affect these other rights, just as universal as that of mobility. Consequently, regulation and choosing a sustainable option that makes these rights as compatible as possible is an obligation for planning because universal rights are all equal. No single right takes precedence over the others, hence their greatness and the difficulty in applying them.

We therefore need a systemic view and to establish solutions via compromises. These compromises have to be found between transport system developments, their impact on the socio-economic system, between different modes of transport as much as within each transport system, and to realise that different forms and organisations affect the mobility through different accessibility levels. This will be broadly explained in Chap. 2. These forms are inevitably determined and vice versa determine the structures of cities and their growth, as explained in the next section.

1.2 City Structure and Its Growth

1.2.1 Urban Sprawl and Socio-economic Transformations

The population growth observed in our conurbations over the last few decades has, together with the increase of car ownership rates, facilitated the spreading of the population over an increasingly extensive territory, constantly going beyond its outer limits and also exhaustively using up all its areas. This expansion of the

territorial scenario sets more challenging requirements for transport infrastructures and services. The movements associated with the internal redistribution of the metropolitan population generally entail the population moving from consolidated urban areas towards traditionally sparsely populated areas. As a result, the density rates for the whole of the territory gradually become more similar, tending to lower the minimum threshold required by a transport system, especially collective transport, in order to be efficient. We are therefore directed towards a situation in which it is no longer efficient to implement any public transport system due to low densities.

This spread of the population throughout the territory also occurs at a local level and is often followed, most particularly in the outskirts of metropolitan regions, by a dispersed model of urban development that reduces the population density even further and makes it difficult to offer transport, even within municipalities, which could compete with the private transport modes, and without strongly subsidising the services. Qualitatively, the movements associated with this redistribution of the population, far from being homogeneous, tend to follow patterns of segregation, both demographic and social, throughout the territory. The result is territorial heterogeneity with the emergence of differing requirements in terms of mobility and transport.

Along with purely demographic elements, other factors of a social nature such as changes in the population's habits or higher income levels also affect the demand for mobility. A population with increasing buying power and an ever greater level of information becomes more demanding in terms of what it wants, when and how. And these greater demands are more spread geographically, with the consequent increase in mobility caused by such an attitude. In summary, we need to serve more people who are more sparsely distributed within a more extensive territorial area. And they must be served while handling a more unpredictable demand in terms of this being variable, as well as more demanding.

With regard to the destination, movements associated with the exit from urban centres to less densely populated areas have not tended to follow exactly the same itineraries as residential migrations and, in many cases, have resulted in a more functional specialisation of the territory, in which some municipalities tend to hold a much larger proportion of jobs than their employed resident population, while others face the opposite situation. Looking further at the municipality level, new land use planning concepts have also favoured in some cases a polycentric structure, where multiple activity centres are developed and around where residential zones expand. Accordingly, transportation systems become essentially developmental factors for the expansion around these centres and for their interaction.

Following the evolution of cities as they are growing today, an even more dispersed development, where activity areas get more specialised (e.g., commercial centres, technology parks, industrial zones, etc.), is observed. This is a consequence of the increasing transport and logistic costs, which suggest firms within a supply chain to obtain benefits from locating near each other. In the case of tertiary activities aimed at the population, especially commerce and leisure, their location at points close to the basic road network but far from residential areas has resulted in

growing voluntary mobility via individual means of transport. A large number of industrial installations, but also tertiary, have often opted to segregate areas away from consolidated urban developments. In addition to the proliferation of such areas of activity, the wide choice offered has also led to a clearly more extensive use of the area occupied so that, on most occasions, the resulting facilities aggravate the problems of inefficient collective transport.

1.2.2 Consequences of Expansion for the Transport System

As we have seen, over the last few years, developments in the location of various socio-economic agents throughout the territory have led to an increase in mobility and in the need for personalised transport modes caused by the need to travel within a context of growing functional specialisation, by the requirements of more demanding agents and the greater possibilities to travel afforded by increased car accessibility.

This rise in mobility has affected the transport system performance, in terms of its capacity, since it must meet a greater demand, its functionality, since it must reach more distant and increasingly out-of-the-way destinations, and also its flexibility, since it must attend to a more variable and changing range of demands.

The demand for expansion is a direct consequence of greater distances covered in travel, which also lead to a growing need for mechanical means of transport in detriment to soft travel modes such as foot or bicycle.

Finally, the demand for flexibility is essentially due to the progressive *individualisation* of travel. Within a scenario where mobility flows tend to be spread among an increasing number of destinations and where commuter travel differs not only from that of the rest of individuals but also from that of the trips carried out by the same individual from one day to the next, movements are tending to display more distinct characteristics: they are tending to become more individual.

This individualisation of travel naturally makes it difficult to collectivise and usually shifts the growing pressure on the transport system onto individual means: the supply of collective transport cannot meet the growing diversity of demands and remains limited to those where minimal volumes of users guarantee its efficiency, so that the rest of travel is mainly left to means adapted to a weak demand as "dial-a-ride", car-pooling, taxi or to the last solution, the individual modes.

Most of this increased demand for mechanical transport entailed by the reduction of trips made on foot will not be met by collective modes of transport but will mainly be passed on to individual means. Even though the relative weight of the former is increasing in some municipalities, principally in those with railway networks, the gains made by individual modes of transport are greater in almost the whole metropolitan region. In other words, whoever has more and better collective transport uses it the most.

Public transport efficiency is therefore a powerful instrument to influence the city structure and growth, as much as the latter determines the potential demand for

public transport. Public transport strongly determines the accessibility of an area and therefore its attractiveness for locating activities depending on the resulting land uses and prices, which in turn creates mobility needs. This completes a cycle, where the same accessibility of the different activity locations determines car dependency (or independency), trip frequencies, modal splits and ultimately the observed spatio-temporal traffic patterns and passenger flows, which finally determine the travel costs for each transport subsystem.

On the other hand, improved accessibility increases travel speeds and therefore reduces the gradient of the accessibility line, as conceptually sketched in Fig. 1.1. Improved accessibility therefore means the same distance can be travelled over a shorter period of time, but it also means that a greater distance can be travelled over the same period of time.

This twofold consequence of improved accessibility is interpreted differently by those responsible for designing mobility policies, which make decisions following the first consequence and therefore going from t_1 to t_2 while d_1 remains the same, and by end users, who take advantage of the second derived effect (with t_1 remaining the same but going from d_1 to d_2). This behaviour has been found to be independent on the geographical or temporal differences among travellers and has been first established by Zahavi in the '60s as one of the invariants in average travel behaviour: on average, people devote a fixed proportion of their daily available time to travel. According to this author, there is a "travel time budget" similar to all citizens regardless of their income, which was reported to be about 1–1.2 h by different empirical investigations.

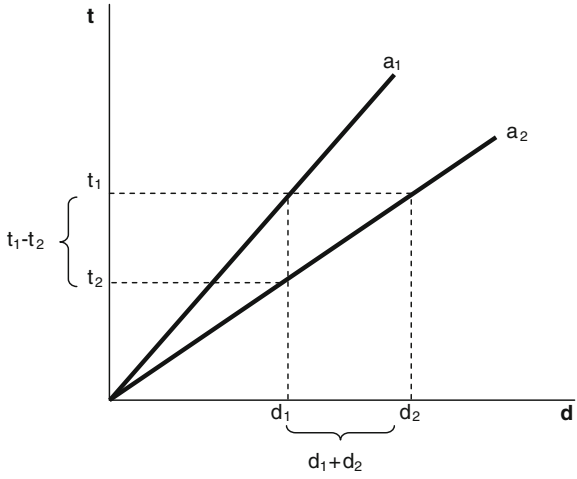

Fig. 1.1 The effect of improved accessibility on users

The identification of stable daily travel time budgets brought home to Zahavi the importance of travel speeds. He realised that travellers did not save time as a result of increases in travel speed, but that they applied the time saved from some trips for additional travel. This led him to explain the resistance of travellers to switch from private cars to slower public transport modes and to conclude that public transport could attract significant numbers of passengers from private cars only by offering higher door-to-door speeds.

Other invariance principles were also found in different other studies, which are based on the share of disposable income (about 10–15 %) and on the average number of trips done in a day (about 3–3.5). Conversely, the number of kilometres travelled each day has been found to be not stable, and to be a reliable proxy measure of urban expansion. These principles allowed developing predictive models for urban development based on the proposition that travellers tend to maximise distance travelled within their constraints of time and money.

Taking these principles into account, territorial planning policies help to establish a limit on urban expansion thanks to progressive improvements in accessibility, as well as limiting urban development beyond a certain distance from the basic transport network, thereby converting improved accessibility into savings in time.

Consequently, limiting expansion in terms of how a territory is occupied represents a significant contribution, not to the rise in the number of trips but to their growing trend to be via individual modes of transport; it is progressive dispersal between various destinations.

Moreover, the increase in population, either occupying more territory or increasing density, leads inevitably to a higher demand for transport. Accordingly, supply as well as its complexity must increase, so that for even relatively small population area, a correct assessment of the adequacy of supply to current or predicted demand can only be carried out by means of modelling, of which the demand assignment on the network is a relevant chapter. The present document is intended to be a contribution in this field.

1.3 Energy Consumption and Efficiency

It is not possible to discuss the role of public transport in sustainable urban development without discussing the energy consumption associated with the different means of transport. Targets for urban expansion and for the accessibility of cities and regions are needed to limit the increasing daily number of kilometres travelled by transportation users, which, in combination with modal split target, determine the total energy consumed/wasted by the mobility and the related negative externalities caused by it.

As stated by the Brundtland Commission of the United Nations in 1987, "*sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs*". In accordance with this definition, the higher the consumption of non-renewable fuel, the less sustainable the mean of transport should be considered. To consolidate this idea, this subchapter compares the energy consumption of different modes of transport and shows the advantages of public transport in this area.

By final energy, we intend the amount of energy consumed in the vehicle's tank or at its connection with the electricity grid. From an energy point of view, railway transport always consumes less per ton in movement than road transport, since it has less mechanical resistance. In comparison with road transport, the railway energy problem lies in the use of greater tares for equivalent transport, especially in passenger rail, notably increasing gravity resistance (gradients) and acceleration resistance and, to a lesser extent, mechanical resistance. Establishing a unit of consumption that allows us to compare railways with the other modes of transport is not easy. The different types of resistance that affect final energy consumption are, in turn, affected by the layout of the line, how it is used, the vehicle characteristics (e.g., vehicle category, age, fuel type, tare), features provided for passengers (auxiliary equipment, air conditioning), driving behaviour (e.g., aggressive style), the meteorology (e.g., rain, wind speed), etc.

1.3.1 Beyond Movement

In addition to the energy consumed by the vehicles while in operation, a thorough analysis of energy consumption should also include the energy employed in their construction, commercialisation, operation, maintenance and disposal, as well as the energy costs invested in constructing the corresponding infrastructures. Energy consumption due to the manufacture of mobile material is not insignificant if we take into account the amount of materials used and the operating life of the different modes of transport in terms of distance covered. For a utilitarian car with a life cycle of 200,000 km, the energy cost of its manufacture could represent 20–30 % of the total energy cost (manufacture and operation). Table 1.1 shows indicative values for the different modes of transport.

Table 1.1 Energy consumption in urban zones (in MJ/person-km), taking into account the average occupancy of the vehicles in urban zones (*Source* UITP)

	Construction	Operation	Total
Bicycle	0.5	0.3	0.8
Tram	0.7	1.4	2.1
Bus	0.7	2.1	2.8
Heavy rail	0.9	1.9	2.8
Petrol car	1.4	3.0	4.4
Diesel car	1.4	3.3	4.7

1.3.2 Primary Energy and Fossil Fuel

Having analysed transport energy-related impacts, including movement, the transformation of energy from its natural state and vehicle construction, it is useful to distinguish between *primary energy* and *fossil fuel primary energy*, making the following observations:

- The greatest problem for energy comes from the use of non-renewable energy sources (coal, gas, petroleum) or the use of technologies that entail great risk or uncertainty (nuclear), not in the consumption of the primary energy *per se*.
- In the transport industry, the biggest energy inefficiency occurs in the combustion of non-renewable resources (coal, gas, petroleum), be it directly within internal combustion vehicles or in the power plants supplying the energy for railways, trolley cars, cable cars, etc. This combustion is responsible for climate change, local pollution and the depletion of non-renewable energy sources.
- Connecting vehicles to the electricity grid reduces the consumption of fossil fuel primary energy thanks to the *energy mix,* but it also involves an increase in demand that must be handled, for example, by using more efficient modes of transport.
- Renewable energy sources can replace energy from non-renewable sources or sources that pose a threat to humanity. Mediterranean countries receive a significant amount of solar radiation and sufficient wind to meet their current energy demands.
- Non-renewable fossil fuels have taken millions of years to form and, with today's rate of extraction, are economically and environmentally unviable in the medium and long term.
- In the case of vehicles with heat engines, it should be noted that all the energy consumed for movement is from fossil fuels, while vehicles that run on electricity will only depend on fossil fuel energy depending on the energy mix.

To determine the actual energy impact of transport activity, it is necessary to identify the entire energy chain, from its state in nature (primary energy) to consumption (final or consumed energy). The transition from primary to consumed energy is not straightforward, and it relies on a series of transformations and transport of the energy which involve additional energy amounts. These are represented in as *energy transform*.

Figure 1.2 shows the energy consumption (in MJ/person-km), by mode of transport, taking into account the average occupancy of the vehicles in urban zones. As can be seen, the high consumption of private vehicles in the city is between 2 and 4 times greater than public modes of transport.

It is therefore a primary goal for sustainable development to incentivise a shift from individual modes of transport to collective modes, given the enormous savings achievable per individual trip. This applies generally also to the externalities caused by transport, which are described in the next section.

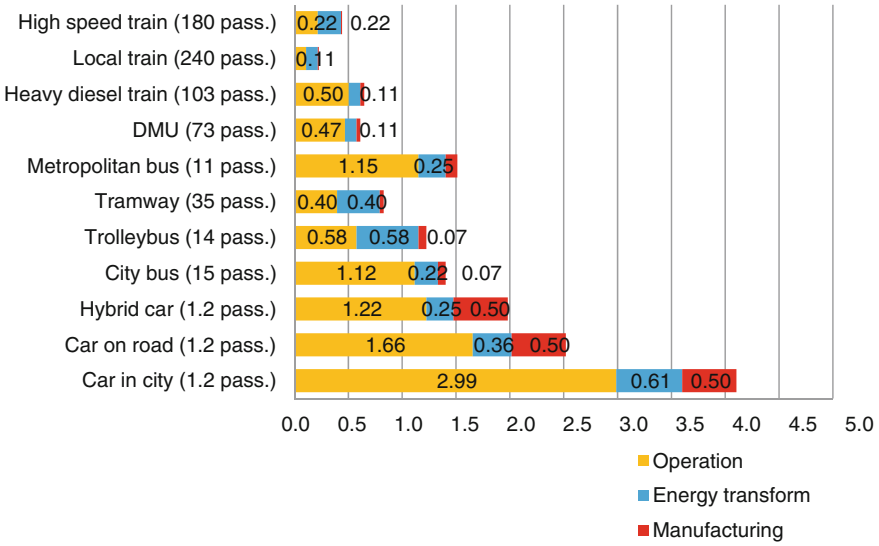

Fig. 1.2 Energy consumption (in MJ/person-km), by mode of transport, taking into account the average occupancy of the vehicles in urban zones

1.4 Externalities

1.4.1 Greenhouse Gas Emissions

Greenhouse gas emissions are widely recognised as being responsible of the global climate change. Among them, CO_2 contributes with the 85 % of the total greenhouse gas emissions in terms of produced volume (Fig. 1.3). However, there are other gases as methane (CH_4), nitrous oxide (N_2O) and chlorofluorocarbons (CFC), which contribute with 8, 5 and 2 % to the total greenhouse gas emissions, respectively. (Source: Inventory of US Greenhouse Gas Emissions and Sinks, 2008, provided by the Environmental Protection Agency, EPA).

When we think of the external costs of mobility, CO_2 emissions have always been considered as critical, having a very close correlation with the fuel employed in the mobility process irrespective of engine performance and efficiency. Only the railway can modify this value, since the share of production of electrical energy is always cleaner in terms of CO_2.

In a study carried out in the metropolitan area of Barcelona, it was determined that the annual CO_2 rates of variation for the period 2006–2010 fell overall by 1.36 %, a period of time towards the end of which the effects of the crisis were beginning to be felt (Table 1.2). Only motorbikes saw a positive annual change (1.05 %) in that period. This increase could be attributed to the tendency to renew

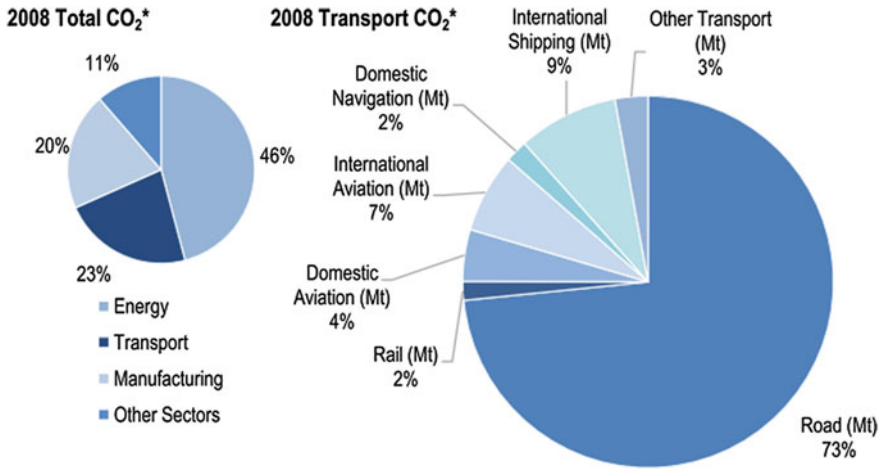

Fig. 1.3 Worldwide CO_2 emissions in 2008 by sector and disaggregated by transport sector (*Source* ITF/OECD 2010)

Table 1.2 Unit emissions of CO_2, PM and NO_x in 2010 by vehicle type and trend compared with 2006 in the metropolitan region of Barcelona

Type of vehicle	CO_2 (g/veh km)		PM (g/veh km)		NO_x (g/veh km)	
	Value 2010	Δ% *06–10*	Value 2010	Δ% *06–10*	Value 2010	Δ% *06–10*
Car	173.43	−0.63	0.042	−0.95	0.58	−3.77
Motorbike	86.47	1.05	0.050	−6.90	0.14	0.44
Bus/coach	1039.13	−0.80	0.306	−9.74	8.78	−6.30
Railway	767.92	−10.02	0.210	0.00	1.01	0.00
Total	222.32	−1.36	0.069	−4.10	0.99	−5.02

motorbikes with more powerful ones and, consequently, with higher relative emissions. As a consequence of the greater share of natural gas consumption in the bus and coach segment, relative emissions increased between 2006 and 2010 since higher energy consumption results in greater CO_2 emissions.

1.4.2 Other Pollutant Emissions

Greenhouse gas emissions cause negative externalities at a global level, i.e., the total volume produced must be regulated. Other pollutants cause health issues through direct exposure. The most relevant in the transport sector are particulate matters (PM) and nitrogen oxides (NO_x). In addition, other pollutant emissions such as carbon monoxide (CO), sulphur dioxide (SO_2) and some volatile organic

compounds (VOC) as benzene (C_6H_6) can have an important impact. The European Environmental Agency publishes reports about air quality in Europe, where these pollutants are monitored by country.

While the individual car and truck have become more efficient, in general both the global emissions and the local ones such as PM and NO_x increased because of the rapid growth in vehicle kilometres travelled. Barcelona, however, has reduced emissions between 2006 and 2010, primarily caused by the economic crisis. Analysing the energy consumption in urban zones (in MJ/personkm), while taking into account the average occupancy of the vehicles in urban zones (Source UITP), both PM and NOx reductions show a sharper trend compared with the variation in mobility; it is therefore useful to analyse the trend in emissions in relation to that of mobility.

Among the different types of vehicles, buses are the vehicles with the overall highest rates and the highest relative reductions. In the case of PM emissions, these segments of vehicles have seen an annual reduction with an average variation of 5.07 % in the case of trucks and 9.74 % in the case of buses. Since the aim of the EURO standards is to apply more restrictive limits on PM emissions by diesel vehicles than petrol, those categories where almost all vehicles are powered by diesel (heavy goods, buses and light goods) have seen a sharper relative reduction in their emissions. As a consequence of the EURO standard policy, PM emissions in 2006–2010 fell by 9.7 % for buses and coaches and 0.95 % for cars. With regard to the trend in NO_X emissions, the relative reduction over the same period ranges from 6.3 % for buses and coaches to 3.15 % for light goods vehicles. The only case where this average annual variation is positive is in the motorbike segment, where once again the tendency is towards more powerful bikes, leading to a rise in emissions.

1.4.3 Noise

According to the EU, noise in cities, towns and their surrounding areas is a growing phenomenon; 80 % of this noise comes from traffic. About 31 % of European citizens believe that noise is the main environmental problem in their cities. Currently, in Europe, one hundred million people are exposed to noise levels above those recommended by the WHO (World Health Organisation), namely 55 dB(A), causing discomfort and having a harmful effect on sleep and quality of life. Moreover, around 40 million people in the EU are exposed to noise levels above 65 dB(A), the threshold where serious effects on health appear. The European Commission's *Thematic Strategy on the Urban Environment* recommends reducing the volume of urban traffic, greater fluidity and reduction at source to reduce noise pollution in European cities.

1.4.4 Congestion

The *European Commission's White Paper* on transport reports that congested networks represent a highly significant threat to economic competiveness and productivity. According to its data, the externalities of congestion attributable to road traffic in 2000 represented 0.5 % of the EU's GDP (Gross Domestic Product). The European Commission attributes this phenomenon of growing congestion to the current situation of mobility, where transport users do not pay the costs they actually generate. This failure to *internalise the external costs* generated by each mode of transport results in a modal imbalance (75 % of passengers and 44 % of goods travel by road, while only 6 % of passengers and 8 % of goods do so by rail) which, together with deficiencies in modal exchanges, gives rise to congestion. It should be noted that the concept of congestion is important as, while in congestion, every car involved increases the cost for the rest of the users. This results in appreciably higher costs as car drivers who cause congestion do not compensate the people they are affecting.

1.4.5 Consumption of Public Space

The existence of an infrastructure in the territory entails the occupation of space that could be occupied by any other activity. This means that space allocated to new building work has an opportunity cost. However, there is another intangible but also important cost, namely the occupation of public space: a space which vehicles must share with citizens. Given the fact that people travel on average about 1 h daily, according to Zahavi's paradigm, it is clear that the largest majority of time vehicles are not used and therefore simply waste public space. In the European Commission's *Thematic Strategy on the Urban Environment*, Eurobarometer data are provided that show that 51 % of European citizens believe traffic is the main environmental problem in cities, highlighting the fact that the occupation of public space by traffic impairs quality of life and makes it difficult for them to feel they belong to a neighbourhood or local community. Historically, debate regarding priorities in the use of public space has been more focused on urban design than mobility, but this is an area that also affects mobility policy and to a great extent. Reclaiming public space for pedestrians and bicycles in detriment of private vehicles has not only been debated for some time now in Europe but is also a desire expressed by its citizens.

1.4.6 Safety and Security

The European Commission's document from 2004: "Towards a thematic strategy on the urban environment" complains about the unacceptable externality caused by traffic accidents in urban zones of the European Union. According to the OECD

database on road traffic and accidents, the cost of accidents represents 2 % of the EU's GDP. Two-thirds of traffic accidents occur in urban zones and 50 % of the deaths are also located in these zones.

The role of public transport (PT) in improving safety is fundamental. Figure 1.4 gives an idea of how PT usage and traffic fatalities are found in inverse relationship. This figure shows also how different national policies determine a systematic difference between the distances travelled by PT users in Europe as compared to other countries worldwide such as USA, Canada and Australia. An interpretation of this figure is that policies aiming at fostering PT usage may have great effects in terms of safety, especially if one looks at Northern European statistics. Another interpretation may be that Northern European countries invest significant efforts in providing safe PT services.

The impact of PT on safety is clearly positive if one looks at the statistics such as in Fig. 1.4 and by using both interpretations. These relations are certainly affecting decisions at the city and regional level, but they also have an effect on the individual travellers, as public transport is generally seen as a significantly safer mode than other motorised modes.

Security issues are also carefully analysed and addressed by municipalities. Vehicle safety, as well as security in stations is determinant for the feeling of satisfaction towards a service. In turn, feeling of personal insecurity is seen as a barrier to public transport use.

Chapter 3 will discuss how ITS can help in improving safety and security aspects, and how past experiences have already provided quantifiable benefits in terms of users' stated satisfaction and overall increase in PT usage.

In conclusion, safety and security represent factors that certainly play a role in the overall perception of quality of services. Nevertheless, a direct relation between these quality elements and travellers' mode choice is not easily quantifiable and normally these are included, together with other factors, such as comfort in its broad definition, in the so-called alternative-specific constants, which will be introduced in Part II of this book.

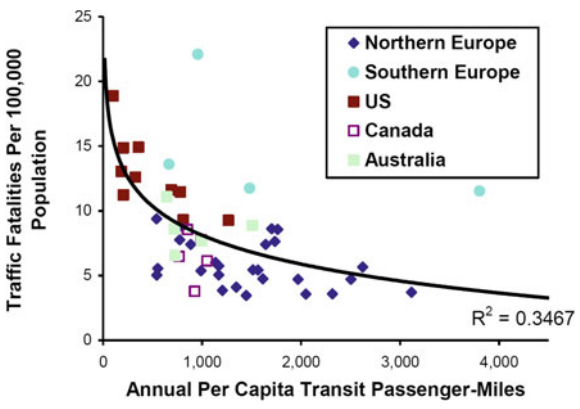

Fig. 1.4 Relationship between distances travelled by PT and fatalities (*Source* Litman 2014)

1.5 Unit Mobility Costs

The costs for mobility are not easily quantifiable and interpretable, given the complex mixture of monetary, time-related and psychological elements. If costs are presented in an aggregate manner, they may be subject to incorrect interpretations and in turn suggest policy-makers to adopt measures that are contradictory or not effective.

Table 1.3 gives an example of how costs for mobility, if presented without sufficient insight, may lead to wrong interpretations. This table presents the costs of transport as quantified by the metropolitan region of Barcelona in 2010. These figures come from several studies commissioned by Barcelona's Metropolitan Transport Authority (ATM) and refer to the metropolitan region. Strictly speaking, therefore, they are only applicable to this territory although, with little variation, they would also be valid within a European context.

The goal of this table is to answer the most general question possible that might be asked regarding the costs of travel. How much, in total, does it cost to travel via public transport and via private transport, which sectors of society meet this cost and to what extent? The internal costs (in bold) consists of monetary and time-related costs. External costs, also shown in bold, are then added to the internal ones and together they represent the total costs.

Note that only two items have been taken into account for *internal* costs, corresponding, respectively, to the economic value attributed by users to their *time* (the so-called value of time, as introduced and discussed more extensively Part II of this book) and the *monetary costs* with pecuniary repercussions. The estimation of time cost has been carried out from a basis of the average speed of a door-to-door journey and a time value of 9.50 €/h, usual value in this kind of studies, in either public or private transport. The different time value that appears in the table comes from the different average duration, bigger in public than in private one. Regarding *external* costs, all those quantifiable elements detailed in the preceding subchapter on externalities have been included.

By a first look at the internal costs, it is striking to observe that while monetary costs are somewhat a more objective factor, time-related factors would point at a policy for which travel time savings in terms of distances travelled using the private transport system would imply a higher gain for the traveller. This would, for example, suggest policies aiming at investing on public transport infrastructure or rolling stock that would overall reduce the total travel time. This means that policies

Table 1.3 Unit costs of transport in the metropolitan region of Barcelona in 2010

	Public transport (€/traveller-km)	Private transport (€/traveller-km)
Monetary	0.205	0.354
Time	0.360	0.295
Internal	**0.565**	**0.649**
External	**0.027**	**0.102**
Total	**0.592**	**0.751**

aiming at incentivising PT use especially for this class of users would certainly reduce the gap between time-related costs. This interpretation may certainly benefit from the insight acquired by using transit assignment techniques. Thanks to a correct modelling of travel behaviour and service performances, predictions on how the above unitary values would change are necessary.

1.6 Mobility and Public Transport in European Metropolitan Areas

This chapter is concluded with an in-depth analysis of the transport systems and their performances in different European metropolitan areas. Understanding the state of practice in this context is a prerequisite to define pertinent policies from case to case, based on the concepts described so far in this chapter.

Decisions on public transport affect the daily lives of millions of people; the investment and operation costs of complex systems often amount to millions of euros, if not billions, and have a decisive impact on the economic dynamism and environmental quality of urban areas. To achieve this end, it is important to:

- Define pertinent territories, corresponding to the reality of the mobility of people. Analysis should capture the actual land use and the resulting mobility patterns, beyond the administrative boundaries of local authorities or transport companies.
- Determine a set of key indicators that should be collected and reviewed regularly so as to have a clear view of the main trends underway.
- Take into account not only public transport but also mobility in a broader sense, naturally including trips involving private cars but also taxis, bicycles and walking.

These requirements motivate the cooperation within an international association of local authorities, the association of European Metropolitan Transport Authorities (EMTA), with the goal of collecting data and publishing its *Barometer* every year, which gathers together the most relevant indicators in each metropolitan area. The latest issue, summarised in the next subsections, is for 2011.

1.6.1 The EMTA Association

The association of EMTA brings together the public authorities responsible for planning, coordinating and funding the public transport systems of 28 of the largest European metropolitan areas, plus Montreal (Canada).

One of the main deliverables of EMTA is the *Barometer*. This document offers an overview of what public transport in Europe is today, through its main metropolises. The raw data, as provided by EMTA members, are included in the master at

the end of this chapter, and the comments and graphics shown in the next few paragraphs have been produced based on figures taken from these tables.

1.6.2 Some Mobility Indicators in Metropolitans Areas

According to the latest Barometer, *car ownership rates* are twice as high in some cities as in others (603 cars per 1000 inhabitants in Turin vs. 327 in Budapest and 354 in Copenhagen).

In Fig. 1.5, different clusters appear and they seem to show that several wealthy metropolitan areas have a relatively low car ownership ratio (under 450 cars/1000 inhabitants) and lower use of private cars. In other words, public transport authorities have growing responsibilities in metropolitan areas to offer attractive public transport services to a less car-dependent community whereas, on the other hand, other equally wealthy areas yield high rates of public transport share. This seems to prove that the car ownership ratio is becoming a relatively less significant variable to explain the modal split. Other factors such as urban density, family size, the existence of efficient public transport systems or the cost of using and parking cars can lead to lower car ownership rates.

Public transport accounts for more than 29 % of all trips (48 % considering only motorised trips) in the densest parts of most European metropolitan areas (in the main cities), illustrating its fundamental economic, social and environmental role in large urban territories. Soft modes (walking and cycling) account for 39 %, this falling to 32 % for the rest of motorised modes (mainly private cars).

As one can observe in Fig. 1.6, most of the main cities achieve more than 60 % of modal share, which is generally used as threshold for what we can consider as

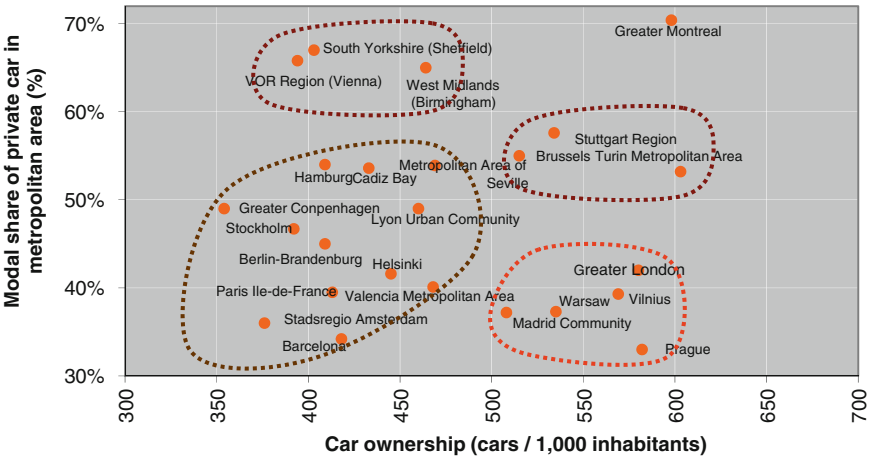

Fig. 1.5 Car ownership ratio versus modal share of private car: not a clear relationship

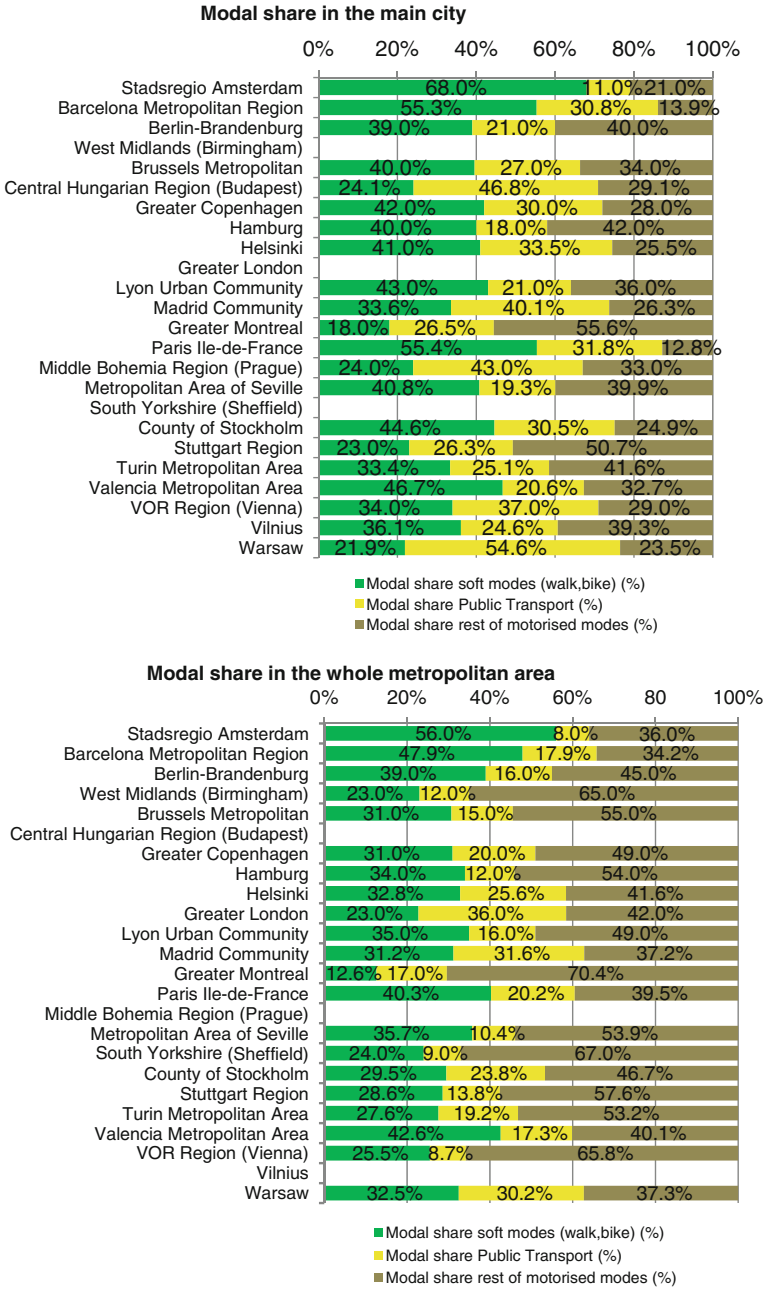

Fig. 1.6 Modal share in the main city versus in the whole metropolitan area

"sustainable mobility" (as the sum of public transport and soft modes). Amsterdam, Barcelona, Budapest, Copenhagen, Helsinki, Madrid, Paris, Stockholm, Vienna and Warsaw stand out with a rate over 70 %, illustrating the very dense public transport systems irrigating the heart of these capital cities and the deep-rooted habit of walking and/or biking in European cities. These metropolitan areas, together with Greater London, Berlin–Brandenburg and Valencia, have a clear predominance of sustainable modes over the private car, if one considers this predominance to be related to car ownership rates. Greater London, Madrid Community and Warsaw are the metropolitan areas among those surveyed where public transport accounts for the highest modal shares of all trips (between 36 and 30 %).

Generally, there is a gap between modal share in the main city and modal share considering the whole metropolitan area where public transport accounts, in average, for 18 % of all trips (27 % considering only motorised trips). Figure 1.7, which has remained relatively steady over the last few years, embodies one of the main challenges facing public transport authorities and operating companies in the coming years: to develop public transport in the suburbs and the less dense parts of the metropolitan areas.

Looking at the demand for public transport in Fig. 1.7, each inhabitant makes more than 250 journeys per year on public transport, more than one trip every

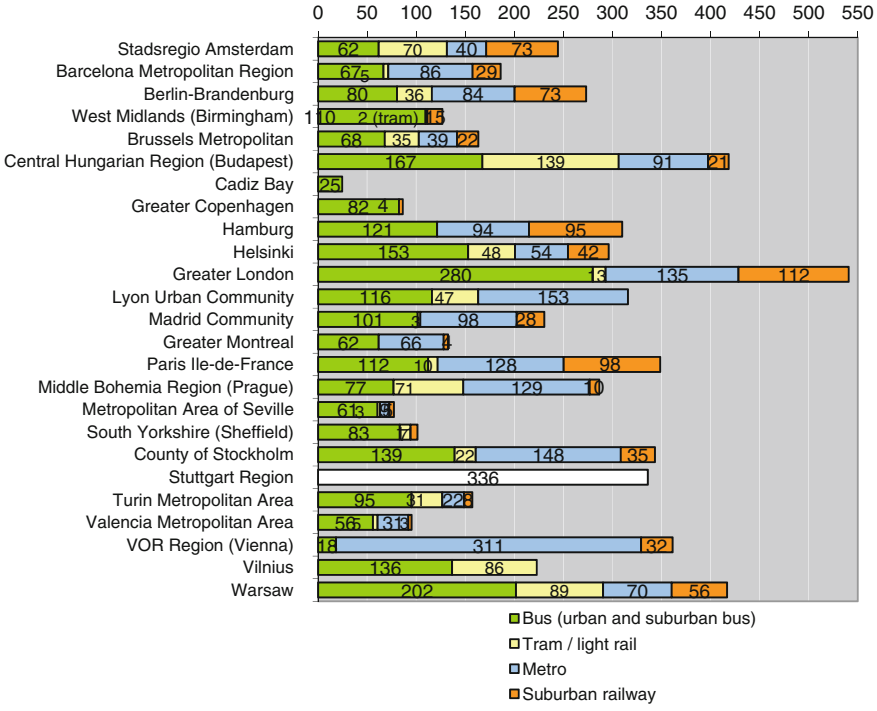

Fig. 1.7 Number of public transport journeys/year by mode

working day. In some cases, the total demand is over 400 journeys as in Budapest, Greater London and Warsaw. In half of the metropolitan areas, the share of the bus mode is still dominant. Over the years, the increase in public transport demand reflects the effort being made by authorities and operators to offer a high-quality public transport system with accessible vehicles and stations, using ITS (Intelligent Transport System) technologies to guarantee reliability and safety in operations and real-time information and contactless tickets to users to promote the use of public transport and make it more competitive compared with private vehicles.

The monthly pass fare in the main city compared to GDP per capita is a good indicator to show how expensive public transport is in a city as the general price standards are cancelled out. The annual GDP in a city divided by 12 gives an average ratio of 2.2 %. In particular, cheap are the monthly passes in Brussels, Copenhagen, Paris, Prague and Warsaw (1 %) as opposed to the highest prices in Sheffield (5.7 %), Birmingham (5.2 %) and London (4.6 %), all situated in the UK (Fig. 1.8).

1.6.3 Public Transport Subsidies

The rate of operational costs covered by fare revenues also varies greatly among the EMTA (Table 1.4), with some cities covering more than 50 % of operating costs with fare revenues but others being far from this figure. On average, among the surveyed metropolitan areas, the operating costs of public transport in 2011 were

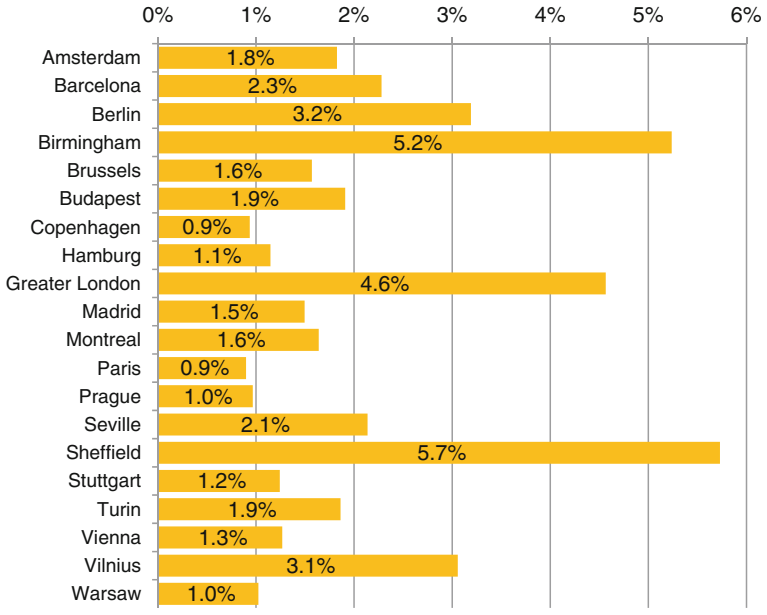

Fig. 1.8 Monthly pass fare in main city/monthly GDP per capita (%)

Table 1.4 Coverage of operating costs by fare revenue

	Population 2011	Population trend (2001 → 2011)	Annual GDP per capita in main city	Ratios of modal split of motorised trips						Total journeys on public transport/inhab./year	Coverage of operating costs by fares	Coverage of operating costs by subsidies	Petrol—price per litre	Monthly pass/single ticket in main city	Monthly pass fares in main city vs. monthly GDP per capita	Single ticket fares in main city vs. petrol price per litre
				Metropolitan area			Main city									
				Modal share soft modes (walk, bike)	Modal share public transport	Modal share of motorised modes	Modal share soft modes (walk, bike)	Modal share public transport	Modal share of motorised modes							
Units	Inhab.	%	€	%	%	%	%	%	%	n.	%	%	€	€/€	%	€/€
Stadsregio Amsterdam	1,424,137	8.4	28,000	56.0	8.0	36.0	68.0	11.0	21.0	244	53.9	46.1	1.600	28	1.8	93.8
Barcelona Metro Region	5,029,000	14.6	0	47.9	17.9	34.2	55.3	30.8	13.9	186	45.7	50.4	1.320	35	2.3	109.8
Berlin-Brandenburg	5,997,507	0.4	28,956	39.0	16.0	45.0	39.0	21.0	40.0	273	46.2	53.8	1.720	32	3.2	139.5
West Midlands (Birmingham)	2,738,100	7.1	22,777	23.0	12.0	65.0	na	na	na	127	na	na	1.653	49	5.2	123.2
Brussels Metropolitan	3,234,475	10.3	39,000	31.0	15.0	55.0	40.0	27.0	34.0	163	46.3	44.4	1.600	28	1.6	112.5
Central Hungarian Region (Budapest)	2,079,331	na	22,755	na	na	29.1	24.1	46.8	29.1	418	37.6	56.3	1.479	31	1.9	80.0
Cadiz Bay	707,245	9.4	na	36.6	9.8	53.6	na	na	na	25	72.4	26.4	1.000	35	na	100.0
Greater Copenhagen	2,491,090	5.4	57,649	31.0	20.0	49.0	42.0	30.0	28.0	86	49.8	50.3	1.642	14	0.9	196.4
Hamburg	3,405,000	28.9	50,300	34.0	12.0	54.0	40.0	18.0	42.0	313	68.0	32.0	1.520	34	1.1	92.1
Helsinki	1,131,372	10.3	na	32.8	25.6	41.6	41.0	33.5	25.5	298	49.7	50.3	1.560	17	na	160.3
Greater London	8,174,000	14.0	35,326	23.0	36.0	42.0	na	na	na	541	52.7	25.4	1.653	na	4.6	na
Lyon Urban Community	1,277,777	9.4	na	35.0	16.0	49.0	43.0	21.0	36.0	316	30.5	24.7	1.500	34	na	106.7
Madrid Community	6,489,680	20.8	38,183	31.2	31.6	37.2	33.6	40.1	26.3	230	44.0	56.0	1.320	48	1.5	75.8
Greater Montréal	3,777,499	10.5	38,679	12.6	17.0	70.4	18.0	26.5	55.6	132	50.3	43.2	0.952	24	1.6	229.0
Paris Ile-de-France	11,866,900	6.1	75,439	40.3	20.2	39.5	55.4	31.8	12.8	349	39.6	20.2	1.500	33	0.9	113.3
Middle Bohemia Region (Prague)	1,831,255	11.0	29,673	na	na	33.0	24.0	43.0	33.0	288	25.6	74.4	1.316	26	1.0	69.2
Metropolitan Area of Seville	1,457,428	15.1	17,405	35.7	10.4	53.9	40.8	19.3	39.9	80	47.9	35.7	1.320	26	2.1	90.9
South Yorkshire (Sheffield)	1,343,600	6.1	19,294	24.0	9.0	67.0	na	na	na	101	na	na	1.653	77	5.7	72.5
County of Stockholm	2,091,473	13.7	na	29.5	23.8	46.7	44.6	30.5	24.9	343	38.9	41.7	1.369	22	na	256.4
Stuttgart Region	2,439,664	2.7	53,403	28.6	13.8	57.6	23.0	26.3	50.7	336	58.9	41.1	1.520	28	1.2	131.6
Turin Metropolitan Area	1,556,805	5.5	0	27.6	19.2	53.2	33.4	25.1	41.6	157	32.1	67.9	1.720	32	1.9	58.1
Valencia Metropolitan Area	1,800,614	15.2	0	42.6	17.3	40.1	46.7	20.6	32.7	95	43.2	53.3	1.320	28	na	106.1
VOR Region (Vienna)	2,813,000	7.5	42,600	25.5	8.7	65.8	34.0	37.0	29.0	361	na	na	1.500	23	1.3	133.3
Vilnius	838,852	−1.4	0	na	na	39.3	36.1	24.6	39.3	223	45.2	54.0	1.320	56	3.1	43.2
Warsaw	2,435,350	na	23,649	32.5	30.2	37.3	21.9	54.6	23.5	417	31.3	60.0	1.228	25	1.0	65.8

45.9 % covered by fare revenues and 45.8 % by subsidies. The remaining 8.3 % encompasses all the atypical revenues not included in the fare box, like publicity, for example. It must be noted that, to assess such a crucial indicator, only operating costs are taken into consideration and not investments which, in some good years, can hugely exceed the overall amount spent on operations.

Public transport funding is a constant concern for public transport authorities, to such an extent that the Barcelona ATM commissioned a comparative study among six EMTA members that were studied in depth. Here are the main findings:

- Multijourney tickets and seasonal passes are the most widely used type of ticket. This trend is the result of policies implemented by the authorities to boost loyalty among users, encouraging the use of longer duration passes rather than single journey tickets. From the users' point of view, passes offer a number of advantages in terms of the features of the ticket and the discount on the fare.

- Regional and local administrations are tending to increase their involvement in terms of public transport funding, while states have reduced this. Within the context of public spending cuts, Paris stands out as an example of good practice in funding, where a specific transport charge has been implemented (Versement transport).
- Authorities are working on qualitative and quantitative improvements in public transport services that are pushing up operating costs. One clear example is the widespread implementation of electronic ticket validation systems which provide more information and control on the demand for public transport in order to establish policies in line with the system's needs.

1.7 The Future of Transport and Mobility in Europe: Smart Cities and Communities

A more sustainable usage of our planet's limited resources and the aspiration of mankind for a better usage of time can be further realised today through information and communication technologies, in particular where high concentrations of people allow for large-scale economies of investments.

In this sense, a city can have enormous benefits through the development of operative systems able to collect, analyse and elaborate flows of data acquired, possibly in real time, from a network of sensors and stakeholders deployed for different reasons and for different applications, with the aim of making more efficient all services that are essential for the life of citizens and companies, typically provided by the municipality through operators.

In this context, the EU has started an initiative in 2011, where a new policy framework, the *Smart Cities & Communities Industrial Initiative,* has presented the strategies for the future investment in low-carbon technologies, with specific climate and energy targets to be met by 2020, and a road map spanning until 2050 in order of reduction of greenhouse gas emissions.

Smart city will foster investments on different application domains, which include traffic, mobility, logistic, safety, security, water, sewerage, heating, energy, health, school, tourism, in-land protection and bureaucracy. Public administration at different levels should be able of monitoring, providing and controlling services for these critical topics.

Public transport will play an important role in the smart city concept. Data taken directly from public transport services (e.g., electronic ticketing, videos, automatic vehicle locations, etc.), which will be discussed more in detail in Chap. 3 of this book, will be integrated with the abundant private transport monitoring and information systems, both for real-time and offline applications. The adoption of vehicle-to-vehicle, infrastructure-to-vehicle and user-to-vehicle (in brief, V2X) communication systems will allow a more ample spectrum of multimodal options, an overall improvement of both services and overall a more efficient use of the transport systems capacity.

Broadband, computer science, geographical positioning and personal nomadic devices constantly connected and in communication with data collection systems are the technological drives for potential change. But the availability of massive data is not enough; to fully achieve the goal of smartness, we need, on one side, customisation and aggregation at different levels to extract useful information and, on the other side, models and algorithms for forecast and prevention. This means that despite the clearer and more complete vision of transport and mobility across individuals and within communities, models like the ones described in this book will be essential.

1.8 Reference Notes and Concluding Remarks

The climate challenge, urban sprawl and other factors have caused the following question to be posed: how can PT attract more passengers from the car and thus contribute to reduced CO_2 emissions? The answer is higher frequency, faster travel with PT, better comfort (both on-board and while waiting) and of course the fare level matters. Fares thus play a pivotal role in how much of the PT costs should be covered by the users and how much should be covered indirectly by the tax payer? As explained in this chapter, this is a very complex decision that goes beyond the pure economic aspects. Looking ahead, it seems that PT must increase its modal share in cities. It is a rather general common opinion that this will be achieved only by carefully integrating the services into a seamless multimodal transport system. Within this view, the fare level and fare structure will be more important in the years to come. Transit modelling should be able to answer questions such as What are the effects of integrated fares across all modes? What are the effects of 50 % reduction in fares? For certain groups? At certain times? What are the effects of a *Zero-Fare* policy? What are the effects of allowing bikes free on buses, trams and metro?

To answer these questions, the development of passenger flow models, able to incorporate transit assignment processes that are sensitive to the above factors and that are able to incorporate the effects of new technologies, is of paramount importance. This motivates the development and focal aspects of the next chapters in this book.

References

ATM (2006) Territori, població i localització d'activitats. Escenaris del context territorial i socioeconòmic de la regió metropolitana de Barcelona. Institut d'Estudis Territorials

ATM (2007) Pla Director de Mobilitat de la Regió Metropolitana de Barcelona. Criteris de sostenibilitat. Estudi Ramon Folch

ATM (2012) Revisión crítica de datos sobre consumo y energía y emisiones de los medios públicos de transporte. Fundación de los Ferrocarriles Españoñes. Ricard Riol Jurado

ATM (2012) Emissions de gasos efecte hivernacle i la qualitat de l'aire de la mobilitat de la Regió Metropolitana de Barcelona. Seguiment de l'evolució de les emissions 2006–2010. Institut Cerdà

Beirão G, Sarsfield Cabral JA (2007) Understanding attitudes towards public transport and private car: A qualitative study. Transport Policy 14, 478-489

de Dios Ortuzar J, Willumsen LG (2011) Modelling transport. Wiley, West Sussex

EMTA (2012) EMTA barometer 2011. European Metropolitan Transport Authorities, Paris

Enoch M (2012) Sustainable transport, mobility management and travel plans. Ashgate Press, Surrey

EU (2011) Transport 2050: the major challenges, the key measures. MEMO/11/197, Brussels, Belgium

Kenworthy J, Laube F (2000) Millennium cities database for sustainable transport. European Environmental Agency, Institute for Sustainability and Technology Policy, Brussels, Belgium

Litman T (2014) Safe travels. Victoria Transport Policy Institute, Steven Fitzroy and Associates, Victoria, Canada

Polak JB, Heertje A (2000) Analytical transport economics. Edward Elgar, Cheltenham

Saleh W, Sammer G (2009) Travel demand management and road user pricing: success, failure and feasibility. Ashgate Publishing Group, Abingdon, Oxon

Schafer A, Victor D (2000) The future mobility of the world population. Transp Res A 34:171–205

SENER (2006) Estudi dels intercanviadors modals a la Regió Metropolitana de Barcelona. Autoritat del Transport Metropolità, Barcelona, Spain

Wegener M, Fürst F (1999) Land-use transport interaction: state of the art. Berichte aus dem Institut für Raumplanung 46. Institut für Raumplanung, Universität Dortmund, Germany

Wilson NHM, Nuzzolo A (2009) Schedule-based modeling of transportation networks. Springer, New York

Zahavi Y (1980) Stability of travel time components over time and regularities in travel time and money expenditure. Transp Res Rec 750:19–26

Chapter 2
Public Transport in the Era of ITS: Forms of Public Transport

Kjell Jansson, Ingmar Andreasson and Karl Kottenhoff

This chapter describes the characteristics of public transport systems, seen as a *system*. With public transport system, we mean mainly the technical system of different modes of transport including vehicles and infrastructure as well as their characteristics, such as capacity in various traffic concepts. Maybe, most importantly, the concept of a system means that the various lines and modes are not self-contained entities, but they are used in combinations, where the travellers regard alternative ways of going between points. These combinations give rise to positive network effects, which can be taken care of in integrated systems.

A major distinction between PT and other forms of transport is on the provision of both facilities (vehicles and/or infrastructure) and services. Except for some hybrid types of PT [e.g., on-demand transport, dial-a-ride (DAR)] which are introduced in Chap. 8, most public transport services are scheduled. This means that operators publish timetables for various times of the day and various times of the year.

The integration of components forms the system. It is important to discuss first, in this chapter, aspects on organisation of public transport, as organisational forms determine the level of service, costs, fares structure, possibilities for service integration and for introducing ITS for integrated planning, ticketing, etc.

Forecasts of how many people will travel by public transport in the future are normally based on the development of population, social and geographical conditions, as explained in Chap. 1, but also on the structure of public transport, standard and passengers' valuations of the supply. In order to understand how to model and then forecast passenger flows, we need to have a basic understanding of how the service is designed, especially the forms of PT networks implemented in practice,

I. Andreasson
LogistikCentrum AB, Osbergsgatan 4A, 426 77 V Frölunda, Sweden
e-mail: ingmar@logistikcentrum.se

K. Jansson (✉)
Riksrådsvägen 22, SE 12838 Skarpnäck, Stockholm, Sweden
e-mail: kjsek@hotmail.com

K. Kottenhoff
Dept of Transport Science, KTH, SE-100 44, Stockholm, Sweden
e-mail: karl.kottenhoff@abe.kth.se

G. Gentile and K. Noekel (eds.), *Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems*, Springer
Tracts on Transportation and Traffic 10, DOI 10.1007/978-3-319-25082-3_2

and how to assess the performance of such networks. In order to plan for good service, we need to know in fact how people value travel, frequency, comfort, etc. If we start with the perception of public transport, it is not certain that we perceive the supply in the same way or even as it is in reality.

This chapter starts with the discussion on organisational forms and gives an overview of the products of public transport available in practice. An overview of the vehicle technologies available completes the description of the product types. Then, the chapter describes the network forms and design dilemmas arising from the interaction between products offered and urban and regional mobility patterns. Finally, it concludes with a high-level description of how infrastructure, network and vehicle technology choices translate into supply capacity offered to the PT users.

2.1 Organisation and Products

2.1.1 Regulation Versus Deregulation

Local and regional public transport is, at least in industrialised countries, owned or supervised by a local or regional public authority. The operation can be public or private, in the latter case, under free competition or procured by the authority through competitive tendering.

Three common types of organisational forms are as follows:

- Traditional: The authority defines and runs the services.
- Restricted competition: Authorities define the services that will be delivered by operators, who compete to get contracts (competition for the market).
- Deregulated system: The operators can freely establish PT services and thus compete for customers (competition on the market).

Some European countries still apply a traditional system where the authority itself operates the services (Fig. 2.1). This classical European solution is characterised by

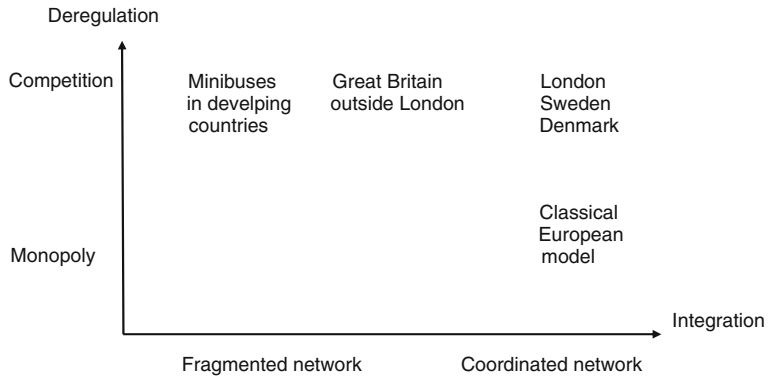

Fig. 2.1 Monopoly versus deregulation and fragmentation versus coordination

monopolies with good coordination, but often also with high operation costs. Other countries, and London, apply restricted competition through competitive tendering. UK, outside London, has a deregulated system of local and regional public transport. Even in Britain, there is restricted competition in cases where commercial operators do not provide service that the authority considers essential.

In local and regional transport, there is a variation with respect to fare competition. In some local or regional areas, there is one authority in charge of all public transport, allowing the passengers to transfer for free between any lines and modes. In other regions, several operators may compete. There may be one metro operator, one commuter rail operator, several bus operators, etc., between which transfers are not normally free of charge. This situation is common, for example, outside London (UK) in some other European regions, and especially in developing countries. There are examples where the regional authority has engaged in trying to make operators cooperate. In UK, there are examples where passengers can buy one pass for travelling with one operator, and another, more expensive pass, for travelling also with other operators.

The trend in Europe is to introduce more competition. A concession is a contract that is granting the right to operate a subsidiary business. It can be a form of less restricted competition, similar to a deregulated system, but with just one operator in each area.

Introducing deregulation does not always increase the efficiency and overall satisfaction from the PT users. It brings issues to the management of the whole system (i.e., guaranteeing standards and quality of all its components have reliable information of all sub-systems, etc.), economical (fare revenue allocation, subsidising), technical (duplication, reliability and permanence of services, services for special user groups, etc.), and societal (safety conditions may differ from company to company). The two sections below describe some experience from UK and Sweden, where deregulation has been most fostered.

2.1.1.1 Deregulation in UK

In 1984, the government of UK approved the London Transport Act, in which it was established that any transport company outside London city was allowed to obtain a bus service licence and therefore offer public transport service to the citizens. With this act, UK introduced for the first time the abolition of road service licensing and allowed for the introduction of competition on local bus services. The main motivation for introducing such a liberal policy was to remove the inefficiencies of the regional and interregional systems due to over-regulation, and due to lack of competition.

According to this act, two bus service types could be provided: commercial and subsidised. In 1985, the operation of new private bus services was established through a tender for the first time and, for a number of years, buses holding a variety of different colour schemes operated alongside those still operating in the traditional red paint. The variety of services was found to be confusing to users expecting to find red-painted buses and, after lobbying from the tourist board, it

became a requirement when contracts were retendered that bus colours must be predominately red.

The integration of the different services was under the responsibility of the London Regional Transport, which in 2000 was replaced by the current Transport for London agency, as a consequence of a revision of the Act in 1999 (Greater London Authority Act), after an assessment of the previous deregulation principles was performed, which showed many structural issues.

Having no control on the commercial services, many lines, which were giving low profits to the companies, were discontinued, causing major issues for the citizens as geographical coverage was getting more and more uneven. On the contrary, tracks where high demand was observed were served by many companies, which were competing to offer the fastest and most profitable service. This was the case of the interregional service between London and Manchester, where buses of different companies were engaged in famous *bus wars* (fake inspectors sent in competitors' buses to push customers away, speeding to overtake each other, etc.), which required the intervention of the authorities to avoid the snowball effect of the dangerous situations happening.

Beyond the above issues due to competition, the results of deregulating the PT system were not positive as expected. Although operating costs were lowered by 42 %, PT users decreased dramatically (−27 %) and fares generally increased (+17 %). Negative outcomes were also observed in system performance measures, such as increased distances travelled due to more commercial services using minibuses.

Nowadays, with the Act of 1999, a softer approach is applied. There is first of all a "Quality Partnership" between road authorities and PT operators: the operator agrees to run or improve a service according to some targets and, in return, gets benefits in terms of infrastructure concessions, expansion and renovation investments. Such contracting sets also specific contractual terms aimed at offering multimodal services and assesses performance in terms of total passenger flow served, which forces services to cooperate and coordinate with each other.

2.1.1.2 Deregulation in Sweden

Since January 2012 in Sweden, the market for PT in regional transport has been opened for commercial operators who can compete with the operations in charge of the regional authorities, as consequence of the new Public Transport Act ("Den nya kollektivtrafiklagen", prop. 2009/10: 200). This Act means that any operator is allowed to compete with the operator in charge of the regional public transport authorities. The main objective of the Act is to benefit the passengers by providing a supposed variety of services.

In one sense, this new Act may be seen as following the British deregulation from 1986, but might also be seen as even more liberal, at least when comparing with the current applications of deregulation in UK. New commercial operators may announce entry or exit of the market only three weeks in advance. There are no requirements at all concerning fares integration.

The question is, whether this new scheme may be financially viable for the commercial operators and whether it may be a good idea from a welfare point of view. The Swedish Ministry of Enterprise, Energy and Communications has given to the government agency Transport Analysis (Trafikanalys) the task to evaluate the evolution. It was found from a simulation study that very strong conditions must be available for commercial lines to be profitable and provide a positive welfare effect. Moreover, the impact of new bus operation was found very marginal: with respect to ordinary regional public transport, there appeared in fact only one non-subsidised commercial line in operation, a bus line in Stockholm operating from February 2012 but laid down in January 2015 due to low demand.

Principles for a level playing field with respect to access to infrastructure—stops, transfer points, depots, etc., as well as models for more effective contracts, are being worked on in the industry. Nevertheless, tentatively the Swedish form of deregulation in local and regional public transport will probably be no great success, especially since private operators have to compete with subsidised services in the hands of regional authorities.

2.1.2 Integration Issues

The different forms of regulation/deregulation made it necessary to deal with the problem of integrating the different services. First and foremost, integration is needed to avoid competition on high-demand routes, at the expense of a sufficient level of geographical coverage. A second important reason is to guarantee that the users see PT as a seamless service, in order to increase the shift from the private car mode. PT integration is nowadays not only about sharing information about timetables, and coordination of services, but has also been extended to include the fares system. This is done to provide the users with a unified fare system and a unified image and in turn increase the efficiency of PT use and increase the total ridership.

Integration brings, however, a number of questions, which are not easily answered and depend on the specific case:

- How to determine an effective integration in terms of responsibilities, coordination, synchronisation, fair operating ratio, etc.?
- How to guarantee a reliable commitment between companies?
- How to assess the quality of each operator? How to control the data concerning their performance, and which indicators and metrics could be unbiased?
- How to fairly redistribute revenues?

It is easy to understand how complex the problem of assessing the performance of an integrated system and of redistributing the revenue is, especially in complex systems aiming to integrate local and regional services. One can think, for example, of their main functionality, which for some company can be to cover the main connections (e.g., interregional services), while for some other can be to act as complementary service (e.g., local bus lines). It is clear that the good performance

of the latter strongly affects the attractiveness of the former, so the high passenger flows carried on by the main connections may not be achieved without the good performance of the local operators.

After ten years of deregulation in the UK, integration made a big return with the new Labour administration of 1997, but the progress towards seamless and integrated transport has not been smooth.

It has been argued that the short-term, non-strategic model of competition adopted in the UK inherently acts against integration. This suggests that current levels of integration are sub-optimal and that the implementation of integration measures would generally be beneficial.

Currently, the vision of many European countries (e.g., France, Sweden, Germany, Denmark, Belgium) is to split up the integration problem into two sub-problems, managed by two different authorities:

- Tariff and infrastructure integration is left to the infrastructure managers, in charge of distributing and allocating infrastructure capacity to the undertakings, who charge the users, elaborate safety regulations and technical standards, and analyse and assess the passenger flows to evaluate the efficiency.
- Operations are left to the operator managers, which provide commercial services to the customers according to tendering processes and agreements with the infrastructure managers, which are normally reviewed regularly (1–5 years).

2.1.2.1 The Asymmetric Demand Aspect

Apart from the integration issues, which arise in a competitive market introduced by the deregulation policies, the asymmetric demand makes the coordination of lines a very complex problem. Here is an illustrative example.

Assume that a welfare maximising public transport authority (PTA) and a private profit-maximising operator, respectively, are considering to invest the same resources on a new line, either line 1 (Fig. 2.2) or line 2 (Fig. 2.3). Assume also that the investments result in the same reduction of generalised cost, from G^0 to G^1, leading to demand increase from x^0 to x^1. For simplicity, prices are assumed to be the same on both lines. The operators' choices depend on expected demand response, partly due to transport substitutes.

Apparently the profit-maximising operator will invest in line 2, where the profit, equal to the dark rectangle, is much larger for line 2 than for line 1.

The authority on the other hand will invest in line 1. The reason is that a welfare maximising authority takes into account the sum of the profit area and the consumer surplus, the light area, and this sum is much larger for line 1 than for line 2.

Besides the welfare issue, there may also be distribution consequences. The private operator will invest in a line where passengers are sensitive to level of service. This may concern areas where many have access to car and may shift to public transport if the line is improved. The authority on the other hand invests in areas where passengers are less sensitive to level of service. The reason for which can be low car ownership or low incomes.

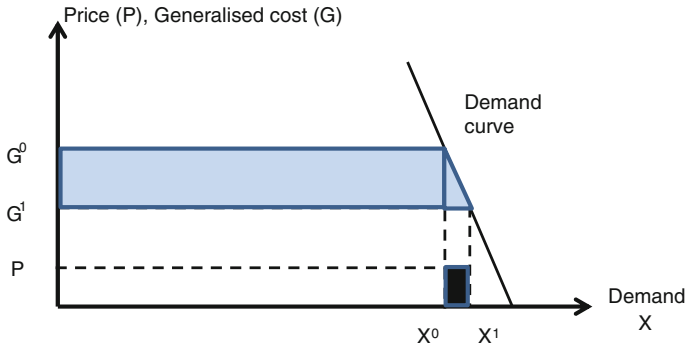

Fig. 2.2 Line 1 with low elasticity with respect to generalised cost

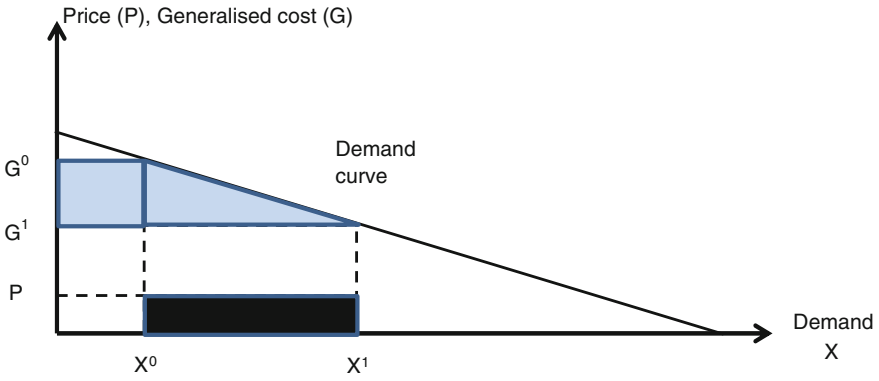

Fig. 2.3 Line 2 with high elasticity with respect to generalised cost

2.1.2.2 Competition Among Lines on a Network

Two examples are provided to explain issues that arise by connecting lines in a network.

With on the road competition, a private operator may leave out a section of a line it could operate profitably. Assume it operates *a–b* and *c–d*, but does not operate between *b* and *c* (Fig. 2.4). The PTA may find that this gap in service is not good from a welfare point of view, and decides to procure the section *b–c* by use of competitive tendering. It is then probable that the operator will win the bid rather than its competitors due to the advantage of already operating *a–b* and *c–d*. And its profit may be even higher than if they operated the section in the first place under "on the road competition".

Some people have argued that private profit-maximising operators would create a better functioning public transport than would a public authority. Nash (1978) questions this by comparing a welfare maximising and a profit-maximising operator, concerning one line. Here, the issue is analysed for a network, even though a

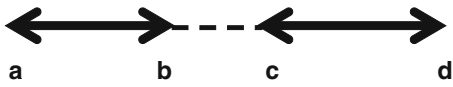

Fig. 2.4 A section left out of service

simple one, where the interaction between demands for various lines becomes a crucial matter.

We assume there is originally an existing line, E (which may be public or commercial and be thought of as part of a network), operating between points *a* and *c*, via point *b*. The question is, whether the introduction of a new commercial line, N, would improve welfare. This line N is partly "parallel" with line E, operating between points *b* and *c*. Some of the passengers, group E2, who travel on the section between point *b* and *c* on line E, are attracted by the new line, giving line N the demand N2. Figure 2.5 illustrates the network.

By use of a numerical example, Jansson (1997) found that the entry of the new line N may reduce the welfare level and yield profits for the two operators together that are smaller than the profits of line E before the entry of line N, even in the case the new line N would make a profit. This holds both if line E is public or commercial. The simple reason why this may occur is that when line N has "stolen" passengers from line E along the section *b* to *c*, the operator of line E finds it optimal to reduce the frequency of this line. This will harm the level of service for those who use line E and also reduce its revenues more than its costs.

2.1.3 Public Transport Products

A public transport product is a term that focuses on the offering that is put on the transport market. This offering includes a public transport supply with certain characteristics for the customers. Example of products is easily seen in the rail sector where *InterCity, Thalys* and *X2000* all are brand names for specific products. In local and regional public transport, the names often denote wider products such as *express, Metro, bus rapid transit* (*BRT*) and buses with high level of service

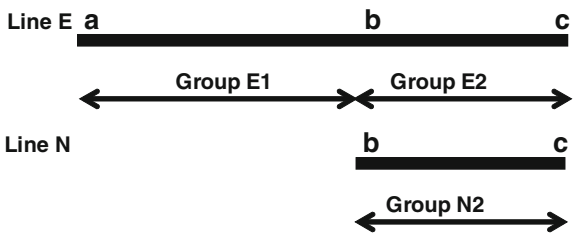

Fig. 2.5 A simple network with *lines* E and N

(*BHLS*). For example, *Metro* is an abbreviation of metropolitan, and the name of many products and services related to urban areas, especially public transport systems.

The PT system consists of vehicles which operate on an infrastructure; this consists of road or railway tracks, with stops/stations (Fig. 2.6). The operator offers its customers a combination of this system and the service and information provided to a certain level of quality, represented by the operating speed, the frequency of a service, etc. Finally, the service is offered at a certain price and complemented with additional services such as static and/or real-time information.

When a public transport product should be designed, there are four dimensions service providers need to define. These are set by answering these questions:

- Which vehicle technology to invest in?
- What infrastructure is optimal/feasible?
- Which network of lines and connections should be developed?
- Which service level to offer to the customers?

These questions are not easily addressed, and they represent the building blocks of the *design dilemmas*, which will be described more in detail in this chapter. Usually, these dilemmas are solved by local (national) authorities, and often solutions are country-specific.

2.1.3.1 Local and Regional

The local and regional public transport market includes, for example, trips to school, work, purchasing and health care. These markets aim to cover two broad categories of trips:

- Local trips, e.g., trips within the same urban area, typically shorter than 10 km and with an average trip duration of no more than 30 min. Typically, everyone takes at least one local trip daily. In European cities, the distribution of local trips is shared between car, public transport, walking and cycling.

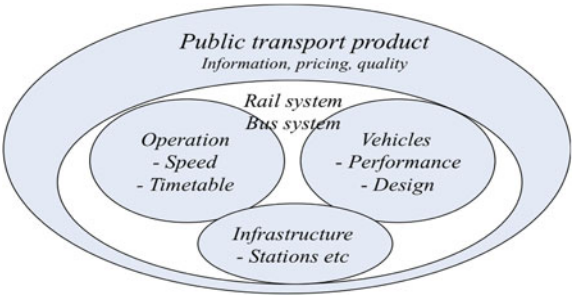

Fig. 2.6 Public transport products, See (Kottenhoff 1999)

- Regional trips are trips from one urban area to another within a region. These are trips that are almost always less than 100 km with a maximum of 1-h travel time. In broad terms, about 40 % of the population in Europe makes a regional trip daily. The average travel length is 20–30 km and the car dominates as mode of transport in this range. The bus dominates among public transport, but the rail has a strong position in some areas where train, metro or light rail services are attractive.

In medium and big European cities, each citizen makes about 200–400 public transport journeys per year (look back at Fig. 1.7). The capitals often have about 350 journeys, one way, per inhabitant and year, but some Swiss cities have even higher public transport use. The public transport share is higher to and from the city centre than for other journeys in the urban region. For example, in Stockholm the share for public transport in comparison with all motorised journeys is almost 80 % for journeys towards the inner city, in the rush hour. For other journeys in the Stockholm region, it is much lower. In some cities, there is significant competition from biking, for example, in Holland and Denmark and biking share is growing in many European cities. There are also attempts in Europe to integrate different forms of transport, for example, bike and public transport or car sharing and public transport. For this reason, so-called Mobility Centres are established in some places.

2.1.3.2 Interregional

The interregional (long distance) market can be defined in different ways: according to the actual demand, according to case, geography, competition, etc., the market can be defined on the basis of the following:

- Travel need, for example commuting, services, leisure.
- The person's time budget, which imposes restrictions.
- The transport systems that provide opportunities to satisfy travel needs in different ways.

Interregional travel can be defined as travel that takes more than an hour and almost always involves journeys of more than 100 km (about 300 km on average), and journey time is usually 3–6 h, which means that no one makes an interregional journey daily. The distribution of interregional trips by transport mode varies both with the traffic base and with the standard of public transport in the various counties. Even if there is a general correlation between increasing size of population centre and increasing travel by public transport, there is also a wide variation depending on supply, tariffs, etc.

The car is the most widely used means of transport due to the high proportion of leisure and holiday journeys, but public transport has a strong position over longer distances, where also rail, air and bus compete with each other. The train is the predominant form of public transport. Interregional travel by public transport is mostly recreational and business travel. Interregional public transport mainly serves

travellers a few times a week or a few times per year. The distances are longer, and the vehicles used are different from those used for local/regional transport. Here, we have intercity trains, high-speed trains, coach services and airlines with higher speeds and fewer stops. Dependent of purpose of journey, stop pattern, valuation of travel time and available income, travellers choose mode or combinations of modes.

To manage an interregional journey over a day and have time to do a work-related activity at the destination lasting at least 6 h (e.g., a meeting during office hours), the trip must take no longer than 6 h in order for all this to be done in the time one is awake is usually awake, i.e., between 06.00 and midnight. An interregional journey over a weekend may take slightly longer and a holiday of course even longer.

Interregional public transport is sometimes privately owned and commercial. But even if it is private and commercial it is, as mentioned, a system, irrespective of whether the various operators of lines and modes compete or sometimes cooperate. Cooperation may mean that operators offer combination tickets.

2.1.4 Multimodal Transport

2.1.4.1 Competition Between Modes

The choice of transportation for local, regional and interregional travel varies greatly; the choice of the right vehicle technology to invest on is clearly dependent on the length of the journey. Variations are also due to supply, price, quality, service, etc. Quite naturally, the train and air travel increase their market shares the longer the distance, while the car decreases. The bus has its largest market share over medium distances. In long-distant transport, the various public transport modes, such as rail, coach and air often compete with each other for customers. There are also often several operators within each mode that compete with each other. Competition can appear in terms of ride time, comfort, frequency of service and fare. With respect to fares competition these are often based on revenue (yield) management, which means that fares can vary substantially between days and even within days.

The so-called path-time diagram can illustrate the competition between modes in terms of speed. In a path–time diagram, the path is represented by the distance on the horizontal *x*-axis and the journey time by the time on the vertical *y*-axis. By adding in typical terminal times and the average speeds of different modes of transport, we can see which mode is the fastest over different distances and also in a simple way illustrate changes in supply in the transport system.

Competition for interregional travel between 100 and 600 km is broadly shown in Fig. 2.7. At the starting point, the train travels in the backbone network at an average speed of 85 km/h, and we estimate 50-min total access time to get to and from the train. This means that in 3 h, one can travel 200 km between origin and destination. The car has a very short access time, i.e., only 5 min to fetch and leave

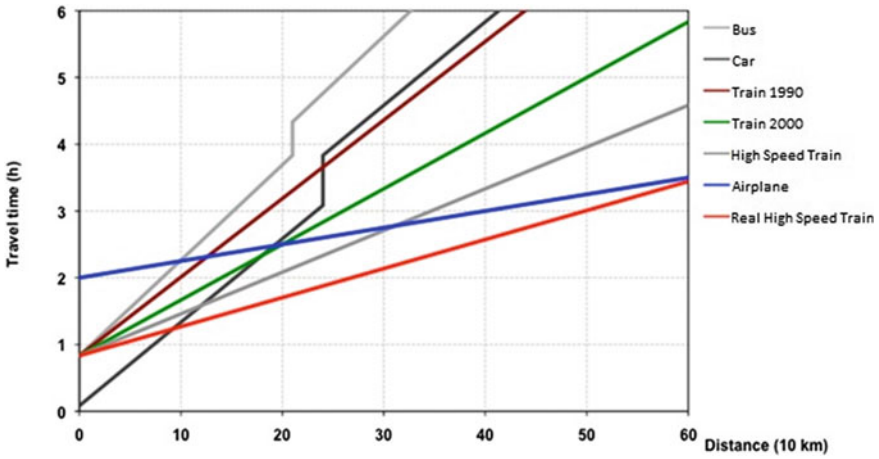

Fig. 2.7 Journey time competition between modes of transport as function of the distance

the car in a parking space and then drive at an average speed of 80 km/h over a certain distance. In 3 h, one can thus travel 240 km. After 3 h, however, we add a break of 45 min before resuming the journey; here the train can thus overtake the car over longer distances. By air, access time is very long; a total of 2 h, of which 1 h 15 min at the origin with feeder transport, watching the time and checking in, and 45 min at the destination. Once on the plane, however, the journey is fast, averaging 600 km/h, which means that one can travel about 400 km in 3 h, or twice as far as by train. We can also see that planes are relatively insensitive to distance; another 600 km can be covered in 20 min. For interregional journeys, the bus has the same access distance as the train; access time is then 50 min, and the average speed 70 km/h if the bus stops a few times along the way. After 3 h, we add a break of 45 min. It is clear that the bus is the slowest means of transport; after 3 h, we have travelled only 150 km. One can also see that the car is the fastest means of transport over distances of up to slightly more than 200 km.

Figure 2.8 presents a similar competition but on within a local area. Price, frequency of service and comfort, which all vary between the different means of transport, are also important. Journey times, however, are crucial to establishing travel habits. It is often the fastest means of transport that generates new travel—once it was the railways, then the car and, most recently, air transport. The fastest means of transport also become market leader most easily, which means an opportunity to charge sufficiently high prices and thus also achieve good profitability.

There seems to be a stable relation between travel time by train and the rail–air market share. A selection of 105 rail–air routes was identified in Europe (and Japan) in a study, which has a maximum speed of at least 200 km/h. Of these, 30 could be identified with market shares from which a statistically significant regression curve was produced. This analysis confirms that there is a very strong correlation between

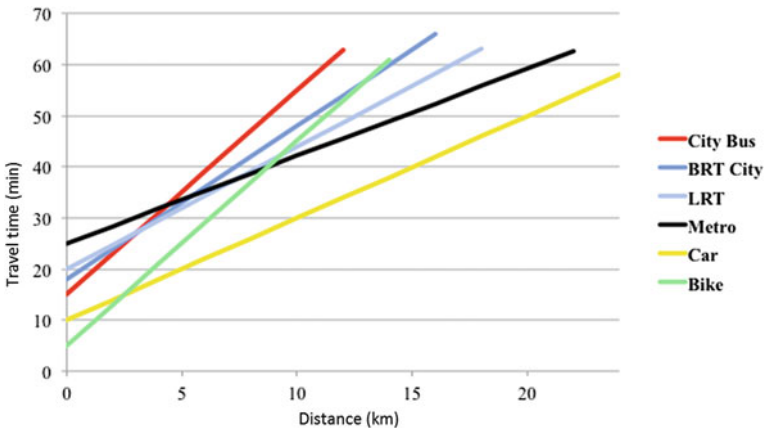

Fig. 2.8 Local journey time competition between modes of transport, as function of distance

the train's absolute journey time and its market share. The relative journey time between rail and air, the journey time ratio from city centre to city centre, was also analysed and showed a clear correlation. A three-dimensional analysis also showed that there was also a correlation between relative journey time, distance and the train's market share. The correlation between the train's frequency of service and price was also studied but proved to be weak. The conclusion is that journey time from city centre to city centre is by far the most important factor when choosing between train and plane.

Introduction of high-speed rail in many countries have changed the competitiveness of the rail mode quite dramatically. More up-to-date figures for the air–rail mode split show that where rail journey times are reduced below 4 h, rail share of the rail–air market increases rapidly with further journey time reductions, and rail tends to have a market share of at least 60 % and sometimes effectively drives air out of the market when rail journey times are below 3 h. Future trends are found to depend on a wide variety of factors including the introduction of environmental charges on air transport and trends in air and rail costs.

2.1.4.2 Intermodality

Intermodality integrates two or more transport modes on the same journey. Aim is to make this interchange as seamless as possible with common information, an integrated ticket and a multimodal station where passengers feel safe, secure and comfortable. If successfully implemented, intermodal passenger transport will give more options to the traveller, is user-friendly and adds to the overall efficiency of the transport system. To achieve seamless intermodal travel, many transport stakeholders have to cooperate closely, which is not evident in a system of

increasing competition. Past studies found that the fragmentations of transport operation occurred by privatisation made the seamless inter-regional transportation difficult.

The importance of intermodality in general public transport systems means that models applied should be able to take all combinations of lines and modes into account, not only main modes, but also to regard different forms of public transport as a seamless system. This also means that more sophisticated choice models need to be developed, which are able to handle different service types, different modelling paradigms (e.g., scheduled services, frequency-based systems, etc.) and their respective connections, plus access and egress costs. These models will have a central role in Part 2 and 3 of this book.

2.2 Vehicles

2.2.1 Trains

Passenger trains can be divided into three major groups:

- *Long-distance high-speed and Intercity trains*: connecting cities in the fastest time possible, bypassing intermediate stations.
- *Fast trains/Interregional trains*: calling at larger intermediate stations between cities, serving large urban regions.
- *Regional trains*: calling at many intermediate stations between cities, serving all inside communities. There are also regional express trains with higher average speeds.
- *Local trains*: calling at all intermediate stations.

There are also local rail systems that are not referred to as train, but rather metro, light rail and tram as well as some guided bus and beam transport systems.

The distinction between the types can be thin or even non-existent. Trains can run as Intercity services between major cities and then revert to a fast or even regional train service to serve communities at the extremity of their journey. Long-distance trains travel between many cities and/or regions of a country and sometimes cross several countries. They often have a dining carriage or restaurant carriage to allow passengers to have a meal during the course of their journey. Trains travelling overnight may also have sleeping carriages.

2.2.1.1 High-Speed Rail

One notable and growing long-distance train category is high-speed rail. Generally, high-speed rail runs at speeds above 250 km/h (155 mph) and often operates on dedicated track that is surveyed and prepared to accommodate high speeds. Japan's

Shinkansen ("bullet train") commenced operation in 1964 and was the first successful example of a high-speed passenger rail system.

The fastest wheeled train running on rails is currently the French TGV (Train à Grande Vitesse, literally "high-speed train"), which achieved a speed of 574.8 km/h, under test conditions in 2007 (Fig. 2.9). The TGV runs at a maximum revenue speed of 300–320 km/h, as does Germany's ICE trains, Italian ETR and Spanish AVE trains as well as the TGV-based EuroCity service trains from London. The largest high-speed network as well as the highest speed currently attained in scheduled revenue operation is in China with 350 km/h.

In most cases, high-speed rail travel is time- and cost-competitive with air travel when distances do not exceed 500–600 km (311–373 mi), as airport check-in and boarding procedures may add as many as two hours to the actual transit time. Also, rail operating costs over these distances may be lower when the amount of fuel consumed by an airliner during take-off and climb out is considered. As travel distance increases, the latter consideration becomes less of the total cost of operating an airliner and air travel becomes more cost-competitive.

Some fast and high-speed rail equipment employs tilting technology to improve stability in curves. Tilting is a dynamic form of superelevation, allowing both low- and high-speed traffic to use the same tracks. Examples of such equipment are the Pendolino, the N700 Series Shinkansen, Sweden's X2(000) services, Amtrak's Acela Express and the Talgo (Fig. 2.9).

In order to achieve much faster operation over 500 km/h (310 mph), innovative magnetic levitation (Maglev) technology has been researched for years. The Shanghai Maglev Train, opened in 2003, is the only commercial operation, operating at speeds of up to 430 km/h. The conventional high-speed technology is cheaper, it can also reach very high speeds, and it is compatible with existing rail infrastructure.

The European Commission is currently also in discussions with the rail sector concerning a major initiative for research and innovation for rail during the coming financial period (2014–2020), called Shift-2-Rail. The European Rail Research Advisory Group (ERRAC) has identified seven priority areas for the future

Fig. 2.9 European high-speed trains: TGV duplex (*source* Wikipedia) and Talgo (*photograph* Kottenhoff)

development of the European rail sector, for example, intelligent passenger information systems, increased the energy efficiency of trains, increased safety speeding up product approval procedures, improved interoperability and attractiveness, cost management and forecast models and finally developing less costly infrastructure maintenance methods. Among the many EU projects can be mentioned as an example the MODTRAIN project, which worked on standardising the numerous components that make up a train.

The EU Framework Programme for Research and Development is contributing to the development of high-speed rail in Europe, focusing on all elements of the system: vehicles, power supply, track bed and sub-structures, tunnels as well as traffic management and signalling. The challenge is to modify a highly unstandardised system due to different technologies and power supply systems at the national levels, into a seamless network. The development of a European Rail Traffic Management System (ERTMS) is co-funded by the European Union and will inter alia contribute to improve interoperability for cross-border traffic, which represents an important (actual and prospective) market for high-speed rail traffic.

2.2.1.2 Local Rail Systems

Different forms of rail-based technologies are designed especially for the urban and regional markets. The most popular are *city rail*, *metro/undergrounds*, *trams* and *cable rails*. Both Metro and tram systems most often have the steel track—steel wheel technology, but even concrete track—rubber wheels exist. These vehicle (and infrastructure) technologies differ significantly from each other in terms of operational speeds, acceleration/deceleration power, maximum slope, minimum technical distance between stops, etc. Most of light rail transit (LRT) systems have a separated infrastructure, apart from the tram, which often shares its infrastructure with other PT types, and sometimes with private cars.

There are today about 150 metro systems and even more LRT systems in the world. Just in Europe, there are more than 170 systems and the number is increasing.

Modern rail systems can be driverless. This is especially practical for metro and other fixed rail systems including cable rail and other rail systems operating in tunnels and on elevated beams.

2.2.2 Buses and Coaches

Buses are the most popular PT service in cities, as it well operates on conventional road infrastructures and offers different opportunities in terms of size/passenger capacity, speeds, combustion technology, etc. The latest trend is specially designed vehicles for “BHLS”. Double-decker buses are also increasing their market share where either seats comfort or capacity is important (Fig. 2.10).

Fig. 2.10 Example of bus types. A double-articulated BHLS bus, a regional double-decker and a long-distance tourist coach (*photographs* Kottenhoff)

Buses can be guided by various technologies, either mechanical guidance in concrete or by steel tracks or electronic guidance by magnets, camera and GPS technologies. Examples of mechanical technologies are the German "Spurbus" and the French "Tram-on-Tire" (It can also be called tram).

Long-distance bus/coach services run in many national and international relations. Private companies most often operate these services. Beside these scheduled interregional public transport services come a great number of long-distance chartered coaches. Companies that run these charter services often also run scheduled bus services.

2.2.3 Aircrafts

The air mode is growing mainly on longer distances, while rail serves as feeder systems to the airports. The market has been developed and changed by low-price airlines.

There are numerous variants of aircrafts, for regional, national and international flights. There is no important reason to go into detail on these in this context. While old aircrafts had rather high relative energy consumption, modern aircrafts are down at levels similar to cars if there are two persons in the car and the aircraft has a high-load factor. The CO_2 emissions are still problematic for a sustainable future.

2.3 Infrastructures and Networks

There are modes that operate:

- on streets (buses and trams);
- partially separated on ground (light railways or light busways);
- fully separated, either elevated or on the ground (heavy railways or heavy busways); and
- underground (subways).

The runway categories show various infrastructures and their position as public transport technologies in cities. In Table 2.1, X denotes *applied* and (X) *applied to some degree*.

The public transport categories and infrastructures are unorthodox and open for further discussion. An example is the common capacity distinction in transports (light vs. heavy), where heavy means always complete or full separation regardless if it is a bus or rail system, whereas light attribute is for systems that anchor on partially separated, or are partially on street or partially fully separated (Fig. 2.11).

2.3.1 Right of Way

Local and regional public transport systems are traditionally planned for the commuting market. Rush hour travel data are used. The pressure on the transport systems is hard during rush hours with congestion in both private transport modes

Table 2.1 Public transport systems and infrastructures

Mode	Runway				
	On streets in mixed traffic	Dedicated lane on streets	Partially separated on ground	Fully separated on ground or elevated	In tunnel or underground
Bus system	X	X	(X)		
Busway system		(X)	(X)	X	(X)
Tram	X	X	(X)		
Light rail transit (LRT)	(X)	(X)	X	(X)	(X)
Heavy railway (HRT)			(X)	X	(X)
Subway/metro				X	X

Fig. 2.11 Lundalänken in Lund, Sweden, is a partially separated "light busway", here passing under an ordinary road. There is a parallel bikeway (*photograph* K Kottenhoff)

and public transport, at least in bottlenecks. For this reason, PT systems can be designed with different right-of-way (RoW) rules:

- RoW A: fully controlled without any grade-level or shared access: these types have no influence of other modes, therefore guaranteeing high-reliability and high-quality services.
- RoW B: longitudinally separated from traffic, with at-grade crossings, normally physically separated by curbs barriers. These are partially influenced from grade crossing with traffic or pedestrians. There is limited interference from other modes, normally occurring at the intersections.
- RoW C: shared surface streets with mixed traffic and reserved lanes. These are strongly influenced by traffic affecting travel time reliability, and vice versa they reduce the capacity of car lanes as often stops are placed on the carriageway.

Clearly, the transport systems with an own right of way can better compete to the performance of private cars. Travel times for going to work with car or public transport show that it mostly takes longer to use public transport. Usually, it takes 1.5 times longer or more, but as mentioned when there is congestion in the streets, public transport with its own RoW can reduce or eliminate this gap. For example, fast regional trains can offer travel times lower than a car journey if the origin and destination are not too far from the stations. The same is true for, e.g., express buses in prioritised lanes.

2.3.1.1 Target Lanes

If car traffic is dense, it can be hard to follow a timetable. There is often an advantage if public transport can be given its own RoW, for example, in the form of reserved target lanes. Very common examples in cities are the bus lanes (e.g., Fig. 2.12). Curb lanes can be used somewhere, but will be almost worthless if cars are parked in them. Curbstone lanes work quite well on streets without berthing. For tram operation, there will be a total stop if anyone parks in the way of the tram.

Fig. 2.12 Public transport reserved lanes in the middle in one direction at Sturegatan, Stockholm. This particular lane also serves as a "queue jumper" (*photograph* K Kottenhoff)

Reserved lanes in the middle of the street are often better for public transport, due to the decreased risk of being impeded by other traffic. Stop islands makes so the vehicles do not need to pull into the dock at curb stops. The trip will be faster and more convenient and faster than for curbstone stops.

A *bus jumper* or queue jumper is a reserved bus lane short before a crossing where there is regular congestion. By having a reserved bus lane, the buses can pass ("jump") the car queue and be at the crossing before or in front of the cars. It is also known as a bypass lane or queue bypass. The buses are often let first by the help of traffic signal priority.

2.3.1.2 Bus Streets and Busways

Bus streets are used for different reasons. In some cases, it is to make the line more direct and thus faster, cheaper and more attractive. In other cases, it is to serve areas that are closed to automobile traffic, for example, in residential areas. To prevent car traffic, sometimes only traffic signs are used, but often are automatic booms, gauge barriers or such used. Busways can be seen as a "runway" for the bus. Cities that have invested in high-level service busways are Almere, Amsterdam, Runcorn.

Figure 2.13 shows three ways to serve a district by a public transport line. The first sketch shows the most efficient route that can serve the area. It often requires a rail- or busway straight through the quarter. The third sketch is the most expensive way to lie out a route for the quarter. It was often designed for cars, for traffic safety reasons.

Bus and tram streets can be built in city environments. Entire roads can be designated as bus streets, such as Oxford Street in London as well as in Amsterdam

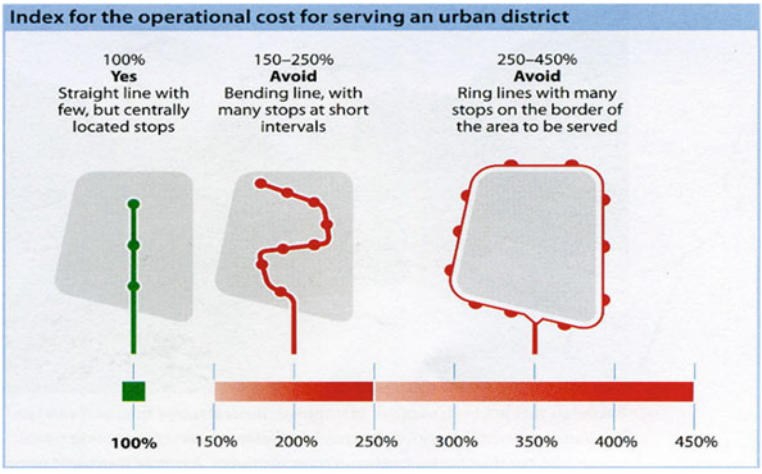

Fig. 2.13 Short and central lines save costs (*Source* Hitrans 2)

Fig. 2.14 Busway in suburb, south of Amsterdam (*photograph* K Kottenhoff)

(Fig. 2.14), allowing buses, taxis and delivery vehicles only. Sometimes bikes and buses are mixed, but a safer solution is to have separate bike lanes.

Bus lanes are normally created when the road in question is both likely to be congested and heavily travelled by bus routes. *Contraflow bus lanes* can allow buses to travel in the opposite direction to other vehicles. Some bus lanes operate at certain times of the day only, usually during rush hour, allowing all vehicles at other times, and it is common to have bus lanes in only one direction, such as for the main direction of the morning rush hour traffic, with the buses using normal lanes in the other direction.

A special form of "public lane" is seen on American highways in urban areas, i.e., the *HOV—high-occupancy vehicle lanes*, where buses and cars with at least 3–4 people may run in a reserved lane.

2.3.1.3 Prioritisation at Traffic Signals

Traffic signals make it difficult to obtain high average speeds and to run by a timetable which assumes that all trips take the same time, as delays encountered are highly variable. Prioritising in traffic signals is an effective measure to increase public transport speed and regularity.

Bunching is a problem to bus, and to some extent tram, operations; if a bus has a delay due to traffic, for instance at signals, there will be more passengers waiting at the next stops, resulting in longer dwell times, which make the bus later and so on. The following bus will have a shorter headway and therefore fewer passengers at the stops and may even catch up with the late bus. The bus bunching leads to an ineffective use of the bus fleet and causes delays and crowding for the passengers. Much would be gained if the initial small delay could be handled before it has grown and caused bunching. Conditional bus priority only giving priority to late buses can reduce the bunching significantly with relatively small delays for other traffic.

PT priority is easier to arrange in isolated signalised intersections than in coordinated systems, where the "green waves" may be disrupted and it may be necessary to make green time compensations. However, compensations made locally in the signal controllers may have negative impacts on the performance of the coordinated system if they are not carefully timed. Bus priority can have a negative impact not only to traffic that crosses the prioritised bus route, but also to other movements following the bus route.

PT signal priority can be passive or active. In passive signal priority, signal timings can be set favouring approaches with PT, and the speed of the progression can be adapted to typical PT—speed including time at stops. Passive priority is often good when the PT frequency is high and the dwell time at stops is short and predictable. Active priority measures can give larger benefits to PT vehicles than passive measures since they are only applied when they are needed, and can therefore be allowed to give more negative short-term impacts to other road users.

Active signal priority can be Unconditional or Conditional. Unconditional priority is mostly used for trams and will always favour the PT vehicles without consideration of negative impacts for other road users. Conditional methods may apply PT priority so that the overall intersection delay is minimised. It can also be arranged by:

- setting limits for extension lengths;
- restricting the priority to uncongested time periods;
- only giving priority to late/not-too-early buses, which will also improve PT regularity.

Too aggressive signal priority can also be counterproductive, if the negative impacts to other traffic cause oversaturation, the queues may spill back and block the buses. There are two main methods for conditional active signal priority:

- local signal timing adaptations with restrictions in a fixed-time system;
- self-optimising methods that minimise an objective function (e.g., minimises total road user costs).

It can in some cases be adequate to give priority only to some public transport vehicles, or different priority to different vehicles, in order to minimise the impacts on other traffic or to handle conflicting requests for priority. In Stockholm, signal priority is only given to trunk bus routes (blue buses) and not to local (red) buses. To this end, all buses in Stockholm are equipped with an AVL system that keeps track of the bus position and scheduled position according to timetable. The system is used for real-time information to passengers and gives the drivers and bus dispatchers, dynamic information on the timetable adherence. Buses that are more than 2 min ahead of schedule, according to the AVL system, will not call for priority. Conflicting calls for priority are handled according to the first come, first served principle.

Here are the four examples of active priority:

- Extension of the ongoing green phase (often the most effective);
- shortening of the red phase;

- insertion of an additional green phase (for public transport);
- switching back to green.

There are different ways to detect buses at traffic signals. The most important are as follows:

- Detection using loops in the road;
- passive loops, for example, long-loop detectors that sense big low vehicles;
- active loops that is receiving a signal from the vehicle;
- infrared detection or other form of transponder;
- computer–radio communication, which sends information to the steering device if a bus is approaching. The position can be given by, for example, GPS; and
- radar detection.

The methods that are developed for the prioritisation of public transport must take into account other traffic modes so that the traffic does not get congested entirely. There are different ways to do this. PT priority settings are usually conducted on the basis of traffic engineering experience. Microscopic traffic simulation is often applied to assess the impacts.

2.3.2 Nodes

2.3.2.1 Bus/Tram Stops

The location of stops should connect to the natural footpaths. At the same time, there are trade-offs with the requirement for a fast and straight bus route. Winding bus routes are neither convenient nor fast, and they are difficult to understand for those who do not use them every day. Slow bus routes make public transport more expensive. The footpaths to stops have importance for the accessibility, especially for children, the elderly and the disabled.

In rural areas and at streets with 70 km/h or higher speed limit, the stops are designed as pockets, where the bus stops are beside/outside of the roadway. For bus stops on highways there are specific design rules.

Curb stops at the sidewalk (footway) are simply arranged by making marks for the stop near the curbstones. With demand on accessibility for disabled, this is a serious matter. The sidewalk should be rebuilt with a proper height and constructions like the “Cassel Curb” can help the driver to “dock” tightly into the curbside. There should also be facilities for sight-impaired people at stops of all kinds.

For many reasons, it might be better with *bulb stops* (Fig. 2.15), where the sidewalk is built out into the street so that the bus does not need to pull into the stop. Comfort for the travellers is improved, the required stopping time decreases slightly and the bus calls adjacent to the curb. Another advantage is that the possible parking distance becomes longer and the risk of parking cars wrongly decreases. The capacity for all traffic on the street is, however, decreased. If priority for public transport is a goal, this makes it.

Fig. 2.15 A bulb or heel stop in Stockholm with a prototype VanHool/Scania BHLS vehicle (*photograph* Scania)

2.3.2.2 Stations and Terminals

The proportion of journeys that involve interchange ranges from less than 10 % to more than 50 % looking at different cities. Terminals are located at key interchanging sites in the traffic network or to key destination points in the region or the urban area. There might sometimes be reasons, including environmental concerns, not to let regional buses into the central part of an urban area. The terminals are in these cases located where interchanging sites give good connections in the local traffic system.

Major bus terminals are often organised in order to have a common stop for clustering alighting operations from all lines. After that, the bus goes to a parking space, which should be close to the staff facilities. Shortly before departure—normally 5 min before—the bus drives up to the boarding stop to take on passengers. Each boarding stop can serve multiple routes, which departure times are coordinated for a maximum utilisation of space.

The lamella terminal is a traditional form, which may have some traffic safety and logistic problems. A terminal type that is much used in Europe is the *Island Terminal*. It is traffic safe if one comes to the island without passing over the roadway. *Docking Terminals* provide short walking distance and high convenience for travellers, but are forcing buses to back out from stop (Fig. 2.16). They are used more often for buses in long-distance service. There are some good examples of underground docking terminals in, for example, Helsinki, Madrid and Stockholm.

Bus terminals often require a fairly big space. One reason is that each route often has its own *post*, a fixed departure stop. One way to get a more compact terminal is to use ITS with dynamic information signs that allow a dynamic allocation of stop points so that buses can depart from the stop locations where currently there is space. Ideally, all buses and all rail stops should if possible be arranged at one island (berth) or platform, which mean transfers over the platform in the same level are preferred. A closed passenger area allows for boarding of vehicles without

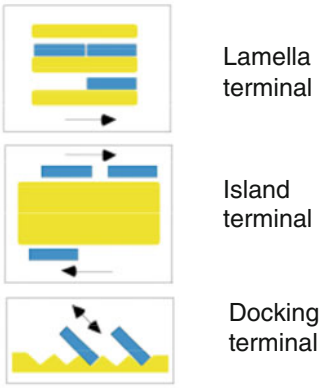

Fig. 2.16 Different types of layout of a bus terminal; Lamella, Island and Docking

validating your ticket once again. This is the typical system in metros and can be used also in advanced bus systems.

At terminals with connection between rail and bus services, the coordination should be in real time between trains' arrival and the departure of buses, which adds to the timetable coordination. A different integrated system with information for the bus drivers shows the number of minutes to the train's actual arrival, allowing the bus driver to determine whether or not he can wait for a delayed train.

2.3.2.3 Multimodal Hubs and Travel Centres

Multimodal terminals or hubs have an essential role in achieving a significant modal split to multimodal journeys. The origin of multimodal terminals results from the fact that transport operators were originally private and not at all interested in collaborating with other operators. Once public administrations started to coordinate public transport and integrated transport tickets consequently appeared that hubs became key elements in public transport networks.

Hubs allow for coordinated action on the part of different modes of transport with differing capacities, both the trunk and branches of the same tree. The various modes should be achievable within convenient distance and without passing any car traffic. Another advantage of integrating several modes is that travellers have, in a very little space, all the information about all the modes.

A good example of this design concept is the multimodal terminal of La Défense, Paris, which handles among the largest volumes of passengers, around 450,000 people per day. It is used both by commuters going into the business centre of La Défense and also by others using the station as an access point for the city of Paris and its public transport network. Also to be highlighted is the *observation centre* of La Défense, which controls the movements of the different modes of transport operating there and can act quickly in the case of disturbances.

As natural extension/generalisation of the multimodal hub concept, *travel centre* has become a name on a main terminal in the city, mostly railway station combined with a bus, tram and/or metro terminal, taxi station and bicycle and car parking. A travel centre should have closeness between the various modes. The stations are nodes in the entire travelling networks. Short walking distances are important. At travel centres, extra services should be added such as restaurant, the opportunity to rent a car, post office, bank/ATM and tourist information/hotel reservation. The aim is to gather all possible service for travellers, but also to create a complement to the town centre.

2.3.2.4 Park and Ride Facilities

Park and Rides (P&R, or P + R) are specific kinds of hubs where passengers change between their private vehicle and public transport. These nodes are designed to provide easy access to modes of public transport for those people living in areas far from stations within the transport system, thereby dissuading them from using their private vehicle when going into the city. In most cases, Park and Rides are associated with railway or bus stations with fast, high-frequency services. In addition to the spaces allocated for parking in general, they can also have areas reserved for those with limited mobility, to drop off and pick up passengers (*Kiss and Ride* and *Wait and Ride*) and to park motorbikes or bicycles.

In general, Park and Rides are owned by public administrations and, in fewer cases, by the underground or railway operators; however, it is always each country's public administration that determines the location and properties of the Park and Ride; hence, it is chiefly responsible for the investment. The purpose of the Park and Ride is to encourage travel by public transport for passengers with long walking distances from home to the nearest stop. Incentives include the following:

- A Park and Ride facility is locating normally outside of the inner city and is in direct connection to the public transport by which the onward journey to the city can be undertaken (e.g., located close to railway stations). To make the park and ride attractive, the walking distance between car park and PT stop should be short (preferably <300 m).
- Normally, car parks are reserved for those using the public transport. However, on certain occasions, to guarantee economic viability or obtain permits from the local administration, not all parking spaces are reserved for public transport passengers as some are also permitted for residents, workers in the area, etc.
- Fee for the parking lot is very low (ideally, zero). Some charges are integrated within the transport tickets. Users are generally offered the chance to buy seasonal tickets at a lower rate than daily ones, and rates charged for parking are lower than those of car parks in the city centre.
- Travel time with PT should be less than by car (due to, among other things, the congestion on the road for the car to the inner city), and frequencies of the public transport should be high.

- Signage and guidance to Park and Rides is considered to be one of the most effective ways to attract Park and Ride users. Some cities use dynamic information panels to indicate the closest facility, as well as the car parking spaces available.

2.3.3 Topological Structures

2.3.3.1 Regional and Urban Structures

A metropolitan area may in principle have one (or more) city centres, sometimes called central business districts (CBD).

In places where the rail services development is recommended, the planning should be focused on locations with good opportunity to organise effective connections between different public transport services. In the first instance and in all the smaller station locations, the goal should be that the station should be reachable by pedestrians or cyclists distance. Medium-sized towns can be linked with fast regional trains.

In many European agglomerations, new or upgraded rail lines and services trigger regional expansion. Even high-speed trains with top speeds up to 200–300 km/h are being used for work commuting. The radius can be 100 km or more with on-board travel times of up to 1 h.

It is difficult to get enough load and utilisation rates of the public transport if settlements and roads do not fit. Therefore, buildings should lie around corridors or "bands" (Fig. 2.17, left). An appropriate structure is, for example, the linear city, with businesses and residents concentrated in a not-too-broad band or corridor. Through this structure, one can create one line (or more lines) with high frequency and satisfactory economy.

In the USA, an old planning idea has come back in form of "TOD"—transit-oriented development. It means that planning of (dense) settlements, working places, etc., is made in public transport corridors and that the environment is nice for walking and biking. This is to encourage people to use public transport and to walk to the stations. For example, some cities permit commerce and multi-storey apartment buildings only within one block of train stations and multilane boulevards and accept single-family dwellings and parks farther away. In

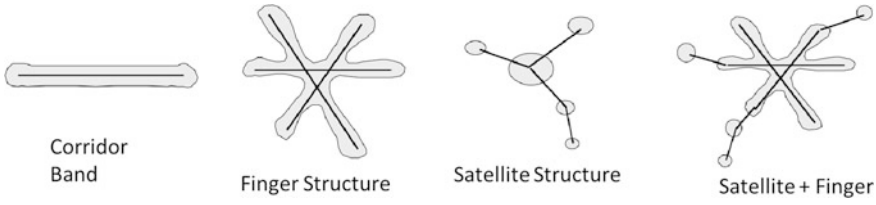

Fig. 2.17 Examples of well-suited metropolitan structures for public transport

Europe, TOD has been used for long. Good examples can be found in many European cities and in special projects like Vällingby in Stockholm, Runcorn in Britain, Almere in Holland and in Freiburg, Germany.

New dwelling and work areas should, if possible, be built in existing public transport corridors, not in between them. In some cases, the expansion of the local rail network and the metropolitan structure has been tightly linked. Such an example is the Stockholm region with a satellite and finger structure (Fig. 2.17, right). The Stockholm metro corridors form the fingers, while the satellites have been located around local and regional railways. An idea that may work well for public transport in mega cities is to complement the central core with a number of regional cores. This policy is encouraged by, for example, the Stockholm region.

2.3.3.2 Network Structures

The main public transport structure is in many large European conglomerations traditionally formed by rail lines (railway, metro, tram), while the bus systems are more complementary. In these cities, buses act as feeders to rail, for direct lines as trunk lines to areas without a close rail corridor. Medium and smaller cities on the other hand are often dominated by bus systems. It is more unusual in large cities in Europe that bus forms the core public transport. In principle, it is possible to build very heavy bus corridors, as have been made in the South American BRT systems. One example is the BRT system in Bogota, where a peak hour capacity of over 40,000 passengers has been reported.

The design of a PT system therefore begins with the identification of the most suited main network structure. Figure 2.18 shows the typical categories of structures used in practice. Examples of European cities with these network structures are as follows:

- *Diametrical system*: Helsinki, Warsaw
- *X-system*: Brussels, Amsterdam, Stockholm
- Radial/cross-system: Rome (Stockholm)
- Circle-radial system: Madrid, Moscow
- Intermeshed/ grid system: London, Paris, Berlin
- *Air-bladder system*: Lille, Rotterdam, Nuremberg.

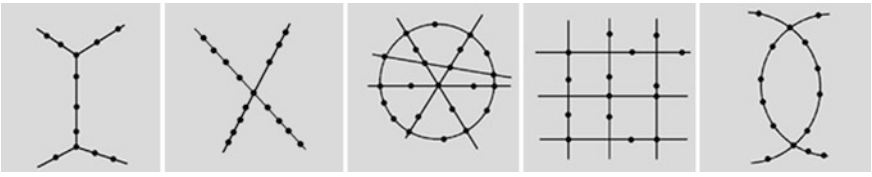

Fig. 2.18 Examples of rail network structures: diametrical with branching (X-system), radial (cross-system), circle-radial system, intermeshed-system and air-bladder system

This classification should not be interpreted as strict, while it gives an example of different ways of designing network structures.

Once the most opportune main network structure is identified, a number of design parameters needs to be defined to optimally use the PT system. These parameters will affect the functionality and performance of the PT system, how easily it will be able to be integrated with complementary services (e.g., the bus network), and how the customers will use it.

A first fundamental design parameter is the *node frequency*, expressed, for example, in terms of average distance between stops. This parameter is clearly dependent on the vehicle technology used, as explained previously in this chapter, but also on the overall system performance. A higher density of stops has in fact a positive impact on PT system accessibility, but it has a negative effect on the total travelling time for the travellers, as well as it affects the reliability of the system as the number of vehicles stopped for boarding/alighting operations increases with respect to those running in between stops.

An interesting second aspect is the centrality of different nodes and stations in the network; in some network structures, one can reach many destinations from many origins, while in other structures more interchanges and even detours are needed. Figure 2.19 shows schematically the design dilemma represented by the *network density of lines*. Higher densities of lines as in the left example imply on average a lower number of transfers between lines for the travellers to reach any node, thus lowering the total travelling times. Providing this service, however, requires the introduction of a significant amount of extra lines for the same connectivity, and in general each line will be used by a relatively smaller flow of passengers.

The benefit of an integrated network is illustrated by the connectivity effect illustrated by the two networks in Fig. 2.20.

In the left network, every line runs in parallel to the other lines. In the right part, twice as many lines form a network. If interchange is possible in every crossing, the travel demand will be 550 % greater than when lines are operated fully independent without shaping a network.

A third design dilemma is on the *line density*, i.e., on the trade-off between increasing frequency of a line, and increases the number of lines. This is shown with the schematic example in Fig. 2.21, where lines and frequencies are traded off. In the first case (left picture), the crossing corridors can be served by just one route

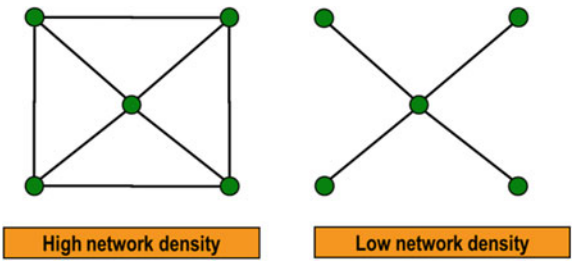

Fig. 2.19 Network density dilemma

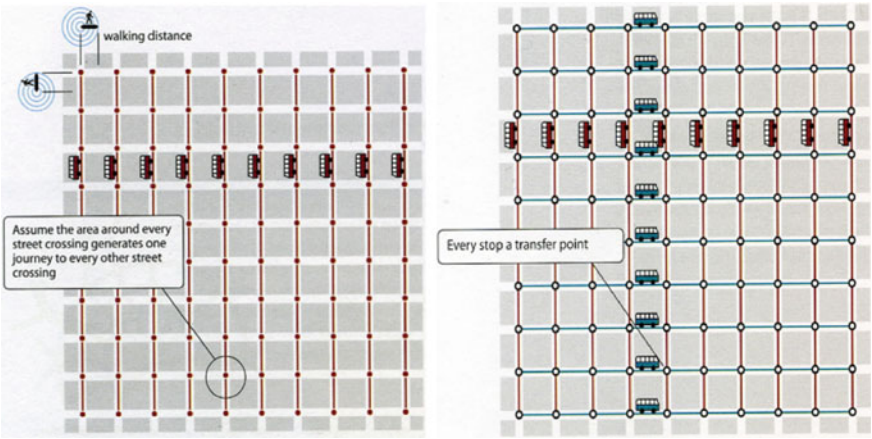

Fig. 2.20 The "network effect" (HiTrans 2)

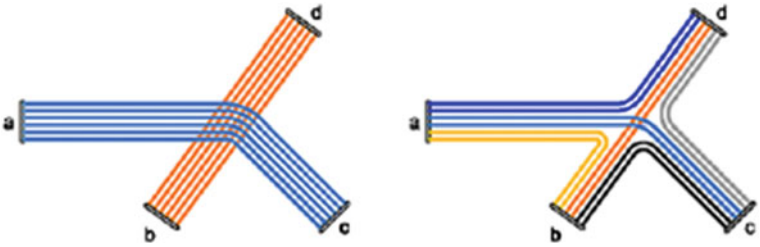

Fig. 2.21 Two network principles for crossing public transport corridors (HiTrans 2)

each. This enables a regular and stable operation on each of the two branches, but many passengers have to make an interchange. The alternative is to make several routes creating direct connections between destinations in different corridors. This, however, reduces the line frequency of each line and travellers may experience longer waiting times if they want to avoid a transfer. The crossing-lines system is often used in highly reliable systems, where synchronisation of vehicle arrivals at the stations is possible, thus reducing the disutility of transferring operations.

Systematically synchronised timetable operations are used both in interregional operations by railways, for example, in the Bahn 2000 system in Switzerland, and in local public transport. Often smaller- or medium-sized cities have a central node—bus terminal—where all lines meet at regular times (e.g., Fig. 2.22). To facilitate even further these operations, transfers are very often done at adjacent platforms. Corridors with heavy travelling exist in most cities. A good example is Wuppertahl in Germany with ordinary rail, bus and *Schwebebahn* along the corridor. The city can also be planned with a number of corridors radiating from the centre in a so-called finger structure. If there are spaces in the corridor it becomes a pearl-finger structure.

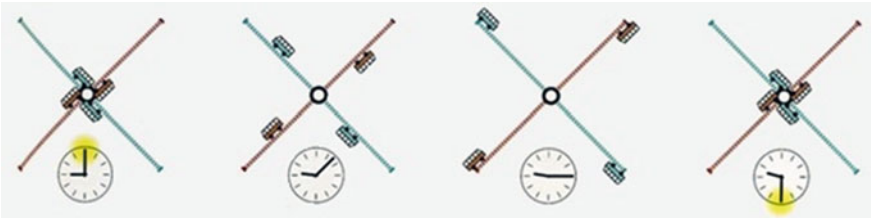

Fig. 2.22 Principle for systematic (pulse) timetable operations (HiTrans 2)

The above design dilemmas provide indications of qualities such as *spatial* and *temporal availability*. Space availability is related to the localisation of stations or stops, while temporal availability has to do with the period of time in which the service is offered. Frequency and headways are always very important indicators for users to evaluate the level of service and quality of a transit system, and partly determine the temporal availability.

Attractive headways are always dependent on the time on board the vehicle: short headways are satisfactory for short trips, while for long trips, lower frequencies are accepted. At this point, in the case that the demand is not high, a trade-off must be defined: whether to give priority to the space accessibility and the number of stops, or to the frequency. The mentioned trade-off offers two opposite limits: the traditional transit service with a high number of stops but large headways, and feeders, serving few points with high demand. Experience determines that high frequency is preferred over access time to a pickup point.

2.4 Service Performances

Public transport can, as we have seen, be run with different types of vehicles, on road, rail, but also beam and water. Seen from the operational point of view, there are more relevant ways to assess the performance of public transport, and to evaluate whether the choices of infrastructure, vehicle types and network type are efficient for the potential demand for PT.

2.4.1 Service and Stop Capacity

Many city buses do have an average speed of 12–18 km/h. In suburban operation, the average speed can be higher, but it is seldom over 25 km/h if not operated on a motorway. Traditional trams on streets are also quite slow, but when operated on their own RoW the average speed can be 25–30 km/h or more. Metro with a station

spacing of one kilometre has an average speed of about 35 km/h (e.g., the red and blue lines in Stockholm).

Capacity is often a main decisive factor for the choice of a mode. When choosing mode capacity, the network design is indirectly chosen, while high-capacity lines often need feeder lines.

The definition of capacity for PT services may have different facets. The two main definitions are as follows:

- *Vehicle capacity*: maximum number of seats and standing places offered.
- *Line capacity*: total number of passengers for all vehicles of a line per unit of time.
- *Passenger capacity*: total number of passengers between two points considering all lines.

These two definitions are related by the simple formula: $C = C_v \times n \times f_{max}$, where C is the line capacity, C_v the vehicle capacity, n the vehicles per unit of time (e.g., 1 h) and f_{max}, the line frequency (TCRP 2003).

Other important terms to distinguish capacity concepts are *offered capacity* in relation to the actual *utilised capacity*. These two terms differ mainly because of the actual load factor/capacity utilisation coefficient, i.e., the percentage of offered capacity that is actually utilised. Finally, during the design of a line, trade-off is done between *peak capacity* and *practical capacity*.

Table 2.2 shows that most transport modes can be designed with low, medium or high capacity. Generally, it is possible to get more capacity on a line with rail, because rail vehicles are easy to connect to trains. An expression that is sometimes used for heavy public transport (rail) systems (in USA) is "mass transit". Buses can haul a trailer, but this is rarely done nowadays. Instead bi-articulated buses are made with one or two joints and up to 25 m in length.

However, rail systems tend to be expensive in low capacity applications. On the other hand, the bus mode needs large-scale corridors with its own right of way, as in South American full BRT systems, to reach high-capacity levels. It should be mentioned that European cities are often more compact and "narrow" than cities in

Table 2.2 Public transport modes and capacity levels

Mode	Capacity		
	Low capacity	Medium capacity	High capacity
Rail modes	Tram in mixed traffic	Light rail with ROW Local train	Metro Regional express trains Interregional trains
Bus modes	Bus in mixed traffic City bus, Suburban bus Express bus on motorway	Bus on busway (ROW) BHLS BRT Light	Full BRT (e.g., Busmetro, Istanbul)
Other modes	Bus-on-demand STS	Personal rapid transit	

America and comfort demands by Europeans are higher than for people in newer economies in South America and Asia. Therefore, achievable capacity levels are in practice lower in Europe.

Regular buses have two axles and are about 12 m long with about 6 m wheelbase. The number of seats in a normal bus varies from about 20 to 50 depending on use and local practice/tradition. Italian city buses have very few seats and many doors, while the regional and interregional coaches in most countries usually are filled with about 50 (rather cramped) seats. The permitted number of standing passengers is restricted by weight and space limitations as well as safety rules. The permitted axle load restricts the length of a bus. Two axle buses can be built to over 13 m length, but often a so-called bogie (two interlocking rear axles) is needed for buses over 12 m. Bogie buses of 14.5 m take nearly as many seated passengers as articulated buses. One problem is that they require more space in curves.

Short buses, *mini- and midi buses*, are relatively rare in many western countries because they are almost as expensive in operation as normal buses. Based on car or truck components, the buses can be cheap but the bus driver cost makes the traffic almost as expensive as for normal-sized buses (Fig. 2.23).

In old, crowded, cities in Europe, sometimes shorter (mini or midi) buses are used for access reasons.

Articulated buses (with one joint) are about 18 m long and the number of seats is often around 60–70. In other countries, there are even larger articulated buses with two joints. Volvo has, for example, produced double-articulated, 25 m long, buses for many modern bus systems (Fig. 2.24). Another way to increase capacity is to use buses with two decks, i.e., the *double-decker*. Often the number of seats is increased significantly, while the capacity of total number of travellers is not so much higher than of an ordinary bus with one deck.

The vehicle capacity figures vary with a number of factors: seat/standing ratio, if it is a low floor design, the permitted weight restriction and more. For vehicles with new propulsion systems, the weight restrictions might be limiting the permitted vehicle capacity. The given space per standing passenger also affects a lot. This can vary from 4.0 to 6.0 passengers/m^2 and even more, at least in practice in metros, commuter trains and buses to high transport demand in mega cities. Table 2.3 shows figures from Stockholm local transport.

Fig. 2.23 Battery-operated small bus in Paris on narrow streets (2003) and Midibus in Leeds (that offer free trips in the city), 2006 (*photographs* Karl Kottenhoff)

Fig. 2.24 Double-articulated bus in Switzerland (*photograph* Segunda-Feira, 2011)

Table 2.3 Maximum capacity for various public transport vehicles

Mode	Capacity		
	Seated	Standing	Max per vehicle/train
Normal bus, 12 m	32–42 (35)	35	67–77 (70)
Bogie bus, 14.5 m	46–56 (50)	45	91–101 (95)
Articulated bus, 18.5 m	44–68 (55)	60	104–128 (115)
Tram/LRV, 30 m	80	105	185
Metro train, full length, Stockholm	400	800	1200
Commuter train, full length, Stockholm	750	850	1600

Source SL Planning Handbook, 2008, available in Swedish, with rounding in parenthesis by Kottenhoff

Practical capacity is far lower with maximum capacities where about 50 % of the standing places can be used in an average rush hour, as a rule for planning. In practice, certain vehicles will be crowded and over-filled if one tries to plan for 100 % capacity, because of statistical variations, daily variations, weekday variations, seasonal and weather variations, part of the route variations, passengers changing from high-capacity rail systems, etc. In the rail modes, some carriages may be laid out to have more standing room than seats, or to facilitate the carrying of prams, cycles or wheelchairs. Some countries have double-decked passenger trains for use in conurbations. Double deck high-speed and sleeper trains are also becoming more common in mainland Europe.

The space and time allocated for stop operations are in relation to the bus frequency, to the average number of passengers that are observed in the boarding and alighting operations, and to the number of accesses that the vehicles offer. If buses arrive with high frequency and many passengers board at the stop, more than one bus is likely to be served simultaneously. With even higher capacity demand, different lines should have different stop areas, or stops. Maybe all lines do not have to serve all stops. For example, express services and trunk lines may skip some stops. This is also a method used for rail services, metros and local rail operations.

2.4.2 Systems Speed—Boarding, Alighting and Travel Times

Speed is a key indicator in the modal share of a territory. However, a person deciding to make a trip compares the travel time among the available alternatives. As a result, it is not only the speed involved in the time spent on a trip, but also the design of the lines (routes), the localisation of stops and the frequency of service. Short travelling times require high average speeds including stops. Low-speed systems are used mostly for local transport, medium-speed systems for regional and high-speed systems for interregional services (Table 2.4).

Boarding and alighting operations vary significantly in Europe. While British buses often have just one door, in the front, Italian city buses sometimes have four wide doors along the side of the bus. These conventions have an impact on the internal distribution of the passengers in a vehicle, as well as their circulation within the bus (Fig. 2.25).

The choice of the type of boarding and alighting system depends also on the ticketing system and control. In some countries, it is possible to purchase tickets

Table 2.4 Speed levels of various modes

Mode	Speed		
	Low speed	Medium speed	High speed
Rail modes	Tram in mixed traffic	Light rail with ROW metro Local train	Regional express trains Interregional trains High-speed trains
Bus modes	Bus in mixed traffic City bus	Suburban bus Bus on busway or bus track (ROW) BHLS, BRT	Express bus on motorway
Other modes	Local ferries	Personal rapid transit Ferries	Air

Fig. 2.25 City bus with five wide doors. Presented by MAN in the EBSF project (European bus systems for the future 2011)

from the driver or by a vendor at the front entrance, while in others the purchase is allowed only in vending services outside of the vehicles. Clearly, the first system requires on average longer boarding time operations.

Rail modes mostly practice boarding and alighting at the same doors. There are often many doors in trams, metro cars and local trains. Most long-distance trains have coaches with two doors for a 26-m coach. Buses are different. The number of doors varies between countries and traditions. For example, Italian 12-m city buses have up to four doors, while their British analogues have just one or two doors.

Service speed and boarding/alighting times are main determinants of the travel time of a vehicle. From the operator's perspective, travel times must add also operations that are not perceived by the customers (*deadhead times*), but are ought to be included to define the rolling stock, i.e., the total number of vehicles used during operations, while in maintenance, stopped for checking operations, etc.

2.4.3 Reliability, Punctuality, Regularity and Robustness

There is no doubt that the transit operator must define a trade-off to assure an attractive service, while assuring it is economically efficient. A set of variables should therefore be designed by the supplier: localisation of stops or stations, routes, RoW category, speed and capacity are certainly the main performance measures that define the overall level of service. Passenger, however, value qualities that are less related to the form of PT offered, such as flexibility, safety and security, costs and attractiveness. Considering the quality of service perceived by users, other variables that have to be previously determined are comfort, convenience, security and safety, and the cost. Quantitative features may indirectly measure some of them.

Literature has determined reliability to be an essential feature of the quality of a transit service. Reliability can be defined in PT services as the degree of trust in the service. This is a significantly different concept than punctuality, which is often the quantitative indicator to evaluate reliability and is measured as per cent of vehicle arrivals within a previously defined period of after the schedule time. Customers depending on the headway of the service differently perceive punctuality: reduced waiting times due to high frequencies lessen the inconvenience of delays. However, low punctuality is often related to longer times on board due to traffic conditions in transit services with no exclusive lanes or pathways.

Stability and regularity can be seen as the analogous of reliability and punctuality, but from the operators' view. Stability can be seen as a measure of the capacity to effectively dispatch the scheduled services in the planned time. High stability of the system implies that the service has also high regularity, i.e., adherence to maintain the frequencies/headways planned. Highly irregular services are often characterised by frequent bunching of vehicles from the same line arriving at a stop.

If reliability and punctuality are qualities that are often related to a specific service (e.g., a line, or the total PT system), robustness refers to the design and operational strategies adopted by the PT service to cope with incidents or network disruptions. Currently, there is no universal definition of robustness, as it still remains a relatively fuzzy concept in transit systems. Increasing robustness has to do with reducing the negative effects of an unpredicted event to the system, and this can be achieved in different ways. For instance, this can be achieved offline by carefully designing timetable in order to absorb the effects of delays or avoid that unpredicted delays propagate along lines and between lines, or online by preparing flexible strategies and backup plans to apply once a disruption occurs (e.g., rerouting vehicles, providing emergency services like substitutive buses in case of a rail disruption, etc.).

2.5 Conventional and Unconventional Services

Conventional services are conceived for a wide public of users, ranging from commuters to students, from inhabitants to tourists. They are organised in lines that serve a fixed sequence of stops with a given frequency and/or a published timetable. Lines using traditional vehicles for public transport on different infrastructures are possibly integrated in a transit network, where some stops are shared. Short pedestrian transfers may be required among different stops and intermodal terminals.

In unconventional services, some of the above assumptions do not hold. For example, they may be devoted to a particular category of users, or may be operated with non-typical transit vehicles, or without a fixed route, or they may be conceived for a particular function.

2.5.1 Complementary Services

2.5.1.1 Feeder Lines

In the case corridors are served by rail feeder buses, they often complement them. But there are also pure bus corridors. Two methods for network planning, which take into consideration both trunk bus lines and so-called feeder lines from smaller counties/towns, can be found in the literature (Fig. 2.26):

- **Trunk-Feeder technique**—larger buses utilise lines in the main network, while smaller buses utilise the more detailed lines in the local network. The main network changes to a more detailed network at the terminals where even the travellers are forced to transfer. The upside with the "Trunk-Feeder technique" is that one can adapt the bus size to the flow of travellers, while the downside is that the travellers are forced to transfer in many cases.

Fig. 2.26 Two techniques for network planning: "Feeder-Trunk technique" and "Convoy technique" (HiTrans 2)

- **Convoy technique**—more bus lines run parallel to the main network and then they branch out to the city's periphery/suburbs. The advantage with the "Convoy technique" is that the main network is utilised by many bus lines, resulting in a high frequency of service and few transfers for those travelling from the network periphery. The disadvantage is the risk for the service on the main network to become over-dimensioned and difficult to understand.

Direct lines, often with buses, take passengers from an origin area to a destination area directly, without a lot of stops in between. Such lines often complement trunk lines by rail (Fig. 2.27). Wrongly made, this competition threatens the economy of scale for the transport system. Properly made, for example, when planned by a PTA, it can increase the travel standard and social economy.

Sometimes direct lines express lines and motorway buses are mixed with respect to terminology. Even newer concepts like BRT are confused with the other terms.

2.5.1.2 Shuttle Services

In many occasions, there are clearly defined patterns of trips for which feeder and shuttle services are the best option to deal with the demand requirements. These requirements sometimes come from the excessive number of automobiles in a train station or transfer point car park, or excessive car trips on the same route from an origin to a destination at specific times of a day. Shuttle buses commonly serve trips between an origin and destination characterised by a high demand with no stops in between them and are usually designed along a fast route. When the destination is a transfer or intermodal point, shuttle buses serve as feeders of other modes.

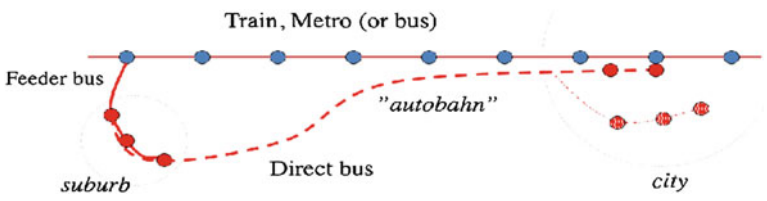

Fig. 2.27 Trunk line with feeder bus and direct bus: the trunk line can be rail or bus, e.g., BRT

The fleet size of a circular shuttle route is assessed taking into account the road network attributes, the average travel time for a vehicle to make a round journey and the time of the day, since the demand distribution may fluctuate. Once the fleet requirements have been considered, the vehicle chains must be designed. The next phase deals with the route strategy. The extreme shuttle service is designed with no stops but the origin and destination. However, this restriction may be relaxed as much as it is required.

Fixed or flexible routes and schedules, together with direction (fixed or bi-directional routes) and the possibility of short turns and cuts (direct shortest path to the destination or back to previous stops instead of an established number of stops and a fixed path) lead to strategies comprised of combinations of the mentioned characteristics. The strategies could successfully be implemented as long as the capacity restrictions are considered and the potential demand has been assessed.

All in all, the aim of the previous phases is to achieve minimum waiting and travel times with the minimum number of vehicles. In the case of a shuttle service to a long-distance bus/train station, the on-time or established arrival should be assured.

2.5.1.3 Differentiated Route Networks

Contrarily to rail-based systems, bus networks can adapt to changing needs rather quickly. Earlier, in medium-sized cities, there was a kind of basic networks that would cover most travelling needs by public transport. Differentiated bus service networks for different customer groups started during the 1980s. Therefore, new forms of network have arisen.

We may speak of a "differentiated networks" with different kinds of routes created to satisfy different needs. Some examples of such route types are as follows:

- *Base (trunk) routes*: fixed routes which run a large part of the day and night.
- *Peak routes*: routes run in the peak hours, in (partial) other relationships.
- *Off-peak routes*: often operating during evenings and holidays instead of base routes.
- *Night service routes*: often a few routes at low frequencies.
- *Express tours (routes)*: go in parallel with other routes but avoiding intermediate stops.

Trunk routes can also describe the biggest and most important routes among the base routes. The most extreme form is BRT routes in dedicated corridors on exclusive right of way.

Some types of routes are tailored to needs of specific groups:

- *Working trip routes*: sometimes created between large residential and work areas.
- *School trip routes*: connect to schools on school hours.
- *Service routes and flex routes*: meet mainly the needs of disabled people.

The concept "service routes" appeared during the 1980s in Sweden. Mini- or midibus networks connect areas where elderly live and where they have to go: city, shopping, hospitals, etc. A newer travelling variant for the elderly is "flexi routes". Here, the small buses drive past a predetermined location. These more specific services are described later in this chapter, together with other non-conventional frequency- or schedule-based services.

2.5.2 High-Level Bus and Rail-Like Systems

2.5.2.1 Buses with a High Level of Service (BHLS)

Of particular interest in recent years has been the creation of a trunk network for buses. The intention has been to get part of the rail service qualities but to a lower cost. An example is the Stockholm's "blue buses", which are serving as trunk. Route 4, the heaviest route, had in 2005 about 60,000 passenger journeys a day.

The following elements can be employed to construct a trunk network:

- Reserved lanes and priority at traffic signals;
- boarding and disembarking at all doors with another type of tariff control;
- real-time display of waiting times, delay information and "next stop" signs;
- straight and direct routes linking areas of high base;
- slightly longer stop distance and sparser networks with higher frequency in return.

Many of the elements are already familiar—priority for buses in traffic, higher quality vehicles, improved comfort at stops, improved information to passengers, integrated ticketing, intelligent transport systems to improve operations management and planning, etc. However, BHLS differs from the conventional approach in three main respects:

- The elements are combined in a holistic way, to achieve a total product improvement rather than just improve specific aspects.
- The BHLS is usually packaged as a concept, given a distinct label and marketed to the high market.
- The BHLS usually serves urban and transport policy or strategic objectives, and there are not just technical or operational improvements.

BHLS means buses with a high level of service which is not a fully developed BRT system but a French/European version suiting the European cities. It may be called light BRT. CERTU (2007) defines BHLS as "BHLS is a public road transportation concept for the structuring services of the network that meet a set of efficiency and performance criteria, coherently integrating stations, vehicles, circulation lanes, line identifications and operating plans in an on-going manner".

2.5.2.2 Bus Rapid Transit (BRT)

BRT is a system operating on its own RoW infrastructure either as a full BRT with high-quality interchanges, integrated smartcard fare payment and efficient throughput of passengers alighting and boarding at bus stations; or as a system with some amount of dedicated RoW (light BRT) and lesser integration of service and fares. This system can cost four to 20 times less than the LRT system and 10–100 times less than a metro system. BRT combines the flexibility and lower cost of bus, and the speed and reliability of rail.

BRT originated from Latin America. During the 1970s, there was a rapid growth in the urban centres placing a high demand on the transport sector for public mode of transport. With a fast population growth and limited resources, people were dependent on the public transport only which led to the development of a rail-based infrastructure in the form of BRT. Examples of highly efficient bus services, BRT, are seen in several places in South America, e.g., in Curitiba, Brazil and Bogota, Colombia but also in Ottawa, Canada, and in a growing number of Asian cities. It has in some cases reached very high capacity in traffic, up to 40,000 passengers per direction in rush hours.

There is no system yet in Europe with 100 % heavy busways. BRT in the urban model and in practice in Europe revolves around light busways.

BRT can assume different forms, ranging from dedicated bus lanes to a complete network of BRT lines in corridors. The infrastructure design must encompass a wide range of system components, including busways, stations, intermediate transfer stations, terminals, depots, control centres, traffic control signals, integration facilities, public utilities and landscaping. Likewise, depot areas must be designed to handle a range of tasks including refuelling, cleaning, maintenance and repair, and vehicle parking. A control centre allows system controllers to ensure a timely service to the customer as well as the ability to respond to any problems or emergencies.

Radial BRT traffic is substituted in large cities lacking a metro system. We can take the example of Curitiba, where a number of radial corridors were developed. More examples are available in other developing countries. Even in America there are certain cities almost completely without rail-based traffic, like Ottawa. Department of transportation (DoT) in USA has issued a report, which highlights a form of BRT system which was called “Light Rail Lite” and described as a cheap alternative of modern light rail. Light Rail Lite is run with lower average speeds and lower passenger volumes.

In cities with a well-developed (rail) transit system which accounts for most of European big cities and exists a need for better tangential and feeder traffic. Amsterdam, Paris and Sydney have developed integrated tangential “rail” corridors based on the BRT concept. In medium-sized cities, the complete transit system can be based on BRT. This idea is being followed by some French and Dutch cities such as Rouen and Eindhoven.

Fig. 2.28 Volvo concept bus for the EBSF project (Volvo Bus) and Translohr rubber wheel tram, "Tram on Tire" in Padua, Italy (*photograph* Kottenhoff 2008)

Many BRT systems are built on trunk lines in corridors with feeder lines. This can prove to be a good solution if the trunk lines are much faster than bus services without the need for transfer.

Both BRT systems as well as the European BHLS concept favour new vehicle designs. The applications vary from ordinary, but distinctively designed buses to buses on tracks and "trams on tyre" (Fig. 2.28).

2.5.3 Demand-Responsive Services

The dominant form of public transport is schedule-based. Buses, trams or trains are running—or trying to run—according to a published timetable. Timetables are designed to meet expected demand (during periods of high demand) or to offer a minimum level of service (at low demand). This means that operators publish timetables for various times of the day and various times of the year. Passengers plan their trips based on the timetable information and/or information on perturbations. If the headways are small enough (below 5–10 min), passengers may not bother about timetables.

In contrast, the idea of demand-responsive traffic is to operate when and where there is demand. Their purpose in relation to scheduled traffic is usually one of the following:

- saving cost in areas or at times of low demand;
- improving access by coming close to points of demand; and
- offering individualised service of high standard.

2.5.3.1 Dial-a-Ride

DAR services are the alternative for regular trips starting or heading to those areas that traditional public transport does not serve due to high private car ownership or

inadequate conditions of roads or streets. Together with other paratransit services such as route deviation or flex-routes, DAR is designed to fulfil the lack of service of regular transit and to provide a cheaper alternative compared with taxis. It commonly serves rural areas, towns and peripheral neighbourhoods.

This hybrid mode is a system operated by a company that owns the vehicles (cars or small buses) and serves a predefined area. In most cases, users reserve a trip by telephone, informing about their transport needs, but systems exist where users arrange the trip with the driver. DAR is also an alternative for regular and arranged trips. An increase in the demand of DAR causes a decrease in flexibility by increasingly fixed routes.

Compared to taxis, DAR is a cheaper service as a result of trip sharing but, on the contrary, inconveniences come from excessive waiting times or uncertainty of arrival time to destination.

DAR is without any doubt a demand-responsive transit (DRT). The hybrid nature of this type of transit leads to difficulties when intending to efficiently meet demand requests. Literature defines two types of routing: many to many or one (or several) to many. This last option is expected to operate as a shuttle/feeder service. Advanced technology such as AVL and GIS provides instantaneous data that reduce the routing-scheduling problem when the fleet exceeds the possibility of manual planning.

DAR is the optimal mode to supply mobility to disabled persons, in areas where regular transit does not serve, at off-peak times of the day or in small cities and towns.

New Paratransit or hybrid modes are based on many-to-one DAR services where routes are fixed. In this case, if the vehicle can deviate from the route to pick up passengers by demand, then the service is called route deviation. However, services may not operate with a schedule but by request: on-call services, which may provide mobility as a feeder service to a train station.

2.5.3.2 Bus-on-Demand

Bus pickup ordered by telephone goes by different names, such as Dial-a-bus, Ruf-Bus, Tele-Bus or Flexi Line. They exist in many European countries. The concept has been around at least since the early 1970:ies with manual order-handling and order-dispatching. Now, of course, these functions can be made automatic.

Common applications include pickup and distribution to/from a major destination such as a service centre or a transit terminal. Pickup stops are ordered (not destinations) and all orders are for the next scheduled tour. Distribution trips from the terminal need no order, and the driver usually does the route planning to verbally requested stops. Without orders, the next trip will be cancelled.

Bus-on-demand is typically introduced in areas of low demand. The area served can be a sector from a terminal or a service centre or it can be a corridor between two nodes. The route can be fixed but the concept allows additional possible stops

over an area since not all of them need to be visited. Advantages are cost reduction from cancelled trips and improved area coverage and short walking distances. The vehicles are often minibuses or contracted taxi vehicles.

With handicap-adapted vehicles and stations close to buildings, this scheduled service can serve some demand which otherwise would have required more expensive transport for handicapped.

2.5.4 Paratransit

Paratransit is a term for forms of transport that lie between ordinary public transport and conventional taxi. It denotes services that do not follow fixed routes or schedules. They vary considerably in the degree of flexibility, and the term is used differently in different parts of the world. It can denote jitney and minibus services in developing countries, handicap services in North America or generally Bus-on-Demand services.

Vuchic (2007) defines Paratransit as urban passenger transportation service mostly in highway vehicles operated on public streets and roads in mixed traffic; it is provided by private or public operators, and it is available to certain groups of users or to the general public, but it is adaptable in its routing and scheduling to individual user desires in varying degrees.

The oldest Paratransit mode to be known is the taxi but other hybrid modes have been designed to meet semi-public transport demand: jitney, DAR and subscription commuting service. Their main attributes are flexibility and demand-responsive performance, whereas regular transit services lack these characteristics. Among the alternatives of Paratransit are the following:

- Exclusive-ride taxi
- Shared-ride taxi
- Children's services
- Subsidised general public transportation
- General public DAR
- Fixed-route feeder services
- People with disabilities
- Subsidised medical care receivers
- Other social service clients.

Four characteristics are claimed to define Paratransit modes: the type of usage, the fleet ownership, the routing and the accessibility. This service may be supplied either to the general public with a regulated fare or to people involved in an organisation (school, factory, office). The provision of vehicles may come from a transport agency, an individual owner (taxi or jitney driver) or any other organisation. Hybrid modes of transport may provide door-to-door service, a fixed route or an alternative between them. From the availability point of view, the service may follow a prearranged schedule or it may be designed to provide specific service to

individuals after a phone call. Needless to say, the combination of this group of attributes leads to different capacity constraints.

Among semi-public Paratransit modes are vanpools, subscription buses and car sharing. Public alternatives are taxis, jitneys, DAR and other hybrid services.

In short, Paratransit involves specific mobility and low passenger volumes. Therefore, besides the specific situations where hybrid modes are suitable, these also offer short-term solutions to deal with mobility requirements before the implementation of a regular service. These modes should be considered part of the mobility in coordination with other public transport systems. Regulation of quality of service, security and safety is greatly recommended for these modes to appeal to the entire community and to avoid the use of private vehicles in those towns or neighbourhoods with high rates of car ownership. However, marketing is essential and information should be clearly delivered. To conclude, commuters, schoolchildren, air passengers and people with disability (PwD) are the first targets of the various modes involved in Paratransit.

The denotation demand-responsive transport does include not only road transport but also guided systems like many PRT (personal rapid transit) solutions. In the following, some forms of demand-responsive traffic are described.

2.5.4.1 Vanpools

Commuter vanpools consist of medium to big vans driven by one of the commuters, and the driver is free to use it for personal purposes too. Unless this driver wants to purchase the van, the operating and maintenance costs are shared among the commuters involved. This mode is recommended for large factories or working destinations which are not served by regular transit.

Subscription buses are used in a similar way as vanpools but are provided by the company or institution or by a transport agency. Subscription buses require a higher number of participants and can provide several runs during the day but are generally operated on fixed route and schedule.

Charter vans as in Buenos Aires offer subscribed luxurious commuting from distant suburbs to the city. Air-conditioning, working space and parking in the city are attractive features.

2.5.4.2 Car Sharing

Car sharing is an option for non-regular trips and people who do not own a private vehicle but need one in specific situations where the transit conditions do not meet their needs. As a result, they are designed to complement public transport. Users have to subscribe and are given a personal key. Whenever such service is required, it can be booked online or by making a call.

2.5.4.3 Taxi Services

Taxis are complete individualised modes of public transport and are offered either at specific stops or on the streets. Vehicle assignment is possible by radio and automatic vehicle location (AVL) systems. The main characteristic of this mode is its availability: everyday and anytime service which provides rides to any destination. Consequently, taxi is a very attractive alternative at specific moments, situations or destinations. Regulation of taxis has traditionally been a matter of discussion. Therefore, fares, stop locations and entry conditions for new licences may or may not be controlled by the public administration. Non-regulated services generally lack security. However, even with regulated fares, there is always an uncertainty of the amount of money a trip would cost since it is a function of the travel time and time-of-day.

2.5.4.4 Jitneys

Jitneys stand for popular taxis or minibuses and have been an alternative of public transport in many developing countries. The vehicle is driven by its owner and has a capacity of as many as 15 seats. This specific mode of transport shows a mixture of advantages and disadvantages over regular transit: whereas frequency and speed are higher, regularity and reliability are not. Jitney performance is good at peak hours, but the level of service is low at off-peak periods of time. Vehicles sometimes are owned by different companies in a city and might be coordinated in routes and schedules (fares are established by public administration). It is a general fact that those cities in which jitneys are not regulated end up lacking security and proper conditions and mostly cover the same commonly used routes at the same highly demanded periods of time. The major inconvenience with jitneys is that even when there is a great number of vehicles around a city, they do not provide transport as a system. Consequently, road occupancy is high due to these vehicles. Moreover, if this service is not under any control of the public authorities, they will never be eliminated when a broader and more organised service is required to be implemented. International experience shows that jitneys succeed when regular transit fails to provide convenient services.

2.5.4.5 Paratransit in Developing Countries

Paratransit systems in many developing world cities are operated by individuals and small business (Fig. 2.29). An entrepreneur often owns a number of vehicles, which are leased to individual drivers who have to earn a sufficient amount every day. Therefore, the minibuses wait until they are full. The fragmented nature of the industry makes government regulation and control hard. Government authorities have cited problems with unsafe vehicles, etc.

Fig. 2.29 Paratransit vehicles in Jakarta, Indonesia (*photograph* Kottenhoff)

2.5.4.6 Corporate Transport Services

Large organisations such as factories, hospitals, universities and schools may prefer to arrange services for their commuting members to and from the workplace. These services are called "corporate/company services", "service vehicles" or just "services". They can operate on a fixed schedule and route. Each bus has a predetermined path to follow and pick up points along this path. In the mornings, the service usually collects the registered members (i.e., employees, pupils) from these certain pickup points to bring them to the work place. In the evenings, after work, it takes them to their home zones.

Modifying the system by introducing more flexibility may offer promising benefits with the help of contemporary IT solutions. Rather than a fixed list of members, opening the service vehicles to the use of non-registered employees of the company will attract more demand. Secondly, rather than a fixed schedule, a more dynamic timetable, in line with varying demand, will also make the system more attractive. Thirdly, rather than a pre-booked registration, designing an IT-based infrastructure which receives bookings electronically over Internet or mobile phones will greatly improve the system (as they can determine the coordinates of the user and calculate the optimum path to serve the optimum number of users, and send messages to the driver of the service vehicle).

2.5.4.7 Special Transportation Service (STS) for Elderly or Handicapped

In Sweden, Canada, USA and some other countries, there are special transportation service (STS) for the elderly or the PwD. This demand-responsive public transport is run in part by taxi and partly by the local PT companies. In some cases, taxi cars are being used, and in other cases, small buses are being used.

In some cities, disabled passengers who cannot use ordinary buses are entitled to transport with special vehicles and assisting drivers. Sometimes, two persons are needed to assist a passenger in stairs. These services are often contracted out to companies with special vehicles and trained drivers/assistants. These services are labour-intensive and expensive, so their scheduling and routing need to be efficient.

A comprehensive STS is much more expensive than ordinary public transport. In Stockholm, it costs more than 15 € per trip.

Known trips are scheduled or pre-ordered the day before the trip. That allows consolidated routes to be planned during the night with multiple passengers picked up and delivered. The time of return trips is often not known in advance, so they need to be scheduled in real time, typically by inserting them into planned routes. Several computerised systems are used for this problem.

As a measure to cut high cost, SL in Stockholm had introduced midi buses on-demand for elderly people (Fig. 2.30). These attract people that otherwise would have used taxi vehicles, but the midibus services are also open to the general public. Each midibus covers a predefined part of the city and runs every hour from 9 to 16 on weekdays. The driver can plan her/his route in real time with limited help of basic ITS, so there is potential for product development.

Service routes run mostly by special buses for 10–20 persons. They are built with low floors. The same vehicle can also be used for "flexi lines", which are a kind of service lines that do route deviations to pick up people at specific pickup locations.

To get low entrances, all or part of the buses and trains today are built with a low floor. The buses that have only partially low floor are usually called *low-entrance buses*, the other are called *low-floor buses*. The low floors make the wheel arches protrude into the passenger compartment and the buses are harder to furnish. Low-floor buses are often difficult to furnish and when the engine is in a cabin, even more seating options disappear.

The platform heights still vary a lot between countries and even between lines and stations in each country, but standardisation has been initiated, for example, in Sweden (550 mm for interregional lines).

Fig. 2.30 Vehicles used by the STS in Stockholm: low floor minibus, Taxi and Midibus. The latter for bus-on-demand services (*photographs* Hans Adeby, SL Stockholm, Public Transport Authority SLL)

2.6 Automation and New Transport Systems

Automation can help in making demand-responsive transport and in making systems where the drivers' role is limited. Automated metro systems are in use in many places. These give possibilities to run shorter trains and denser services in off-peak times without driver costs. There are also forms of automated smaller systems like automated light rail systems and people movers.

2.6.1 Advanced Control for Rail Systems

Railway safety systems are based on four major components. Track-free detection system (track circuits or axle counters), monitors the occupation and releases of track sections. The infrastructure and interlocking safety requires that movements are done only if all points (switches) are properly set and locked, conflicting routes set and locked, flank protected, and requested tracks are clear. The occupation status of block sections ahead is transmitted to trains by a signalling system, among which conventional trackside signals are the most common ones. Automatic train protection (ATP) guards against driver errors, by supervising operations and intervenes in case of driver errors. Those basic functionalities are implemented in traditional legacy systems that greatly differ between countries.

Control of rail operation is also done by manual intervention. In principle, this intervention function should not at all affect the safety function of the traffic management system, but the systems are to some respect integrated. A big challenge for European rail is that each national railway has its own technical solution, even when the supplier of the systems can be the same in some countries.

The ERTMS is the Europe-wide standard to enhance cross-border interoperability of train operations. The overall ITS scheme aims at improving safety, capacity and performance of a railway network. Main components of the ERTMS are the European Train Control System (ETCS), a standard for signalling, control and train protection system, and communication standards among which GSM-R.

ETCS is specified at four different levels:

- L0: ETCS compliant vehicles interact with non-ETCS compliant trackside equipment and signalling
- L1: ETCS is installed trackside (possibly together with legacy systems) and on board (cab signalling); data are transmitted from track to train at prespecified locations via ETCS balises
- L2: as level 1, but ETCS data transmission is continuous; the currently used data carrier is GSM-R
- L3: as level 2, but train location, track-free detection and train integrity are determined on board rather than by trackside equipment.

ETCS L2 is the most advanced system currently available for deployment. ETCS L2 is a fixed block system with track-free detection that the interlocking system uses to set routes.

The interlocking receives track-free status to set routes and communicates those to a Radio Block Centre (RBC), which then translates this information into Movement Authority (MA). The MA, together with a track description, is sent to the ETCS L2 computer on-board trains. The train determines a dynamic speed profile based on the MA, train characteristics and all track speed restrictions. The MA and speed profile are displayed to the driver in the cabin, without the need for trackside signals. The train sends its position and speed to the RBC, including routes requests.

The on-board computer supervises that the permitted speed at each location is within safe braking; an emergency mechanism handles the case when the driver does not react promptly and exceeds the braking curve.

2.6.2 Automated Rail and Metro Systems

A further step forward might arise from automatic train operations (ATO). Vehicles will compute the best speed profile to match the current MA. Such an arrangement is currently used mostly for closed systems such as subways. Automated metro systems are in use in many places worldwide.

The earliest ATO were used in operations in the 1960s in metros, and more modern services are continuously developed and applied to metro systems worldwide. Among the different arrangements, the automation can vary from driver supervision, to driverless operations but with a driver or attendant on board, and going as far as being remotely controlled, or completely automatic. There are also forms of automated smaller systems such as automated light rail systems and people movers. For example, a cable-drawn minimetro operates in the small city of Perugia with headway of 1 min.

Benefits of ATOs consist in an extremely high level of reliability in realised departure, running and arrival times. This results in a much higher capacity of the link, as the distance between successive vehicles can be controlled with a high level of reliability, and headways can be safely lowered to 90 s or even less. These give possibilities to run shorter trains and denser services in off-peak times without driver costs. In fact, the cost structure of a driverless system is completely different than when needing a driver. To adapt to variations in demand, a variable amount of vehicles can be sent and operated for a smaller headway time and increased frequency. Applications of ATO in metro have a target reliability result of 98 % of less than 2 min delay; up to 99.4 % service availability has been reached in Copenhagen metro.

2.6.3 Cable-Propelled Transport (CPT)

Vehicles pulled by cables have been around since the nineteenth century. Initially, running on rails at grade but later also elevated and suspended. An example is Perugia automated Minimetro in Italy running on rails partly elevated and partly tunnelled. LaPaz "Linea Roja" opened in 2014. Most recent applications use a continuously running cable with detachable grips. Cable-propelled transport (CPT) lines can have speeds up to 45 km/h, although typically 15–25 km/h. Vehicle capacities and headways range from 12 passengers every 20 s to 180 passenger trains every 2.5 min. Typical transport capacities range from 2000 to 5000 persons per hour and direction.

2.6.4 Personal Rapid Transit (PRT)

The idea of PRT is to offer "car quality" within a public transport system. PRT offers automated taxi service over a network of dedicated rights of way. PRT can run on elevated guide-ways, in tunnels or at grade between fences. Each vehicle takes 2–6 passengers. According to some visions, it would be possible to construct beam transport systems with both PRT and scheduled public transport on the same infrastructure.

Transport is on-demand and non-stop along the quickest path to each passenger destination. Since trips are individual or made with a chosen company, vehicles can be small, allowing slender guide-ways. Stations are offline so that stopping vehicles do not delay passing traffic. Typical PRT travel speeds are 35–45 km/h which is often faster than car traffic. Safe headways as low as 3 s between vehicles allow for line capacities up to 1200 veh/h. Ridesharing is encouraged for passengers with the same destination. Guide-ways are one direction with merges and diverges but no intersections.

PRT complements scheduled transit in many respects:

- serves areas as opposed to only corridors;
- vehicles wait for passengers instead of the opposite;
- many stations are possible without slowing trip times;
- travel times are typically half of those with scheduled transit; and
- automation allows all 24 hour service.

One factor that may be crucial for traffic on elevated beams to get a major breakthrough is whether it can fit in the visual environment. It is thus very important to get sleek beams and columns. The biggest problem is often terminals and stops, which may be bulky and ugly in the urban environment with its raised position. So far four PRT systems are in operation; in Morgantown (USA), Heathrow (UK), Masdar (UAE) and Suncheon (South Korea) as of April 2013 (Fig. 2.31).

Fig. 2.31 Personal rapid transit systems (Vectus) in Suncheon S Korea and at Heathrow airport (*photographs* Yonrap News and Kottenhoff)

2.6.5 Automated Road Transport

Development of self-driving vehicles is ongoing, and these may be allowed on at least some motorways within some 10 years. It may also be possible within cities or certain areas in cities. An example of automated minibuses already operates on a route in a business area in Rotterdam (Fig. 2.32). These operate on a closed busway with a number of small stations. Similar technologies can also be used for full PRT systems.

The full use of automated cars can be attractive for some people who get stressed when driving personal vehicles. Private cars can adopt some of the advantages for bus and rail. They could become more like mobile offices in which people can text, talk by cell phone, send emails, or sleep without worrying about the dangers of distracted driving. Many transit passengers today use transit because they can engage in electronic communications without worry of causing an accident. Much of the advantage public transit now enjoys in attracting choice riders would be gone with automated cars.

Perhaps the biggest, and no doubt the most controversial, question would be how it would affect the position of bus drivers. If they didn't have to drive the vehicle, would their role become more of a customer service person, providing passenger

Fig. 2.32 Automated minibus in regular service on special busway in Rotterdam (Kottenhoff, June 2014)

assistance, information and security? Is it possible that the position of bus driver will not be necessary?

Self-driving cars can be used in taxi fleets (aTaxi) offering cheaper rides than manually driven taxis. Routing, ride-sharing and empty running can be optimised with ITS. Depending on ownership, regulation and pricing a Taxi can be a complement or partial substitute for other public transport.

2.7 Reference Notes and Concluding Remarks

Organisational forms have changed over the last decades. Different forms of deregulations have in many countries gradually replaced PT services based on pure monopoly, i.e., one transport operator, normally linked to the PTA. As consequence of these deregulations, lack of integration has become crucial, especially how it affects the users in terms of an easy understandable and seamless view of PT and operational efficiency. Intermodality and a seamless journey are of great importance for the public transport travellers.

Rail system types that increase their market share in Europe are high-speed trains and regional train services with modern EMUs (Electric Multiple Units). Even Metro and LRT are expanding. Modern rail systems can be driverless. This is especially practical for metro and other fixed rail systems. But now technology for driverless operation without conventional tracks is emerging with a first driverless minibus line in Rotterdam. Even older technologies like cable-propelled transport may get a new application in driverless minimetros.

Right of way for public transport is of utmost importance for public transport efficiency and attractiveness. Various forms of bus lanes, busways and dedicated infrastructure with and without rail can increase average speeds and regularity. Public transport nodes can made efficient and customer friendly by good design and by using modern traffic management and information systems. Close integration of rail and bus connections enhances the ease of use.

The regional and urban structures are to some extent integrated with the needs of efficient train and public transport system designs. This is better in Europe than in, for example, the USA, where planners start making “TOD”. Rail or BRT systems are today integrated with development of the cities. For example, cities, or parts of them, can be built in public transport corridors like linear cities. This partly solves the dilemma of having many direct lines with lower density or fewer networked lines with higher density. Today a system with many lines may be hard to understand, but with IT system, it can be easier to communicate the system design and offer real-time travel information.

Regarding capacity, different technologies and different system designs have different maximum capacities. An ordinary bus line in the city may have a capacity of about 1000 passengers per hour, but a BRT line with dedicated infrastructure, as line 34 in Istanbul’s BusMetro has a practical capacity of almost 20,000 passengers per hour and direction. This is similar to many European metro systems. Rail

systems also have a great span, varying from about 2000–50,000 passengers per direction and hour in a corridor. The local and cultural demand on comfort and privacy affects how many people should be transported together in a public transport car, and this also affects the practical capacity.

The infrastructure design and standard not only affect the capacity but also the speed for various modes. Regular buses and trams have low speeds, while Metro, BRT and LRT should reach medium speeds. In relative scales, regional express trains and interregional high-speed trains have competitive door-to-door travel times with car and air.

Public transport should also be reliable and punctual, and departures should be regular. Besides good infrastructure design, passive and active traffic management systems are used both for rail and bus operations. Efficient traffic signal priority is crucial but also systems for efficient use of rail and bus terminals and stops. For local rail and bus traffic, there is a variety of active traffic control and management systems. Information can, for example, be presented to the driver continuously, showing how his/her vehicle is positioned in relation to the vehicles is front of and behind. This is especially important for high-frequency trunk services as Metro, LRT, BRT and BHLS. For bus operations in corridors, the trunk-feeder and the convey technique both put demands on active traffic control for coordination and regular operation. Shuttles and a variety of route types, for example peak services that take people to work on routes that operate only a few times in the morning and afternoon, complement the trunk corridors and routes. The magnitude of differentiation is always a trade-off between simplicity and customer demand adaption. Rail, BRT and BHLS are examples of the former, while demand-responsive services represent the other end of the scale. In between, there are scheduled local minibus services and special services for old or impaired travellers. There are also various forms of DAR and bus-on-demand.

There exist so-called Paratransit services of various forms. In developing countries, Paratransit is operated by smaller or bigger private companies operating Jitneys on semi-fixed routes, while in USA and Europe, some Paratransit systems are heavily assisted by an ITS for transport management and control. There are also vanpools, car sharing systems, taxi and corporate services.

In rail operation, computer-controlled safety systems are everyday technology on most lines, and advanced traffic management systems have also been in use for long. The next step is to integrate national systems into a European system ERTMS. Full ATO are not yet on the agenda for long-distance rail, but already in operation for metros and other local rail systems. Higher operational quality is expected. It then also becomes economically feasible to operate off-peak periods with higher frequency.

The idea of PRT is to offer as many as possible “car quality” within a public transport system, such as no waiting times and journeys without transfers. PRT offers automated taxi service over a network of dedicated rights of way. Small “podcars” can run on elevated guide-ways, in tunnels or at grade between fences. In the near future, we will probably also see route-fixed and not route-fixed on-demand services with automated small-bus-like vehicles. These technologies are very dependent on advanced ITS including a high level of traffic safety control.

References

EBSF Objectives (2012) EBSF Homepage. www.ebsf.eu/
Jansson K (1997) Welfare aspects of the organization of passenger transport. Int J Transp Econ 24:11–33
Kottenhoff, K., (1999) Evaluation of Passenger Train Concepts, Div. of Traffic and Transport Planning, Royal Institute of Technology, KTH, Dissertation ISRN KTH/IP/FR-99/48-SE
Nash CA (1978) Management objectives, fares and service levels in bus transport. J Transp Econ Policy 12:70–85
Stahl A (1998) Service routes or low floor buses? In: Proceedings of 8th international conference on transport and mobility for elderly and disabled people, Perth, Australia
SL (2008) RiPlan 2008. Handbook for public transport planning in Stockholm, Stockholm public transport administration, Sweden
Lag (2010:1065) om kollektivtrafik, Svensk författningssamling (2010)
TCRP (2003), Transit Capacity and quality of service manual, TCRP (part 4)
Vuchic, VR (2007), Urban Transit Systems and Technology, ISBN: 978-0-471-75823-5, John Wiley & sons, March 2007

Chapter 3
Public Transport in the Era of ITS: ITS Technologies for Public Transport

Andrés Monzón, Sara Hernandez, Andrés García Martínez, Ioannis Kaparias and Francesco Viti

Mobility started to grow continuously from mid-nineteenth century thanks mainly to the development of rail transport services, and this trend increased even more rapidly with the growing fleet of cars and buses. The potential of those mechanised transport means made us more mobile. Getting around from place to place is essential to human engagement and endeavour.

As shown in Chap. 1, urban mobility has increased not only in the number of daily trips per person but also in distances travelled. According to the Eurostat Panorama of Transport, the average daily distance travelled during 2006 in land transport means in Europe was 31 km per person. Car trips account for 26 km, which is therefore the largest share of the daily mobility. The increase of problems associated with the urban mobility such as the steady growth in car ownership, increasing congestion and rising energy consumption has led to the system to move away from sustainability targets. The response to those major challenges cannot be limited to traditional measures, and the innovation plays a major role in finding appropriate solutions. New tools and solutions are currently conceived to achieve sustainable mobility in an efficient way.

A. Monzón (✉) · S. Hernandez · A.G. Martínez
Transport Research Centre (TRANSyT-UPM), Universidad Politécnica de Madrid, ETSI Caminos, Canales y Puertos, Profesor Aranguren s/n, 28040, Madrid, Spain
e-mail: andres.monzon@upm.es

S. Hernandez
e-mail: sara.hernandez@upm.es

A.G. Martínez
e-mail: andres.garcia@upm.es

I. Kaparias
City University London, Northampton Square, EC1V 0HB, London, UK
e-mail: kaparias@city.ac.uk

F. Viti
University of Luxembourg, 6 rue Coudenhove-Kalergi, 1359, Luxembourg, Luxembourg
e-mail: francesco.viti@uni.lu

G. Gentile and K. Noekel (eds.), *Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems*, Springer Tracts on Transportation and Traffic 10, DOI 10.1007/978-3-319-25082-3_3

Beyond the aspects related to the forms of public transport (vehicle technology, infrastructure, network of lines and level of service), described in Chap. 2, one key driver to improve public transport (PT) performance, and its competitiveness to less sustainable modes of transport, is exploiting the support of information and communication technology (ICT). According to the International Union of public transport UITP, ICT is crucial for the study, design, development, implementation, support or management of computer-based information systems, particularly software applications and computer hardware. ICT deals with the use of electronic computers and computer software to convert, store, protect, process, transmit and retrieve information data.

Taking into account the technical and technological evolution of the new ICT tools, it is possible to provide real-time and accurate information to travellers and to better monitor and manage the public transport system. The application of ICT for transportation systems is usually referred to as intelligent transportation systems (ITSs). Several ITS solutions are currently being deployed to improve PT operations and to offer the PT users a more effective information of the (real-time) status of the scheduled services. Moreover, ITS has an important role for integrated ticketing as well as for enhancing the use of public transport within a multimodal journey.

According to the above potential contributions of ITS to PT systems, this chapter provides an overview of the current ITS solutions in public transport in four main aspects. We focus first on the ITS technologies for improving PT operations. These cover aspects related to individual PT vehicle trajectories, to the line or general service regularity and delays recovery solutions. Then, we explore the current technologies adopted for optimising the traffic infrastructure to the PT services, and in particular to the prioritisation of PT within the supply system. A third aspect covered in this chapter is the introduction of smart ticketing systems, which allow faster and easier transactions, and simultaneously enables the collection of useful information on the transport demand. A fourth aspect is then focusing on the use of ITS for information purposes, i.e., how sensors and data collected from PT services as well as from the smart ticketing are used to dispatch useful information and guidance to the travellers. These aspects are finally linked to the current international deployment of ITS and are consolidated through a survey of the European ITS market.

Since the development of ITS is changing rapidly, the examples presented in this chapter are assumed to present the state of the art at the time of writing. It is not possible to predict how technological solutions and standards will evolve in the future. The main goal is to highlight the various evidences of how ITS has an impact on PT operations and utilisation, and to introduce the main aspects that are later needed to understand how to incorporate ITS in the models described in the next chapters. Hence, the chapter concludes with an overview of the variables used in transit assignment, which will be likely affected by the various ITS solutions.

This chapter concludes the first part of this book, which had the scope of providing the basic knowledge for understanding the functional aspects and characteristics of public transport services, and therefore providing the basic building blocks for formulating and developing transit assignment frameworks and models.

3.1 ITS Solutions for Fleet Management

Public transport fleet management has the objective of improving the efficiency, reliability and the environmental impact of PT systems. ITS brings several advantages to PT fleet management through a number of ICT enablers, which allow, for example:

- Real-time tracking, location, monitoring and visualisation of PT vehicles;
- collect data for analysis of performance and for planning purposes;
- high-quality real-time passenger information services both on-board and off-board (via, e.g., portable devices and information panels);
- dynamic control and advisory systems through on-board communication;
- improved punctuality of bus/tram services through coordinating lines and transfers;
- transit Signal Priority (TSP) at traffic lights; and
- equipment diagnostics and maintenance planning and scheduling.

The best way to monitor and operate bus or train fleets using ITS technologies is through a centralised control centre, which is the current state of practice in some European cities. The system operates based on a two-way communication protocol (Fig. 3.1).

The vehicle provides real-time information to the control centre (via automatic vehicle location—AVL-systems), which in turn produces guidance instructions to each bus, either through the driver display or by information panels along the line. There are different electronic data interchange (EDI) protocols to support this continuous communication between the control centre and each vehicle in

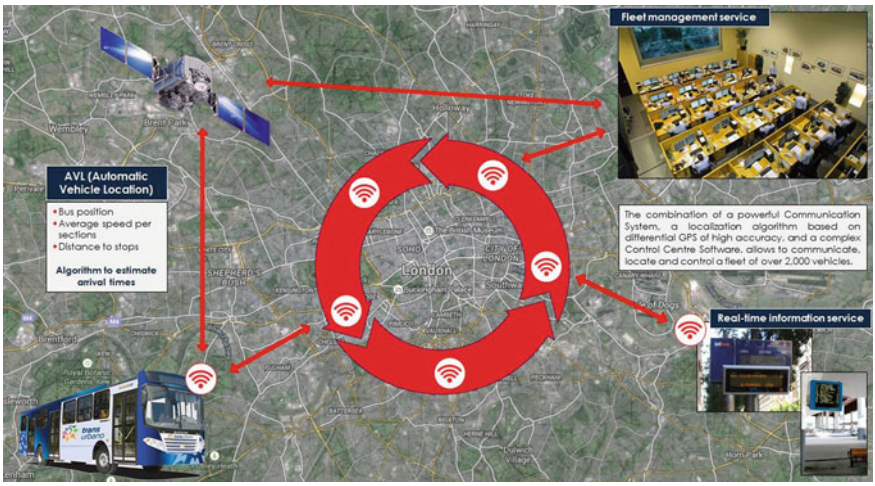

Fig. 3.1 Communication network for public transport fleet management

operation. This communication allows to maintain headways between vehicles and to inform of any incident in the service. The consequence is that drivers are not the only responsible of providing a good and reliable service because they receive the support from the control centre. Even more, any emergency or accident could be solved rapidly or quick actions can be taken to overcome road-obstacles in the vehicle itinerary.

The information collected by these tracking systems could also be used to deliver real-time information to the travellers while at the stops or stations, such as waiting time for the next services and reports on incidents. Some experiences based on the EBSF-FP7 project show that users perceived a clear improvement of service quality when bus stops are equipped with real-time information displays.

Nowadays, large cities already have this type of technology deployed in their PT systems. Smaller cities in size are now investing on fleet management systems, once their positive benefits have been verified. In this regard, an example of fleet management can be found in Craiova (Romania), where a fleet global positioning system (GPS) and a driving efficiency monitoring system were introduced. Other cities, as San Sebastian (Spain), Lodz (Poland) and Forlì-Cesena (Italy), have implemented a fleet monitoring system. The advantages of its implementation result on better service planning, improvement in security and the possibility to monitor consumption in real time, with the installation of additional devices. All these examples are explained in depth within the next sections.

3.1.1 Infomobility Tools for Sustainable Fleet Management (Craiova, Romania)

Craiova is the 6th most populated city in Romania, with approximately 270,000 inhabitants (as of 2011). Currently, the public transport system consists of 3 trolley tram lines and 17 bus lines. It is operated by the Regia Autonomă de Transport Craiova (RAT Craiova), a corporation run by the City Hall. One ticket is around 0.5 €, and the total fleet consists of 342 buses and 49 trams serving the city.

Craiova is also a major railway centre and is connected to all other major Romanian cities, as well as local destinations, through the national Căile Ferate Române network. Craiova is served as well by Craiova Airport, which has recently been modernised.

The EU's quantitative objectives for 2020 include reducing greenhouse gas emissions by 20 % of 1990 levels, reducing energy consumption by 20 % of the projected 2020 levels and increasing the share of renewable sources of energy to 20 % of total energy generation. In order to comply with these objectives, the municipality of Craiova developed and implemented a plan for sustainable development. One of the measures initiated in this area is the introduction of infomobility tools for fleet management (Fig. 3.2). The main objectives of this initiative are to

Fig. 3.2 Ticket counter in bus fleet in Craiova

increase residents' trust in public transport by providing a reliable, predictable, comfortable and safe service, to reduce traffic levels by encouraging people to use public transport instead of private cars and, simultaneously, to reduce the levels of pollution and fuel consumption. One of the measures taken to achieve these objectives is the implementation of a complex management system within the public transport company. This system utilises global positioning system (GPS)-based tracking components in order to increase the efficiency of public transport and to optimise energy consumption.

The basic function of the system is vehicle tracking. Besides locating the vehicles, the GPS transmits to the software application information about speed, direction of travel, etc. The information gathered about the positioning of the vehicles can also be used with the information panels installed in a number of bus stops to make public transport more predictable and attractive, and also for better route management during rush hours.

With the aim of further reducing the energy consumption, the public transport operator RAT rolled out a system that monitored the energy performance of drivers. It was designed to complement a modest capital investment in chopper technology that had been made on the braking systems of nine trams.

The system includes equipment to acquire, store and analyse data with the hope of saving on energy consumption and freeing up funds to improve travellers' comfort. The system allowed a breakdown of energy use on different sections of track via monitoring equipment connected to different power supply stations. This gave a detailed picture of not only drivers' performance, but also each individual's energy profile along different parts of the network.

Together with the new chopper technology installed on the nine trams, the monitoring measure was shown to produce energy savings of up to 40 %. These results clearly show that expanding the measures to cover the rest of the city fleet would be much cheaper than buying new trams.

3.1.2 Monitoring and Planning of Public Transport Systems (San Sebastian, Spain)

San Sebastian is a city situated in the north of Spain, on the Bay of Vizcaya coast and 20 km from the French border. The city is the capital of the region of Gipuzkoa, in the Basque region. The municipality population is about 186,500 inhabitants (as of 2013).

The bus is the primary mode of public transportation system of the city. The service offers 40 lines, plus a special service to the Igara industrial site. In addition to the above, the public transportation system offers four lines of minibuses, a taxi bus service to the districts that conventional buses cannot serve and 10 night lines running all Fridays and Saturdays. The rail service connects different parts of the city with many towns in Gipuzkoa and other important Spanish cities, and including a TGV service towards Paris.

Within the CIVITAS Archimedes project, the Tramway Company of San Sebastian (CTSS-DBUS) has defined and implemented a new expert planning system for the bus fleet and a fleet monitoring system. It is installed on a server and can be used from any computer via online. In this system, it is necessary to calculate the buses' schedules and the drivers' timetables. Before this, it is required to input all the trips of the buses offered to the travellers on every line. These data can be introduced easily in the system using a special input application. Another set of inputs are the labour restrictions and the bus network: lines, line routes, bus stops and points for driver replacement. Once all the inputs are filled in, the system calculates the best solution for the drivers' timetable, optimising the total number of shifts and service hours. Spreading the solution of the different day types for the whole year, the system gives the schedule for every driver about the working shifts for every day.

An important element of quality is the feeling of security of the bus passengers. For this reason, CTSS installed 22 security cameras in the buses. The security camera system on-board of the buses consists of 4 cameras located in different strategic points. The 4 cameras are connected to an advanced on-board computer that administrates and manages and records all the videos. With the security camera system, CTSS-DBUS has improved the security and the physical integrity of the drivers, travellers and material equipment of the public transport.

At the same time, as implementing the security camera system in the CTSS-DBUS fleet, a HSDPA-3G communication system between the buses and the control centre was also implemented. The security camera system has been a solution to reduce vandalism problems (Fig. 3.3). In the first year of operation, CTSS-DBUS detected that vandalism was reduced significantly in the buses with the security camera system, and several accidents were solved.

Fig. 3.3 Security camera system in San Sebastian

3.1.3 CCTV Monitoring System on Public Transport for Security Purposes (Lodz, Poland)

Lodz is the third-largest city in Poland. It is located in the central part of the country, and it has a population of 742,387 (as of 2009). Moving around Lodz with public transport is simple and straightforward, since an extensive, modern bus and tram network is spread throughout the city.

In order to improve the quality of the service in regard to security, Lodz public transport authority decided to install CCTV monitoring systems at both stops as well as on the vehicles. The aim of the project was to improve passengers' safety on city public transport in Lodz by CCTV monitoring system installation. It was a 1 M € project to be co-financed by the Regional Operational Programme (RPO) of the Lodzkie Voivodeship for the years 2007–2013. The project plan was to install 1264 CCTV cameras on 183 vehicles, and 20 other video cameras installed on selected stops and terminals. The system is operated from a dedicated monitoring centre. On some vehicles, there are extra cameras on the outside of the vehicle, either on the front or on the back.

The notification of any incident can be done by phone or by SMS. In the case of a short message, the description has to include identification number of the vehicle and the kind of problem. The observing centre can get the views from all the cameras in the system in real time (Fig. 3.4). The views can also be forwarded to city guards or police. The announcements are encouraging passengers to react as it is the quickest way to achieve a quick reaction from the support and security services. Additionally, there is a GPS unit on every vehicle, which allows to automatically locate the incident and, if needed, send the proper services to the exact location. As all the views from the CCTV cameras are recorded, they can be used as the proof material for the legal procedures.

The CCTV monitoring system has improved security on buses and trams of Lodz. In the first 3 months of functioning, no dangerous incident was recorded.

Fig. 3.4 CCTV monitoring system in Lodz

3.1.4 Consumption Monitoring and Ecodriving Training (Forlì–Cesena, Italy)

Forlì–Cesena is a province of the region of Emilia-Romagna, Italy. Its capital is the city of Forlì. The region has 30 municipalities, and it has an area of 2377 km^2 and a total population of 358,525 inhabitants (as of 2001). In order, cities with more population are Forlì, Cesena, Cesenatico and Savignano sul Rubicone. The Azienda Trasporti Regionale (ATR) public transport network connects Forlì, Cesena and the other main towns by bus. The intertown public transport system in Romagna is over 1600 km long and is divided into 4 routes through the valleys (of which 3 are integrated in the suburban valley routes), 4 routes across the plain, 7 secondary routes in the mountains and 5 routes across the plain.

In this scenario, a modern, environmentally friendly and alternative fleet does not directly mean lower energy consumption and CO_2 emissions. Vehicle procurement, vehicle monitoring and especially proper driving styles are the key elements to achieve real benefits for the environment. Being conscious of the high green impact of combining energy consumption monitoring with improved driving performances is a fundamental rule for the automatic vehicle monitoring (AVM) managers. It also creates the best work environment for employees and their active involvement in greening services.

AVM has implemented a wide awareness campaign on ecodriving issues among its drivers. All drivers received a tailored code of practice with tips and tricks for ecodriving a bus. The FLEAT pilot action was also implemented by an ad hoc test on a specific route. The redesigned Line 6 of the urban public transport service in the Municipality of Cesena is constantly monitored in its energy and environmental performances.

Services are provided with a fleet of 8 buses. Line 6 was monitored from December 2008 to August 2009 for those aspects related to fuel consumption, service interventions, anomalies, driver shifts, etc. Drivers have been informed every week about their actual fuel savings by a wall chart in their meeting and restroom.

After the FLEAT training and awareness campaign, the average aggregated CNG consumption of the line was 3.4 % lower compared to the pretest performances. By using a proper monitoring scheme and enhancing the ecodriving attitude of drivers at Line 6, about 4000 kg of CNG could be saved per year.

3.2 Integrated Management of Traffic and Public Transport Prioritisation

Buses usually use the same road/street infrastructures with cars and duty vehicles, or partly interact at junctions. The coordination of both ITSs for controlling bus services and traffic is very important (Fig. 3.5). It requires a common platform where traffic and public transport control centres are linked and coordinated. The coordination of control centres could produce a number of benefits, such as priority to buses in traffic lights and intersections, green waves for bus lanes and corridors, anticipating information on congestion or incidents in the bus line to allow changing itineraries or rescheduling.

One of the main outcomes of this kind of coordination is to facilitate reliability in bus services, which is the backbone of the quality of public transport and its acceptability. The coordination is also very important for emergencies, accidents and any kind of issue.

3.2.1 Transit Signal Priority

Traffic controls are often considering some form of *TSP*, i.e., additional responsive control policies that aim to improve the efficiency of PT operations at signals.

Fig. 3.5 Dedicated bus lane and different types of signalisation for giving priority

TSP improves PT operations via temporary traffic signal timing adjustments. Active TSP is one of the most efficient and cost-effective measures to improve the efficiency of PT operations. By contrast, lack of appropriate TSP leads to unreliable bus arrival time prediction displayed at bus stops.

Generation and processing of priority is taken care by two logical processes called priority request generator (PRG) and priority request server (PRS). The former determines the necessity for generating a request and communicates the request to the PRS. The PRS selects and sends requests to signal controller for implementation. PRS contains the priority control algorithm that specifies the level of priority that a PT vehicle can receive depending on various criteria. Signal control includes preferential adjustments such as early green, green extension and phase rotation/insertion. The first two are the most commonly used.

TSP consists of three components, a detection system, a signal control system and a communication system that links the two components. The latest (third) generation of TSP is based on automated vehicle location (AVL) as detection systems. This allows developing conditional (differential) active priority, in which the system verifies if the approaching vehicle meets the criteria for granting the priority. These criteria depend on the PT operating policy (typically schedule- or headway-based). In the former, punctuality is used as a criterion, while in the latter, it is regularity. In the future, TSP will include vehicle occupancy criterion—possible thanks to the emerging automated passenger counting (APC) systems. The introduction of AVL allowed to switch from conventional point-based detection (e.g., AVI loops) to zone-based methods. The latest zone-based technique uses infrastructure-less GPS-based virtual detectors, which allows placing and reviewing priority calls at any time, while the vehicle approaches the intersection.

CV technology brings the opportunity to increase efficiency and reduce negative side-impacts of TSP. Therefore, US Department of Transportation listed DSRC-based TSP as one of the high-priority mobility applications

The advances in the communication component of TSP are driven by the fact that currently used communication technologies are restricted to the transmission of small messages. Recent developments based on the IEEE 802.11 standard allow to bypass this limitation and to enable continuous exchange of update messages between vehicles and traffic signals.

These types of ITS solutions encourage the use of public transport, as they reduce travel times, the number of stops at traffic light and consequently the waiting times at stations or stops. In Toulouse (France), the average bus waiting times at traffic lights was reduced by 52 % with this kind of systems. Other cases, as the dedicated bus–tram lane in Warsaw (Poland), and the “greenway scheme” in Edinburgh (Scotland) have also shown that this type of measures yield big results, at a comparatively low cost. These examples are treated below, accompanied by the full integrated management of all public transport services in Madrid. This case is a good practice of how a transport network could be globally coordinated.

The control centre manages all modes of transport and public ones and also interacts with the traffic control systems of the city and the road network in the metropolitan area.

3.2.2 *Bus Priority System (Toulouse, France)*

Toulouse is the capital city of the department of Haute-Garonne, and of the region Midi-Pyrénées, in the south-west of France. The city has 1,250,251 inhabitants (as of 2011), and its metropolitan area is the fourth largest in the country, with 5381 km^2. The public transport network of Toulouse contains 77 urban bus lines, two underground lines and a tramway line to two neighbouring towns. In addition, a city-wide bicycle rental scheme called VélôToulouse was introduced in 2007, with bicycles available from automated stations for a daily, weekly, monthly or yearly subscription.

Toulouse aims to improve the service quality and to achieve a modal shift towards public transport. The bus priority system can be considered as one of the most effective solutions for improving the quality of the bus network service in the city and therefore for fostering this modal shift.

Within the CIVITAS MOBILIS project, two bus lines were equipped with a priority request system in order to assess the advantages and constraints of the system for both public transport and private car flows (Fig. 3.6). The radio priority system enables a permanent contact between the bus and the traffic junction. It calculates the time when the bus will arrive at the traffic junction and therefore can give green light to the arriving bus.

Surveys in a test phase showed that bus regularity and journey time improved and average bus waiting times at traffic lights was reduced by 52 % which is in average 9 s per traffic light equipped. This ITS solution is particularly adapted to the most congested lanes. Bus priority system permits to smartly deal with conflicts on several lanes on a same traffic junction.

Fig. 3.6 Bus equipped with a priority request system

3.2.3 *Revolutionised Public Transport with Dedicated Bus–Tram Lane (Warsaw, Poland)*

Warsaw is the capital and largest city of Poland. It is located on the Vistula River, 260 km from the Baltic Sea and 300 km from the Carpathian Mountains. Its population is estimated to be 1.7 million residents (as of 2011) within a greater metropolitan area of 2.67 million. The area of the city covers 516.9 km^2. Public transport in Warsaw includes buses, trams, metro, light rail, urban railway, regional rail and bicycle-sharing systems. Bus service covers the entire city, with approximately 170 routes totalling about 2603 km, and with some 1600 vehicles.

As regards the notice published by ELTIS (European Local Transport Information Service), the "Trasa W-Z" route ("W-Z", eng. east–west) with its total length of 6760 m, and intense public transport traffic (trams and buses) is one of the busiest central thoroughfares of Warsaw, connecting its west and east districts and assuring access to Warsaw underground from the right bank of Vistula. Until the route was reorganised in its middle section (about 3.2 km long), there were two lanes in each direction, accessible for all traffic participants. The tram track was not separated from the road traffic and thus shared a single lane. As a result, trams were stuck in traffic jams during peak hours.

In 2007, it was decided to separate the tram lane from the public one by means of a painted line. Two years later, the tram lane was made accessible also for buses (since 2011, 4 bus lines run the route), creating multimodal interchanges along the route (Fig. 3.7).

Before the new solution was brought in, there were on average 1700 cars per hour running on one direction. Since there was only one public lane left, the number of cars decreased by 40 %, whereas tram passenger numbers have grown by 250 %. Additionally, the transport authority has increased frequency on the four tram lines that operate along the route from 55 up to 60 departures per hour during peak hours.

The separation of the tram track has been met with severe criticism among the users of private cars. However, the city authorities have decided to keep the tram lane active as it has led to great improvement in the service for both the tram and buses that use the route.

Fig. 3.7 Public transport with dedicated bus–tram lane in Warsaw

The application of common bus/tram lane on the "W-Z" route, the first of this type in Warsaw, has shown how to improve the functioning of public transport significantly and how to encourage passengers to use trams and buses instead of their own cars, at low costs and in a short time. Such a solution has the potential to be replicated across other city routes, where tram tracks are embedded in the roadway, leading to loss of time for the passengers.

3.2.4 Bus Priority, the "Greenways Scheme" (Edinburgh, Scotland)

Edinburgh is the capital city of Scotland and its second most populated city. Its population is 487,500 (as of 2013). The city, despite having a metro system, has a *dende* bus network served by 110 bus lines. There are no integration forms with private transport, but Edinburgh has implemented a unitary fare system for different transport modes with an electronic ticketing system.

The problem of traffic congestion in Edinburgh has been acutely felt in the 1980s and 1990s, causing unacceptably high pedestrian accident rates, unpleasant conditions for pedestrians and cyclists, noise and pollution, pressure on parking space, higher business costs and a slower journey into work for many commuters. Increasing dependence on private cars has led to more congestion for public transport in a vicious circle which grinds down the public transport level of service for those without a car.

The "Greenways" scheme, along with other public transport initiatives, was introduced with the aim of restoring the balance of car use and public transport. The "Greenways" primary aim is to improve the reliability and speed of bus services. It aims to cut bus journey times by at least 10 % and thus to encourage more people to abandon the car in favour of the speedier and more reliable bus services. Key elements of the scheme include the establishment of new parking and loading bays, provision of more cycle lanes, more bus shelters and bus stop information or the installation of more pedestrian crossings and raised level crossings. In addition, an electronic detection system has been installed in the road surface at 25 traffic signals.

Before the implementation of these measures, buses were encountering fewer delays while using the extensive bus lanes, which are no longer blocked by illegal and dangerous parking. The eradication of delays should result in shorter journey times and a more reliable service. Improved facilities at bus stops, including new shelters and planned passenger information panels, are assisting both current and potential passengers.

Greenways are an example of a public transport service improvement, in combination with improved conditions for pedestrians and cyclists, at a comparatively low cost.

3.2.5 *Speed Advisory Based on Signal Phase and Time (SPaT) Information*

Therefore, there is a great potential for reducing the unnecessary stops at traffic lights, hence reducing fuel consumption and tailpipe emissions of PT vehicles by investing on Driver Advisory Systems (DAS).Current advisory systems guide the PT drivers towards more efficient and eco-friendly driving profiles and support the adherence to schedules by monitoring the progress along the route.

A new class of DAS systems is possible thanks to the access to signal phase and time (SPaT) real-time data. In SPaT-based DAS, drivers use signal control information to optimise their trip. Two SPaT-based DAS solutions are using Green Light Optimal Speed Advisory (GLOSA) and Green Light Optimal Dwell Time Advisory (GLODTA). The former facilitates the operations and the access through the signalised intersection by providing a vehicle with speed guidance, while the latter advices additional dwell time at the bus stop preceding the signals (near-side bus stop). For instance, the GLOSA system developed by Audi within Compass4D project receives SPaT data from city's central traffic server using cellular communications technology. Benefits of GLOSA are around 7 % in the average fuel consumption, 10 % in trip time, 90 % in the average stop time, and reductions in CO (80 %), NOx (35 %) and particulate matter emissions (18 %). GLOSA for buses was evaluated in the COSMO project, where it was shown that as soon as bus drivers follow speed recommendations in at least 50 % of signals, the average speed is increased, while fuel consumption and CO_2 emissions are reduced up to 20 %. The Dresden DVB tram north–south corridor is currently the most advanced implementation of a PT GLOSA.

The main objectives of GLOSA can be complemented with additional dwell time advisory (GLODTA), i.e., if a stop at the signals cannot be avoided uniquely by speed advisory, the potential waiting time at the signals is shifted to the near-side bus stop. The main advantages of the GLOSA and GLODTA are that while they improve bus performance with respect to traffic signals, unlike TSP, they are non-intrusive (i.e., do not modify signal timings). At the moment, these systems are being developed and implementations in practice are not available.

3.3 Intermodal Services Coordination and Interchange Facilities

An additional field of coordination is among different public transport modes and operators within the same mode, e.g., different bus operators in the same corridor or within the same city or region. These kinds of ITS developments are rather new, but they are providing rather good results. In many countries and cities, the travellers are no longer specific operator customers, but they are PT system customers.

They use flat-rate travelcards that allow them to use any means of transport within a geographical zone.

In this field, the most advanced level of service coordination happens in the interchanges or intermodal centres which serve as public transport hubs (Fig. 3.8). Therefore, ITS and ICT applications are a key element in the design of interchanges to assure good quality services.

By tracking the basic events taking place, it is possible to create a complete idea and collect up to date information of the activities in the interchange. These basic events (people walking, standing, lying, dropping, etc.) may lead, through the use of some intelligent knowledge-based systems, to the identification of more complex events such as people passing by or using the interchange for non-transport-related activities such as meeting points, how people use the waiting time (stay in the facility or exit it), how people might get lost at the terminal or need some assistance, either critical (medical, security, etc.) or non-critical (just needing some information). In addition, the professional users of the terminal may also take advantage of the technology for receiving complementary information of the status of current and future transport units (trains, buses, etc.) at the terminal (e.g., combining sensoring and advanced decision-making systems), instant information provided to drivers and operators (e.g., by means of the use of the communications embedded in Cooperative Systems, instant information of transport approaching and people flux for vendors, emergency situations for the security services, synchronisation of vehicles and units in time, etc.

Apart from the well-known vehicle-to-vehicle (V2V) and infrastructure-to-vehicle (I2V) communication systems, there are technology allowing new ways of cooperative models: Infrastructure-to-Users (I2U), where particular procedures need to be followed for providing local information of the terminal status and each independent user trip options and/or alternatives in case of incidents. This information can be easily accessible from smart terminals and phones through conventional communication links and data channels for providing advanced information-based ITS mobility services to travellers at the transport interchanges.

In order to facilitate intermodality between different modes of transport, numerous smartphone apps try to compile all public transport services aiming to

Fig. 3.8 Interchange station in which converge several public transport systems

supply valuable information to travellers. Other recent strategies are about to offer facilities to integrate bicycles (publics and privates) with all other modes of transport. Consistent with this type of ITS, the Municipality of Almada (Portugal) made available to the public a transport guide, with information of all modes of transport, as well as details of fare schemes, operating times and routes. In Sofia (Bulgaria), a Website was developed, with real-time information of public transport network, parking and cycling. There are also examples in public bicycles systems, as Call-a-Bike in Germany, in which bicycles can be integrated in a trip chain with public transport.

3.3.1 Integrated Public Transport Guide (Almada, Portugal)

Almada is located to the south of the city of Lisbon, bordering the Tagus River and its estuary, which lies between both cities. The city covers an area of 70.2 km^2, and its municipal population is 164,844 inhabitants (as of 2012). The existing public transport network in Almada consists of buses, trains, boats and a new tram line completely finalised on 2008. This network was slightly fragmented with gaps between certain origins and destinations. The introduction of the new tram in Almada significantly enhanced the public transport network, connecting existing bus, train and boat routes. To encourage and make it easier for people to use the new network, a detailed guide was developed containing the new routes, multi-modal links, timetables and fare information.

The Municipality of Almada and all public transport stakeholders were all jointly involved in the development of the new public transport guide, from its conception, design to final production. This included the main public transport operators, as well as other smaller operators.

The involvement of all public transport stakeholders allowed a coherent overview of all available transport modes, details of various fare schemes, operating times and routes to be included. It also included example routes, for various trip purposes, and contained travel times and different mode options for these trips.

The underlying philosophy for the guide's development was to produce a guide that was sufficiently detailed to provide users with the correct amount and most useful information to allow them to accurately plan their journeys, although, in a manageable form so as not to overburden them with information overload. An online version of the guide was also produced, containing the same practical information as well as an online route calculator (time, cost, etc.) and an interactive overview of pedestrian routes from the individual tram stations to local places of interest.

3.3.2 The Urban Mobility Website: Information About Public Transport on Site (Sofia, Bulgaria)

Sofia is the capital and largest city of Bulgaria. It occupies a strategic position at the centre of the Balkan Peninsula. The city has a population of around 1.25 million people (as of 2012). Public transport in Sofia is well-developed with bus (2380 km) network, metro (31 km of track), tram (308 km) network and trolleybus (193 km network). Despite the extensive offer of public transportation, car ownership has grown by 50 % since 2002.

Public Transport used to be the norm some time ago, but now many people use their cars, even in those cases where it may not be the most appropriate—especially during the rush hour. The urban structure of Sofia was not planned to deal with such a high level of traffic, and therefore, some solution had to be found to make sustainable travel planning easier.

The Sofia Urban Mobility Centre's Website was created exactly for this purpose. Inside one integral and easy-to-use interface, it provides real-time information relating to public transport, parking and cycling. This grouping of services also reflects the organisational responsibilities of the Sofia Urban Mobility Centre itself.

The use of public transport is made easier by an online timetable. Users can either search by route or station, or can use the interactive map, in which they can enter the start and finish points. All these are accompanied by detailed information about service schedules.

These interactive services are accompanied by lots of additional information, and Website users are encouraged to contribute or interact through the online discussion forums, or by telephone. The Website is also optimised for mobile devices, and in this way, all the travel information can be easily access through smart phones.

3.3.3 Call-a-Bike: Public Bicycles in Germany

Call-a-bike is a commercial public bicycle service that is offered by DB Rent, which is a subsidiary company of Deutsche Bahn (DB, German Rail). The service started in October 2001 in Munich. It has expanded to other German cities and is now available in Berlin, Cologne and Frankfurt. 4200 specially designed silver-red bicycles are available for rent in these cities from spring to fall (Fig. 3.9).

The scheme is designed for one-way trips. The bicycles are not bound to a rack but can be left at the nearest crossing in a defined core area, as they have a lock mechanism installed at the bicycles themselves. Therefore, they can be integrated in a trip chain with long distance rail or regional and urban public transport.

To obtain access to the Call-a-Bike service, users have to register once and need to provide their credit card information or give a direct debit authorisation. After

Fig. 3.9 Call-a-Bike public bicycle system in Germany

registration, the public bicycles can be unlocked by using a code that the user receives via cell phone. Call-a-Bike uses an advanced technology for the checkout and returning process of the bicycles. Registered users call by mobile phone a number that is displayed on the bike. They receive a four digit code which is entered on a touch screen to release the lock, integrated in the bike. At the destination, the traveller leaves the bicycle at a crossroad, locked to a fixed object and submits a return receipt code that appears on the display by mobile to DB Rent. The user has to provide also information about the location where he or she leaves the bike. The utilisation fee is charged on the user's credit card or automatically withdrawn from his/her bank account. Call-a-Bike is a so-called smart bike, which enables to track the user of the bike, which reduces the risk of theft.

Currently, the Call-a-Bike service is not financially self-sustaining. However, it is not the goal of DB to make a profit of the service. It is rather aimed at a break-even and at the attraction of rail customers that use the service in a trip chain. Bike sharing is therefore designed as complementary (*last-mile*) service for longer-distance trips where the main mode is train or bus. For the same reason, a similar service was recently established in the Netherlands (the *OV-fiets*).

In 2004, Call-a-Bike had approximately 71,000 clients in Germany (+40 % users compared to 2003), and around 380,000 trips had been made with the bicycles (+19 %). Main users tend to be morning commuters who extend public transportation trips with a bicycle. Bike services tend to peak on sunny days and weekends.

DB Rent is still expecting an increase in the number of Call-a-Bike users. A further expansion of the scheme to other large German cities is possible. The user group of multimodal travellers who are willing to combine different modes in a trip chain is growing.

3.3.4 Multimodal Travel Planners

Multimodal travel planners are front-end–back-end computer systems which provide travellers the best itinerary, according to several parameters characterising and affecting an intermodal passenger transport journey. These systems usually supply timetable, routing and other travel information, as the best mode choice or traffic incidents. Several data are required, as the system needs to know about public transport services, transportation networks and private transportation.

Multimodal travel planners provide better modal integration and more sustainability by enabling travellers to select the most suitable combination of transport modes for the journey and could lead to an increase use of public transport, cycling or walking in urban environment. In case of congestion events, travellers can receive accurate information of alternative routes, allowing better use of existing transport infrastructure.

Currently, the European Union calls for better multimodal travel planning solutions. Simultaneously to the 10th European ITS Congress in Helsinki, the European Commission released the analysis on the state of the art of multimodal travel planners and plans for the way forward. The Commission Staff Working Document towards a roadmap for delivering EU-wide multimodal travel information, planning and ticketing services identifies the major challenges to overcome to create a framework supporting more comprehensive services. It also presents the advantages of multimodal travel information and planning services, before the document suggests an integrated approach in the coming years. The intention is to establish a framework for EU-wide multimodal transport information and respond to the need for further integration of the different modes of transport to make mobility more efficient and user-friendly.

Examples of multimodal travel planners currently being used in Europe are “Resrobot” in Sweden, “SITkol” in Poland, “DELFI” in Germany, “SCOTTY” in Austria, “TransPOR” in Portugal, “INFOTEC” in Belgium, “BilRejseplanen” in Denmark or “Rutebok” in Norway.

3.4 Ticketing

One of the most deterrent factors for the use of public transport is the need of different tickets for different transport means. Integrated ticketing is therefore a key issue for the use of public transport and acceptance of intermodality. It leads to the overall increase of the public transport system usage and better use of intermodality and interchange points. Electronic payment results are extremely important for making the system easy to use and to foster intermodality (Fig. 3.10). ICTs new Near field communication (NFC) protocols provide contactless payment systems for all transport modes incorporated to the platform and even to other services in the city, such as parking and citizens’ services.

Fig. 3.10 Payment process by smart card

These intelligent payment cards are also very convenient for designing different price schemes: special groups (elderly, children, etc.), different times of the day (peak, off-peak) or intermodal services avoiding the penalty of transferring to other mode.

They provide useful information for transport managers and planners, in such flexible way that could be specific for one line, or stop, or type of persons, or for the whole network. Thus, it is possible to avoid survey costs or counting of number of passengers using time-consuming and labour-intensive data collection methods.

At the present time, transport authorities are immersed in a process of renewal of existing ticket validation systems, in which they opt for new contactless payment systems. In this direction, Norwich (UK), Brescia (Italy) and Lisbon (Portugal) have strived to upgrade the e-ticketing system to a new modern one, which has better positive impacts on travellers. Barcelona (Spain), on the other hand, is an exemplary case in the fare integration of several titles and different modes of transport. These advances in fare integration and new ticket validation systems imply improvement of the public transportation system.

3.4.1 On-Street Ticket Vending Machines (Norwich, UK)

Norwich is the regional administrative centre and county town of Norfolk. The built-up area of Norwich had a population of 213,166 (as of 2011). This area extends beyond the city boundary, with extensive suburban areas on the western, northern and eastern sides.

Bus services in the city area operate via the radial road network to and from the city centre. For orbital trips, it is necessary to change services in the city centre, although a few services do provide through-city links. The rail network, even

Fig. 3.11 Ticket vending machine in Norwich

though being considered a mode for long journeys, it can be used for local trips, as there are some stations in Norwich.

Traditionally, the majority of passengers using bus services in Norwich and many other UK cities have purchased their ticket from the driver on boarding. This project is innovative in delivering a comprehensive solution for roadside bus ticket sales suitable for the deregulated environment applicable in the UK outside London. Roadside ticket vending machines (TVMs) have networked communications links to remote monitoring and revenue management systems and enable customers to choose between different operators and tariffs when buying a ticket. Sixteen TVMs are installed at different locations including Norwich Bus Station, Norwich Railway Station and Castle Meadow. Servers for remote monitoring and revenue management systems are located at County Hall (Fig. 3.11).

Initial findings indicate that dwell times at bus stops have decreased due to the contribution of the ticket machines. The machines have generally performed reliably, and the use of them has increased steadily since they were installed.

3.4.2 Development and Upgrade of the E-Ticketing System (Brescia, Italy)

Brescia is a city and municipality in the region of Lombardy, in northern Italy. It is situated near the Alps, with a population of around 194,000 inhabitants (as of 2014). The public transport system in the city includes metro and buses. There are bus routes through all of the main streets. The Brescia Metro is a rapid transit network that opened on 2 March 2013. The network comprises one line, 13.7 km long, with 17 stations between, of which 13 are underground.

Intermodality in the urban area of Brescia was only for suburban travel involving the train station and the main stations of the suburban bus stops, before the development of the measure. Fare integration between the various companies was possible only for students of the suburban area in possession of integrated passes.

The main goal of this project was the introduction of a new contactless card integrating various transport systems in terms of technology and fares. This was intended to encourage the use of collective systems such as local PT, bike sharing, car sharing and the future Metro, and thus enhance Park & Rides.

First of all, research and development activities were carried out in order to design the technical features that would characterise the new e-ticketing system. The second phase began after the new cards were purchased. This involved scheduling the validation test for the software and the analysis of its integration with NFC technology.

The integration of functional services was realised from a technical point of view in a "virtual" way. In order to manage the card and maintain the assets of the existing technology systems, thus limiting investments and development time, the original technologies of each system were implemented and improved.

3.4.3 The Viva Smart Card System (Lisbon, Portugal)

Lisbon is the capital and the largest city of Portugal. The city lies in the western Iberian Peninsula on the Atlantic Ocean and the River Tagus and has a population of 547,631 (as of 2011) within its administrative limits on a land area of 84.8 km^2. Lisbon's public transport network includes metro as its main artery, connecting the city centre with the upper and eastern districts, and reaching the suburbs. Bus, funicular and tram services have been supplied by the Companhia de Carris de Ferro de Lisboa (Carris), for over a century.

At the end of 2001, the *Metropolitano de Lisboa* installed a new ticketing and access control system to the city's metro network. The change was from an open access system to a closed one with control lines and access channels equipped with doors. The doors are commanded by the reading and validation of data stored in tickets. This new access system required a major change in the ticketing system, involving the introduction of magnetic tickets and the contactless card—"Lisboa Viva", which replaces the traditional pass and is intended mainly for regular public transport users in the Lisbon region (Fig. 3.12).

The Lisboa Viva card has an embedded chip and antenna which works by holding the card over a validator, located at entrances of a station. The validator reads and validates the data loaded in the chip and, provided the card is valid, enables access to the networks—presently the Metro and Carris networks. The procedure used for entering the system is the same used to leave the system. The card allows the loading of fares exclusive to each associated operator, multimodal fares and combined fares. The 7 Colinas card is also based on contactless technology, intended to be loaded with multimodal tickets for urban and suburban trips in the Carris and Metro networks.

Since the introduction of the Lisbon Smart Card, the gate access provides greater security and revenue protection, the ticketing is faster, and there is a better knowledge of origin-destination flows within the underground network. In addition,

Fig. 3.12 The Lisbon Smart Card

systematic data collection is possible regarding entrance points of the bus network, so as the Smart Card provides non-frequent users an integrated ticketing system between bus and metro.

3.4.4 The Use of Ticket Validation for Transit Planning Purposes (Barcelona, Spain)

Barcelona is the second largest city in Spain and the largest metropolis on the Mediterranean Sea, located on the coast between the rivers Llobregat and Besòs. The city has a population of 1.6 million (as of 2012) inhabitants within its administrative limits on a land area of 100 km^2. Beyond its urban area, the metropolitan region congregates 5 million inhabitants on a land area of 3240 km^2.

The public transport network includes metropolitan and urban buses, metro, local trains and a tram network (since 2004). There are different operators, which manage these services (Transports Metropolitans de Barcelona, Ferrocarrils de la Generalitat de Catalunya, Renfe and Tramvia Metropolità). Metropolitan Transport Authority (ATM) is an interadministration partnership between administrations with transport service rights, created to coordinate public transport within the Barcelona Metropolitan Region, with the support of the Government of Catalonia and local administrations. Some of its missions consist of coordinating the services offered by public and private operators, taking responsibility of the fare policy, finance the system by the various administrations and execution of infrastructure projects. In this sense, ATM performed in 2001 a fare and ticket integration among the entire metropolitan region. To make it possible, the region was divided into zones on the basis of geographical rings (6 concentric bands), as shown in Fig. 3.13. All transport tickets form part of the system, and the ATM is responsible for pricing. The price is based on the number of zones through which a passenger passes, and it is related with the trip frequency and length. The maximum number of zones to be paid is 6, and transfers are not penalised.

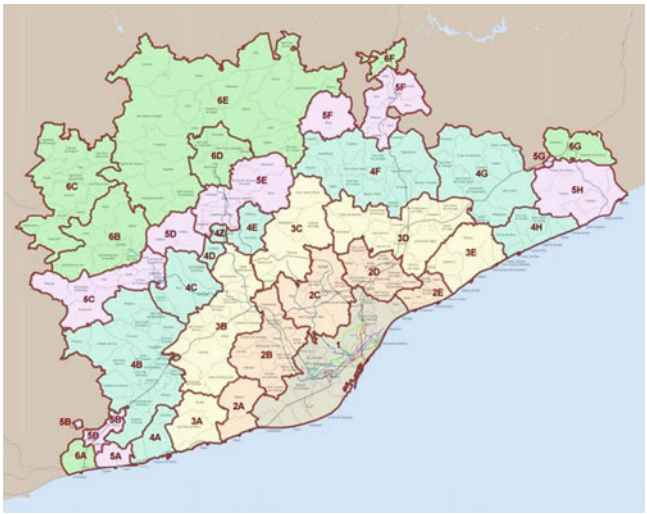

Fig. 3.13 Zoning of the Barcelona faring system

All this information about ticket validations must be treated to further compensate transport operators. Thus, a daily transmission of all validations made in the transport system is carried out in the integrated fare management system. The main problem occurs when a traveller takes a multimodal trip. To solve this issue and to reconstruct the different stages of each trip, a monitoring bit is activated. This is a code that identifies the card, and when this is validated, it leaves a "trail" in the different validating machines and allows to establish the distribution of revenue among the transport operators.

3.4.5 Using Ticketing Data for Improving Transit Planning and Scheduling Services

Commuters tend towards more flexible work scheduling, as telecommuting is more and more spread in working environments. Transit agencies should determine whether routes can benefit from weekday exceptions. The impact of exception schedules on ridership was assessed by the automated fare collection data, which revealed a lower ridership on Fridays (by 4.7 %) which allowed a 7.4 % reduction in vehicle hours operated. The processed data gives hourly ridership (boardings) as the annual 85 %-tile and the proportion of the deviation from average weekdays. The assessment of service change is based on service frequency (min and buses/h), operation (pass/bus, utilisation ratio, veh-hours changed) and crewing impacts (assigned operators per day).

Other observations also rely on automated data collection. The case of London Overground network investigates the passenger incidence behaviour with respect to published timetables. Such timetables could be modified in real time, and thus, the passenger incidence could be estimated as a relative frequency of passenger incidences over normalised incidence headway of services.

The impact of holding strategies on level of service and transit performance is analysed based on AVL data, which inform driver and may lead to a next time point where the bus may be held. Measures of effectiveness include average waiting and standing time per passenger, % of bunching, average holding time and delay per bus. Automatic vehicle location (AVL), automatic passenger counting (APC) and automatic fare collection systems are also used to assess transit system performance and reliability, estimate dwell time and estimate passenger waiting times as related to service reliability.

3.5 Real-Time Information Services

Most of the previous applications rely on the exploitation of ICT and ITS services for improving the planning, management and operation of PT services. ITS real-time data can be also used to improve the information offered to the service stakeholders, for instance:

- Information to travellers throughout the total journey, in the trip planning phase and during the trip especially at the interchange points (Fig. 3.14).
- Online integrated information given at the interchange points including incident information.
- Dialogue between information systems of various operators. For the traveller integrated information should appear on the screens in the vehicles, stops and terminals as well as available through mobile equipment (Fig. 3.15).
- Ticket purchasing systems, especially smartphone-based solutions.

Fig. 3.14 Real-time information panel at a bus stop

Fig. 3.15 Real-time smartphone app

- Emergency and daily incident information concerning all players involved, demanding immediate action from the responsible bodies, all operators and guidance and management of the passengers and other customers.

ITS technology allows users to better react to possible incidents or delays in the PT system. Generally, in large cities, bus stops or stations already have screens with real-time information. This type of ITS is spreading to other smaller cities, which contributes to improve the quality of public transportation system. In recent years, this technology tries to position itself even closer to the users through new formats. In this sense, the more requested format is via smartphones, followed by household computers, attending to the fact that this real-time information should be accessible at any time. The current trend shows that transport authorities make all existing data on public transport available to the general public, especially developers, so they can program their own applications (e.g., Moovit, described in Sect. 3.5.4).

Other authorities, however, argue for developing their particular applications (case of London, UK). In the first case, it is necessary to keep in mind that potential developers are equally users of the service, so it has the added value that they can contribute with new ideas and features.

3.5.1 Real-Time Countdown System (London, UK)

London is the capital city of England and the UK. It is the most populated city in the UK with a metropolitan area of over 13 million inhabitants (as of 2013). The city transportation system is famous for its red buses. Approximately 8500 of these iconic red buses carry more than 6 million passengers each weekday on a network serving all parts of Greater London. The service was previously served by a

fixed-time bus information system called Countdown. In October 2011, Transport for London (TfL) introduced its new Countdown service, providing real-time bus information for all 19,000 of London's bus stops. It is the largest real-time bus information system in the world. This information is available via the Internet, smartphones and by text message. In addition, 2500 new-generation bus stop displays are being installed replacing old signs. Via the internet on a computer or smartphone, the new system can provide not only live bus arrival times but service disruption information and links to London Underground service updates.

Bus stops can be searched for by route number, street name, postal code or via a map. Users can save their five most used stops. The smartphone mobile Web service is available to all mobile phones with the Internet connection without the need to download an app. To use the text message service, passengers text a bus stop code (displayed on each bus stop or available from the TfL Website) to TfL to receive live information for that stop.

New Countdown signs at bus stops comply with the latest disability guidelines. They use amber LED displays with a black background. Live information is transmitted using TfL's state of the art AVL, radio and on-bus passenger information display and announcement system known as iBus, which is installed on all its buses. This reduces operational costs and allows signs to be sited in places where the current system could not be used and makes Countdown predictions more accurate.

A new communication network to the Countdown signs has resulted in more reliable transfer of bus information, leading to greater accuracy and improved availability of information. Predictions are now 95 % accurate. Operational costs have been reduced, with communication costs now up to only 20 % of the previous level.

3.5.2 Real-Time Passenger Information at Bus Stops (Lille Métropole, France)

Lille is a city in the north of France, the fourth-largest metropolitan area in France after Paris, Lyon and Marseille. Lille is situated in French Flanders, near France's border with Belgium and has a population of 226,827 (as of 2009). The Lille Métropole has a mixed mode public transport system, which is considered one of the most modern in the whole of France. It comprises buses, trams and a driverless metro system, all of which are operated under the Transpole name. The metro system has two lines, with a total length of 45 km and 60 stations. The tram system consists of two interurban tram lines, connecting central Lille to the nearby communities of Roubaix and Tourcoing, and has 45 stops. 68 urban bus routes cover the metropolis, 8 of which reach into Belgium.

The municipality intended to install some screens to improve the real-time information in the public transport system (Hubacher 2012). The first information screens installed in 2003 ran on a battery with a lifetime of 6 months. The screens did not have a sound function, and the font size was quite small, making them inappropriate for use by people with visual and hearing impairments. It was also important to increase the number of screens at bus stops as part of the development of a high-quality bus service. The technology had to be improved, particularly in relation to the battery and the accessibility of the displays.

In 2008, a new call for tender was written with the help of Transpole, the Lille Métropole operator. The industrial company Serelec was the chosen candidate because the technology it had developed matched the goal of implementing more efficient screens. The new screens were installed in 2009 and 2010. The batteries of the new screens are charged at night through public lighting. They have an acoustic support, and the size of the font is bigger and more visible. The new screens are more efficient than the previous ones, and the supervisory control enables the daily operating of the screens to be monitored.

3.5.3 VAO, Traffic Information Austria

The mission of Traffic Information Austria (VAO) is to create information service for all of Austria with consistently high quality that covers all traffic developments (for cyclists, pedestrians, public transport, motor vehicles and Park&Ride).

By highlighting alternatives, the options available to switch to more environmentally friendly means of transport become attractive and greater awareness is ensured. According to the results of the research project ITSworks, the potential for shifting traffic from the car to more environmentally friendly modes ranges up to six percentage points.

VAO can be made available directly, but it can also be used as basis for the project partners' traffic information services. What results is a milestone in data quality and comprehensive information. But the administration, too, is given completely new options in active traffic control and management and in the provision of up-to-the-minute information relating to ongoing traffic developments.

Within VAO II, the project Traffic Information Austria will be further improved: additional data will be collected; detection of traffic data and real-time data will be optimised; and new mobility services (sharing concepts) will be integrated. Usability and performance of end-user services will also be optimised. VAO II cooperates closely with numerous projects such as GIP.at, GIP.gv.at, FCD-model region Salzburg or Testfeld-Telematik.

3.5.4 Two-Way ICT Communications Through Crowdsourcing Data Collection

New generation of personal travel planners is one step beyond real-time information systems. Both of them are usually based on crowdsourcing and open data. However, in contradistinction to the latter, new personal travel planners supply information bidirectionally. In other words, passengers do not only receive real-time information, as they interact with the system providing new data from the passenger side, making it possible to enhance the available information. In this sense, for example, the procedure to calculate the arrival time of a bus to a certain stop is not only based on its GPS coordinates, but also on other users that are travelling in the same route.

These systems are rather new, as they combine existing applications which provides real-time information in one direction (operator–passenger) and social media. They allow to choose the easiest or quickest route as desired by the user and supply information on schedules and last minute incidents. The new social function provides data to other users and operators as well. By this way, users can transmit anonymously their location and vehicle speed when they travel. It is also possible to participle by qualifying drivers, the cleaning of vehicles and the routes according to their experience in their daily commute.

Besides the above, another interesting characterising is that it is allowed to perform campaigns for data collection, obtaining valuable information. This function permits to evaluate as a whole the public transport system or know the opinion of users to specific events (incident management, evaluation of service in peak and off-peak hours.)

One of the most popular applications in this field is Moovit, which supports 41 countries and 400 cities. Another well-known example is OpenTripPlanner, which is also offering its service worldwide.

Moovit uses crowdsourced information to provide real-time transit information for passengers. Since the company began its slow roll-out in January 2012, Moovit's core goal is to enable users to find real-time data about their public transit, finding quicker routes and sussing out the peculiarities of cities' transport routes. The application has the following characteristics:

- Live map: The application gives the user the possibility to view nearby transit stops on a live map, to see lines that stop in the surroundings and check upcoming arrival times.
- Plan a trip: This feature allows to compare different trips from A to B anywhere in a city and select the most efficient and convenient route.
- Navigate the city: Moovit gives step-by-step directions guide to the final destination, including walking segments and transit transfers. This is combined with ride mode, which permits the user to get dynamic estimated time of arrival (ETA) to its destination.

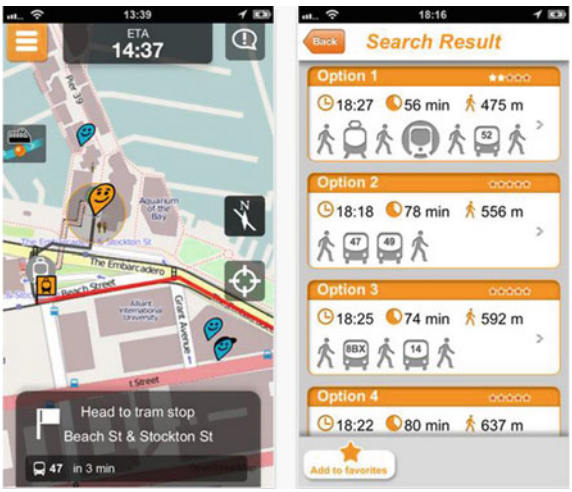

Fig. 3.16 Moovit interface

- Favourites: This characteristic customises usual locations (home, work, etc.) and lines for quick and easy access.
- Service alerts: Users can get personalised alerts and updates on service changes, route alterations or station closures.

Besides these features, users can interact with each other. As shown in Fig. 3.16, it is possible to qualify routes in terms of punctuality, travel time, driver's behaviour and many others. This two ways ICT communication is gaining ground over simple real-time information applications, as it offers more accurate real-time information, based on users' real-time experiences.

3.6 Development and Maturity Level of ITS in Europe

In previous sections, a descriptive review on ITS types and case studies across Europe was carried out. Building on these, the objective of this section is to investigate the degree of policy integration of ITS for public transport in Europe. Indeed, at a focus group held in 2010 by the FP7 CONDUITS project (2009–2011) and featuring representatives from 16 European cities, changing the modal split in favour of public transport was identified as a key priority of nearly all city authorities, with ITS playing an important role (Zavitsas et al. 2010). In fact, ITS were identified as offering potential solutions to many of the cities' problems, with the vast majority of the cities having implemented ITS technologies in terms of providing information to the public and facilitating public transport, alongside road traffic management, or planning to do so in the near future.

The present section, therefore, goes into more detail with regard to analysing the level of maturity of the available ITS applications in Europe and the extent to which

these are integrated in public transport policy making and decision making. This is achieved by means of two survey exercises: the first one has been carried out as part of the FP7 CONDUITS project, and its objective is to provide a broad overview of the state of deployment of public transport ITS in Europe using data collected from 33 cities directly with the help of a purpose-developed questionnaire; and the second one has been conducted as part of COST Action TU1004, where a subset of five cities are looked at in more detail.

3.6.1 Broad Overview of the State of Public Transport ITS Deployment in Europe

Within the FP7 CONDUITS project, a questionnaire aiming at collecting detailed data and feedback from cities on a series of best practices in order to create an extended database on public transport ITS policies and technologies implemented has been developed in 2010 (Zavitsas et al. 2010; Bell et al. 2012). The questionnaire covers several areas, such as general statistics of the transport systems, organisational structures, monitoring and forecasting, provision of information and demand management. It has been attempted to identify a group of about 50 key cities who would complete the questionnaire, and cities were chosen from thorough research in the literature and were based on the recommendations of experts, interviewed using the well-known Delphi Method (Linstone and Turrof 1975). The final result offers a diverse group, in which not only European metropolises are included (such as Paris, London, Rome, Istanbul and Athens), but also cities known for their virtuous attitude towards innovative transport systems (such as Trondheim, Karlsruhe and Turin), as well as medium-sized cities (Southampton and Stuttgart) and emerging realities (Kocaeli, Haifa, Funchal). The overall response comprises completed questionnaires from a total of 33 cities, which are shown on the map of Fig. 3.17.

The cities that have participated in the survey range from small towns to large metropolises (Table 3.1): towards the lower end, the sample includes the cities of Funchal and Trondheim, with populations of less than 200,000, while towards the upper end, Istanbul has over 12 million inhabitants. In terms of modal split, the average private transport share is slightly lower than 50 %, while the share of public transport ranges from 10 % in Trondheim to more than 40 % in Prague and Zurich. It can also be noted that the Turkish cities of Istanbul and Kocaeli have a large share of walking, though this is not the case for Ankara. No observable patterns can be identified for cycling, with high shares of cycling occurring in cities of varying sizes.

As can be further seen in Table 3.1, most cities of the sample (24 out of 33) have a concrete and concise 10–20 year strategic plan in place, while another seven state that they “may have one in the next five years”. By reviewing the strategic plans, it can be noticed that some of the main future concerns of cities regard safety,

Fig. 3.17 CONDUITS study participating cities (Bell et al. 2012)

Table 3.1 Key characteristics of the CONDUITS participating cities (Bell et al. 2012)

City	Metropolitan population	Modal split (%)				Strategic plan	ITS architecture
		Private trans.	Public trans.	Walk/cycle	Other		
Ankara	3,890,000	N/A				Yes	No
Athens	3,840,000	54	37	9	0	Yes	Yes
Barcelona	4,930,000	16	33	51	0	Yes	Yes
Berlin	5,970,000	38	27	35	0	Yes	Yes
Bologna	920,000	39	26	28	7	Yes	Yes
Brescia	190,000	N/A				Yes	Yes
Brussels	3,000,000	56	15	26	3	Yes	Yes
Bursa	1,820,000	N/A				Next 5 years	Yes
Edinburgh	470,000	55	20	21	4	No	Yes
Frankfurt	5,500,000	39	23	38	0	Next 5 years	Yes
Funchal	100,000	52	31	17	0	Next 5 years	Yes
Haifa	1,020,000	57	15	18	10	Next 5 years	Yes

(continued)

Table 3.1 (continued)

City	Metropolitan population	Modal split (%)				Strategic plan	ITS architecture
		Private trans.	Public trans.	Walk/cycle	Other		
Istanbul	12,600,000	24	14	49	13	Yes	Next 5 years
Karlsruhe	1,500,000	44	18	38	0	Yes	Next 5 years
Kayseri	900,000	N/A				Yes	No
Kocaeli	890,000	26	33	41	0	Next 5 years	Next 5 years
London	7,570,000	38	38	23	1	Yes	Yes
Milan	3,080,000	52	36	12	0	Yes	Yes
Munich	2,600,000	37	21	42	0	Yes	Yes
Paris	11,600,000	15	29	56	0	Next 5 years	Next 5 years
Prague	1,490,000	33	43	24	0	Yes	N/A
Rome	4,200,000	66	28	6	0	Yes	Yes
Sheffield	1,200,000	54	28	9	9	Yes	Yes
Southampton	230,000	N/A				No	Yes
Stockholm	1,980,000	46	28	26	0	Yes	Yes
Stuttgart	2,700,000	N/A				Next 5 years	Yes
Tel Aviv	3,150,000	51	23	19	7	Yes	Yes
The Hague	1,000,000	56	27	17	0	Yes	Next 5 years
Thessaloniki	800,000	72	22	2	4	Yes	Yes
Trondheim	170,000	58	11	31	0	Yes	Yes
Turin	N/A	56	15	28	1	Yes	Yes
Vienna	3,000,000	34	35	31	0	Yes	N/A
Zurich	1,080,000	36	48	16	0	Yes	Yes

sustainability, efficiency, pollution and reliability. Most cities' strategic plans analyse the objectives, though some focus only on the cities' targets. For example, Brussels has set a concrete target of reducing car traffic by 20 % in terms of vehicle-km, while The Hague has the broader objective of promoting sustainable transport modes. Additionally, several cities stress in their strategic plans the need for more efficient management through the application of ITS.

Many cities already have or are developing public-transport-oriented systems, in order to realise their long-term target that aims at a modal shift from private to public means. Looking at the infrastructure present in the sample (Table 3.2), the bus is clearly the most common public transport means, present in all 33 cities, though with differing network sizes in terms of total length and number of lines,

Table 3.2 Public transport and ITS infrastructure in the CONDUITS participating cities (Bell et al. 2012)

City	Public transport system			Features			
	Bus	Tram	Metro	Priority measures	Integration forms	Unitary fares	Electronic ticketing
Ankara	Yes	Yes	Yes	No	No	No	No
Athens	Yes	Yes	Yes	Yes	No	Yes	In 5 years
Barcelona	Yes	Yes	Yes	Yes	Yes	Yes	No
Berlin	Yes	Yes	Yes	Yes	No	Yes	In 5 years
Bologna	Yes	No	No	Yes	Yes	No	No
Brescia	Yes	No	No	Yes	Yes	Yes	No
Brussels	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Bursa	Yes	No	Yes	No	Yes	No	No
Edinburgh	Yes	No	No	Yes	No	Yes	Yes
Frankfurt	Yes	Yes	Yes	Yes	No	No	Yes
Funchal	Yes	No	No	Yes	No	No	No
Haifa	Yes	No	Yes	Yes	No	In 5 years	No
Istanbul	Yes	Yes	Yes	Yes	No	Yes	In 5 years
Karlsruhe	Yes	Yes	Yes	Yes	Yes	Yes	In 5 years
Kayseri	Yes	No	No	Yes	Yes	Yes	Yes
Kocaeli	Yes	Yes	No	Yes	Yes	Yes	In 5 years
London	Yes	Yes	Yes	Yes	Yes	No	Yes
Milan	Yes	Yes	Yes	Yes	Yes	No	Yes
Munich	Yes	Yes	Yes	Yes	No	Yes	No
Paris	Yes	Yes	Yes	Yes	No	No	No
Prague	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Rome	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sheffield	Yes	Yes	No	Yes	No	No	In 5 years
Southampton	Yes	No	No	Yes	No	No	No
Stockholm	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Stuttgart	Yes	Yes	Yes	Yes	No	No	Yes
Tel Aviv	Yes	In 5 years	No	Yes	Yes	No	No
The Hague	Yes	Yes	No	Yes	Yes	Yes	Yes
Thessaloniki	Yes	No	In 5 years	Yes	Yes	No	Yes
Trondheim	Yes	Yes	No	Yes	No	Yes	No
Turin	Yes	Yes	Yes	Yes	No	Yes	No
Vienna	Yes	Yes	Yes	Yes	No	Yes	Yes
Zurich	Yes	Yes	No	Yes	No	Yes	In 5 years

ranging from 150 km in The Hague and 175 km in Zurich to 9300 km in London, and from 18 lines in Brescia to 683 lines in London. It can also be observed that cities of similar size can have considerably different bus network lengths, depending on the presence of other public transport modes in the city; for example, Bologna, Haifa, The Hague and Zurich all have populations of around 1 million, but the lengths of their bus networks are 464, 2140, 150 and 175 km, respectively, possibly due to the fact that in the former two buses are the sole transport means available, while the latter two also have extensive tram networks. Specifically on light rail/tram systems, these are very common (present in 23 of the 33 cities), again with varying network size. This ranges from less than 10 km in Trondheim and Ankara to as much as 530 km in Barcelona. Noteworthy examples are Prague's and Zurich's very dense tram networks (over 100 km of track served by up to 34 lines), which can be attributed to the fact that both these cities have tram-oriented public transport systems, as opposed to most other cities in the sample, which are rather bus-oriented. Finally, out of the 33 cities, 20 have metro networks, which due to the high construction and operational costs are naturally much less extensive than bus and tram networks. Almost all cities of the sample with over 1.5 million inhabitants have metro systems, with the exception of Kayseri, Kocaeli and Tel Aviv. Metro network lengths range from 5 km in Haifa to 402 km in London, while Paris has the network with the most lines (16).

An important element of road-based public transport infrastructure is the provision of priority measures and systems. The most widely used public transport priority measure is the use of bus lanes, present in 30 cities (all except Ankara, Bursa and Karlsruhe), whose aim is to improve the reliability of bus schedules. Systems granting priority at traffic signals to buses and trams over private transport are also in place in most cities of the sample (26). Regarding the underlying detection methods of the priority systems, the most common are loop detection and dedicated signals, used in 16 and 11 cities, respectively.

Besides the infrastructure itself, integration schemes with other modes are an additional important component of a public transport system. These can be classified as "traditional", which include schemes such as car pooling and car sharing, or as "dynamic", which include integrated public transport and feeder services. In the sample used in this study, 16 out of the 33 cities have integration forms with private transport. Finally, unitary fare systems for different transport modes and electronic ticketing are very common methods to simplify and encourage the use of public transport. Unitary fare systems exist in 19 cities. Electronic ticketing, on the other hand, is less widespread, with actual implementations having been deployed in 13 cities (e.g., London's "Oyster" card), but with another 7 planning to introduce it in the next five years.

Twenty-three out of the sample's 33 cities have also provided information as to their current and future uses of ITS in public transport demand management. As such, it is found that 15 cities use ITS for the purposes of improving access to public transport, such as journey planning software and real-time information on expected times of arrival, and 3 more intend to join them in doing so in the near future. On the other hand, few cities (5) make use of ITS for the management of

pedestrians and cyclists (e.g., during big events). Looking at the situation of traveller information provision, almost all cities provide information to the public. The types of information provided, though, as well as the methods of dissemination vary. Most cities (30 out of 33) provide information about planned events in public transport. Less common, but definitely of importance, seems to be the supply of alternative routes to public transport users, which is done in 19 cities. 17 cities even go as far as suggesting walking and cycling routes to road users, as an attempt to promote these two sustainable travel modes, while 10 cities also provide travel-related weather forecasts. Finally, considering the methods used to inform the public, most cities have a Website with traveller information in place, with 32 cities out of the sample's 33 doing so. Many cities use television and radio broadcasts to disseminate traveller information (25 out of the 33), as well as information boards which are used in 24 cities. Sixteen cities have a telephone line for traveller information, while 6 cities also use mobile phones, either in the form of SMS text messages, or through smartphone apps.

3.6.2 More Detailed Insight of Public Transport ITS Deployment in Selected European Cities

This aim of this second study is to investigate the state of deployment of public transport ITS in more detail for a smaller number of selected cities. For this purpose, an ad hoc survey has been designed as part of COST Action TU1004 in 2010 to assess the degree of maturity of public transport ITS in five selected European cities in relation to objective indicators, such as population and GDP per capita. The survey looks into the provision of public transport ITS applications in the five cities in more depth and more specifically: the number/percentage of stations/stops equipped with real-time information boards and/or covered by systems disseminating traveller information through mobile phones; and the number/percentage of public transport vehicles equipped with real-time information tools, such as AVL and APC systems.

The five cities selected are Barcelona, Madrid, Stockholm, Thessaloniki and Edinburgh, offering an adequately diverse sample in terms of population, with larger (e.g., Madrid) and smaller (e.g., Edinburgh) cities; wealth, with cities with higher (e.g., Stockholm, €34,500/inhabitant) and lower (Thessaloniki, €19,000/inhabitant) GDP per capita; car ownership, with cities with higher (e.g., Madrid, with 457 vehicles/1000 inhabitants) and lower (e.g., Edinburgh, with 310 vehicles/1000 inhabitants) car ownership rates; public transport usage, with cities with higher (e.g., Madrid, 38 %) and lower (e.g., Edinburgh, 20 %) public transport modal split shares; and soft mode travel, with cities with higher (e.g., Barcelona, 52 %) and lower (e.g., Thessaloniki, 2 %) walking and cycling modal split shares.

Regarding the state of deployment of public transport ITS applications in the five cities, Barcelona has installed real-time passenger information panels at all urban

track-based public transport stops (metro and light rail), but only at half of the suburban rail stations and only at a subset of the road-based public transport stops (9 % of the metropolitan bus stops and 4 % of the urban bus stops). Trip planning for all public transport modes is available, but there is no intermodal transport management of public transport ITS applications. In comparison, Madrid has installed real-time passenger information panels in the vast majority of the track-based public transport stops/stations and also offers trip planning. Intermodal transport management of urban buses, metro, light rail and suburban rail is carried out under an integrated system of public transport ITS applications, and e-ticketing is provided in all urban buses, metro and suburban rail, in half of the metropolitan buses and in a quarter of the light rail fleet. Video surveillance is also available in the metro and the suburban rail network. The public transport ITS applications in Barcelona and Madrid are shown in Table 3.3.

Looking at public transport ITS applications in Stockholm, Thessaloniki and Edinburgh (Table 3.4), metropolitan and urban buses in Stockholm are equipped with real-time information devices and AVL. Also, all stations/stops are covered with mobile phone traveller information systems, but only 10 % of urban bus stops are equipped with real-time information panels. APC systems are also in place in small parts of Stockholm's both road-based and track-based public transport networks. On the other hand, Thessaloniki and Edinburgh appear to have a lower deployment level of ITS applications in their entirely road-based public transport

Table 3.3 ITS applications in public transport in Barcelona and Madrid

ITS applications	Units	City	Metro bus	Urban bus	Metro	Light rail	Suburban rail
Number of intersections with traffic light priority	Units	Barcelona	0	0	All	85	All
		Madrid	0	0	All	2	All
Real-time information to travellers	Yes/No	Barcelona	N/A	N/A	N/A	N/A	N/A
		Madrid	No	Yes	Yes	Yes	Yes
Number of public transport stops with real-time information	%	Barcelona	9	4	100	100	50
		Madrid	9	8	100	93	70
Automatic vehicle location (AVL)	% fleet	Barcelona	95	98	N/A	N/A	N/A
		Madrid	0	100	N/A	N/A	N/A
Trip planning	Yes/No	Barcelona	Yes	Yes	Yes	Yes	Yes
		Madrid	Yes	Yes	Yes	Yes	Yes
Intermodal transport management	Yes/No	Barcelona	No	No	No	No	No
		Madrid	No	Yes	Yes	Yes	Yes
E-ticketing	% fleet	Barcelona	0	0	0	0	0
		Madrid	50	100	100	25	100
Video surveillance in the public transportation system	Yes/No	Barcelona	N/A	N/A	N/A	N/A	N/A
		Madrid	No	No	Yes	No	Yes

Table 3.4 ITS applications in public transport in Stockholm, Thessaloniki and Edinburgh

ITS applications	Units	City	Metro bus	Urban bus	Metro	Light rail	Suburban rail
Stations/stops equipped with real-time information devices	%	Stockholm	70	10	100	50	100
		Thessaloniki	N/A	10	–	–	–
		Edinburgh	15		–	–	–
Stations/stops covered with SMS/mobile information systems	%	Stockholm	100	100	100	100	100
		Thessaloniki	55	100	–	–	–
		Edinburgh	100		–	–	–
Vehicles equipped with real-time information devices	%	Stockholm	100	100	0	0	0
		Thessaloniki	N/A	100	–	–	–
		Edinburgh	100		–	–	–
Automatic vehicle location (AVL)	%	Stockholm	100	100	N/A	N/A	N/A
		Thessaloniki	N/A	100	–	–	–
		Edinburgh	100		–	–	–
Automatic passenger counting system (APC)	%	Stockholm	12	12	0	20	10
		Thessaloniki	99	N/A	–	–	–
		Edinburgh	0		–	–	–

networks, consisting solely of metropolitan and urban buses. Specifically, Thessaloniki has achieved complete coverage of all urban bus stops with mobile-phone-based traveller information systems and has also equipped all its vehicles with real-time information panels and AVL systems, but has only equipped 10 % of its bus stops with real-time information boards. Thessaloniki's metropolitan bus system does not have all of these features, but it does provide 55 % coverage of its stops with mobile-phone-based traveller information, as well as near-complete equipment with APC systems. Edinburgh, in comparison, also offers complete coverage with mobile-phone-based traveller information, in-vehicle real-time information and AVL, as well as 15 % of bus stops equipped with information boards, but no APC system.

3.6.3 Discussion and Outlook of Public Transport ITS Maturity and Deployment in Europe

Appraising the results from the two survey studies, certain trends with regard to the state of deployment of ITS in Europe can be extracted. Specifically, from the first exercise, it can be observed that most of the 33 cities surveyed have rather private-transport-oriented networks, which are accompanied by corresponding modal split figures, with only few exceptions of more public-transport-oriented (e.g., Zurich, Prague) and soft-mode-oriented cities (e.g., Barcelona, Paris). Yet, the

vast majority of the cities are mainly interested in improving the efficiency of their transport network by achieving a modal shift away from private means, and with infrastructure investments being less attractive, ITS can play an important role.

Indeed, ITS have contributed to the implementation of more advanced public transport management schemes and strategies and have increased the range of techniques available. Examples of these include the granting of priority to public transport vehicles at signalised intersections and the provision of e-ticketing, both of which have become possible through recent advances in the ICT field. ITS have also broadened the field of data collection and information provision, with the latter being disseminated to the public through various means. The information provided mainly concerns delay times and alternative routes, and travellers can be kept up to date through Websites and telephone information lines prior to their travel, or by information panels, radio broadcasts and mobile phones en route.

With respect to the deployment level of public transport ITS, as assessed through the second survey exercise in five selected cities, it appears that this is correlated with the population, but also with the indicative wealth of the city, as expressed by the GDP per capita index. In other words, it seems that cities with a greater population and a higher GDP per capita index (such as Madrid, Barcelona and Stockholm) are likely to have a greater degree of deployment of public transport ITS than smaller and less wealthy ones (Thessaloniki, Edinburgh). This may be explained by the fact that ITS are associated with both initial and maintenance costs, and even though these are substantially lower than corresponding infrastructure investment costs, smaller and less populated cities may still appear reluctant to undertake them at a large scale. It is encouraging to see, however, that the numerous advantages delivered in managing fleets, planning trips or improving accessibility, are gradually acknowledged by the majority of the cities, which has led to a progressive proliferation of public transport ITS in recent years.

The analysis of the results shows several interesting conclusions. Firstly, it is observed that the total surface of a city is directly related to the motorisation rate. In the case of Thessaloniki, with a total surface of 1456 km^2, it is more difficult to provide adequate coverage of public transport to outlying zones, and thus, the use of private vehicles increases, as there are no other transportation alternatives. Similarly, the GDP also affects directly to the motorisation rate. Cities with a high GDP stand out to have high motorisation rates.

Regarding the modal split, it is perceived that cities with larger population (Madrid, Stockholm and Barcelona), and a higher GDP, are also those with a lower percentage of private car use in daily mobility. This is due to the fact that public transport services are more viable, from the financial point of view, in cities with less population. Therefore, in these areas, the percentage of trips by public transport, walking or cycling is higher.

However, it draws attention to the case of the city of Edinburgh. Of the five cities studied, it is the one with the smallest population, and one of the lowest GDP values. Despite of this fact, Edinburgh has the highest percentage of trips made by public transport. This is because the city has a total surface of 264 km^2, being its

developed area 120 km^2 (one-third of the developed area of Madrid), which facilitates the implementation of a quality system of public transport.

As for the deployment of ITS services in public transport, they are implanted in a greater proportion in Madrid, Stockholm and Barcelona. This is caused by the fact that they have more population and a higher value of GDP than Thessaloniki or Edinburgh. Generally, ITS services have high costs of implementation and maintenance, and thus, only considerable cities can undertake their installation. However, these services provide numerous advantages when managing fleets, planning a trip or improving accessibility, which have led to a progressive proliferation of ITS services in recent years.

3.7 Including ITS Factors in Transit Assignment

To conclude this chapter, we provide a more direct link to the next book chapters, which will mainly focus on the modelling aspects related to transit assignment. In particular, the ITS solutions applied to real-time information and to ticketing

Table 3.5 Relation between ITS tool and transit assignment variables (input) and impacts (output)

ITS tool	Variables (input)	Impact (output)
Real-time information (at stops) e.g., about arrival times	• VoT for waiting times will change • Reliability is increased • Expected travel time updates	• Less journey times • Reduce uncertainty • Alternative services when delays and alterations occur
Fleet management coordination and control	• Headways (at stops) • Reliability of travel times • Capacity	• Better estimation • Smoother operations • Less delays
(Multimodal) trip planners	• Compliance • Calibration/validation data • Agent-based decisions (micro-level)	• More compliance to advices • Travel time savings • Modal shifts
Traffic control (priority)	• Travel time • Travel time variability/reliability • (Line) capacity	• PT performance • Journey time savings • Modal shifts
E-ticketing	• Monetary expenditures • Total journey times • Boarding capacity • Less delays in bus stops	• Modal shifts • Journey time savings
DSS for operators	• Headways • Line suppression/feed	• PT performance

services should be modelled to influence the travel choices of the PT users, such as departure time and line choice, but also to improve the travelling strategies for more effective intermodal transfers.

Questions we therefore address in this book are, for example, as follows:

- Which ITS tool will affect what input variable or parameter in the assignment models?
- Where do we expect ITS to have an impact?
- How ITS change the traditional formulations of transit assignment?
- Which assignment parameters/factor/variables will be modified or added?
- Which travel behaviour do they affect?

To conclude this overview, a table giving an overview of ITS—variables—impact is given (Table 3.5), where the most relevant decisional variables are related to ITS solutions, and the impact each solution is expected to give to the travellers' choices.

3.8 Reference Notes and Concluding Remarks

Cities are deploying ITS measures for most of their transport services. This is a win–win development: on the one hand, ITS technologies and their applications are cheaper than other technological developments and installations. Nevertheless, they provide a basis for improving quality of public transport services. ITS measures deliver benefits to all actors: operators, city managers and travellers. On the other hand, last generation of ITS services consist of two-way communications systems. That means that facilitating information to users improve the perceived quality of the services, but also those users provide back to the system a very rich information for managing and operating the service. Therefore, once it has been demonstrated its effectiveness and usefulness in the management of transport means, ITS deployment will continue in the coming years.

Fleet monitoring systems are widely used in big–medium size cities, but they are extending to smaller cities. They want to operate their transport networks using also the potential of ITS: real-time information provision, detecting incidents for managing them more effectively, greater safety and quality in the public transport system.

At present, transport authorities focus their attention on ticketing and real-time information. They tend to be more integrated using the same technological platforms: ticket validation systems, based on rechargeable and contactless e-cards, and real-time information system for all modes of transport, either in panels at stops or stations, or via smartphone applications.

The concept of big data is becoming increasingly important in this area, as any developer can access the data to perform transport applications. It is noteworthy that in the new transportation applications information flows bidirectional. These tools simultaneously perform the functions of providing information to the user, while

serving as a social platform for users' interaction, and in turn enable transport authorities to collect additional data from the passengers. In addition, non-users can be incorporated to the system by collecting their preferences and expectations and attracting them into the system.

Another point to highlight is that interchanges are key elements for integrating information, as far as they are transfer nodes. They collect real-time revealed preferences and mode choices in only one place where both transport and other services are harmoniously integrated. Public bicycle systems have been recently incorporated to the trip chain, linked to public transport services. Providing bicycle parkings at interchanges and smart information about the public bicycle services is the way forward to link bikes to PT.

However, there is still room for enhancing the system. First, it is necessary a better harmonisation and standardisation. The different communication protocols should be common for all applications in all cities. They would reduce prices and improve users' understanding. Some cities are making agreements with world-wide servers such as Google Transit and other platforms. Those promising integration initiatives are beneficial for users and for operators. They are normally open to researchers and developers of new applications, providing new and enhanced possibilities.

To sum up, it is clear that the implementation of ITS systems in public transport networks produces an enhancement on security, efficiency, reliability and sustainability in the use and operation in all modes of transport. Users are not only granted with better performance of their demands for transport services, but also receive useful information for their choices on services and routes in an increasingly multimodal and complex transport system.

References

Bell MGH, Kaparias I, Nocera S, Zavitsas K (2012) Risultati di una recente indagine sulla presenza in Europa di architetture di sistemi telematici per i trasporti (Presence of urban ITS architectures in Europe: results of a recent survey). Ingegneria Ferroviaria 67:447–467

Hubacher S (2012) Real-time passenger information at bus stops in Lille Métropole (France). ELTIS: the urban mobility portal

Linstone HA, Turrof M (1975) The Delphi method—Techniques and applications. Addison-Wesley Publishing Company, Boston

Zavitsas K, Kaparias I, Bell MGH (2010) Transport problems in cities. In: CONDUITS Deliverable 1.1, 7th Framework Programme, European Union

Further Reading

Banister D (2005) Unsustainable transport: City transport in the New Century. Routledge, New York, p292

Blaquiére A (2012) Bus priority system in Toulouse (France). ELTIS: the urban mobility portal

Burlacu M (2012) Informobility tools for fleet management in Craiova (Romania). ELTIS: the urban mobility portal
Camus R, Longo G, Macorini C (2005) Estimation of transit reliability level-of-service based on automatic vehicle locations data. Transp Res Rec 1927:277–286
Carrasco N (2012) Quantifying reliability of transit service in Zurich, Switzerland: case study of bus line 31. Transp Res Rec 2274:114–125
Cats O, Larijani AN, Olafsdottir A, Burghout W, Andreasson IJ, Koutsopoulos HN (2012) Bus-holding control strategies: simulation-based evaluation and guidelines for implementation. Transp Res Rec 2274:100–108
Cohen G, Salomon I, Nijkamp P (2002) Information-communication technologies (ICT) and transport: does knowledge underpin policy? Telecommun Policy 26:31–52
Chiffi C (2010) CNG consumption monitoring and ecodriving training in Cesena and Forlì (Italy). ELTIS: the urban mobility portal
City Council Almada (2010) Almada's integrated public transport guide (Portugal). ELTIS: the urban mobility portal
Comunidad de Madrid (2013) CITRAM. Public Transport Management Centre, Madrid, Spain
CONDUITS (2009) Coordination of network descriptors for urban intelligent transportation systems. In: 7th Framework Programme, European Union
DB Rent GmbH (2011) Call-a-Bike: public bicycles (Germany). ELTIS: the urban mobility portal
EBSF (2008–2012) European bus system of the future. In: 7th Framework programme, European Union
ELTIS (2011) Greenways—Bus priority measures in Edinburgh (UK). ELTIS: the urban mobility portal
Epp T (2013) VAO—the collaborative traffic information service for all of Austria. COST TU1004 TransITS Meeting, Stockholm, Sweden
European Commission (2007) Green paper of urban transport: towards a new culture for urban mobility. COM/2007/551, Brussels, Belgium
European Commission (2009) Eurostat statistical books: panorama of transport. Office for Official Publications of the European Communities, Luxembourg
European Commission (2010) Directive 2010/40/EU of the European Parliament and of the Council of 7 July 2010 on the framework for the deployment of Intelligent Transport Systems in the field of road transport and for interfaces with other modes of transport. Official J Eur Union Legislation 53:1–13
Frumin M, Zhao J (2012) Analyzing passenger incidence behavior in heterogeneous transit services using smartcard data and schedule-based assignment. Transp Res Rec 2274:52–60
Gangi G (2011) Development and upgrade of the e-ticketing system in Brescia (Italy). ELTIS: the urban mobility portal
Hammerel M, Haynes M, McNeil S (2005) Use of automatic vehicle location and passenger count data to evaluate bus operations: experience of the Chicago Transit Authority, Illinois. Transp Res Rec 1903:27–34
Heves G (2012) Sofia urban mobility website: information about all means of urban transport on one site (Bulgaria). ELTIS: the urban mobility portal
Horoch W (2012) Revolutionised public transport with dedicated bus-tram-lane in Warsaw (Poland). ELTIS: the urban mobility portal
Kaparias I, Zavitsas K, Bell MGH (2010) State-of-the-art of urban traffic management policies and technologies. In: CONDUITS Deliverable 1.2-1.3, 7th Framework Programme, European Union
Lu A, Reddy A (2012) Strategic look at Friday exceptions in weekday schedules for urban transit. Transp Res Rec 2274:30–51
McBrierty C (2012) Real time Countdown system in London (UK). The urban mobility portal, ELTIS
Milkovits MN (2008) Modeling the factors affecting bus stop dwell time: use of automatic passenger counting, automatic fare counting and automatic vehicle location data. Transp Res Rec 2072:125–130

Monzon A, Hernandez S, Cascajo R (2013) Quality of bus services performance: benefits of real time passenger information systems. Transp Telecommun 14:155–166
NICHES (2007) Facilitating urban transport innovation on the European level. Research and policy recommendations. http://www.rupprecht-consult.eu/uploads/tx_rupprecht/17_Facilitating_Urban_Transport_innovation.pdf
Orfeuil JP (2000) L'evolution de la mobilité quotidienne: Comprendre les dynamiques, éclairer les controverses. Synthèse INRETS 37, Arcueil, France
Thrift NJ (1996) Spatial formations. Sage, London
Palkowska N (2012) CCTV monitoring system on public transport in Lodz (Poland). ELTIS: the urban mobility portal
Roselló X (2013) The use of ticket validation for transit planning purposes. COST TU1004TransITS Meeting, Stockholm, Sweden
Smith N (2011) The Lisboa Smart Card System (Portugal). ELTIS: the urban mobility portal
Spencer G (2012) Monitoring of tram driving efficiency in Craiova (Romania). ELTIS: the urban mobility portal
Tyrinopoulos Y, Antoniou C (2008) Public transit user satisfaction: variability and policy implications. Transp Policy 15:260–272
UITP (2013) Information Technology and Innovation Commission. http://www.uitp.org/tags/information-technology
Urry J (2000) Sociology beyond societies. Routledge, London, UK
Vallejo J (2012) Monitoring and planning of PT system in Donostia/San Sebastian (Spain). The urban mobility portal, ELTIS

Part II
From Transit Systems to Models - Klaus Noekel

Chapter 4
From Transit Systems to Models: Purpose of Modelling

Markus Friedrich, Fabien Leurent, Irina Jackiva, Valentina Fini and Sebastián Raveau

From Part I of the book, it will be obvious that public transport plays an essential role in providing mobility to people, especially in dense urban areas. The social welfare generated by good public transport comes at a price, however. Almost all forms require large investments into infrastructure, vehicles and operation. With limited finance, ideal public transport remains a distant goal, and a lot of effort goes into finding an optimal allocation of budget to investment options. The key question for these decisions is: How big is the total benefit of a proposed investment? To answer it, one needs to predict how the potential users will make use of the hypothetical improved public transport. For responsible decision-making, this prediction should be rational, transparent and accountable. It is no surprise therefore that models are typically used to produce the predictions. These models span the whole range of mobility decisions made by individuals, from long term to short term. Passenger route choice, the focus of Part III, accounts for only a part of the complex decision hierarchy. Before zooming into route choice models, this chapter looks at the planning process as a whole, explains the role of models in

M. Friedrich (✉)
University of Stuttgart, Pfaffenwaldring. 7, 70569 Stuttgart, Germany
e-mail: markus.friedrich@isv.uni-stuttgart.de

F. Leurent
Laboratory on City, Mobility and Transportation, University Paris-East, Ecole des Ponts ParisTech, Champs-sur-Marne, France
e-mail: fabien.leurent@enpc.fr

I. Jackiva
Transport and Telecommunication Institute, Rīga, Latvia
e-mail: jackiva.i@tsi.lv

V. Fini
DICEA—Dipartimento di Ingegneria Civile Edile e Ambientale, Sapienza University of Rome, Via Eudossiana, 18, 00153 Rome, Italy
e-mail: valentina.fini@uniroma1.it

S. Raveau
Department of Transport Engineering and Logistics, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna, 4860 Macul Santiago, Chile
e-mail: sraveau@uc.cl

G. Gentile and K. Noekel (eds.), *Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems*, Springer Tracts on Transportation and Traffic 10, DOI 10.1007/978-3-319-25082-3_4

decision-making and gives an overview of the whole decision hierarchy. The last two sections introduce the general mathematical framework, in which decision models are formulated and set the stage for the description of specific models.

4.1 The Planning Process

Markus Friedrich, Fabien Leurent, Irina Jackiva, Eftihia Nathanail and Klaus Noekel

4.1.1 States and Phases of a Transport Plan

Planning deals with the development of measures which are supposed to change the existing state or current state, to a future state or improved state, which is as close as possible to a desired state, also denoted as target state. Deficiencies can be identified by comparing the current state with the target state (Fig. 4.1).

Planning may be considered as a process which systematically reduces the uncertainty concerning the suitability of potential measures. In this way, planning prepares the decision-making about when and where measures should be implemented so that the planning objectives are achieved in an optimal way taking into account all relevant constraints.

Realistically, optimal solutions can hardly be achieved as planning is always a decision problem with competing objectives requiring some type of compromise.

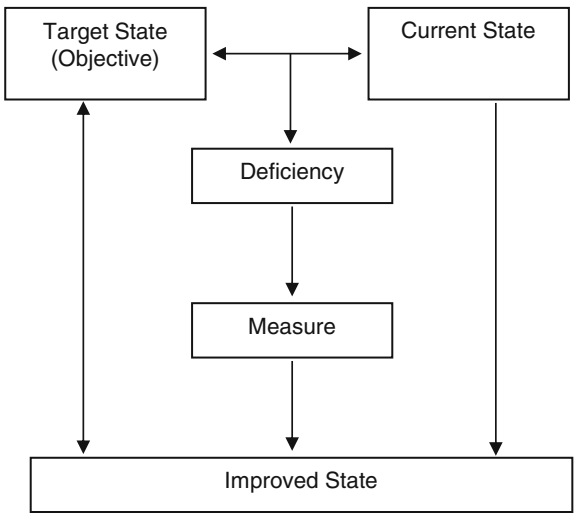

Fig. 4.1 Relationship between target state and current state

The context of planning is usually such that it is impossible to improve the welfare of one or more individuals without simultaneously decreasing the welfare of at least one other individual. In economics, such a state is called a Pareto-optimal state.

Therefore, it is the goal of a planning process to distinguish "worse" solutions from "better" solutions trying to identify a good or the best of all potential solutions. One method for identifying suitable solutions is the Kaldor–Hicks criterion: A measure should be implemented, if from the gains of the winners all losses of the losers could be compensated leaving a surplus. However, the Kaldor–Hicks criterion does not request that the surplus is used to actually compensate the losses of the losers.

The planning process can be described in various ways. Figure 4.2 illustrates one possible way where the planning process is divided into five phases (FGSV 2001):

- Preorientation phase
- Problem analysis phase
- Measures examination phase
- Decision phase
- Implementation phase.

4.1.1.1 Preorientation

The phase of preorientation assesses whether any kind of transport planning is required. Legal requirements, identification of deficiencies by political bodies or by the general public and suggestions for solutions may trigger the need for transport planning. Basic ideas for the solution of transport problems and resulting requirements are discussed in this phase.

4.1.1.2 Problem Analysis

The problem analysis phase aims at identifying deficiencies in the current state. The analysis covers the following steps:

- *Developing an evaluation scheme*:
  The evaluation scheme serves as a benchmark for the evaluation of the current state as well as of the future state. Developing an evaluation scheme includes the selection of objective indicators (e.g., travel time) describing each objective (e.g., minimize time) and the formulation of an evaluation function, e.g., definition of level of services. The evaluation scheme can be more or less formalized ranging from a multicriteria evaluation to cost–benefit analysis. An objective function ebbed in a mathematical optimization model also represents an evaluation scheme.

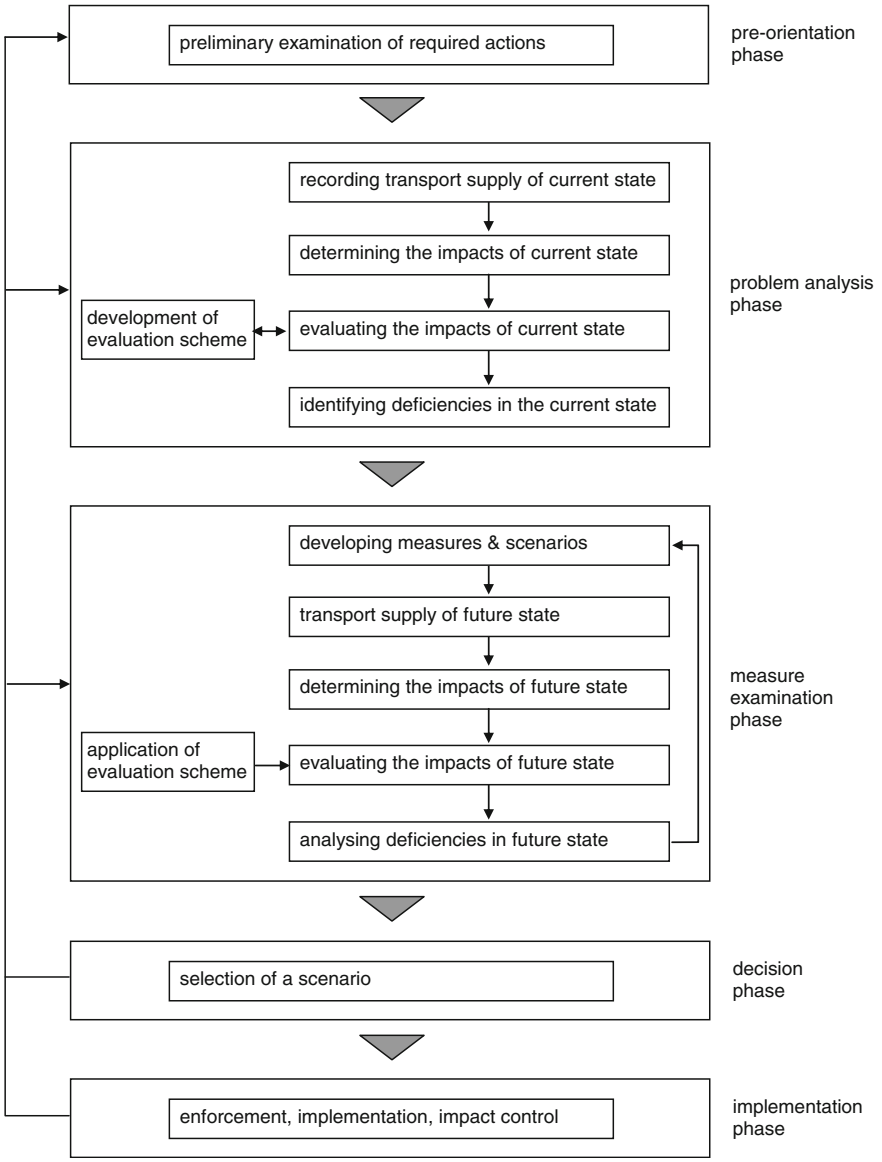

Fig. 4.2 Transport planning process (adapted from FGSV 2001)

- *Recording the transport supply of current state*:
  In this step, transport supply data, land use data, data on mobility behaviour and volumes from traffic counts are recorded.
- *Determining the impacts of the current state*:

The impacts of the current state are determined, i.e., the values of the objective indicators are observed or computed for the current state.
- *Evaluating the impacts of the current state*:
  The impacts of the current state are evaluated by comparing the indicator values with the evaluation function.
- *Identifying deficiencies in the current state*:
  Causes of deficiencies in the current state are identified.

4.1.1.3 Examination of Measures

The examination of measures phase is the core step of the transport planning process. In this step, potential measures are developed and grouped to a scenario which represents a potential solution for a future state. Then, the impacts of the scenario are estimated. A feedback between the different steps of this phase ensures that the steps are repeated until an appropriate solution is found meeting the objectives:

- *Developing measures and scenarios*:
  A set of measures (e.g., modified transport supply or modified land use structure) and assumptions (e.g., future prices) is developed constituting a scenario.
- *Determining the impacts of the proposed measures*:
  The impacts of the future state defined by the scenario are determined, i.e., the values of the objective indicators are computed.
- *Evaluating the impacts of the proposed measures*:
  The impacts of the future state are evaluated by comparing the indicator values with the evaluation function. It is also helpful to directly compare the impacts of the current and the future states.
- *Analysing deficiencies in the future state*:
  Causes of deficiencies in the future state are identified. They provide information for further improvements.

4.1.1.4 Decision

The decision phase deals with the final evaluation by comparing the advantages and disadvantages of possible scenarios by political bodies authorized to make decisions. This phase can generate the following results:

- rejection of all scenarios,
- selection of one scenario,
- decision to examine further scenarios.

4.1.1.5 Implementation

This phase deals with the implementation of the selected scenario. This involves legal steps (e.g., official approval of the plan or development scheme) and the development of financial schemes and organizational structures. The implementation should be followed by an ex post impact study to evaluate the efficiency of the measures.

4.1.2 Public Transport Design

This section explains how measures are defined within the process described above.

The line network and timetable form the basis of the supply in public transport. They determine the competitiveness of public transport in comparison with other transport modes. Of major importance for the traveller are short walking distances to the stops, high service frequencies and short and reliable travel times. When planning line networks and timetables, the following questions need to be clarified:

- *Which is the appropriate transport system*?
  Street-bound transport systems (buses) require only low investment costs, but the vehicle capacity is relatively low (approx. 1000 travellers per hour in case of a 5-min headway). Rail-bound transport systems (tramway, light rail, metro) on the other hand require high investment costs for the track, especially if the route runs underground. Rail-bound transport systems are suitable for high-demand networks.
- *What is the appropriate number of stops*?
  Short distances between stops reduce the access and egress time of travellers. At the same time, short distances increase the running times as the vehicles need to stop more frequently. With higher travel speeds, the time loss per stop increases due to deceleration and acceleration. Therefore, in practice, the distance between bus stops lies between 300 and 400 m, for tramways between 400 and 500 m and for metro lines between 800 and 1000 m.
- *What is the appropriate number of lines*?
  High network coverage with a dense line network reduces the access and egress times. A high line density reduces the number of transfers as the number of direct connections is higher. High network coverage and line density on the other hand increase the length of the line network, thus leading to higher investment and operation costs.
- *What headway and vehicle size are appropriate*?
  A short headway reduces the waiting times at the origin stop and at transfer stops. On the other hand, a short headway increases the vehicle kilometres travelled and thus the costs. High-capacity vehicles are more cost efficient as a smaller number of drivers can transport a higher number of travellers.
- *What are the appropriate running times for the timetable*?

Short running times permit high travel speeds and a high productivity. However, considering the reliability and punctuality, the running time should include buffer times to absorb delays from disturbances.

These questions explain why the planning of line networks and timetables in public transport is an optimization problem. The planner must decide on the weights of attributes to each particular objective. Until now, it is not common practice to develop a public transport supply using a comprehensive mathematical optimization model which directly computes an optimal solution. Since each road section can be part of a public transport line, stop locations may be placed at various locations, and every departure time of a line can be modified, the system has many variables leading to a complex optimization problem which until now can only be solved sequentially. Therefore, the design of a public transport supply generally applies computer-aided procedures. Computer-aided transport planning leads to a division of tasks between the transport planner and the computer. While the transport planner improves the solution step by step based on the current solution state, the computer determines the impacts of the solution (Friedrich 1994). Computer-aided transport planning uses modelling techniques consisting of network models for the transport supply and travel demand models (TDMs) for determining the impacts on the travel demand. Figure 4.3 illustrates a typical computer-aided design process for public transport. The design process is based on the planning process shown in Fig. 4.3. Here, the planner creates a scenario of a line network with timetables. As a good public transport design equally addresses service quality and operational efficiency, the planner may also develop the vehicle deployment plan and even a driver deployment plan as input for a comprehensive cost calculation. For this task, vehicle and driver scheduling algorithms are available. The impacts of the scenario are then evaluated using a TDM with a public transport assignment and a line costing model. Both models determine the indicator values for the evaluation scheme. The design process is influenced by various factors shown on the left side of Fig. 4.3. Typically, the design follows some type of general planning concept. This concept may suggest network hierarchies (e.g., rail, express bus, local bus), a system headway (e.g., a 60-min headway for a train network) and a set of operating principles (e.g., demand responsive). Often it relies on specific vehicle types, and the planner may use templates for lines which already consider typical constraints like layover times and driver breaks. Other important factors are local characteristics ranging from the settlement structure to the road and rail infrastructure and the operators of the region with their specific cost rates. Also, legal requirements (e.g., driver breaks) must be considered.

4.1.3 Scenario Definition

The planning process for public transport design examines scenarios which group together changes to the supply and demand. When the planner defines the scenarios,

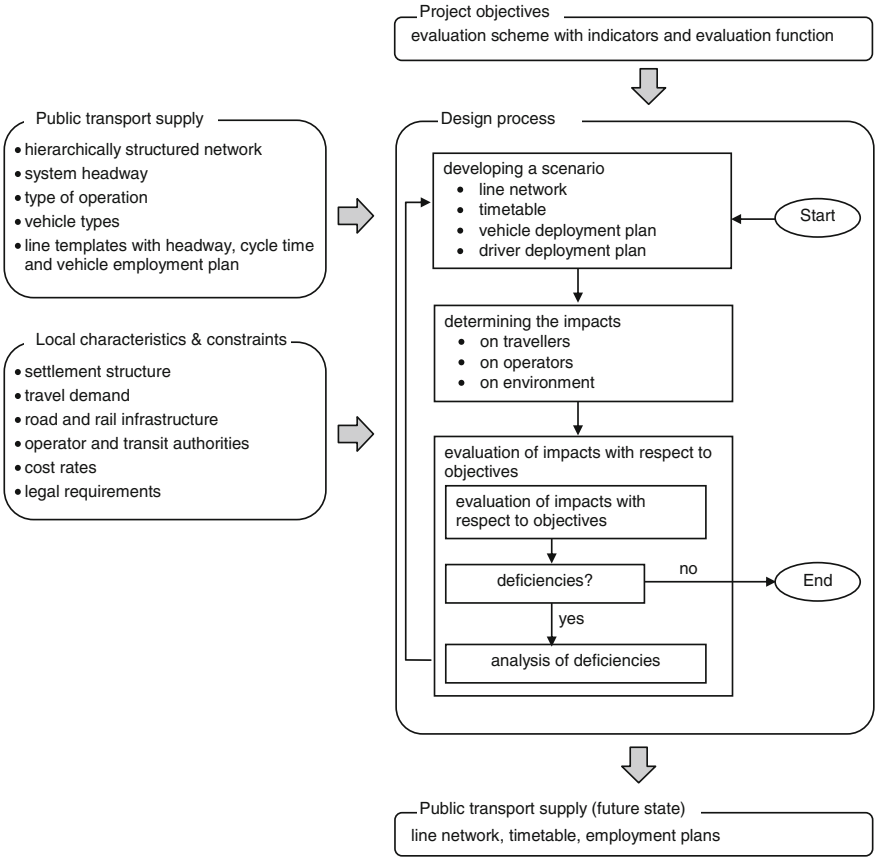

Fig. 4.3 The public transport design process (adapted from Friedrich 1994)

one should take into consideration the specifics of public transport as a complex sociotechnical system.

First, it is important to understand that many different groups of people are involved in some form in public transport each of which has its own set of objectives, often conflicting. Most obvious are the passengers themselves, not only as customers that buy travel services, but also individuals that let themselves be driven in vehicles (i.e., moving machines that take speeds which involve some risk) in which comfort conditions and travel duration are variable. Public transport services are operated by public or private companies who employ diverse staff to deliver the service (drivers, sales, customer care, etc.). Along with the network operator and the population of users, the urban community is affected by the transit system as it is a collective asset for people transportation throughout the city and hence for the accessibility to people, places and activities (homes, workplaces, shops, services, etc.). Thus, the city government regulates the transit system and

often provides funds and land for its development and perhaps also funds for its basic operations. The relationship between the transit operator and its users is grounded on some principles: service information, tariffs, rules of access and use. The relationship between the transit operator and the local government also has its own principles, about the service conditions (the area of coverage, time spans, service grids, quality of service objectives) and also about the ownership of infrastructure and its land, the station buildings and the rolling stock. These relationships constrain the development of scenarios.

Another characteristic feature of a transit system is that the main resources have long to very long lifetimes (20 or 40 years for vehicles, potentially much more for tracks and stations) conditional on periodic maintenance, meaning that the activity is a highly capital-intensive one. So it must be dealt with in the very long term. This long-term orientation must consider the current state of the city, the demography, available land and future technologies in transport. This is especially important in quickly developing countries and cities. In cities with high densities only mass transit systems can provide the required capacity for motorized person transport. In places with increasing income more people can afford a car. Higher car availability may lead to urban sprawl and more car trips, if public transport does not provide an appropriate alternative. And future car technologies with automated driving will challenge future public transport systems.

To prevent such risks to arise and to achieve good city performance and vibrancy, it is of major importance for the city to provide transit services of adequate capacity and quality. Continued planning is in order to maintain the adequacy of urban needs and transit services in the long run. This assigns a long-term horizon to the planning studies, which must deal with long-term scenarios.

A transit network scenario characterizes a set of features that pertain either to the demand side or to the supply side. Most of the system policy belongs to the supply side. However, customer information and pricing are also important instruments to shape the demand.

The first step in scenario composition is to determine the range in both time and space. To deal with long-run stakes requires extending the time range along some tens of years, from at least 20 to 50 years or even more. Along this time axis, a number of stages, or epochs, will be identified, say every 5 or 10 years, so as to simulate the system state at every stage and to specify the potential evolution from every stage to the next one, in a progressive and cumulative manner. The relationship between supply and demand is fundamental for the system state and the evolution of each side: the state characteristics of any given stage are likely to drive the adaptation of each side for the next stage.

The range in space is a serious issue concerning an urban transit network. The extent of the coverage area is a stake of its own, in relation to the institutional arrangements notably between neighbouring municipalities. For the study analyst, the basic requirement is to set a study area that contains both the major supply components, notably so the railway lines, and the main traffic generators. Within the selected perimeter, spatial features must be detailed in line with travellers'

perceptions and behaviours: for example, any element that takes some tens of seconds to travel along it must be identified, since the associated time element is significant to a potential network user.

4.1.3.1 Demand in the Long-Run Perspective

The next step in scenario design specifies demand at each stage, including the evolution of demand from one stage to the next one. The future population, its composition with respect to age, gender, and socioeconomic characteristics including income and number of persons per household will influence future trip pattern and the structure of the origin–destination matrix. The formation of travel demand must be addressed by using a TDM, including at least the basic four steps of trip generation, trip distribution, mode choice and network assignment (see Sect. 4.2).

Important demand characteristics to be specified in a scenario include the following:

- Population demographics: the overall number of persons and their distribution by age and social position make the primary determinant of travel demand. The future population must be forcasted by age for the planning horizion.
- Demographic and social composition of the population: the share of working people by gender and type of occupation (white collar employee, blue collar worker), the share of students and retired people and the household size.
- Location of population activities such as home, work, school, services (health, bank, etc.), shopping, leisure.
- Long-term mobility choices which may be considered exogenous to the model: car motorization, car licence ownership and subscription ownership (by coverage area and tariff scheme) are important drivers of transit demand.
- Population segmentation by income and the growth rate of individual incomes.

4.1.3.2 Supply in the Long-Run Perspective

Many important supply characteristics are likely to be stable between two epochs: the location of railway stations and lines, as well as the amounts of time to travel from one station to the next along a line or to transfer from one line to another in a given station. However, even such prominent features of supply may be subject to change: station access is likely to be improved by station managers, especially so for pedestrians and two-wheelers, while roadway congestion or parking pricing may

penalize access by car. Furthermore, transit congestion may alter the travel time between stations and the travel comfort.

It is thus important to capture the relationship between traffic and quality of service by using a relevant transit assignment model (see Part III) and, on the scenario side, to describe the targeted evolution of the supply conditions.

Important supply characteristics to be specified for each stage in a scenario include the following:

- For each transit line, the location of stations and infrastructure paths between stations, as well as the transit mode and the service routes.
- For each station, the conditions of passenger access and transfer between the service routes that serve the station. Their dependency on the access mode is important, too. Also, the conditions of passenger waiting contribute to the quality of service: seat availability, information, ticketing and shopping opportunities.
- For each service route, the operational conditions for each typical period (e.g., morning peak, evening peak, between-peak) in terms of run frequency or timetable, headway (ir-) regularity and the vehicle characteristics that pertain passenger comfort and above all capacity: overall inside capacity, seat capacity and door capacity for the exchange of passengers between the vehicle and the station platform. This is related to the vehicle fleet that is available at the studied epoch, in line with the equipment strategy of the network manager.

The conditions of travel information and pricing are important not only at the line level but also at the network level: seamless access to a full range of services reduces the transaction costs and the search costs and improves the overall quality of service, thus contributing to transit attractiveness in the modal competition.

The extent of supply offered is the most important driver of costs to the transit supplier: notably the vehicle fleet, but also the service operations (labor cost, fuel cost), the development and maintenance of infrastructure, generate production costs which must be managed in the long run. Therefore it is important to consider the trade-off between costs for investment and operations and the revenues from tickets and other sources (government subsidies, revenues from advertising in stations or on the vehicles).

The financial balance of transit supply determines the range of its strategic development. It is related to the community's ability to invest in and pay for transit services; all the more so as the urban area is heavily and densely populated, with prosperous inhabitants and vibrant activities.

4.1.3.3 Extending the Range of Scenarios

In most cities throughout the world, the urban future is quite an open field. Beyond the historic heritage and its consequences through the existing infrastructure, vehicles, population composition and location, in the long run, there is much room

for many changes to take place. In every big city, the transit system is a major part of the urban fabric: its development interacts with the city development in many ways, so the transit development scenarios must be integrated in, and contribute to, the urban development scenarios.

In order to identify an optimal level of transit service provision, the set of scenarios should span a range from little investment to highly ambitious plans. Here are some instances of ambitious transit planning:

- A well-designed grid of railway lines to cover the central part of any urban area populated by more than several hundreds of thousand citizens.
- Systematic priority to transit vehicles at traffic signals along roadway arterials.
- Providing bus lanes and specifically designed bus stops to speed up operation and minimize travel time variability.
- Well-coordinated transfers at convenient transit hubs in order to minimize door-to-door travel times.

For the most dynamic cities of the developing countries, a high level of ambition should be the rule in transit system development, in order to prevent the quick rise of individual, mostly car-based traffic and its cortège of bad consequences. An ambitious transit system is no luxury but an essential good for large cities.

4.1.3.4 ITS and Supply Operations

Information and communication technology (ICT) has pervaded the supply of transit services in a number of respects. First, the operations of many components have been automated: from traffic signals (along roadways as well as railways) which are essential in vehicle running along transit routes, to vehicle driving, passing by railway track devices, vehicle door opening and closing, ticketing machines and so on. Such automation affects the practical performance in a dramatic way—its results must be mirrored in the supply settings of any scenario. Even passenger comfort is concerned, e.g., through the thermal regulation of vehicles and waiting areas.

Further on, ITS has pervaded the traffic operations along a given transit line; by monitoring the vehicles' positions in real time, and reporting them to the line headquarters, it has become customary to manage the line in real time by feeding instructions back to the vehicle drivers (or driving systems) and to any line device. Thus, the operations are made adaptive and responsive to traffic disruptions, which allows for curbing them more quickly, more efficiently and at reduced cost.

Dynamic traffic information (DTI) extends the scope of dynamic line operations to user behaviour: information provided by radio, variable-message sign (VMS) or the Internet may influence the traveler to adapt his route, his time and mode of travel. It may also affect the choice of a particular car of a train to get less crowded conditions. Such behaviours contribute to load equalization either at the train level or at the level of vehicle runs serving a given transit route, or at the level of transit

lines and maybe even between submodes. Their overall effect should be represented in the traffic model.

4.1.3.5 ITS on the Demand Side

In a transit system, DTI involves the demand side in conjunction with the supply side. The diffusion of individual devices such as smartphones and tablets allows for the customization of information, for personalized route guidance which will multiply the benefits to the traveller, both on an individual basis and through the overall effect on traffic conditions.

Reliable customized information, e.g., provided by journey planners, should help the trip-maker to adapt his travel not only en route or pretrip but also in his habits, by making him more aware of the rich set of opportunities that are made available by the transit network. The traditional instruments for travel information have been quite modest in view of the complexity inherent to a transit system: network maps are rather crude representations since they depict neither the service frequency nor the run times nor the transfer times at stations. Interactive information indeed brings about a revolution in traveller awareness. This should contribute to narrow a traditional gap between the practical world and some major behavioural assumptions in assignment models, making them more realistic.

The travel experience is likely to be transformed by individual devices in yet another way. Such devices greatly ease the performance of other activities such as listening to music, video watching, phone calling, Internet mailing or any work activity related to the treatment of information. In the Paris railway network as of 2013, about one passenger out of two is using a device to perform another activity in parallel to, or better in association with, making his trip. This conjunction is indeed a revolution in trip-making: it amounts to the recycling of the travel time, thus reducing its inconvenience in the daily programme of activities. Such opportunity can be expected to increase the attractiveness of transit with respect to the modes that involve more active participation of the traveller: the car, any kind of two-wheelers, or even walking.

The small size of the most recent individual devices makes a significant advantage over the laptop for use in urban transit: by easing both the use and the carrying of the device throughout the day, whereas a laptop is more suited to seat interurban travel.

4.1.3.6 New Frontiers in the Supply–Demand Interaction

Along the three revolutions in dynamic information, supply awareness and time recycling, several in-depth transformations can be expected due to ITS technologies. Some have already come to some existence in diverse urban networks:

Dynamic traffic management at the network level is more complex than at the line level. It involves joint monitoring and coordinated management of the lines and submodes, considering the passenger flows. Dynamic, network-wide traffic management should be aided by and based on, a traffic assignment model, so as to match the real-time actions to the current pattern of passenger trips in the network.

The development of a truly multimodal travel environment for the individual trip-maker by providing customized information about all means of transport, including parking conditions and prices.

Last but not least, the ubiquity of information enables novel services of car sharing, ride sharing and park sharing to be developed: in other words, intermediary modes that involve the sharing of small-capacity vehicles between information-connected users. Such modes bridge the gap between the traditional car and transit modes, as well as between the private and public operations of transport resources.

Some planning scenarios should give place to the large diffusion of such travel options: all the more so in less dense areas so as to serve them with small-capacity vehicles and to reserve the large-capacity vehicles to those parts of the urban area where they are mostly needed.

4.1.3.7 ITS Implementation as Part of Network Design

More generally, the quick diffusion of ITS-ready individual devices among large populations of trip-makers constitutes an unexpected opportunity for transit supply to re-invent itself. The planning agenda should give room to innovative services, at any level of transit supply, whether within a vehicle, at the line level, at the network level, as well as in the organization of individual mobility by the trip-maker and in the systematic sharing of transportation resources between the trip-makers, possibly with coordination by the transit operator who is or should be an expert in resource sharing.

The stakes are very high, thus resulting in doubling the number of persons per car trip, or halving the parking time per shared vehicle, representing cheap options to double the network capacity.

4.1.3.8 About ICTs in the Design Process

Lastly, ICTs can be expected to transform the design process in itself. The current state consists in computer-aided design by one "designer" or a team of designers, with periodic meetings with the main stakeholders on the basis of planning documents, maybe also of computer-based visualization.

A future state of the design process is likely to involve many additional "special effects" of virtual reality, by the association of a much larger computing power to

much more advanced functions of visual and audio simulation. The representatives of the main stakeholders and also of diverse user groups will be invited to join the design team for some time and share an online, virtual experience of network performance. Functions that have been standardized in some video games about city simulation and planning can be expected to develop in the specialized area of transit simulation and planning, meaning the individual implication in an actors' game, eventually in a networked version.

4.1.4 Evaluation

There are different definitions of "evaluation", a famous one pointing out that "evaluation can be defined as a set of activities to conveniently arrange the information needed for a choice in order that the various participants in the choice process are enabled to make this choice as balanced as possible" (Nijkamp et al. 1990). Meyer and Miller gave the following definition: "Evaluation is the process of determining the desirability of different courses of action and of presenting this information to decision makers in a comprehensive and useful form" (Meyer and Miller 2001).

Three main questions can be distinguished when a system is evaluated:

- Efficacy: Does it actually work?
- Efficiency: Is it built and operated as cheaply as possible and at the necessary quality?
- Effectiveness: Does it meet the stakeholders' objectives and gives them or society positive net benefits?

Different stakeholders have different perspectives on each aspect, and evaluation schemes use diverse indicators to capture this diversity.

Two main types of evaluation are distinguished: ex ante (preimplementation) and ex post (post-implementation) evaluation. *Ex ante evaluation* assesses how well a proposed project will achieve its objectives in the future. *Ex post evaluation* assesses how well a completed project achieved its goals and objectives in reality. In transportation planning, ex ante evaluation is the rule. It is more complex than ex post evaluation because there are many sources of uncertainty: incomplete information, the lack of certain technical or economic parameters of the project, conflicting objectives of different stakeholders or even unforeseen behaviours and unpredictable actions by the agents involved. Ex ante evaluation can be carried out on the basis of:

- expert findings and knowledge (a qualitative assessment);
- comparison with similar projects in other countries or cities (best practice, benchmarking);
- modelling: the proposed project is investigated using a microscopic or macroscopic simulation model.

Modelling is indispensable in the evaluation of transportation projects, because the future performance of the proposed transportation system cannot be observed at the time of planning. The mathematical models described in this book help close the gap. Traditionally, they are used for:

- analysis of the current state (status quo);
- definition of future scenarios;
- forecasting the performance of the alternative scenarios and scenario comparison.

4.1.4.1 Purpose and Subject of Evaluation Based on Transit Assignment Modelling

The data flow chart in Fig. 4.4 connects a description of supply (transport infrastructure) and (total) demand (mobility needs of travellers) to the principal performance indicators for the main stakeholders: passengers, transit operators and

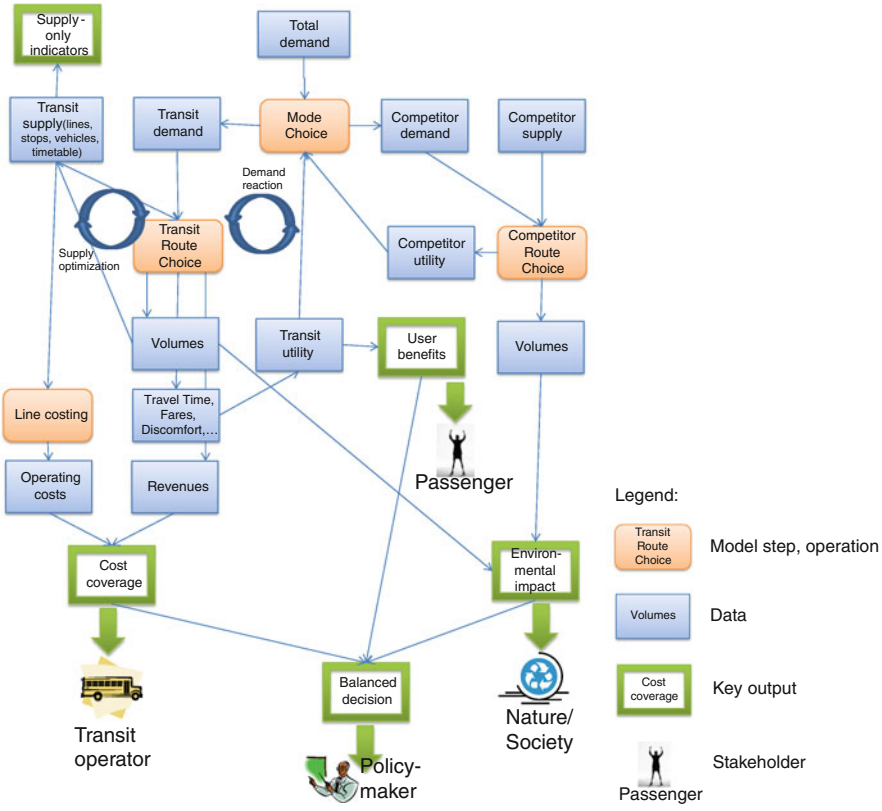

Fig. 4.4 Data, impact assessment and optimization models as parts of a transport model

society/nature/environment. Alternative scenarios within a transport projects are defined as variations of supply and demand (model input), and a cascade of mathematical models eventually yields the indicators by which the scenarios are ranked.

Before the evaluation process and the component models are formalized, consider in qualitative terms the diverse effects brought about by a transport project.

Benefits to the users of some new transit service may include reductions in travel time and cost, improvements in comfort, convenience, security, and safety, etc., which may have knock-on effects, such as inducing new trips to new destinations or at different times. Cost reductions may be balanced by cost increases elsewhere and possibly for other actors: changes in fares, operating costs, and so on. Understanding the importance of these factors in the decision processes of the traveller is key in understanding his mobility choices and forecasting with confidence how well a future transport system will be utilized. If transit users make a new trip, or car drivers switch to transit, they have, by whatever internal reasoning they use, made a decision that they are individually better off. In other words, they see net benefits. A similar logic governs the perspective of other stakeholders, e.g., the operators. They will only invest into new transport infrastructure, if there is a positive return on the investment (taking into account also subsidy payments).

In determining the effects of a transportation project, we typically distinguish between *direct* and *indirect* impacts and what are often referred to as additional economic effects. The *direct* impacts or effects of a transportation project must be sought in the market where the intervention takes place, departing from the identification of all agents affected by it. Therefore, the effects will be determined by the extent used in the definition of the transport project. Examples of *direct* impacts include the following:

- user and business time saving
- reducing the volume of car traffic and congestion
- reducing the noise and pollution.

The *indirect* impacts or effects appear in the markets (secondary) whose products or services have a complementarity or substitutability relationship with the primary market and where there is some distortion that prevents the price being equal to the marginal cost. In many transport projects, it is usual that any intervention in a particular mode affects modal distribution, significantly affecting other transport markets where there can be congestion, externalities and so on. Some *indirect* impacts are as follows:

- providing access to large centres of employment, health care, education establishments and other services
- residential development
- social inclusion
- modal shift to public transport, reducing pollutant emissions and carbon footprint.

Transportation system changes tend to have dual effects: they can improve the public's access to public transport, but they can also result in problems related to greater traffic levels within or near a corridor area. Transit supply changes may significantly affect travellers, often by increasing or decreasing the amount of time required to reach a destination. Noise effects are often the most significant impact on the liveability of an area because they are not confined to the outside environment but intrude into people's homes. Noise may result from a number of sources, including increased traffic or a new transit route, and may affect residents in a variety of ways, including creating sleep disturbances and heightening stress levels.

Even though transportation projects are local, their influence often spreads out beyond the area of implementation. Responding to changes in the road network, traffic will shift from the impacted part of the network to other areas, and the intensity of the shift will depend on several factors, such as road characteristics, demand structure and network configuration. Quantification of the likely changes in transportation benefits and costs associated with the capacity expansion is crucial for policy maker in order to determine the net benefits from capacity expansion projects. This information can be used in the process to select the projects that are most likely to generate the highest return to society.

Output from the TDM can be used to quantify many important costs and benefits for transit users, most notably time savings. These time savings' benefits are based on the consumer surplus theory.

Usually, the proposed investments will not only affect travel times, but also reduce vehicle operating and ownership costs for non-transit users. Because some drivers will instead choose to use transit, there will be fewer automobiles on the road and thus fewer vehicle miles travelled. Aside from reducing congestion and increasing vehicle speeds, lower car miles travelled results in quantifiable vehicle operating cost savings. It may also encourage some transit users to own fewer vehicles. In terms of operating costs, shifting from driving to transit reduces overall vehicle miles travelled, which provides savings in the marginal costs of auto travel (fuel, maintenance and tires). Households that have good transit accessibility and own multiple vehicles are strong candidates to reduce their auto ownership level.

Reductions in vehicle miles travelled lower the incidence of traffic accidents. The cost savings from reducing the number of accidents include direct savings (e.g., reduced medical expenses, lost wages and lower individual insurance premiums) as well as significant avoided costs to society (e.g., second party medical and litigation fees, emergency response costs, incident congestion costs and litigation costs). The value of all such benefits—both direct and social—could also be approximated by the cost of service disruptions to other travellers, emergency response costs to the region, medical costs, litigation costs, vehicle damages and economic productivity loss due to workers inactivity. There is the practice to estimate accident cost savings for each of three accident types (fatal accidents, injury accidents or property damage only accidents). Some studies perform more disaggregate estimates of the accident cost savings, applying different accident rates to different types of roadways (e.g., interstate, highway, arterial) (Table 4.1).

Table 4.1 Categories of costs and benefits in transportation project evaluation

	Direct	Indirect
Benefits	Travel time saving Vehicle operating and parking costs for people who switch private car to public Improved accessibility for certain population groups From reduced emissions From reduced environmental damage	From increased economic activities Increased employment in transit service areas From land use development (property development and increased government taxes)
Costs	Operation costs Infrastructure construction costs	Of noise pollution Costs of traffic delays during construction

The investments can create environmental benefits by reducing air, noise and water pollution associated with automobile travel. In addition, transit travel is usually more energy efficient than auto travel (in terms of energy consumed per traveller), creating benefits associated with energy conservation.

Transit operations are traditionally labour-intensive, and transit expenditures tend to provide more local economic activity than most other transportation investments. Similar to operations and maintenance expenditures, construction expenditures also generate additional economic activity, jobs and employment earnings. These additional economic effects fall into three categories:

- *Direct* impacts from expenditures on construction materials, service and labour.
- *Indirect* impacts from subsequent intra- and interindustry purchases of inputs and production of outputs as a result of the initial direct expenditures/change in output of the directly affected industry.
- *Induced* impacts generated from increases in household spending on goods and services that result from additional employment through the direct and indirect effects on purchasing power.

4.1.4.2 Behavioural Changes and Psychology

Real-time information systems form a category of ITS, which does not directly affect the operation of the transit system, but rather provides knowledge about its current state and expected performance, to the passengers. The aim of such systems is to make transit more attractive to users through provision of higher service quality, to improve mobility and optimize (when advice is also provided) overall performance (user travel time, air quality, energy, etc.). More information about such systems has been presented in Chap. 3.

Understanding the impact of real-time information on traveller behaviour is usually based on the development of models that capture and analyse behaviour in

the presence of information, either by providing the information or by suggesting, for example, the best path to be chosen.

The evaluation of the effectiveness of such systems requires the development of prediction models for traveller behaviour and the estimation of the impact of such behaviour on transit assignment. The objective of the evaluation framework would be to:

- assess the parameters that may be used in traveller behavioural models;
- formulate and develop models to be used for the estimation of the probability of the effect of real-time information on travellers;
- examine and evaluate the degree of accuracy the above models may provide in the prediction; and
- estimate and evaluate the impact alternative information strategies may have on traveller behaviour and consequently on transit assignment.

Usual evaluation techniques are based on comparisons between planned and actual measures, or between the measured against a threshold value, which may be determined by the transit agency. Apart from the direct impact of the ITS systems on transit performance (e.g., bus delays at junctions, travel times, delay variability, schedule regularity) and other indicators (fuel consumption, emissions), other impacts are observed and may be assessed. Such impacts are related to the behavioural changes of the users of the transit system and the transportation system in general. These changes are anticipated to lead to modal shift and itinerary changes, and therefore ridership changes.

The theory of planned behaviour (TPB) is one of the most common methods used for explaining the impact of attitudes on intentions and intentions on behaviour (Ajzen 1991). According to TPB, a person's behaviour is affected by social norms and attitudes and is defined by his/her intention to implement a process. A behavioural intention comprises the cause of behaviour and depicts the willingness to perform the behaviour. Knowledge on such attributes may predict intention and consequently behaviour. A combination of this theory (thus attitudes and intentions) with observable variables (behaviour) would provide the tools for assessing the change of attitudes towards real-time information provision and the impact of attitudes on behaviour and travel choices.

Measuring self-reported and/or observed behaviour leads to the estimation of the behaviour change, which is the objective of ITS in public transport. A minimum requirement would be to conduct before and after the measurement of the behaviour in focus. Apart from the actual measurement (i.e., mode and route selection), the predictors of behaviour, variables that are common to the above mentioned theoretical models should be assessed, such as behavioural beliefs, normative beliefs, control beliefs, personal/moral norm, attitude, subjective and descriptive norms, perceived behavioural control, behavioural intentions and past behaviour.

4.1.4.3 Evaluation Design

Any evaluation requires the prior definition of the measurement variables, the data collection method and the data collection technique.

Evaluation designs can be classified in three broad categories: non-experimental, experimental and quasi-experimental designs (Trochim 2006). Non-experimental design uses before and after measurements of the effect of the intervention on the exposed group. Experimental and quasi-experimental designs assume the formulation of two groups, the intervention group to be exposed to the intervention and the control group to be kept away from the impact of the intervention. In experimental designs, subjects are randomly allocated to different groups, while in quasi-experimental designs, they are not. In both designs, the invention group is compared to the control group, in more than one measurement, before and after the intervention. Subjects belonging to one or the other group are defined before the implementation of the intervention.

Control group ensures strong internal validity that is lacking in single-group evaluation designs, because they provide the means for comparing the effects of an intervention (e.g., information provision) observed in the intervention group with the effects observed in the control group where no intervention was given. In practice, experimental designs are hard to implement when a large-scale application is being tested.

As measurements are based on a sample, the analyst has to extrapolate from the sample to the full population and test whether results are likely due to random factors or to a real relationship. Inferential statistics are used to compare any two of more variables and enable the evaluator to examine the strength of association (Wholey 2004; Logan et al. 2006). To apply inferential statistics, a statistical hypothesis must be specified first. The null hypothesis is that an equality relationship between any two variables of interest is true. If this hypothesis is rejected, there is a statistically significant difference between these two variables and the statistical significance is determined.

When the same person is used for both variables, the t test for dependent means is applied. If the variables are collected randomly, the t test for independent means is chosen instead.

When comparing variables in more than two groups, the statistical procedure followed is the analysis of variance (ANOVA). Analysis of covariance (ANCOVA) is used in order to reduce the error variance and increase the statistical power of the research design by controlling for factors relevant to the (outcome) measurement variable. If the variables are nominal, then the chi-squared test is implemented.

Prediction models correlate behavioural changes with other psychological parameters. For example, simple regression analysis can be used to correlate the dependent variable with one parameter, while multiple regression analysis considers more than one parameter.

4.1.4.4 Evaluation Techniques

Different types of evaluation refer to the time of the measure intervention when evaluation processes take place. To that respect, there are two types: *formative* evaluation and *summative* evaluation.

Formative evaluation is used to provide feedback about whether the intervention may be improved or strengthened. It takes place during the implementation of the measure, analyses the process (process evaluation) and is used to determine whether it is being implemented and working as planned. Both quantitative and qualitative methods may be used.

Summative evaluation is used to review the results of the intervention. As outcome evaluation, it collects or estimates those parameters that are expected to be changed owing to the intervention and may involve ridership, average travel time, vehicle occupancy, etc. and other psychological factors, i.e., attitudes, intentions, beliefs, reported behaviour, etc. These factors are described in more detail in the next paragraphs. Further analysis is used to estimate the impacts of the intervention, in the sense external components of the transit system, as pollution levels, noise and safety.

In efficiency-effectiveness evaluation, the evaluation of a transportation project can be approached from two different perspectives: socioeconomic evaluation, considering all benefits and costs which the project generates for society as a whole, or financial evaluation, just focusing on the monetary revenues and costs generated by the project. Socioeconomic evaluation is performed by comparing the social costs and benefits of a transportation project, once homogenized temporarily through their social net present value (*NPV*), whereas the financial evaluation only compares the income and monetary costs associated with the project, calculating their financial *NPV* (Cascetta 2009a, b, c).

The result of the evaluation supports decisions about the project by comparing the benefits and costs generated over time. If the decision is taken by a private operator, it is only relevant to consider the revenues and costs generated by the project for the operator, especially if there is no external source of funding. If, however, the decision is made by a public agency from the point of view of society as a whole, the comparison must include the social benefits and social costs of all agents affected by the project, even if some do not directly participate in the transport market. Examples are improvements in environmental conditions, public health and safety, or enhancing the attractiveness of the urban environment.

A variety of formal techniques are applied to the evaluation (ex post and ex ante) of transportation plans and projects, all of which follow the general process shown in Fig. 4.5:

- Cost–benefit analysis (CBA)
- Economic evaluation method—social cost–benefit analysis (SCBA)
- Financial evaluation method—revenue–cost analysis methods
- Multicriteria analysis (MCA)
- Cost-effectiveness analysis (CEA)

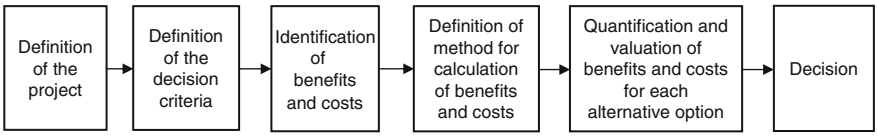

Fig. 4.5 The process of evaluation of projects

- Regional economic impact study (REIS)
- Environmental impact assessment (EIA).

The methods can be grouped into two major categories: the single-criterion methods (monetary approach) and the multicriteria methods (non-monetary approach). Below, we will describe the most well-known method in each category: CBA and MCA. CBA has been used for impacts which are easily quantifiable in monetary terms, while MCA is preferable if impacts are not easily monetized. It should be noted that the choice of evaluation technique may bias the result: the CBA tends to favour strategies that improve indicators with an immediate monetary equivalent (e.g., travel times), while the MCA is able to recognize soft factors (e.g., environmental and safety issues) more easily. A balanced approach should match the evaluation method to the objectives of all stakeholders and consider both quantitative and qualitative impacts.

4.1.4.5 Cost–Benefit Analysis

CBA is a widely applied method for evaluating the "goodness" of public investments as well as for ranking alternative investments. CBA attempts to measure the total cost and benefits in monetary terms. CBA is based on the basic principle of demand and supply and is manifested in utility theory.

It is defined in the Green Book as: "Analysis which quantifies in monetary terms as many of the costs and benefits of a proposal as feasible, including items for which the market does not provide a satisfactory measure of economic value" (HMT 2003). CBA calls for the examination of all costs related to the production and consumption of an output, whether the costs are borne by the producer, the consumer or a third party. Similarly, the method requires an examination of all benefits, regardless of who realizes the benefits. Because the ultimate objective of CBA is the comparison of benefits and costs, they both must be evaluated in the same unit of measurement. The definition and quantification of benefits and costs depend, of course, on the stakeholders for which the analysis is performed. In the context of transportation projects, CBA is applied to alternative scenarios, one of which is the option of doing nothing or the absolute minimum (no-build, status quo). The benefits of each alternative are then valued and compared to their expected costs. The alternative for which benefits exceed costs by the greatest amount is identified as the project alternative to be suggested for implementation. If this should be the no-build scenario, then the project is rejected.

There are several principles of CBA that apply generally to all policy evaluation:

- All significant impacts should be addressed.
- Relative differences between alternative policies are often more important than absolute impacts.
- The distribution of impacts can be more important than their totals.
- A benefit or cost in the future has less value than the same one now.

The evaluation of the project will then consist of an exercise of equilibrium comparisons through which its effects on society can be assessed. Evaluating is, therefore, equivalent to analysing the different levels of social welfare achieved with each build scenario (taking into account all its implications from the beginning until all its effects wear off) compared with the no-build scenario, i.e., what would have happened if the project had not been carried out.

CBA defines various measures for evaluating the benefits and costs (financial and social) of a project. The development of costs C and benefits B, over time can be captured by the following three evaluation criteria:

- *NPV*
- Benefit–cost ratio (*BCR*)
- Internal rate of return (*IRR*)

Each of these summary measures compares the benefits of the scenario with its costs, although there are differences in definition which give each measure a different appeal.

The most straightforward CBA measure is the *NPV*, a single, synthetic criterion, which is the equivalent value in year 0 of the time stream of annual project costs and benefits (the sum of the discounted project benefits less discounted project costs). It can be expressed as the following formula:

$$NPV(r) = \sum_{t=0}^{\eta} \frac{(B_t - C_t)}{(1+r)^t}, \tag{4.1}$$

where

- η is the number of years included in the time stream (lifetime of the project or analysis period);
- r is the applicable discount rate per year;
- B_t benefits in year t;
- C_t costs in year t.

Using *NPV* as a decision rule, a project is potentially viable if the *NPV* is greater than zero which means that the total discounted value of benefits is greater than the total discounted costs. Furthermore, the alternative with the highest *NPV* should be chosen.

The *NPV* is supposed to include all relevant effects expressed in monetary terms. The actual monetary values are derived from the consumers' willingness to pay and do not require the selection of additional weights for the different cost and benefit

categories. The social discount rate used in the economic evaluation of a project should reflect the opportunity cost of the resources used in this project over time.

Decision-making under uncertainty requires explicitly considering its effect on the possible values of *NPV*, which also becomes a random variable. When uncertainty is incorporated into the evaluation, using risk analysis in addition to the expected *NPV*, the decision maker has the probability distribution of *NPV*.

The principal content of the *NPV* calculation consists of the different time-dependent weights attached to the time-displaced benefits and costs by the use of the so-called *discount factor* $(1 + r)^t$, where $r > 0$. The higher the values of r and t, the lesser the added contribution from the discounted values. The actual value of the discount rate is an expression of the emphasis on benefits in the near future as compared with benefits in a more distant future. The discount rate has a profound effect on which projects will appear beneficial. If the rate is high, benefits must be generated in short term or cost be spread out over a longer period, whereas with a low rate, benefits are allowed to balance costs over a longer period.

Discount rates are adjusted to the development of the economy and the financial markets and therefore vary over time and between countries. For example, in Denmark, the discount rate has been changed from 7 % in 2000 to 6 % in 2003. The rate varies across Europe, in the so-called HEATCO project, the rate is set to 3 % for EU assessment projects, whereas the overall rate used for Scandinavian infrastructure projects across boundaries is found to be 4 % (Lyk-Jensen 2007). When conducting a *NPV* calculation, a base year must be determined for the price level. No attention is paid to inflation, but account can be taken of forecasted growth in real terms of some of the benefit components' unit prices.

Using the *NPV* as the decision criterion implies that all projects with a positive *NPV* should be carried through. However, if there are only limited financial resources and not all projects with a positive *NPV* can be implemented, the relative value of these projects must be considered in order to rank them.

The second investment criterion is the *benefit–cost ratio* (*BCR*). It is used in order to perform a project ranking. It is defined as the present value of benefits divided by costs and is given by the following formula:

$$BCR = \frac{PVB}{PVC} = \frac{\sum_{t=0}^{\eta} \frac{B_t}{(1+r)^t}}{\sum_{t=0}^{\eta} \frac{C_t}{(1+r)^t}}, \tag{4.2}$$

where

- *PVB*—present value of future benefits
- *PVC*—present value of future costs

Projects with the *BCR* greater than 1 have greater benefits than costs; hence, they have positive net benefits. The higher the ratio, the greater the benefits relative to the costs. Note that simple *BCR* is insensitive to the magnitude of net benefits and therefore may favour projects with small costs and benefits over those with higher net benefits.

For alternatives which are independent of each other, the *BCR* can be used correctly to rank independent projects as to which are mostly cost-beneficial. Given the usual constraint of a limited budget, projects can be pursued from the highest to the lowest ratio until the budget is exhausted. However, when a selection must be made between competing alternatives that are interdependent, the *BCR* fails. Interdependence occurs when the benefits or costs of one alternative depend on whether or not certain other alternatives are also selected. Interdependence will in its most extreme result in mutual exclusivity—when selection of one alternative precludes selection of any of the others. In these cases, a combination of both the *NPV* and *BCR* approaches can be helpful.

The *IRR* is defined as that discount rate, which equates the present value of the stream of expected benefits in excess of cost to zero. In other words, it is the highest discount rate at which the project will not have a negative *NPV*. To apply the criterion, it is necessary to compute the *IRR* by solving Eq. (4.3) for *IRR* and then to compare it to the prescribed discount rate. If the *IRR* is greater than or equal to the discount rate, the project should be undertaken for its *NPV* that is non-negative. If the *IRR* is less than the rate, the project has a negative *NPV* and should not be undertaken:

$$\sum_{t=0}^{\eta} (B_t - C_t)(1 + IRR)^{-t} = 0, \tag{4.3}$$

Like the *BCR* method, this procedure is mathematically equivalent to the *NPV* method, and it always gives the same result. While the *IRR* method is effective in deciding whether or not a project is worth undertaking, it is difficult to utilize in ranking projects.

4.1.4.6 Multicriteria Analysis

If some aspects are ignored because they cannot be monetized easily, an incomplete and unbalanced evaluation of the project will result. Multicriteria analysis (MCA) attempts to consider the complete set of impacts, even if their analysis may appear more subjective and more difficult. MCA is distinguished from CBA by the fact that CBA uses money values (private, public or a combination) as the aggregation unit, whereas MCA uses a set of weights based on people's responses. These people might be common citizens, experts or political actors.

MCA establishes preferences between scenarios in reference to an explicit set of objectives that have been identified and for which performance indicators (that measure the degree to which an objective is attained) have been defined. The objectives should be

- specific,
- measurable,
- agreed,

- realistic, and
- time dependent.

These multiple objectives are often in conflict with each other and have different relative weights in the overall decision-making process.

MCA follows the same steps as CBA, but evaluates the performance of each option against each objective separately and defines a specific method for aggregating these into a ranking.

MCA methods can be classified according to the nature of the decision problem restrictions (implicit or explicit) and the nature of the results (deterministic or random). The most suitable methods for transport project evaluation are as follows:

- Simple Multiattribute rating technique (SMART);
- Analytic hierarchy process (AHP);
- Analytic network process (ANP);
- REGIME;
- ELECTRE method;
- Multiattribute utility approach.

The following description focuses on the analytic hierarchy process (AHP), developed by Saaty (1990), which is essentially the formalization of intuitive understanding of a complex problem using a hierarchical structure (Yoon and Hwang, 1995).

The AHP is a three-stage method: (i) building the hierarchy, (ii) weighting the indicators by a pairwise comparison and (iii) calculating the final value for the alternatives.

The attribute hierarchy has at least three levels, which include the focus or the overall goal of the problem at top level, multiple criteria that define the performance of project alternatives at the middle level and the competing alternatives themselves at the bottom, as shown in Fig. 4.6. The elements of the hierarchy can relate to any

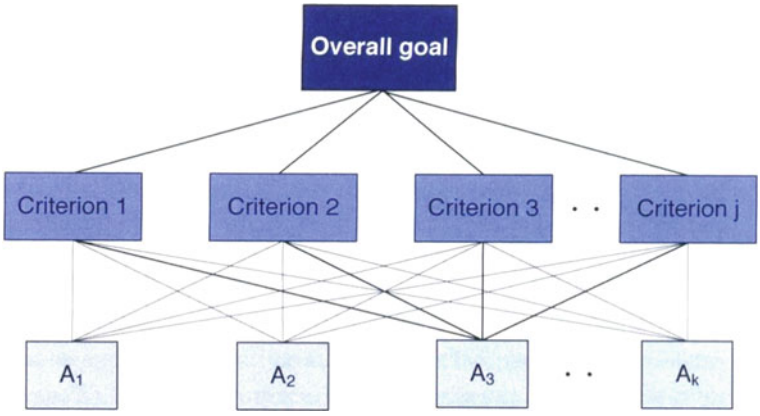

Fig. 4.6 Decision hierarchy in the AHP

aspect of the decision problem (tangible or intangible, carefully measured or roughly estimated, well or poorly understood) anything at all that applies to the decision at hand. When criteria are highly abstract, such as "landscape", subcriteria (or sub-subcriteria) are generated subsequently through a multilevel hierarchy.

Once the criteria and subcriteria have been settled, a set of weights is required. These weights represent the relative importance of the criteria, subcriteria and attributes belonging to a specific nest in the hierarchy. According to the procedure developed by Saaty (1990), these weights are obtained from pairwise comparison matrices, for each nest in the hierarchy. Once weights are available, the hierarchical structure is collapsed, following a folding-back procedure. For each alternative option, performance indicators are evaluated and the results aggregated upwards using the weights, resulting in an overall score. Based on these scores, the best alternative is chosen.

To ensure acceptance of the MCA from all stakeholders, it is important to involve them both in the construction of the hierarchy and in the derivation of weights. This can be done either by interviewing members of the affected parties directly or by examining policy statements that reflect their concerns. Saaty stresses that "special care must be taken when building up the hierarchy such that pernicious double counting of attributes is avoided" (Saaty 1990). The number of criteria should be kept as low as is consistent with making a well-founded decision. There is no rule to guide this judgment, and it will certainly vary from application to application.

4.1.5 Reference Notes and Concluding Remarks

In the chapter above, available literature was reviewed along subject matter of the topic, with the central theme—transit project evaluation. The review of literature below, in addition, aims at providing notes in order to identify useful literature for deeper study.

There are many approaches to illustrate the planning process (e.g., Federal Highway Administration 2007). The text in Sects. 4.1.1 and 4.1.2 adapts an approach described in FGSV (2001) and in Friedrich (1994).

Several researchers working in the Netherlands have made important methodological contributions in the transport project evaluation (especially in the case of MCA) (see for instance Nijkamp et al. 1990; Hinloopen and Nijkamp 1990).

Beria et al. (2012) investigated strengths and weaknesses of the two techniques—MCA and CBA—especially when assessing sustainable mobility (SM) at the neighbourhood scale and investigating the applicability of MCA and CBA to evaluate some relevant SM strategies and policies at neighbourhood scale.

The Victoria Transport Policy Institute (VTPI) publishes a number of references on enumerating and valuing the various benefits and costs of transportation programs and policies (e.g., see Litman 1999).

Also, a number of official guidelines exist, and a set of it was published for applying CBA to transport projects in practice, which was meant to raise the general level of analysis and promote uniformity in the appraisal methods used. It is the basic tool for many countries in the world (PIARC 2004; EVA TREN 2008; World bank 1996; COM—The European Commission 2007) and in Europe (HEATCO 2005; OECD, ECMT 2005). They always refer to one single common theoretical framework.

4.2 Travel Demand Models

Markus Friedrich

4.2.1 Basic Definitions and Notations

The multimodal *network* of transport infrastructures and services is represented by means of a directed graph (N, A), where N is the set of *nodes* and $A \subseteq N \times N$ is the set of *arcs* (ordered couples of nodes), each representing an small segment of user trips. Different assignment models build up the network from input data in specific ways, as shown in Part III.

Land is partitioned into a set Z of *zones*. All socioeconomic activities located in each zone $z \in Z$ are assumed to be concentrated in one single point, called *centroid*, where trips start and end. Each zone centroid in the transport network is associated with one *origin* node and with one (possibly different) *destination* node, which makes up two sets of nodes denoted $O \in N$ and $D \in N$, respectively.

Users are divided into a set G of *classes*, depending on their personal characteristics and their trip purposes. The different ways to perform transport are classified into a set M of *modes*.

The *assignment period* is discretized into η subsequent *intervals* separated by an ordered set T of *instants*. By convention, we refer to an interval as to its initial instant.

For every *pair* $od \in O \times D$ and mode $m \in M$, there is a set K_{odm} of *routes* connecting on the network origin o to destination d. Typically, each route $k \in K_{odm}$ is represented by a *path*, that is a concatenated sequence of arcs $A_k \subseteq A$. Travel demand is represented as the flow d_{odmgt} of class $g \in G$ users travelling on mode $m \in M$ that depart during time interval $t \in T$ from origin $o \in O$ directed towards destination $d \in D$, which is the generic entry of the *OD matrix* $\mathbf{d}_{ODmgt}$. The portion of this flow using route $k \in K_{odm}$ is denoted q_{kgt}.

In static models, the reference to the time interval is usually omitted. If there is only one user class, then the reference to it is omitted. In the reminder of the book, we focus mainly on one mode (public transport), whose reference is possibly omitted.

4.2.2 Models for Transport Planning

Examining current traffic states and assessing the impacts of potential measures (e.g., a new public transport line) and coming developments (e.g., petrol prices, demography) on future traffic states is the core task of transport planning. Based on the fact that traffic in transport networks is the result of various decisions taken by individuals, transport planning models replicate the decision-making process of individuals. Figure 4.7 shows transport-relevant decisions of individuals which finally lead to traffic in transport networks. These decisions range from long-term to short-term decisions. Long-term decisions cover decisions concerning the place of residence and the workplace. These decisions influence subsequent medium-term decisions regarding the purchase of a car or a season ticket for public transport, which then affect later decisions on the activity locations and the transport modes. Short-term decisions on departure time, a certain route or a certain lane are taken within a short time horizon.

For transport planning purposes, various types of models were developed and applied which typically focus on specific decision processes:

- *Land use models* forecast the distribution of the population and activity locations in a study area by modelling location choice processes.
- *Ownership models* determine the share of persons with car availability and with season tickets.
- TDMs reproduce the decision-making processes of individuals leading to movements in the transport network. In person transport, these decisions cover

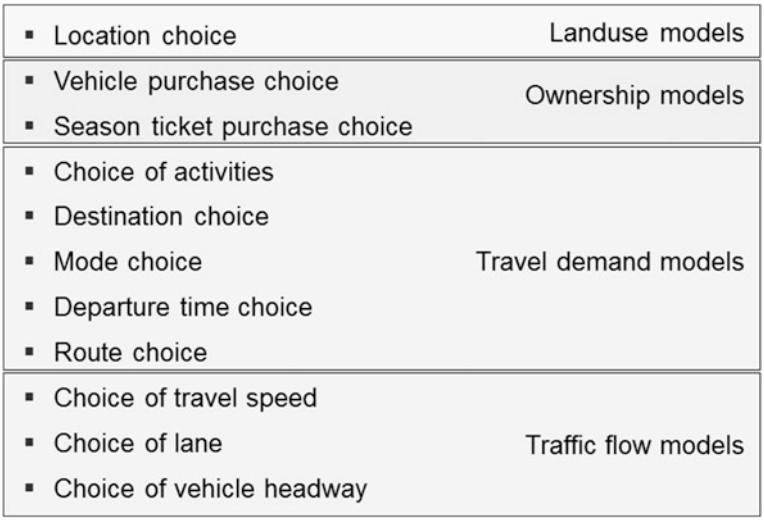

Fig. 4.7 Transport-relevant decisions of individuals

activity choice, destination choice, mode choice, departure time choice and route choice.
- *Traffic flow models* simulate the flow of vehicles and pedestrians considering the interactions between the moving objects. In car transport, traffic flow models replicate decisions of drivers describing the choice of travel speed, lane and preferred distance to the vehicle ahead.

State-of-the-art transport planning models often integrate features covering choices of different model types. For example, a TDM may include a traffic flow model to determine more accurate travel times for destination and mode choice.

Figure 4.8 shows a TDM embedded in a more comprehensive transport planning model framework. The core of the TDM is highlighted in the figure. It is part of the "impact model movement" which comprises the decision-making processes leading to movements. In a broader sense, the data models providing the input data also belong to a TDM:

- The *Data Model Mobility Behaviour* comprises all data describing the mobility behaviour of the population. Primary source of these data are household surveys with trip diaries providing information on mobility pattern of individuals (activity chains, choice of mode, distance travelled, time spent).
- The *Data Model Transport Supply*, often also called network model, comprises the data of the transport supply including the costs for the usage of this transport supply. It contains nodes and stops, links of the road and rail network and public transport lines with their timetables. Control devices such as traffic lights or vehicles with their specific characteristics (capacity, fuel consumption) can also be part of the transport supply.
- The *Data Model Land Use* comprises all data describing the spatial distribution of activity locations (locations of residence, workplaces, schools and universities, shops, leisure facilities). These data can be available on the level of individual buildings or as aggregated data on the level of traffic analysis zones.
- The *Impact Model Movement* models the impacts of a given land use and transport supply on the travel demand. It uses the land use data, the behavioural data and the transport supply data to determine the movements in the network. For this, it replicates the transport-relevant decision processes of individuals.
- The *Impact Model Transport Impacts* determines the impacts resulting from the movements of the travellers. This may include noise and pollutant emissions, accidents costs or revenues from public transport tickets and road tolls.
- The output data of the impact models form the input data of the *assessment models*. Assessment models evaluate the quality of the transport supply from the perspective of the traveller and the impact of transport on transport operators, on society and on the environment. They define a target function for the planning process.

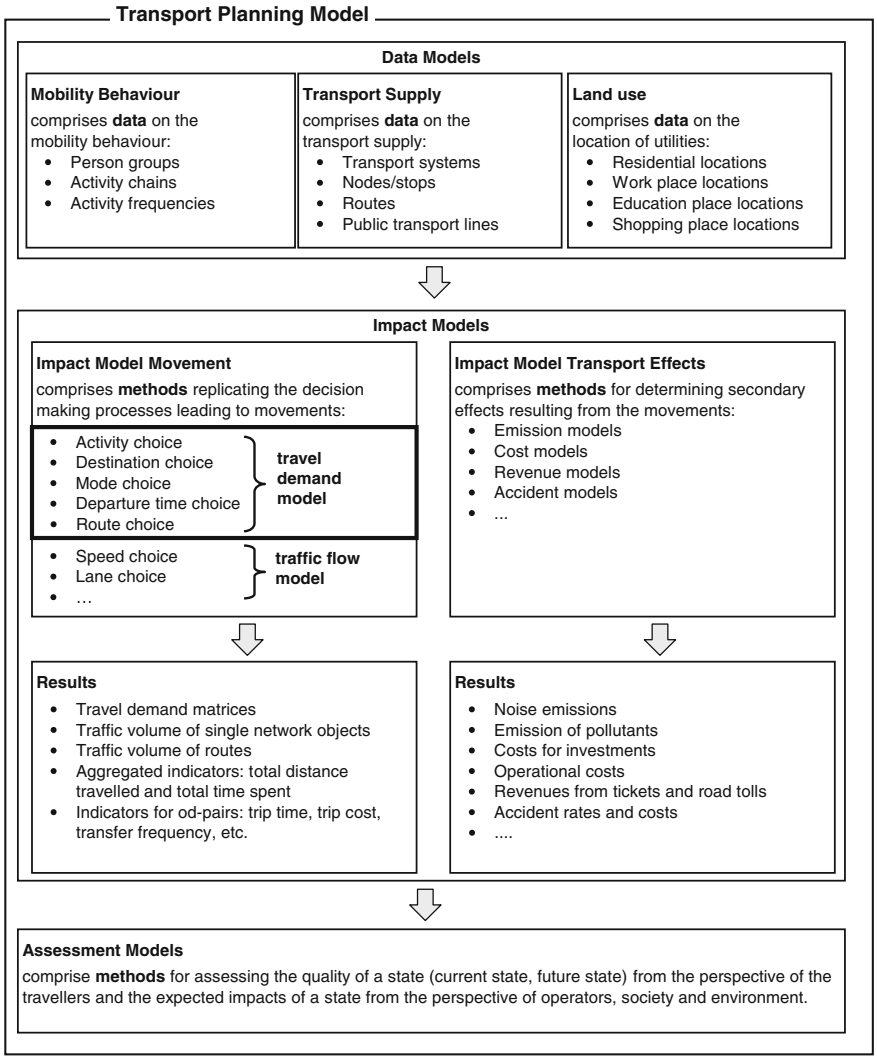

Fig. 4.8 Data, impact, assessment and optimization models as parts of a transport model (Friedrich 2011)

4.2.3 Characteristics of Travel Demand Models

A TDM provides a set of methods for determining travel demand matrices and traffic volumes in the network of a study area. For this, it describes objects and processes of the real world in a simplified and abstract way. The simplification and abstraction are based on assumptions. The underlying assumptions lead to various types of TDMs.

4.2.3.1 Behavioural Assumptions

Most TDMs are based on the following fundamental assumptions:

- Individuals have needs which require movements. Planning and performing a movement require decisions where individuals select from a set of alternatives. They assess each alternative and chose the alternative which they believe is optimal, i.e., they maximize their personal utility.
- The perceived utility of an alternative is influenced by the level of information available to the choice makers. They can have perfect or incomplete information.
- The individual choices can influence the state of the transport network. This is the case when high demand affects the traffic flow and the service quality. As shown in Fig. 4.9, this affects the utility of alternatives and the choices of other individuals.
- The travel demand and the network are in a steady-state over a long time period. This fact permits a learning process where individuals can collect information on the traffic state and adapt their choices, which finally leads to an equilibrium state. In this state, the travellers stick to their choices so that the traffic state and the resulting utility of alternatives are constant.
- Real-time traveller information systems can improve the level of information and influence choices.

4.2.3.2 Four-Step Model

Classical TDMs cover four basic decisions made by the travellers of class g which are represented sequentially in four submodels (Bates 2000):

- *Trip generation models* determine the number of trips produced (or generated) from each origin o and attracted from each destination d, denoted d_{og}^{gen} and d_{dg}^{att}, respectively;

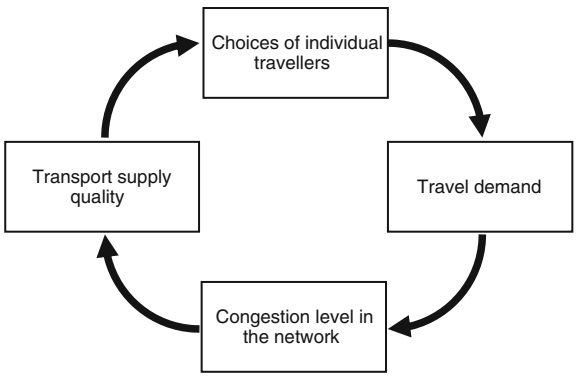

Fig. 4.9 Feedback loop between individual choices, travel demand and supply quality in a transport network

- *Destination choice or trip distribution models* determine the number of trips between origin o and destination d, denoted d_{odg};
- *Mode choice or modal split models* determine for each origin–destination pair (OD pair) the number of trips using mode m, denoted d_{odmg};
- *Route choice or assignment models* determine the number of trips on route $k \in K_{odm}$ connecting zone o to zone d by mode m, denoted q_{kg}.

Before or after the mode choice model, it may be necessary to partition the whole (daily) demand among several time slices, each one referring to a desired departure interval; the assignment model can refer to one or more of such time intervals depending on its nature being static or dynamic. In this section, we refer for the sake of simplicity to a specific assignment period of a static model (e.g., the morning peak).

As destination choice and mode choice depend on the trip purposes and characteristics of the traveller TDMs distinguish user classes with specific travel pattern and modal preferences. A user class g can refer to a specific trip purpose or origin–destination group (e.g., home–shop) performed by a specific person group (e.g., employed persons).

After executing the four steps for all classes, the elementary trip volumes q_{kg} are known for all OD pairs, modes and routes. If the elementary trip volumes traversing a link are aggregated (i.e., summed up and possibly multiplied by a specific equivalency coefficient ω_{ag}), we obtain the traffic volume q_a of this link for all modes of transport.

The result is often presented as a map of transport volumes, in which the line width is proportional to the traffic volumes of the each network element.

Figure 4.10 shows the sequence of the four-step model, the input data of each step and the resulting output data. Skim matrices describe the supply quality (e.g., travel time, number of transfers, service frequency, cost) in the transport network between OD pairs. They are generated in an initial step which is identical to the final route choice step but uses an empty network or a network with an estimated trip table.

4.2.3.3 Combining, Aggregating and Extending Stages

The classical four-step model assumes that the four stages are independent of each other. This is not true. To consider interdependencies between decisions, model stages can be combined. It is also possible to combine stages with the objective of simplifying the model structure. Or one may extend a stage to cover additional decisions taken by the traveller:

- Combining destination and mode choice:
  A traveller choosing a destination will usually reflect the available modes and the mode-specific accessibility of a destination. Simultaneous destination and mode choice models ensure that persons without car availability consider only destinations which can be reached without a car. They model the combination of a destination and a mode as one alternative with a specific utility.

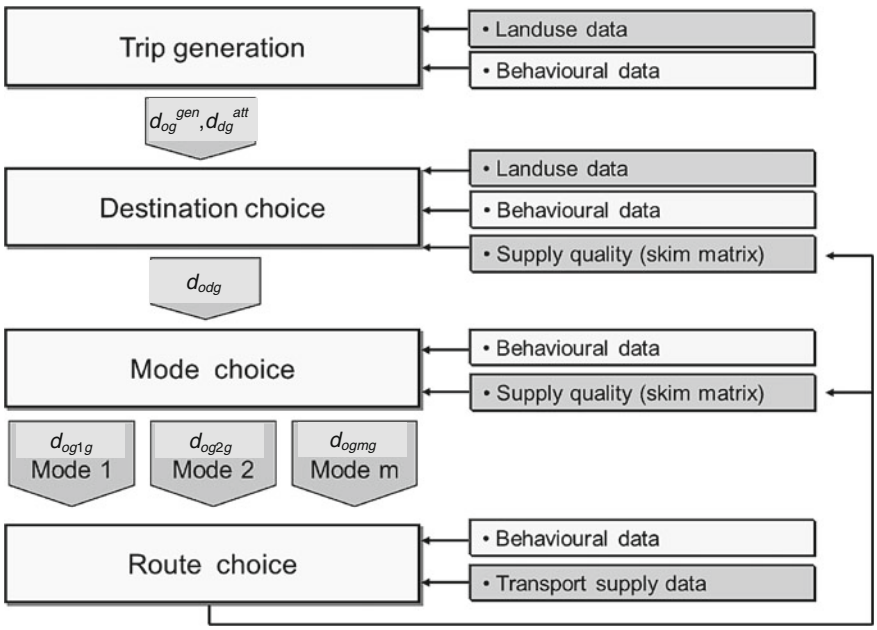

Fig. 4.10 Four-step model with the three types of input data: land use data, behavioural data and transport supply data

- Combining trip generation and mode choice:
  Assuming that the characteristics of a person, especially car availability, are the dominant factors influencing mode choice for longer trips, it can be reasonable to aggregate the steps trip generation and mode choice. This type of model is called trip-end model. It determines the share of each mode on the level of zones depending on the accessibility of the zone and the car availability of the user class. Such a simplified approach can be adequate for national models with only two modes (car, public transport) or for model applications in areas with poor public transport.
- Combining departure time choice and route choice:
  The departure time of a traveller may depend on the time-dependent travel time in the network. In public transport, travellers need to adjust to the timetable, and in car and public transport, congestion during peak hour may lead to a peak spreading, where some travellers adapt their departure time. Dynamic TDMs can cover this effect by adding a desired departure time which is connected with the time-dependent travel time to a combined utility.
- Combining submode choice and route choice:
  Some modes consist of several submodes. The mode public transport, for example, may include the submodes bus, metro and commuter rail. The choice between competing submodes can be modelled in an explicit mode choice step, or it can be integrated in the route choice step.

4.2.3.4 Feedback Between Stages

Travel time strongly influences destination choice, mode choice, departure time choice and route choice. Travel time in car transport increases in congested networks. In public transport, congestion in road networks can lead to timetables with longer running times during peak hour or to delays. Additional waiting time may be also the result of overcrowded vehicles. As the choices of the traveller determine the spatial and temporal distribution of the travel demand in the network, this results in a feedback between choices and travel time as shown in Fig. 4.11. The travel time influences the choices as travellers try to minimize the time spent in the network and the travel time itself is influenced by the travel demand.

To incorporate this feedback in a TDM, the steps trip distribution, mode choice and assignment are computed in an iterative process until a state of equilibrium is reached. A network is in equilibrium when the travel demand and the travel time do not change between two iteration steps. This requires the definition of a threshold value ε. Using a threshold value ε^{abs} for the absolute deviation and a threshold value ε^{rel} for the relative deviation, the following two conditions are tested for all objects:

$$\text{Condition 1 absolute deviation: } |y_n - y_{n-1}| < \varepsilon^{abs}, \text{ or} \tag{4.4}$$

$$\text{Condition 2 relative deviation: } \left|1 - \frac{y_n}{y_{n-1}}\right| < \varepsilon^{rel}, \tag{4.5}$$

where

- y indicator which must be in equilibrium (time or volume);
- ε^{rel} threshold value for the relative deviation (e.g., 5 trips or 10 s);
- ε^{abs} threshold value for the absolute deviation (e.g., 0.01).

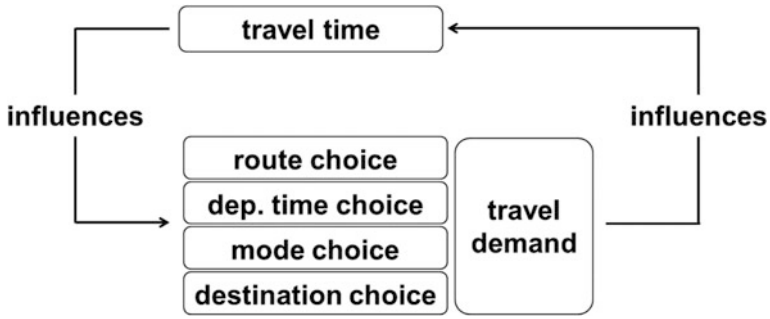

Fig. 4.11 Feedback between travel demand and travel time

The indicator y may refer to the objects links, routes or OD pairs.

Often the indicator values are smoothed to ensure or to increase the convergence in the feedback loop. Exponential smoothing applies a constant smoothing factor α and has the following form:

$$y_n^* = \alpha \cdot y_{n-1}^* + (1 - \alpha) \cdot y_n. \tag{4.6}$$

where

- y_n^* smoothed value for current iteration step n;
- y_{n-1}^* smoothed value for previous iteration step $n - 1$;
- y_n real value for current iteration step n;
- α smoothing factor with $0 \leq \alpha \leq 1$.

For $\alpha = 0.5$, this method successively determines the average between the real value and the smoothed value. A higher value of α increases the influence of the previous iteration. To ensure that the influence of the real value decreases with the number of iterations, α can be defined as a function:

$$\alpha = \frac{n-1}{n}. \tag{4.7}$$

This leads to

$$y_n^* = \frac{y_{n-1}^* \cdot (n-1) + y_n}{n} = \frac{\sum_{i=1}^{n} y_n}{n}. \tag{4.8}$$

This method is therefore known as the method of successive average (MSA).

4.2.3.5 Macroscopic Versus Microscopic Models

The terms "macroscopic" and "microscopic" describe the degree of abstraction and aggregation of objects or states used in a model. Macroscopic models aggregate and unify real-world objects and states. Microscopic models directly replicate real-world objects and states. Usually, the properties of a macroscopic object or state can be explained by properties of microscopic objects or states:

- Persons: A microscopic model explicitly models every person in a study area. A person, often called agent, is described by a set of properties (e.g., home location, workplace, car ownership). A macroscopic model distinguishes person groups and traffic zones. The sum of all agents in a traffic zone should be equal to the number of inhabitants in a macroscopic model.
- Choices, trips and volumes: Agents in a microscopic model make decisions. In a choice situation, they choose exactly one alternative out of a set of alternatives, for example one particular mode. A microscopic model therefore can identify single trips (agent, origin, destination, mode, route). In a macroscopic model, the

alternatives are chosen with a certain probability. It does not compute single trips but volumes between zones and volumes of routes. To determine the volume on a link, a microscopic model adds up all agents and a macroscopic model all routes traversing the link. Volumes produced by a microscopic model are always integer numbers (whole trips). Volumes computed by a macroscopic model are real numbers.

- Traffic state: A microscopic model determines the position and speed of every single vehicle on the link. A macroscopic model determines the mean speed of all vehicles on the link, so that in a given time interval, all vehicles have the same speed.
- Experience and memory: In a macroscopic model, the experienced travel time is stored in a skim matrix on the level of OD pairs or in an assignment on the level of routes. All persons, independent of the person group, have the same experienced travel time. In a microscopic model, every agent may store its travel time experience as a unique property in its memory.
- Transport supply: Network objects describing the transport supply can be modelled with various levels of detail. For example, a public transport stop can be described as one point or as a set of stop points referring to the precise location of a platform. The level of detail is to a certain degree independent from the model type as the network representation mainly influences the quality of the set of alternative choices. Depending on the model application, microscopic and macroscopic TDMs may work with a relatively simple network representation or may require a more detailed network representation.

Macroscopic TDMs require only one model run with feedback to produce a unique equilibrium solution. This is an important advantage compared to microscopic TDMs which, if applied correctly, require a large number of runs. As macroscopic models do not replicate individual agents, they cannot determine position of individual persons in the network and they cannot model decisions taken in the context of a household, e.g., which person uses the household's car. Modelling of delays in public transport resulting in the bunching of vehicles is also impracticable in a macroscopic model.

4.2.3.6 Trip-Based Versus Tour-Based Versus Activity-Based Models

Every day individuals perform a set of activities like working or shopping or visiting friends. They determine the set of activities and decide on the sequence of activities, e.g., first working and then shopping. They leave their home location and perform a tour involving two or more trips before returning back home. A set of activities is defined as an activity chain (e.g., home—work—shop—home). Selecting a location for every activity transforms an activity chain into a trip chain. Activity chains and trip chains can be modelled in different ways. This leads to trip-based, tour-based travel and activity-based demand models.

Trip-based TDMs determine the trips between two zones without explicitly considering the context of a trip within a trip chain. The trip chain is divided into single trips with one origin and one destination. Lohse assigns every single trip of an activity chain an origin–destination group which indicates the context of the trip through the type of activity at the origin and destination (Lohse 2011). Examples for origin–destination groups are as follows:

1. home–work
2. work–home
3. home–shop
4. shop–home
5. home–other
6. other–home
7. work–other
8. other–work
9. other–other

Origin–destination groups 1–6 are home-based groups indicating that one trip end is the home location. Origin–destination groups 7–9 are non-home-based groups. Tour-based models require a specific trip generation step ensuring that the number of produced and attracted trips is equal for each zone. They can produce inconsistent results when a varying supply quality in both directions of a movement leads to different modal split shares.

Tour-based TDMs in contrast regard the entire tour from leaving the home location until returning. They are called activity chain models as the activity chain establishes the foundation of the demand model. They model destination and mode choice in the context of the activity chain. A traveller starting a trip chain as car driver cannot return home by public transport. A trip chain, however, may start with the mode car passenger and end with the mode public transport. In this way, tour-based models intrinsically produce consistent movements.

Activity-based TDMs are extensions of tour-based models. Tour-based models use observed trip chains and their observed frequency as input in the trip generation step. Activity-based models explicitly model the activity choice of activities, their sequence and the duration of each activity (Fig. 4.12).

4.2.3.7 Static Versus Dynamic Models

Static TDMs do not have a time axis. They model one time interval, usually a day or a part of a day, assuming a constant travel demand and a constant transport supply for this period.

Dynamic TDMs have a time axis and distinguish discrete time intervals. Each time interval has a certain demand and a certain supply resulting in a time-specific traffic state. Quasi-dynamic models assume that the traffic state in one time intervals does not influence the subsequent time interval. Fully dynamic models consider this influence.

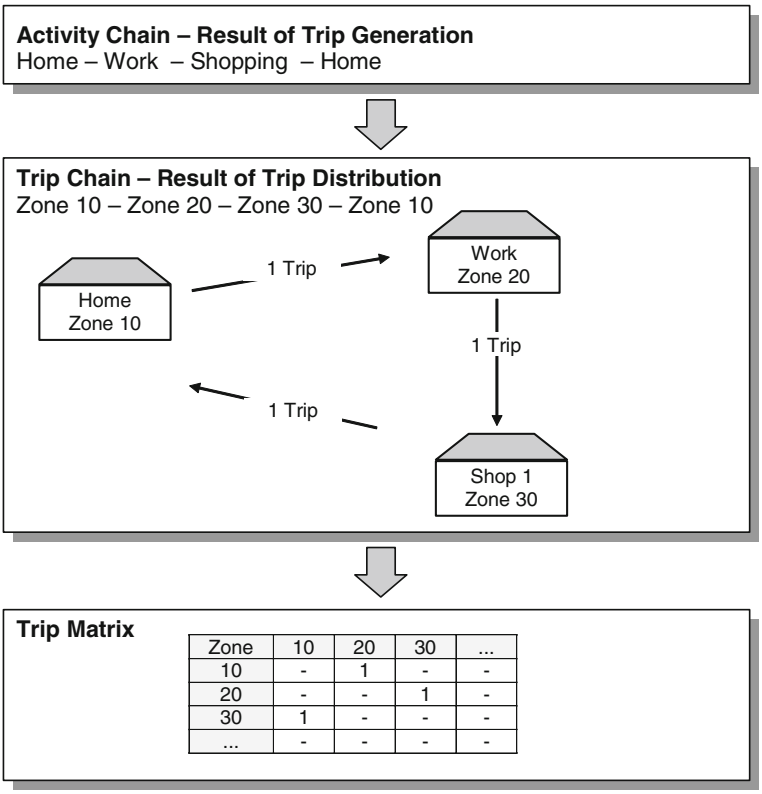

Fig. 4.12 Connection between activity chain, trip chain and trip matrix

Models serve a specific purpose which depends on the model application. The task of a model specification is to collect the technical and functional requirements a TDM shall fulfil. Based on these requirements, a model specification selects an appropriate type of demand model (behavioural assumptions, number of stages, macro or micro, trip- or tour-based, static or dynamic). Additionally, it specifies the level of segmentation for spatial objects (network objects), temporal objects (time axis), user classes and modes.

4.2.4 Model Specification

4.2.4.1 Spatial Segmentation

The spatial segmentation determines the size of traffic zones and the detail level of the link network.

Table 4.2 Number of traffic zones in a travel demand model

Model scope	No. of zones	Reference unit	Inhabitants/zone	Model application
National model	100–500	County or district (NUTS level 3)	100,000 to 1 Mio	National transport plan without assignment
National model	2000–10,000	Municipality/city district	5000–20,000	National transport plan with assignment of motorized modes
Regional model	1000–5000	Part of a municipality	1000–3000	Regional transport plan with assignment of motorized modes
Metropolitan model	500–2000	Part of a municipality	1000–3000	Urban transport plan with assignment of motorized modes
Metropolitan model	2000–10,000	Building block	50–500	Urban transport plan with assignment of motorized and non-motorized transport

- In macroscopic TDMs, trips originate and terminate at traffic zones. The number and size of traffic zones depend on the size of the study area and the requirements of the application. In the vicinity of an infrastructure measure, traffic zones should be small enough to properly feed the demand into the network. The number of zones ranges between 500 and 2000 in typical applications and reaches up to 10,000 in large or detailed applications. The following table shows the number of zones for typical model applications.
- In microscopic TDMs, the origins and destinations may refer not only to a zone but also to a building, a building block or a link.
- A road network may cover the complete road network or only the main roads. The complete network is only important for urban models. Especially in urban models, the network model should contain information on capacity and delays at intersections.
- A public transport network may cover the complete public transport supply or only the long-distance supply rail (rail, air, long-distance busses). Urban models should include all regular lines (Table 4.2).

4.2.4.2 Temporal Segmentation

Usually, a TDM describes the travel demand of a normal working day. If demand and supply vary significantly during the course of the week, as it may be the case in long-distance transport, it can be appropriate to model more than one day or an entire week.

Within a day, the temporal segmentation determines the size of the discrete time intervals in a dynamic TDM. The interval size may vary within a TDM depending

on the type of decision modelled. Destination or mode choice models may consider the traffic state on an hourly basis. Dynamic schedule-based assignment models for public transport or traffic flow models will require much shorter time intervals in the range of minutes or even seconds.

4.2.4.3 Demand Segmentation

In person transport, the attributes of a movement (origin, destination, departure time, mode) are mainly determined by the travelling person and the trip purpose.

In macroscopic models, it is therefore common practice to segregate the travel demand into user classes considering trip purposes and person groups with similar behaviour. Such a disaggregated model approach provides the following advantages:

- Trip generation: Each person group or trip purpose forms a user class with specific trip rates.
- Trip distribution: Attraction potentials and parameters describing the willingness to invest travel time for an activity can be set for each user class. Thus, it is possible replicate that employees are willing to spend more time for a trip to work than for a shopping trip.
- Mode choice: The mode choice can consider the availability of the different modes, especially the availability of private cars, for every user class.
- Departure time choice: For every activity transition (e.g., home–work or work–shop), specific temporal distributions can be applied.
- Assignment: Parameters indicating the influence of costs (tolls, fares) or transfers can be specified for every user class.

Simple disaggregated models differentiate between two home-based trip purposes HB-W (home-based work) and HB-O (home-based other) and one non-home-based trip purpose NHB-O (non-home-based other). The following tables list a rather detailed differentiation into person groups and trip purposes. The differentiation into persons with or without car availability reflects to a certain extent the income situation of the person groups. In many applications, especially outside Western Europe, it may be appropriate to explicitly consider the income situation of the person groups (low, medium and high income) (Tables 4.3 and 4.4).

4.2.4.4 Mode Segmentation

The segmentation of the means of transport into modes of transport determines the choice set in the mode choice step. For many applications on a national or regional level, it is sufficient to distinguish car and public transport. In urban models of Western countries with an integrated fare system for public transport, it is common to distinguish the four modes car, public transport, bike and walk. Separating the mode car into car driver and car passenger solves the problem of converting person trips into vehicle trips. More complex models also consider combinations of public

Table 4.3 Examples of persons groups

Person group	With car availability	Without car availability
White-collar employees	☑	☑
Blue-collar workers	☑	☑
Self-employed	☑	☑
Part-time employees	☑	☑
Unemployed	☑	☑
Houseman/housewife	☑	☑
Primary school pupils		☑
Pupils		☑
Trainees		☑
Students	☑	☑
Retired persons ≤75	☑	☑
Retired persons >75	☑	☑

Table 4.4 Examples of trip purposes

Trip purposes	
Work	White-collar employee
	Blue-collar worker
	Self-employed
	Part-time work
Education	Elementary school
	Secondary school
	Vocational school
	University
Shopping	Daily needs
	Other goods
	Shopping tour
Private business	Doctor, bank, post office
Leisure	Visiting friends at home or in hospital
	Visiting restaurant and cultural activities
	Sport activities, visiting parks
Bring/pickup	Nursery school, primary school

and private systems (Park and Ride, Rail and Fly). As this combination makes the model set-up more complicated and as it is only used by a small number of travellers, it should only be considered for model applications addressing intermodal planning tasks. The segmentation should consider specific local modes with a relevant share of demand. Examples of such modes are taxis, heavy occupancy vehicles (HOVs), three-wheelers or motorbikes and buses for specific person groups (labour buses).

The choice between various means of public transport (bus, metro) is usually not modelled in the mode choice step but in the route choice step. As travellers using

Table 4.5 Examples of modal segmentation

Means of transport	Mode of transport		Bimodal model	Multimodal model					Typical number of trips/workday
			①	②	③	④	⑤	⑥	
Car	Car		x	x	x	x			
	Car driver						x	x	1.5
	Car passenger						x	x	0.5
	Taxi								
Walk Bus Tram LRT Metro Regio Train …	Public transport				x	x	x	x	0.3
Long-distance train			x	x					
Aircraft				x					
Bike	Bike	Slow mode			x	x	x	x	0.3
Walk	Walk					x	x	x	0.7
Car Means of public transport	Park and ride							x	$\ll 0.1$
Number of modes			2	3	3	4	5	6	Σ3.3
①	National model with car and train								
②	National model with car, train and aircraft								
③	Urban/regional model with slow modes								
④	Urban/regional model with bike and walk								
⑤	Urban/regional model car driver and car passenger								
⑥	Urban/regional model with park and ride								

public transport often require more than one means of public transport to reach their destination, they choose between public transport and car and not between bus and car. Competition between various means of public transport is determined in the route choice step of the traffic assignment (Table 4.5).

4.2.5 Basic Model Formulation

4.2.5.1 Model Formulation

A TDM is a mathematical model consisting of various parameters and variables. Parameters are constant input values which are determined during the set-up of the

model. They do not change between base year and future year. Variables are input values which may change. The purpose of the model is to determine the influence of the variables on the model results.

Table 4.6 shows the equations of a simple 4-step TDM (Friedrich 2011). The steps destination, mode and route choice are based on a decision model of type logit using a logsum approach (see Sect. 4.4.1.2). The computation of generalized costs c is shown for motorized transport using a volume delay function. The generalized costs cover the travel time as only component. The example shall demonstrate the interaction between model parameters (in Greek characters) and model variables.

Table 4.6 Parameter and variables of a simple 4-step travel demand model (Friedrich 2011)

Trip generation • Trips d_{og}^{gen} of user class g produced at origin o • Trips d_{og}^{att} of user class g attracted at destination d • Value a_{zc}^{zone} of land use attribute c in zone z • Set C^{zone} of zone land use attributes	$d_{og}^{gen} = \sum_{c \in C^{zone}} \beta_{cg}^{gen} \cdot a_{oc}^{zone}$ $d_{dg}^{att} = \sum_{c \in C^{zone}} \beta_{cg}^{att} \cdot a_{dc}^{zone}$
Destination choice • Trips d_{odg} of user class g from origin o to destination d	$d_{odg} = d_{og}^{gen} \cdot \frac{d_{dg}^{att} \cdot Exp\left(\beta_g^{dest} \cdot w_{odg}\right)}{\sum_{d \in D} d_{dg}^{att} \cdot Exp\left(\beta_g^{dest} \cdot w_{odg}\right)}$
Mode choice • Trips d_{odmg} of user class g from origin o to destination d using mode m • Utility over all modes w_{odg} of user class g to travel from origin o to destination d • Value a_{mgc}^{mod} of mode attribute c for mode m and class g • Set C^{mod} of mode attributes	$a_{mg}^{mod} = \sum_{c \in C^{mod}} \beta_{cg}^{mod} \cdot a_{mgc}^{mod}$ $v_{odmg} = a_{mg}^{mod} + \beta_g^{mod} \cdot w_{odmg}$ $w_{odg} = Log\left(\sum_{m \in M} Exp\left(v_{odmg}\right)\right)$ $d_{odmg} = d_{odg} \cdot Exp\left(v_{odmg} - w_{odg}\right)$
Route choice • Flow q_{kg} of user class g on route $k \in K_{odm}$ from origin o to destination d on mode m • Utility over all routes w_{odmg} of user class g to travel on mode m from origin o to destination d	$v_{kg} = \beta_g^{route} \cdot c_{kg}$ $w_{odmg} = Log\left(\sum_{k \in K_{odm}} Exp\left(v_{kg}\right)\right)$ $q_{kg} = d_{odmg} \cdot Exp\left(v_{kg} - w_{odmg}\right)$
Traffic volume • Traffic volume q_a on arc a	$q_a = \sum_{g \in G} \sum_{o \in O} \sum_{d \in D} \sum_{m \in M} \sum_{k \in K_{odm}} \sum_{a \in A_k} \omega_{ag} \cdot q_{kg}$
Generalized link costs • Generalized costs c_{ag} of traversing an arc a with a limited capacity κ_a for users of class g • Zero flow (uncongested) travel time t_a^0 of arc a • Monetary (non-temporal) cost c_{ag}^{nt} of arc a for users of class g	$c_{ag} = \gamma_{ag} \cdot t_a^0 \cdot \left(1 + \alpha_a \cdot \left(\frac{q_a}{\kappa_a}\right)^{\beta_a}\right) + c_{ag}^{nt}$
Generalized route costs • Generalized costs c_{kg} for user class g on route $k \in K_{odm}$ from origin o to destination d on mode m	$c_{kg} = \sum_{a \in A_k} c_{ag}$

4.2.5.2 Model Parameters

Some model parameters, especially the number of zones, are determined in model specification. All other parameters result from the model calibration. It is a fundamental task of the modeller to determine the parameters:

- Number of user classes: A user class describes a particular part of the travel demand resulting from individuals with similar choice behaviour. User classes usually distinguish trip purposes and where appropriate also person groups.
- Production and attraction rates: The rates determine the number of trips in the study area. Data source are trip diaries from a household survey.
- Sensitivity parameter destination choice: These parameters determine the distance travelled and the time spent. They result from trip distance and trip time distributions derived from trip diaries which are compared with the modelled distance and time distributions.
- Sensitivity parameter mode choice: These parameters determine the modal split. They can be estimated from observed mode choice decisions recorded in trip diaries applying a maximum-likelihood estimation (see e.g., Train 2009a, b).
- Sensitivity parameter route choice: These parameters determine the distribution of the travel demand onto routes. In private transport, the parameters determine the impacts of time, gradient, road class and toll on route choice. In public transport, the parameters describe the influence of various time components (access, egress, in-vehicle, waiting), number of transfers, service frequency and fares. A parameter estimation applying the maximum-likelihood method requires observed route choice behaviour. As this is still difficult to record, especially in car transport, the parameters are often manually adjusted from experience using volume counts for the calibration.
- Parameters of volume delay function: A volume delay function describes the assumed relationship between traffic volumes on a link and the resulting travel time or delay time. The parameters of a volume delay function can be derived from measured travel times during off-peak and peak hours.

4.2.5.3 Model Variables

Model variables comprise all input values which may change due to proposed measures (e.g., a new road, a new settlement, fares) or due to external developments (e.g., higher petrol prices, changes in the structure of the population):

- Transport supply variables: Essential variables of the private transport supply are the network elements (nodes, links, turns) and their properties (speed, capacity, tolls). Important variables of the public transport supply cover the line routes, the running times, the service frequency, the capacity of the vehicles and the fares. These variables influence the quality of the supply, subsequently the generalized costs and finally the travel demand.

- Land use variables: Important variables of the land use are the number of inhabitants, their home location and other activity locations used by the inhabitants. Age and other sociodemographic properties are important to allocate the inhabitants to person groups.

Variables of the current state must be taken from existing data sources or recorded with appropriate methods. For proposed measures, variables of future states are the result of the planning process. Changes of variables resulting from external developments must be forecasted with specific models or taken from existing forecasts, e.g., a population forecast.

4.2.6 The Process for Model Calibration and Validation

The relationship between model variables (input) and computed impacts (output) is determined by the model parameters. The parameter values are obtained by comparing observed and modelled values of the current state. This process in which the model is adjusted to the reality is called calibration. In the calibration process, the set of parameters is adjusted such that the difference between observed and modelled values is minimal. This requires an optimization procedure or a manual iteration. Figure 4.13 shows the general process.

Starting from an initial choice of model parameters and observations of input and output, the model f is applied. The model passes the validation, if the model output matches observations. If not, the model parameters are improved (calibrated), using a parameter estimator f which can be either analytic or a manual procedure.

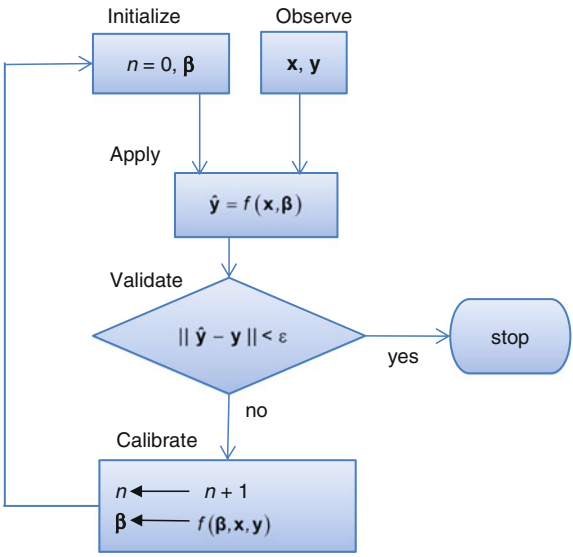

Fig. 4.13 Process for validation and calibration

Calibrating a TDM should be done at the level of each choice step:

- Trip generation: Derive parameter attraction rate by trip purpose and person group directly from household survey.
- Trip distribution: Derive parameter by user classes so that calculated trip distance or travel time distribution is similar to trip distance or travel time obtained from household survey.
- Mode choice: Apply a maximum-likelihood estimation to obtain the parameters of the utility function by user class. A maximum-likelihood estimation compares the supply quality of all modes (travel time, cost, number of transfers) to the observed choice on the level of single-choice situations.
- Travel time: Set the parameters of the volume delay function so that calculated travel times are similar to travel times for selected network sections. Distinguish delay times on links and intersections.
- Route choice: If observed route choice data are available, apply a maximum-likelihood estimation to determine the influence of travel time, length, road category, number of intersections and other factors on the utility of a route.

The model validation examines whether the model does not only replicate the observed values but also verifies and whether the model produces consistent and robust results reflecting the interdependency between transport supply and demand. It should detect fundamental errors in the network model (missing link or turn, incorrect speed, incorrect structural data), violation of constraints (e.g., capacity limits of roads, public transport vehicles or parking facilities) or deficient assumptions (e.g., the impact of prices or the level of segmentation). The following paragraph suggests tests for a validation process:

- Value comparison trip generation: Compare modelled and observed trip numbers for the study area by user class. This requires a projection of the observed trips.
- Value comparison trip distribution: Compare modelled and observed values for total kilometres travelled and total time spend per person of one person group.
- Value comparison mode choice: Compare calculated modal shares with those obtained from surveys differentiated by distance for user classes.
- Value comparison of travel time: On the level of OD pairs, compare travel time of car (free flow, congested) and travel time of public transport. Compute direct speed, i.e., the ratio of direct distance to travel time. Identify OD pairs with extreme values. Compare travel times with those from other sources. This can be traveller information systems or other networks (e.g., OpenStreetMap).
- Value comparison traffic volume: Compare calculated volumes with counted volumes at selected measurement points (e.g., links, turns, stops) or for screenlines, comprising a set of links.
- Plausibility check zones: Verify whether inflow and outflow are identical for all trips and car driver trips.
- Plausibility check network capacity: Check all overloaded network elements.

- Plausibility check spatial segmentation: Analyse the share of travel time and travel distance allotted to centroid connectors.
- Sensitivity test road network: Change the capacity on one link or add one link and compare the results of the model.
- Sensitivity test public transport network: Increase the service frequency of all lines and compare the results of the model.
- Sensitivity test demand: Increase the number of inhabitants by 10 or 20 % and compare the results of the model.
- Sensitivity test prices: Increase the price level significantly (50–100 %) and compare the results of the model.
- Visual test: Analyse traffic flows originating at selected zones or traversing selected links.

4.2.7 Reference Notes and Concluding Remarks

There are many good textbooks on travel demand modelling (e.g., Ortuzar and Willumsen 2011a, b; Cascetta 2009a, b, c; Lohse 2011), on discrete choice modelling (e.g., Train 2009a, b) and on one model calibration and validation (e.g., Cambridge Systematics 2010 or WebTag by the British Department for Transport 2014). The chapter above gives a general overview on travel demand modelling without listing the large number of publications in this field. The text above is adapted from lecture notes (Friedrich 2015) and from Friedrich (2011).

4.3 Psychological Factors Affecting Passenger Behaviour

Valentina Fini and Jens Schade

In this chapter, some basic notions about psychology are firstly given and the decision-making process is analysed. The limits of normative approaches are also outlined, and prospect theory is presented as a descriptive approach. Some applications of (cumulative) prospect theory are also introduced, with a particular interest in the comparison between its performances and those of other theories. Finally, a model of change is presented, and the steps followed by people in the process are outlined.

4.3.1 Some Basic Notions About Psychology

Psychology is the scientific study about people's behaviour and their underlying mental processes. The aims of this discipline are as follows:

- to observe and describe behaviour,
- to explain behaviour,
- to predict behaviour,
- to modify behaviour.

4.3.1.1 Modelling Human Behaviour

Human *behaviour* can be represented as a function of the person (the user of a public transport system) and of the environment (the infrastructure and traffic environment for that system). More specifically, as shown in Fig. 4.14, on the one hand, there are the *personal dispositions*, such as attitudes and routines, which are influenced by past experiences, while on the other hand there are stimulus from the environment, the so-called *situation*, given by the information, by other people's behaviour and so on. Personal dispositions and the situation are then combined in a process to determine behaviour.

Analysing more in depth the current process, Fig. 4.15 represents a simple model of human information processing, composed of these three consecutive phases:

- Perceptual encoding. The outside stimulus is perceived through different perceptual challenges and registered.
- Central processing. Perception, combined with the long-term memory, is the input for the decision-making process.
- Responding. It involves the section of a response and its execution.

As illustrated in Fig. 4.15, the attention resources are fundamental in all phases of human information processing. In particular, the attention resources influence

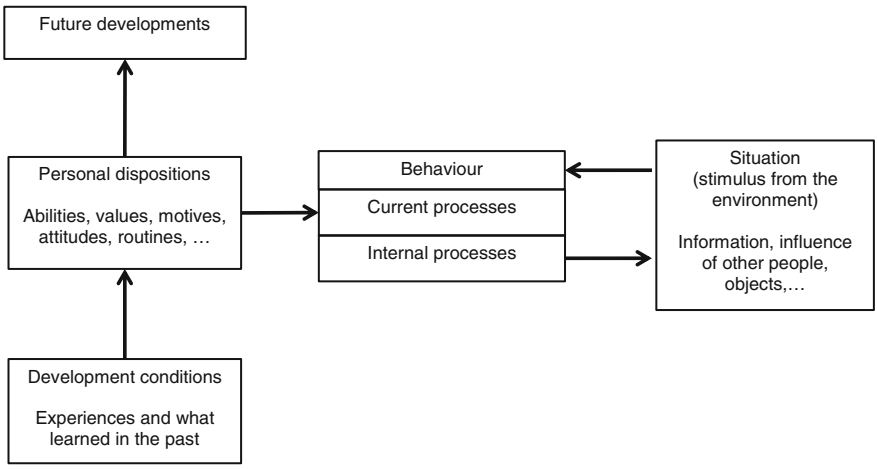

Fig. 4.14 Human behaviour as a function of personal dispositions and stimulus from the environment

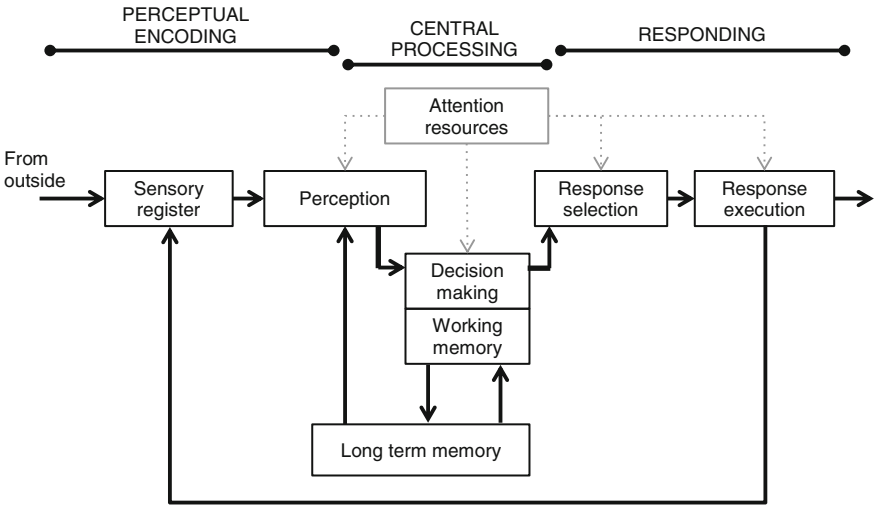

Fig. 4.15 A model of human information processing: current processes (Wickens et al. 2004)

what people are able to perceive. If they are focused on a specific issue, they may not perceive something that happens in front of their eyes. So, as shown in Fig. 4.16, there are two different processes that lead to perception:

- One possible start is the stimulus world, i.e., people perceive what happens around them through their senses. If their attention resources are focused on a specific task, they may not be able to perceive something different from what they are focused on.
- The other possibility is people's experience: If they know that something has to happen, then they can perceive it easily though they are focusing their attention on another aim.

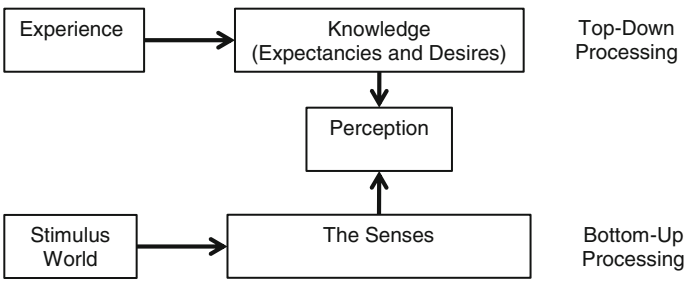

Fig. 4.16 Perception: *bottom-up* versus *top-down* processing

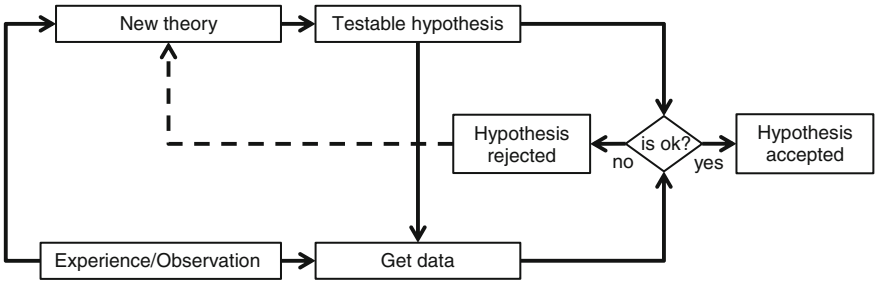

Fig. 4.17 Methodological procedure in psychology

4.3.1.2 The Methodological Approach in Psychology

The methodological procedure in psychology, as described in Fig. 4.17, is composed of the following three main phases:

- Conception of a new theory based on observations from the reality.
- Evaluation of this new theory and especially of its hypotheses, through a comparison with data collected from reality.
- Acceptance or rejection of the theory.

Experiments play an important role in the methodological procedure just described. They can be performed on the so-called experimental group. When an experiment is performed, it is possible to separate a control group from the rest of the group. Concerning the control group, the tested variables (measurement) shall not influence the results. The characterization of an experimental group and a control group can be found also in Fig. 4.18.

The importance of having a control group in the experiment is because it helps in identifying the effects produced by the tested variables and the effects produced by other factors. This in turn helps in the comparison of the "After" of the experimental group and the "After" of the control group, as presented in Fig. 4.19.

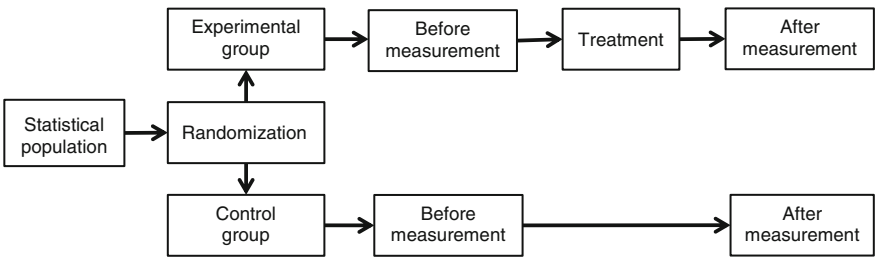

Fig. 4.18 Randomized control: group pre- to post-experimental design

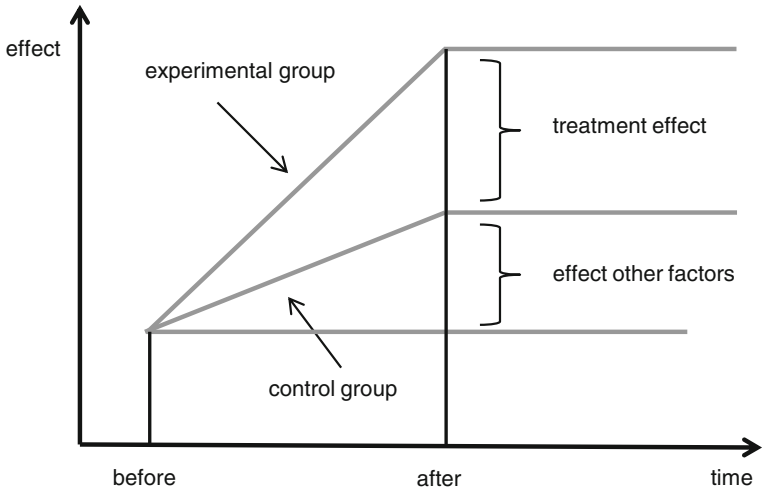

Fig. 4.19 Identifying causal effects

4.3.2 Prospect Theory: A Descriptive Approach to Decision-Making

Modelling human decisions is fundamental in all fields. In the area of transit assignment, the aim is in particular to model the user choice of network route, mode of transport, departure time, etc., for a passenger travelling from an origin to a destination for a given purpose.

Some important models to represent human choices are the expected utility models and the random utility models (RUMs). Some limitations to the expected utility model exist as there is evidence that such models cannot fit reality in the best possible way, as they assume that the decision maker is rational and perfectly informed. In RUMs, the assumption of perfect information is removed, but not that of rationality. Prospect theory is then introduced as a model that can overcome the limitations of these normative theories. The pros and cons of all these models are discussed, and the results of some implementations of prospect theory in the transportation field are also analysed.

4.3.2.1 Decision-Making

The importance of thinking can be seen in everyday life, because people take decisions, believe particular things, plan their lives and establish their aims depending on the thinking process that they have followed. Consequently, the study of thinking is necessary in all fields involving human behaviour and decisions, such

as transportation. There are two different approaches that can be adopted in the study of thinking:

- the normative approach;
- the descriptive approach.

The first approach evaluates thinking in relation to personal goals. The decision of a generic individual is the best one to achieve its goals. The second approach, instead, concerns the way people normally think and decide. The definition of the normative approach is also linked to the concept of "rationality"; a rational thinking is the one that every person, conscious of it interests, would do to reach its aims.

In everyday life, an individual faces the evaluation of the likelihood that a certain event will happen. Hence, the normative approach and the descriptive approach represent the evaluation process differently.

In addition to the thinking process underlying the judgment of people on probabilities, in everyday life, people have to undergo the decision process. Also, in this case, the normative approach and the descriptive one lead to different models.

4.3.2.2 Normative Theory of Probability

The normative theory of probability defines probability as the numerical measure of how much the person believes in a certain statement. There are three different approaches to normative probability:

- The first one examines the probability of an event as the measure of its frequency of occurrence. The drawback of this first approach is that each event may be classified in different ways; so the probability of a same event is evaluated differently depending on the classification criteria adopted.
- The second one refers to logical possibilities. This approach is useful only for exchangeable events.
- The last one evaluates the probability of an event as a personal judgment. The construction of these personal judgments has to follow appropriate methods that ensure also the coherence among them.

4.3.2.3 Descriptive Theory of Probability

The descriptive theory of probability is based on the fact that in reality, people do not follow the normative theory when they evaluate probabilities.

Evidence from reality has shown that people tend to overestimate the probability of less frequent events and to underestimate that of more frequent events. Moreover, researchers have found that in reality, people are overconfident when frequency is high and underconfident when frequency is low.

Three main reasons can explain this overconfidence:

- People fail in thinking of reasons why they could be wrong;
- People stop thinking about alternative possibilities too early, and this leads to insufficient thinking;
- People base their confidence on the evidence that they have, without thinking about their credibility.

As a consequence of the necessity of representing these features in probability judgments, some heuristics have been conceived. The representativeness heuristic and the availability heuristic are two of these; the difference among them is how people judge probabilities: in the first case by similarity and in the second case by thinking of examples.

4.3.2.4 Normative Theory of Choice Under Uncertainty

As indicated in the previous section, the normative theory of choice under uncertainty is used by people to take the best decisions to achieve their goals. The most popular and used normative theory is the expected utility theory, where decision makers are utility maximizers and the utility can be considered as the measure of goal achievement. The expected utility v_k of a certain alternative $k \in K$ can be calculated by the following equation:

$$v_k = \sum_n p_n \cdot v_{kn} \tag{4.9}$$

where n is the generic outcome, p_n is the objective probability (the occurring frequency) of the nth outcome and v_{kn} is the utility of the nth outcome. The decision maker will choose the kth alternative in the set K of all available alternatives if and only if:

$$v_k > v_h \quad \forall h \neq k \in K\,. \tag{4.10}$$

This theory is implied by the following axioms:

- Completeness. Given two alternatives k and h, we must either prefer h to k, or prefer k to h, or be indifferent.
- Transitivity. If a person prefers k to h and h to j, then he must prefer k to j.
- The sure thing principle. If two or more states of the world lead to the same outcome, then the person choice would not depend on that.

The first two points imply the weak ordering principle.

The problem is that these axioms are not often respected in reality, and as a consequence, this model cannot represent human choices in the best way.

Another normative model is the RUM. In that case, decision makers are still utility maximizers, but the utility of each alternative is composed of two different terms: a systematic utility and an error term that can follow different distributions.

The latter term represents the fact that decision makers can make a mistake in judging the utility of a certain alternative. For this reason, RUMs can fit reality better than the expected utility models.

4.3.2.5 Descriptive Theory of Choice Under Uncertainty

The descriptive theory of choice under uncertainty has its major representative in the prospect theory.

Prospect theory can explain that two effects cannot be represented through the normative models, while we observe them in reality:

- the certainty effect, i.e., the fact that people normally prefer a certain and lower win than an uncertain and higher win,
- the framing effect, i.e., the violation of the invariance principle, since the decisions depend on the way alternatives are presented to the decision maker.

Prospect theory was introduced by Kahneman and Tversky (1979), aiming to explain how and why real choices deviate from those predicted by the normative models. Prospect theory is composed of two parts, one concerning probability and another concerning utility. The main idea is that people warp probabilities and judge the utility of an alternative as variations from a so-called *reference point*.

Such distortion is represented by the so-called $\pi(p)$ function, depicted in Fig. 4.20, which expresses the judged probability for a given objective probability p.

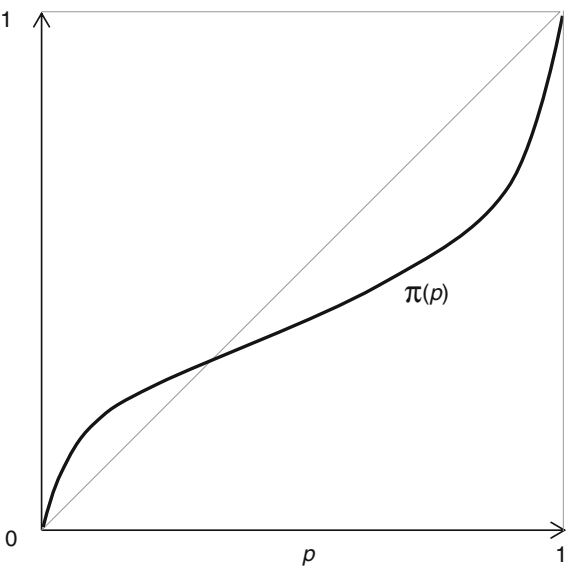

Fig. 4.20 The π function

The most important property of the π function is the overweighting of very low probabilities and the underweighting of the very high probabilities, as normally happens in reality.

This leads to the utility part: the decision maker evaluates the utility of an outcome by looking at the change from a reference point. Changing the reference point, different decisions can result in the same conditions depending on how the situation is presented to the decision maker (framing effect). The value function, which expresses the value for the decision maker of a given loss of gain, is depicted in Fig. 4.21.

The shape of this function represents loss aversion. According to prospect theory, people normally consider a loss much more seriously than the equivalent gain and take particular decisions, to avoid losses and not to obtain gains. Moreover, the fact that this function is convex for losses and concave for gains can represent the tendency of people to avoid risks in the domain of gains and to accept greater risks in the domain of losses.

The following equations represent human decision-making according to prospect theory. The expected utility v_k of a certain alternative $k \in K$ is given by:

$$v_k = \sum_n \pi(p_n) \cdot h\left(v_{kn} - v_k^0\right), \tag{4.11}$$

where v_k^0 is the reference point for the utility of alternative k. In the following, the difference between the utility of the nth outcome and the value of the reference point is denoted by $\Delta v_{kn} = v_{kn} - v_k^0$.

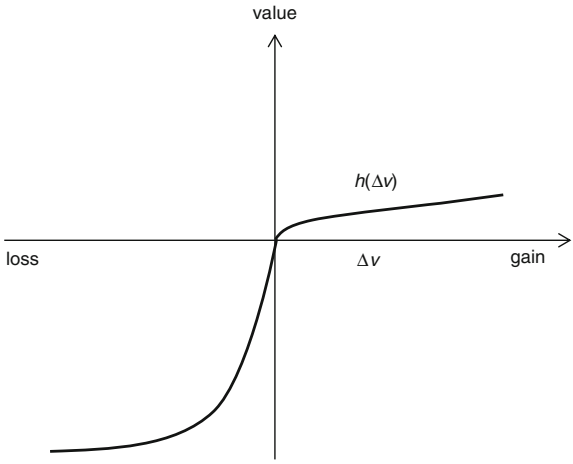

Fig. 4.21 The value function

The function $\pi(p)$ can be expressed by the following equations:

$$\pi^{+}(p) = \frac{p^{\gamma}}{(p^{\gamma} + (1-p)^{\gamma})^{\frac{1}{\gamma}}}, \tag{4.12a}$$

$$\pi^{-}(p) = \frac{p^{\delta}}{\left(p^{\delta} + (1-p)^{\delta}\right)^{\frac{1}{\delta}}}, \tag{4.12b}$$

where parameters γ and δ define the shape of the weighting function (γ in the domain of gains and δ in the domain of losses), capturing the different perception that people have when they treat gains or losses.

The value function $h(\Delta v)$ can be expressed as follows:

$$h(\Delta v) = \begin{cases} |\Delta v|^{\alpha}, & \text{if } \Delta v > 0 \\ \lambda \cdot |\Delta v|^{\beta}, & \text{otherwise} \end{cases} \tag{4.13}$$

where λ is a coefficient ≥ 1 that represents loss aversion, and α and β are two coefficients ≤ 1 that represent the diminishing sensitivity moving from the reference point in the two directions.

Also, in the case of prospect theory, an alternative will be chosen if and only if its utility is higher than the utility of all the other possible alternatives.

Finally, prospect theory can be seen as a decision process divided into two stages:

- the editing phase, in which people identify gains and losses for each option,
- the evaluation phase, in which the decision maker evaluates the outcomes of each alternative and distorts probabilities.

4.3.2.6 The Cumulative Prospect Theory

There are also other descriptive theories that we can see as the evolution of the prospect theory, like the cumulative prospect theory.

The main assumption here is that people treat outcomes jointly and weight their probabilities as cumulative. Moreover, for each alternative, the set of outcomes is partitioned between the set of positive outcomes where $\Delta v_{kn} \geq 0$ and the set of negative outcomes where $\Delta v_{kn} < 0$, and the outcomes are ordered in terms of gain (loss) from $-i$ to j. Then, the utility of a certain alternative $k \in K$ can be calculated with the following equation:

$$v_k = v_k^{+} + v_k^{-}, \tag{4.14a}$$

$$v_k^+ = \sum_{n=0}^{j} w_{kn}^+(\mathbf{p}) \cdot h(\Delta v_{kn}), \tag{4.14b}$$

$$v_k^- = \sum_{n=-i}^{0} w_{kn}^-(\mathbf{p}) \cdot h(\Delta v_{kn}). \tag{4.14c}$$

where w_{kn}^+ and w_{kn}^- can be calculated as follows:

$$w_{kn}^+(\mathbf{p}) = \pi^+\left(p_n + \cdots + p_j\right) - \pi^+\left(p_{n+1} + \cdots + p_j\right) \quad \forall n \in [0, j-1], \tag{4.15a}$$

$$w_{kn}^-(\mathbf{p}) = \pi^-(p_{-i} + \cdots + p_n) - \pi^-(p_{-i} + \cdots + p_{n-1}) \quad \forall n \in [1-i, 0], \tag{4.15b}$$

$$w_{kj}^+(\mathbf{p}) = \pi^+\left(p_j\right), \tag{4.15c}$$

$$w_{k-i}^-(\mathbf{p}) = \pi^-(p_{-i}). \tag{4.15d}$$

The cumulative weighting functions are depicted in Fig. 4.22, where we can observe the difference between the π function for losses and that for gains.

The shape of this function is very sensitive to changes in probability near the end points (0 and 1), while it is quite insensitive to changes in probability in the middle area; this property is called "diminishing sensitivity".

Also, in the case of cumulative prospect theory, an alternative will be chosen if and only if its utility is higher than the utility of all the other possible alternatives.

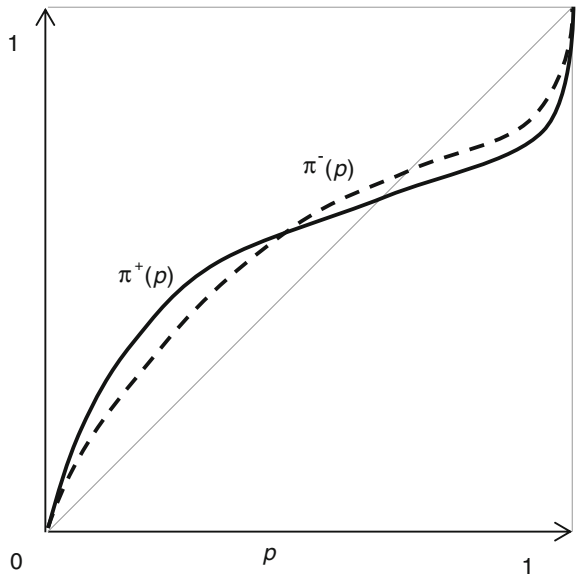

Fig. 4.22 Cumulative weighting function for gains and losses

Tversky and Kahneman (1992) analysed a stated preference data set and applied a regression procedure to estimate the parameters that we have introduced, obtaining

- $\alpha = \beta = 0.88$
- $\lambda = 2.25$
- $\gamma = 0.61$
- $\delta = 0.69$

However, the presented theories have some limitations. One of the problems of (cumulative) prospect theory is that the results of these models strongly depend on the reference point, which is often unclear and not univocal. Another problem of these theories is that their validity is often based on experimental data that could differ from reality and that depend on the willingness of respondents to fully participate in the experiment. Moreover, in these experiments, probabilities and values of outcomes are provided, which is unlikely that such information is available in transport applications. In reality, people have to choose among alternatives without having any knowledge of such information.

4.3.2.7 Regret Theory

Regret theory is another descriptive theory, and it refers to the feeling of regret that people often feel if they learn that a different choice would have led to a better outcome. In general, according to this theory, a person would choose k if and only if

$$\sum_{n \in N} p_n \cdot r_{khn} > 0 \quad \forall h \tag{4.16}$$

where r_{khn} is the regret (or rejoice) of alternative k compared to alternative h.

We just note now that a person, according to regret theory, would choose an alternative that has an intermediate performance, while, according to expected utility theory, he would choose an alternative that is very good in some attributes but not in others (Ramos et al. 2014).

4.3.2.8 Pros and Cons of Different Modelling Approaches

A schematic comparison between the main models that can be used to represent human choices is presented below, to sum up what has been said until now. The regret theory is not considered in this case because of the low number of studies that have been done in this field.

Expected Utility Models

PRO:

- Much research has been done
- Approximate representation

CONS:

- Axioms that in reality are violated
- People in reality have a bounded rationality and they are not perfectly informed

RUMs
PRO:

- Much research has been done
- Better representation than the one given by EUM

CONS:

- People in reality have a bounded rationality and they are not perfectly informed

Prospect Theory/Cumulative Prospect Theory
PRO:

- Loss aversion
- Underweighting of high probabilities/overweighting of low probabilities
- Representation of framing effect
- Representation of certainty effect
- Diminishing sensitivity for gains and losses

CONS:

- Great dependence of the results on the reference point chosen, whose definition is often unclear
- Its validity is based on experimental data (stated preferences); probabilities and values of outcome are provided.

This scheme shows that each model has its pros and cons and no model is ideal. However, the potentiality of the (cumulative) prospect theory suggests investigating this field more.

4.3.2.9 Non-compensatory Choice Models

So far, we have examined compensatory choice models, where a utility is associated with each alternative.

Some non-compensatory choice models:

- conjunctive model: the alternative must meet some minimum requirements;
- disjunctive model: if the alternative does not meet the minimum requirements, it must have features that will compensate the loss;
- lexicographic model: the alternatives are ranked with respect to each attribute and are then compared using the first rankings;
- elimination by aspect model: the attributes are ranked by importance and minimum required values are defined.

4.3.3 Application of Prospect Theory in the Transportation Field

The next step is to focus on some applications of (cumulative) prospect theory in the transportation field.

One of the applications of (cumulative) prospect theory can be found in modelling travellers' route choice; in this case, we mention three possible definitions for the reference point (Gao et al. 2010):

- as the free flow travel time,
- as the habitual travel time,
- as the expected travel time.

This is an example of one of the cons pointed out in the scheme, i.e., the definition of the reference point is not univocal.

One of the studies that have been done about the application of cumulative prospect theory to route choice modelling is the one by Jou and Chan (2013), where a survey among freeway drivers in Taiwan was conducted. The survey was composed of three different parts:

- The first one was about socioeconomic characteristics of the driver.
- The second one was about the driver's trip characteristics.
- The last one was about the experimental scenarios; the area was divided into three sections, with each section having its own switching point.

In the second part of the survey, drivers were also asked to choose between three different values for the average travel time experienced in the three sections. The following information represents the reference point.

This study is particularly interesting because the parameters of cumulative prospect theory are estimated from the empirical data, unlike other studies which use the parameters estimated by Tversky and Kahneman (1992). This estimation has been done through maximum likelihood of a logit formulation (see Sect. 4.4.5) where the systematic utilities are computed through (4.14a, b and c). The results are shown in Table 4.7.

It is possible to conclude from the estimated values that

- Drivers who usually encounter traffic congestion on freeways are more sensitive to travel time savings (gains).
- Drivers who habitually use the freeway are more loss sensitive than drivers on business trips, because they include a buffer time when they plan their trip. If drivers are on leisure trips, they are more relaxed.
- Drivers who require precise predictions of real-time traffic information are more loss averse, while people who usually use the freeway are less loss averse.
- Drivers whose age is over 51 are risk sensitive.
- Drivers whose trip purpose is a visit are more risk sensitive.
- Drivers whose trip purpose is leisure are risk sensitive too.
- Drivers who usually encounter traffic on the freeway are risk sensitive too.

Table 4.7 Estimated parameters

Parameter	Estimate (t-value)
α (gains in the value function)	
Constant	0.92 (2.79)
Traffic usually encountered on freeway	0.86 (1.55)
β (losses in the value function)	
Constant	1.04 (9.71)
Business purpose	−0.34 (−2.40)
Leisure purpose	−0.17 (−1.53)
λ (loss aversion)	
Constant	8.66 (1.99)
Prediction of real-time traffic information required	0.004 (1.78)
Habitually users of the freeway	−4.90 (−1.96)
γ (gains in the weighting function)	
Constant	0.80 (6.97)
More than 51 years of age	−0.67 (−4.39)
Visit purpose	1.43 (2.30)
δ (losses in the weighting function)	
Constant	0.66 (12.96)
Leisure purpose	−0.12 (−2.13)
Traffic usually encountered on freeway	−0.18 (−3.12)
Noise	
Constant	3.35 (2.13)
Feasible log pseudo-likelihood	−4217.4
Log pseudo-likelihood	−2021.31
Samples	3184

In conclusion, the authors have found that this cumulative prospect model, with the successfully estimated parameters, can represent the Taiwanese freeway drivers' attitude very well, while the expected utility theory axioms are violated. Though successful results are obtained, it is important to underline that some estimated values for β and λ do not respect the constraints that some authors, like Timmermans (2010), associate with the value of these parameters.

Another study has been conducted by Ramos et al. (2010). They based their study on a survey conducted among drivers who had to make 40 consecutive choices among 3 different routes, in which real-time information is not necessarily available. They applied the cumulative prospect theory to model the data collected by this survey in two different frameworks, by taking into account heterogeneity and not taking into account the heterogeneity. Heterogeneity is the fact that different decision makers can consider either different factors or the same factors in different ways to make their decisions, thus considering different reference points. The results of this study show that the cumulative prospect model is particularly able to predict route choice behaviour taking into account heterogeneity.

One more study has been conducted by Huang (2013) applying cumulative prospect theory, to model the choices among managed lines and toll-free alternatives. Moreover, a survey was conducted asking drivers their choices among a toll line and a toll-free line. One of the aims of this study was the estimation of users' willingness to pay for time saving. Both a logit model and a cumulative prospect model have been estimated through the collected data. The results show that the descriptive model predicts more than the 65 % of the total choices correctly, while the normative one predicts only 35 %. As a result, prospect theory is shown to perform better than the normative approach.

Xu et al. (2010) presented another interesting research. They conducted a survey about route choices and calibrated a cumulative prospect model defining the values of parameters through the data collected. They introduced for the first time the idea of setting the reference point with an optimization problem, as the definition of the reference point is one of the main problems in the application of prospect theory in route choice modelling. This study could suggest a possible way to overcome it.

The reference point that they wanted to consider was the time people reserved for their trip. However, because of the huge number of OD pairs, paths, travellers and trips, a uniform method was needed for its definition. Given a trip from o to d, the probability p_{od} is the maximum likelihood of achieving an effective trip, which satisfies the following optimization problem:

$$p_{od} = Max(p_{odk}, \forall k \in K_{od}) \quad \text{subject to: } Pr(t_k \leq t_{od}) \geq p_{odk}, \forall k \in K_{od}, \tag{4.17}$$

where $k \in K_{od}$ is a generic path between o and d, t_k is the travel time on path k and t_{od} is the reserved time for the trip (the reference point). If we assume that t_k follows a probability distribution, the optimization problem used to define the reference point is the following:

$$t_{od} = Min(\omega_k, \forall k \in K_{od}) \quad \text{subject to: } Pr(t_k \leq \omega_k) \geq p_{od}, \forall k \in K_{od}. \tag{4.18}$$

It is important to underline that p_{od} is related to the commuter's requirement of travel time reliability. As a consequence, the method proposed by Xu et al. (2011) to set the reference point can also take into account the framing effect and the differences among people, thanks to the assignment of different values to p_{od}.

The authors also used this approach in a simple situation in which people choose between two different routes, with the aim of clarifying the new method: different reference points correspond to different p_{od}. Travellers with a greater reliability requirement reserve more time for the trip and prefer less risky routes.

The final conclusion of this study is that this descriptive model, with the successfully estimated parameters and with the setting of the reference points through the optimization problem, captures reality better than expected utility theory.

Another study is about adapting route choice modelling. In that case, it is assumed that travellers do not choose a priori, a route, but a routing policy, a strategic plan that we can represent as a hyperpath. Gao et al. (2010) created four different models: a cumulative prospect theory routing policy choice model, a

cumulative prospect theory non-adaptive path model, an expected utility model for routing policies and an expected utility model for non-adaptive path choices. The comparisons among the models suggest that the two cumulative prospect models can capture risk seeking under high probabilities and risk averse under low probabilities, while the expected utility models cannot. At the same time, it is also concluded that the cumulative prospect theory routing policy choice model captures the diversions during the trip, differently from the other descriptive model estimated.

We can now focus on the application of the (cumulative) prospect theory in modelling public transport. A first interesting study has been conducted by Binetti et al. (2005). They calibrated two different models based on data collected by survey: a logit model and a cumulative prospect theory–logit model that is a multinomial logit where the probabilities used to calculate the systematic utility are cumulative weighted probabilities. In that survey, people had to choose between two different bus lines, and the various scenarios were different for the travel time, the waiting time and the variance associated with the attributes. The results show that the logit model represents better the transit users' choices. The only scenario in which the CPT–logit performs better is the one created with the intention to emphasize the variability of the attributes.

Another study was conducted by Ceder et al. (2013), and it focuses on the effects that transfer routes have to transit users' behaviour. A survey was proposed to transit users, who had to choose between one direct route and two time-saving transfer routes in different scenarios characterized by different degrees of reliability of the service, different transfer walking times and different transfer waiting times. The analysis of these data showed that people judge a route as the most attractive if it has the lowest variability out-of-vehicle times. They created two different models, a fuzzy logic and a cumulative prospect model, based on the data collected through this survey and excluding the responses given for the direct route. The results show that both fuzzy logic and cumulative prospect theory can perform similar to the reality, though the former performs a little bit better. In that case, they also underline the strong dependence of the results of the cumulative prospect theory on the reference point and of the results of the fuzzy logic on the membership functions and fuzzy rules.

In conclusion, prospect theory has a great potentiality in representing human choice because it can capture features that the normative models do not consider.

Its application in the transportation field is quite new, so further research is needed to understand the role that this descriptive approach can play in transportation modelling. However, we can say, based on studies so far, that prospect theory can represent the reality better than the normative models, especially when real-time traffic information is provided to the users. The use of models in which the normative approach and the descriptive one are mixed should also be investigated more in the future.

On the other hand, few studies have been done to model transit users' behaviour through prospect theory, and the results from the comparison between the prospect model and a different approach have shown that the latter fits reality better. Despite

these results, it cannot be concluded at this moment that the use of prospect theory to model transit passenger behaviour has to be avoided. More studies should be done.

4.3.4 Modelling Human Behaviour: Transtheoretical Model of Change

Representing people's changes in behaviour can be very useful in transportation, to aid in understanding how to control users' demand. For example, everyone knows that too many people use a car to move, but what can we do to change the situation? Of course a model that represents how people change their choices could be fundamental to identify the best policies to be implemented.

The transtheoretical model of change is based on the idea that there are six different stages of change:

1. Precontemplation. At this stage, people are not conscious of the problem and do not want to change behaviour in the predictable future.
2. Contemplation. People have started to understand that changing behaviour could bring advantages, so they want to change in the near future.
3. Preparation. People will change behaviour in the immediate future: their plan of action is ready, also thanks to the previously undergone experiences.
4. Action. The changing process has started.
5. Maintenance. The change needs to be maintained for at least six months.
6. Termination. The new behaviour can be considered as fully established.

Contemplation is of course one of the most important phases of the change process because it is the step in which the person starts to think about a possible change because he or she starts to be dissatisfied with his own behaviour. As a consequence, it is possible to understand that the transition from the precontemplation phase to the contemplation one plays a key role in the change process. So, according to this theory, if we want to change people's behaviour, first of all we have to make them dissatisfied with their own views and practices. At the same time, the attractiveness of the alternative that we want people to choose plays a fundamental role. In fact, during the contemplation, people start to collect information on the alternatives to their current choice: if no alternative is judged well enough, then the change process ends and people do not change behaviour.

As an example, we can now analyse more in depth which aspects can encourage or discourage people from using public transport. In particular, the journey to and from the bus stop/terminal plays a key role in the attractiveness of the alternative we are talking about; it is possible to identify these four different factors that influence people's willingness to walk:

- Accessibility. The greatest length that can be considered as acceptable for a walking journey is around 1 km, so it is really important planning for pedestrians trying to reduce their deviations.
- Convenience. To make the journey more convenient, there is the necessity of creating a pedestrian network that is attractive and lacking of unnecessary obstructions.
- Safety. People have to feel safe and secure during the journey and also at the terminal or bus stop; they have to walk and wait, neither with the fear of being involved in a crash, nor with the fear of being attacked by someone.
- Aesthetics. Transport facilities should be a pleasant place where to spend time (e.g., waiting at stops). They can even become artworks, like for the case of Naples underground.

Returning to the importance of the transition from the precontemplation phase to the contemplation phase, an effective message can play a fundamental role in making this change happen. In particular, there are these seven elements that make a message effective:

- Attractiveness. The message needs to be captivating.
- Clearness. The message needs to be easily understandable.
- Consistency. The message needs to be used over and over again.
- Credibility. The message has to be believable.
- Persuasiveness. The message needs to be able to make people change behaviour.
- Pertinence. The message needs to concern a real problem.
- Trustworthiness. Everyone should be able to achieve the suggested alternative.

In conclusion, when there is the necessity of making people change behaviour, the first steps are identifying the target group, studying the expectations and needs of people who belong to the identified group, and making them aware and dissatisfied with their current choice. In this contest, the transtheoretical theory is a useful tool to understand the change process and how to make people change habits.

4.3.5 Reference Notes and Concluding Remarks

This chapter aims at giving some information about human psychology and introduces two new theories that represent human behaviour:

- the prospect theory that represents the decision-making process according to the revealed features of human's behaviour,
- the transtheoretical theory that represents the entire change process.

In both cases, the aim is to represent reality as best as possible, though human behaviour is often irrational and contradictory. As a consequence, these theories can

be also very useful to understand how people choose and how they change their choices.

The following reviews the literature on which the chapter is based.

In particular, the books "Thinking and deciding" (Baron 2008), "Choices, values and frames" (Kahneman and Tversky 2007) and "Thinking and reasoning" (Manktelow 2012) investigate human psychology and talk about prospect theory. The transtheoretical model, instead, can be found in the book "Treating addictive behaviour" (Prochaska and Di Clemente 1986).

The other reference notes that can be found below are the articles introduced in this chapter written by researchers in this field, where the results of their experiments and of the use of prospect theory to model human behaviour are presented. In particular, the use of prospect theory to model route choice can be found in Jou and Chen (2013), De Ramos et al. (2013) and Huang (2013), Xu et al. (2011). On the other hand, Binetti et al. (2005) and Ceder et al. (2013) deal with the use of prospect theory to model and better understand transit users' behaviour.

4.4 Discrete Choice Models

Sebastian Raveau, Emilio Picasso and Maria Nadia Postorino

This book is not intended to provide an exhaustive review of RUMs, but only to introduce the reader to the most popular modelling approaches. We refer to Ben-Akiva and Lerman (1985) for a more extensive elaboration of the models introduced. Furthermore, the basics of the estimation and validation of discrete choice models are presented.

Obtaining the necessary parameters for the application of the models is the first step in urban public transportation planning. The contents and methods presented in this chapter can be extended and applied to any modelling approach covered in this book, in particular mode choice and route choice models. In Sects. 4.4.1–4.4.1.3, a microeconomic approach to aid in understanding travellers' decisions and the most widely used discrete choice models are presented. In Sect. 4.4.1.3, some extensions to the traditional discrete choice models to address correlation and heterogeneity are described. Finally, in Sect. 4.4.5, the basic methods to validate the models are discussed.

Let us consider a rational individual u that deterministically chooses an alternative k, within a discrete choice set K_u maximizing his own perceived utility. The utility u_{uk} that he/she associates with the generic alternative $k \in K_u$ is typically a function of objective attributes related to the level of service (e.g., travel times, fares and transfers) and socioeconomic characteristic of the individual (e.g., income level, gender and age). It is also assumed that the analyst, who is just an observer without perfect information, is only able to define/observe a systematic utility v_{uk}. Thus, it is necessary to associate an error term ε_{uk} to each alternative, as shown in Eq. (4.19):

$$u_{uk} = v_{uk} + \varepsilon_{uk}. \tag{4.19}$$

As the individual is assumed rational, alternative k will only be chosen if $u_{uk} \geq u_{uh}$ for all alternatives h that belong to K_u. Therefore, as the analyst can only observe the systematic utility, only a choice probability can be obtained. As shown in Eq. (4.20), this probability p_{uk} depends on the distribution of the error terms ε_{uh}. Different assumptions regarding the distribution of the error terms will result in different discrete choice models, yielding the following probability for choosing alternative k over all other alternatives:

$$p_{uk} = Pr(v_{uk} + \varepsilon_{uk} \geq v_{uh} + \varepsilon_{uh}, \forall h \in K_u). \tag{4.20}$$

The expected value w_u of the maximum perceived utility is called *satisfaction*:

$$w_u = E(Max(v_{uk} + \varepsilon_{uk}, \forall k \in K_u)). \tag{4.21}$$

Probabilities and the satisfaction can be expressed in general through the following integrals, given the systematic utilities and the probability density function $\varphi(\boldsymbol{\varepsilon})$ of the jointly distributed error terms:

$$p_{uk} = \int_{E_{uk}} \varphi(\boldsymbol{\varepsilon}) \cdot d\boldsymbol{\varepsilon} = \int_{\varepsilon_{uk}=-\infty}^{+\infty} \cdots \int_{\varepsilon_{uh}=-\infty}^{v_{uk}-v_{uh}+\varepsilon_{uk}} \varphi(\boldsymbol{\varepsilon}) \cdot d\boldsymbol{\varepsilon}, \tag{4.22}$$

$$w_u = \sum_{k \in K_u} \int_{E_{uk}} (v_{uk} + \varepsilon_{uk}) \cdot \varphi(\boldsymbol{\varepsilon}) \cdot d\boldsymbol{\varepsilon} \tag{4.23}$$

$$E_{uk} = \{\boldsymbol{\varepsilon} : v_{uk} + \varepsilon_{uk} \geq v_{uh} + \varepsilon_{uh}, \forall h \in K_u\}, \tag{4.24}$$

where E_{uk} is the set of points in the space of error terms where alternative k has the maximum utility.

Traditionally, the systematic utility is proposed to be a weighted summation of the different attributes that influence the decisions of travellers. Let us denote the different attribute values by a_{ukc} where c represents a particular attribute. This way, the systematic utility can be expressed by Eq. (4.25), where β_c are the marginal utilities of the attributes (parameters to be estimated).

$$v_{uk} = \sum_{c \in C} \beta_c \cdot a_{ukc}. \tag{4.25}$$

By substituting Eq. (4.25) into Eq. (4.20), it can be seen that the probability of choosing a particular alternative k will depend on

- the level of service of all alternatives, measured through all a_{ukc},
- the parameters β_c and
- the assumptions regarding the distribution of the error terms ε_{uk}.

In the following, for the sake of simplicity, we will omit the reference to the individual u when not necessary.

4.4.1 The Logit Model

Some properties of the logit model are briefly described here. Principles and features of this model are described in much more detail in other literatures such as Ben-Akiva and Lerman (1985) and Louviere et al. (2000).

The logit model belongs to the family of RUMs that have been used extensively to deal with travel decisions. For computational reasons, the logit model is the most frequently used among alternative RUMs. The RUM framework assumes a decision structure with choices among a number of (potentially large) discrete travel alternatives that are mutually exclusive. RUMs are consistent with the assumption that every traveller maximizes his own utility, but allow for random influences on travellers' decisions that are outside the scope of the model. The result is that the model predicts the probability by which a certain alternative is chosen and not the choice itself. An estimate of the number of times an alternative is chosen is given by aggregating the choice probabilities over decision makers.

4.4.1.1 Basic Formulation

The simplest version is the standard multinomial logit model (MNL). The specific MNL assumption is that $\boldsymbol{\varepsilon}$ is an independently and identically distributed (IID) Gumbel random variable for all alternatives. Then, the familiar logit model equation can be derived (see Ben-Akiva and Lerman 1985, for more details):

$$p_k = \frac{Exp\left(\frac{v_k}{\theta}\right)}{\sum_{h \in K} Exp\left(\frac{v_h}{\theta}\right)}, \tag{4.26}$$

where θ is a scale parameter which is proportional to the standard deviation of ε_k (in the MNL, the error associated with each alternative $k \in K$ has a same distribution, although independent):

$$\theta = SD(\varepsilon_k) \cdot \frac{\sqrt{6}}{\pi}, \tag{4.27}$$

If the systematic utilities in Eq. (4.25) are substituted into the probabilities given by Eq. (4.26), it can be seen that all parameters β_c will be scaled by θ. Here, the subscript c ranges over different attributes contributing to utility, e.g., time and cost. It is not possible to estimate both types of parameters separately, but only their ratio. Therefore, the scale parameter θ is omitted and included in each parameter β_c.

An interesting property of the MNL model is that the probabilities do not depend directly on the absolute value of the systematic utilities, but rather on their absolute differences:

$$p_k = \frac{1}{1 + \sum_{h \in K-k} Exp\left(\sum_{c \in C} \beta_c \cdot \left(a_{hc} - a_{kc}\right)\right)}. \tag{4.28}$$

The following example illustrates the basic principles of the MNL model. More extensive case studies are described in Sects. 4.5.4 and 4.5.5.

The objective was to estimate the demand for a park and ride facility in the northern access to Buenos Aires city in Argentina. A stated choice experiment was implemented via online survey. Commuters were presented with four alternatives:

- car-to-destination,
- train,
- charter (small bus with comfortable service),
- park and ride (large parking facility located in the border of the city, connected to the centre via subway).

Each alternative was characterized by three attributes:

- price (toll, fare and parking rate) measured in Ar$/trip,
- running cost (fuel and car maintenance) measured in Ar$/month,
- travel time measured in minutes/trip.

The sample size was 150 interviews, each one containing 15 scenarios with varying attributes following a predesigned experimental plan. The structure of the MNL model includes an alternative specific constant (ASC) for each mode (except car-to-destination, selected as the reference) and a generic coefficient for each attribute. The systematic utility definitions for the four modes are shown in Eqs. (4.29a, b, c, d):

$$v_{car} = \beta^{price} \cdot a_{car}^{price} + \beta^{run} \cdot a_{car}^{run} + \beta^{time} \cdot a_{car}^{time}, \tag{4.29a}$$

$$v_{p\&r} = \beta^{price} \cdot a_{p\&r}^{price} + \beta^{run} \cdot a_{p\&r}^{run} + \beta^{time} \cdot a_{p\&r}^{time} + \beta^{p\&r}, \tag{4.29b}$$

$$v_{train} = \beta^{price} \cdot a_{train}^{price} + \beta^{run} \cdot a_{train}^{run} + \beta^{time} \cdot a_{train}^{time} + \beta^{train}, \tag{4.29c}$$

$$v_{char} = \beta^{price} \cdot a_{char}^{price} + \beta^{run} \cdot a_{char}^{run} + \beta^{time} \cdot a_{char}^{time} + \beta^{char}. \tag{4.29d}$$

Table 4.8 Results of MNL model estimation

ASC park and ride	$\beta^{p\&r}$	−0.211
ASC train	β^{train}	−1.740
ASC charter	β^{char}	−0.576
Price	β^{price}	−0.0267
Running cost	β^{run}	0.000114
Time	β^{time}	−0.0166

The results of maximum-likelihood estimation are shown in Table 4.8.

The negative ASCs mean that commuters prefer the car-to-destination to all other modes (all other attributes being equal). Park and ride comes second, and train is the least preferred. Price and time coefficients are statistically significant and negative, as expected for disutility causes. The coefficients for attributes cannot be compared to each other because they are in different units. Each Ar$ of price values 0.0267 units of utility, hence it takes 65 Ar$/trip price difference to persuade the average individual to travel by train instead of car. The running cost coefficient is not statistically significant, meaning that either of the commuters does not care about it, or this basic MNL model is not able to detect its influence. The WTP (willingness to pay) for the travel time can be calculated by dividing the coefficients for time and price: 0.62 Ar$/min.

The scale parameter is not identified. All coefficients include it as an unidentified factor. The model can be used to forecast the probability of choice for park and ride in different scenarios. The demand curve is obtained by forecasting it at different rates keeping the rest at actual values.

This model provides basic understanding of the problem; however, it has various limitations. It assumes that all individuals have the same preferences and alternatives are not correlated, whereas car-to-destination and park and ride are probably closer to each other than to public transport. These shortcomings can affect prediction. Sections 4.4.2 and 4.4.3 describe refinements of the basic MNL to address these limitations.

4.4.1.2 The Logsum

The logit model produces not only measures for probabilities but also measures for expected generalized cost. The satisfaction is in this case represented by the logsum:

$$w = \theta \cdot Log\left(\sum_{k \in K} Exp\left(\frac{v_k}{\theta}\right)\right), \tag{4.30}$$

The difference in consumer surplus between two alternative public transport scenarios is represented by the difference between the logsums of these scenarios.

The interpretation of the logsum to represent the composite (expected maximum) utility is useful also for economic assessment. The logsum can be rescaled into monetary units by dividing the logsum by the cost parameter. The value of a change in infrastructure to an individual can then be described by the change in the rescaled logsum.

4.4.1.3 Marginal Rates of Substitution and Elasticity

When analysing individuals' preferences, it is of interest to analyse the marginal rates of substitution between the different attributes in the utility function. This is how much of one attribute are they willing to sacrifice in order to improve another attribute by a certain amount. The marginal rates of substitution correspond to the ratio of the marginal utilities of each attribute. Although marginal rates of substitution can be obtained for any pair of attributes, in practice the most important marginal rate of substitution is the value of time (marginal rate of substitution between travel time and monetary cost), also called willingness to pay for a travel time reduction (WTP). For a MNL model, the marginal rate of substitution between attributes a_{kc} and a_{kj} is given by Eq. (4.31):

$$MRS_{cj} = \frac{\partial v_k / \partial a_{kc}}{\partial v_k / \partial a_{kj}} = \frac{\beta_c}{\beta_j}. \tag{4.31}$$

While the marginal rates of substitution are an indicator of trade-off between attributes, the model's elasticities are an indicator of trade-off between an attribute and the demand. Elasticity is defined as the ratio of the percentage change in demand to the percentage change in an attribute. We can distinguish between direct elasticities (effect on the demand for an alternative when one of its attributes changes) and cross-elasticities (effect on the demand for an alternative when an attribute from another alternative changes). For a MNL model, the direct elasticity is given by Eq. (4.32) and the cross-elasticity is given by Eq. (4.33):

$$ELA_{kck} = \frac{\partial p_k}{\partial a_{kc}} \cdot \frac{a_{kc}}{p_k} = \beta_s \cdot a_{kc} \cdot (1 - p_k), \tag{4.32}$$

$$ELA_{kch} = \frac{\partial p_k}{\partial a_{hc}} \cdot \frac{a_{hc}}{p_k} = -\beta_c \cdot a_{hc} \cdot p_h. \tag{4.33}$$

The logit model is then characterized by the following features:

- The direct elasticity is proportional to the scale. This means, for example, that the elasticity is 10 times higher for a journey of 100 km than for a journey of 10 km.

- The cross-elasticity is uniform. This means, for example, that if the generalized cost of rail is reduced, then the demand of the other modes air, bus and car is reduced by the same percentage.

4.4.1.4 Independence of Irrelevant Alternatives (IIA)

By dividing the probability of alternative k by the probability of alternative h, we find that

$$\frac{p_k}{p_h} = Exp\left(\frac{v_k - v_h}{\theta}\right). \tag{4.34}$$

where only the utilities of alternatives k and h are included in the expression. This is called the independence of irrelevant alternatives (IIA) property. An advantage of this property is that additional alternatives can be easily included. In such a case, equal shares of all alternatives will be attracted to the new alternative. However, if added alternatives appear more similar to some of the existing alternatives than to others, this seems more like a disadvantage. A standard example of this is the so-called red–blue bus problem. This example demonstrates that in a car–bus choice model (where the utilities are the same and thus result in equal choice probabilities), an added identical blue bus alternative will attract as much demand from the car alternative as from the red bus alternative. Intuitively, one would expect more demand to come from the blue bus alternative. Again, this issue can be resolved by adopting more sophisticated model formulations than the basic MNL, such as those presented in the following sections.

4.4.1.5 Overcoming the IID Assumption

The root of the red–blue bus problem is the IID assumption on the random component ε. Because the red–blue bus alternatives are so close, they would be expected to share all unobserved utility (except for the colour). In that case, the assumption that the random components are independently distributed is not valid. There will be a very high correlation between the random components of the red–blue bus alternatives. The most common way to deal with this problem in mode choice applications is to apply the so-called nested logit model. In this model, alternatives that are perceived to be more similar are grouped in separate nests, each being a separate MNL model.

Another way to avoid this simplification is to adopt the probit model (see Ben-Akiva and Lerman 1985, for a full description of this model), where the error terms are normally distributed. A multivariate distribution can be used, admitting correlation among the error terms of different alternatives. The probit model is highly flexible, but computationally costly beyond four alternatives, and is left out of this book for simplicity's sake.

4.4.2 The Nested Logit Model

Although the MNL model is widely used due to its simplicity, its principal limitation is that it does not incorporate a structure for capturing correlation between alternatives. In some cases, the assumption of independence between alternatives can be unrealistic and can lead to modelling problems. Different extensions of the MNL model have been proposed to introduce correlation between alternatives, like the nested logit and the mixed logit models.

The idea behind the nested logit model (NL) is to introduce correlation between blocks of alternatives (called "nests") while maintaining the simplicity of the MNL. All the alternatives that belong to a nest will have the same coefficient of correlation and will be uncorrelated with alternatives from different nests. Let us define the utility of alternative k that belongs to nest m as

$$u_k = v_k + \varepsilon_m + \varepsilon_{k|m}, \tag{4.35}$$

where $\varepsilon_{k|m}$ are independent and identically distributed Gumbel variable with null mean and parameter θ_m. The summation of $\varepsilon_{k|m} + \varepsilon_m$ is also distributed as a Gumbel variable with null mean, but with parameter θ_0. The correlation between the alternatives inside a nest comes from the common error term ε_m, which is said to distribute as a logistic function (the difference of two independent Gumbel distributions results in fact in a logistic distribution).

Each nest can be characterized by the expected maximum utility w_m of the alternatives that belong to it. Let K_m be the set of alternatives that belong to nest m, and the satisfaction of nest m is given by:

$$w_m = \theta_m \cdot Log\left(\sum_{k \in K_m} Exp\left(\frac{v_k}{\theta_m}\right)\right), \tag{4.36}$$

The ratio w_m/θ_m is called the logsum of nest m. Based on Eqs. (4.35) and (4.36), it is possible to obtain two distinct MNL probabilities, the probability of choosing nest m, among the different nests [given by Eq. (4.39)], and the conditional probability of choosing alternative k in nest m, among all alternatives in the nest [given by Eq. (4.38)]. Let M be the set of nests. The total probability p_k of choosing alternative $k \in K_m$ is given by the product:

$$p_k = p_m \cdot p_{k|m}, \tag{4.37}$$

where

$$p_{k|m} = \frac{Exp\left(\frac{v_k}{\theta_m}\right)}{\sum_{h \in K_m} Exp\left(\frac{v_h}{\theta_m}\right)}, \tag{4.38}$$

$$p_m = \frac{Exp\left(\frac{w_m}{\theta_0}\right)}{\sum_{j \in M} Exp\left(\frac{w_j}{\theta_0}\right)}. \tag{4.39}$$

Given expressions (4.39) and (4.38) for the choice probabilities, the nested logit model can be estimated using traditional methods, such as maximum likelihood (see later Sect. 4.4.5). As with the MNL model, some normalizations must be made for identification: it is not possible to estimate all scale parameters, but only their ratio $\delta_m = \theta_m/\theta_0$. These ratios must be positive and lower than one. The coefficient of correlation among the alternatives in a nest is $1 - \delta_m^2$. Manipulating (4.39) and (4.38), it can be proven that if all ratios are equal to 1, the NL model is equivalent to a MNL model and there is no correlation between alternatives in any nest.

An example for long-distance travel could be IC and X2000 trains, which are both train alternatives. For these two alternatives, unobserved variables like the service operator may be the same. In a nested model, an MNL model would be used for the choice between these two alternatives, and then a composite train alternative would be formulated in a MNL model for the choice between train and the other modes. The NL model offers a simple way to formulate such a composite alternative.

The property of the Gumbel distribution implies that the log of the denominator of an MNL model (the so-called logsum) represents the expected maximum utility of the alternatives in the model (the IC and X2000 trains in our case).

The properties of the logsum can be further illustrated by the following example:

Assume initially a single alternative, $K = \{1\}$, with (observed) systematic utility v. In this original situation, the composite utility w according to the logsum is simply

$$w = \theta \cdot Log\left(Exp\left(\frac{v}{\theta}\right)\right) \equiv v. \tag{4.40}$$

Assume now that the number of alternatives doubles so that there are two alternatives with the same v. The new composite w is then

$$w = \theta \cdot Log\left(2 \cdot Exp\left(\frac{v}{\theta}\right)\right) \equiv \theta \cdot Log(2) + \theta \cdot Log\left(Exp\left(\frac{v}{\theta}\right)\right) \equiv v + \theta \cdot Log(2). \tag{4.41}$$

The change of the composite w is thus $\theta \cdot Log(2)$. By adding n alternatives with the same v, the new composite w would increase by $\theta \cdot Log(n)$.

This example shows that

- As there is (unobserved) perceived utility, increasing the number of alternatives implies a better chance to find an alternative with a higher total utility. Even if an alternative with a lower utility were introduced, there would be a gain (but to a lesser extent).

- The gain from increasing the number of alternatives depends on the variance of the unobserved utilities—the larger the variance, the higher the scale factor θ and the higher the increase of the expected maximum utility (if there is no unobserved utility, then the scale parameter θ will approach zero, and there will be no gain by increasing the number of such alternatives).
- The increase does not depend on the (observed) systematic utility v.
- When considering a size variable for a given alternative in the utility function, it should appear as the logarithm of the attribute.

In the classic nested logit model, each alternative belongs to one nest. Generalized nested models have been developed where each alternative can partially belong to several nests. One of those is the cross-nested logit (CNL) model, to be discussed in Sect. 4.5.3.3.

4.4.3 The Mixed Logit Model

The NL model overcomes two limitations of the MNL model: different variance and correlation among alternatives can be represented. This is useful in modal choice problems to represent the public and private modes in different nests and also in route choice problems to represent routes with different degrees of overlap. Nevertheless, other limitations of the MNL model remain unaddressed. The heterogeneity of individuals can hardly be represented. Even if demographic variables can be entered in the model measuring the effect on utility of differences among individuals, this method only represents systematic heterogeneity. Highly relevant factors like individual variations in price sensitivity or impatience cannot be captured by systematic modelling.

The mixed logit (MXL) model has a Bayesian structure capable of representing each individual as he/she is. The random utility specification is similar to the one for the MNL model, but the parameters are modelled as random variables with specific (mixing) distributions. Hence, the probability of choice conditioned on the parameters is the same as for the MNL, and the unconditional probability of choice is calculated by averaging through the mixing distributions:

$$p_k = \int \frac{Exp\left(\boldsymbol{\beta}^T \cdot \mathbf{a}_k\right)}{\sum_{h \in K} Exp\left(\boldsymbol{\beta}^T \cdot \mathbf{a}_h\right)} \cdot \varphi(\boldsymbol{\beta}) \cdot \mathrm{d}\boldsymbol{\beta}. \tag{4.42}$$

An equivalent specification of the MXL model includes error components instead of random parameters. Error components are random variables with zero mean added to the utility functions. They can be used for different purposes: adding randomness to a parameter by multiplying a characteristic, creating correlation by linking several alternatives, representing heteroscedasticity by adding variance to some alternatives, etc.

Mc Fadden and Train (2000) have demonstrated that any RUM can be approximated as close as desired by a MXL model with an appropriate mixing distribution, showing its great flexibility to adapt to any behavioural setting.

The calibration of the model produces estimates for the mean and variance of the mixing distributions (hyperparameters). The value of the parameters for each individual can also be estimated in a second phase. The absence of a closed form for the probability of choice complicates the estimation of the MXL model. A high-dimensional integral has to be calculated in each iteration of the likelihood maximization process. However, thanks to its logit kernel, a simulated version of the maximum-likelihood method can be implemented (Train 2009a, b), and it is computationally tractable with current technology. Alternatively, hierarchical Bayes estimation can be employed.

Panel data, where individuals make a series of decisions under varying circumstances, make the full use of the MXL model. The simpler models would consider each decision as a different "pseudo-individual", ignoring the correlation among responses of each individual. Building panel databases in a real context [revealed preference (RP)] is costly; however, hypothetical contexts can be experimentally created [stated preference (SP)]. These choice experiments consist of a series of scenarios with alternative travel modes or routes to choose from. The characteristics of each alternative (time, cost, risk, etc.) are varied according to an experimental plan, predesigned with the purpose of enriching the information for a more accurate model calibration. The realism of the choice experiment is key to prevent hypothetical bias. On the other hand, RP provides higher external validity but not without disadvantages. The researcher only observes the chosen alternative, whereas the rest of the choice set considered by the traveller is hypothetical, inducing measurement error. Omitted variables can induce endogeneity problems in RP modelling. Another advantage of SP is the capability to elicit preferences for non-existing alternatives, like a new transport mode or route under evaluation. Combining RP and SP is a smart empirical strategy. RP provides external validity, and SP allows the exploration of extended range of alternatives and characteristics (travel times, costs, etc.). However, both sources of data cannot be merged straightaway because the (unidentified) variance of utility can be different. Specific data fusion techniques have been developed for that purpose (Louviere et al. 2000).

The data collected in the choice experiment on the park and ride facility in Buenos Aires (example in Sect. 4.4.1.1) can also be represented with a MXL model. The structure of the utility equations is identical to the MNL model; however, the coefficients are random: heterogeneity among commuters is allowed. Table 4.9 shows the results of the maximum-likelihood estimation. Two structural parameters correspond to each variable: the mean and the standard deviation of the corresponding coefficient.

The ASCs were modelled as normally distributed. Their means can be compared to the ASCs in the MNL model. They follow the same pattern of preferences; however, they are all higher in magnitude. The MXL model is more accurate than the MNL. The lower variance of the errors is reflected by a higher scale parameter impacting the scale of all the coefficients. The coefficients of price, cost and time

Table 4.9 Results of the mixed logit model estimation

Variable	Prob. law of coefficient	Structural parameters	
		Mean	SD
Park and ride	Normal	−0.962	1.48
Train	Normal	−6.59	3.15
Charter	Normal	−3.34	2.03
Price	Lognormal	−3.327[a]	0.855[a]
Running cost	Lognormal	−5.956[a]	1.118[a]
Time	Lognormal	−3.490[a]	0.838[a]

[a]Mean and standard deviation of the logarithm of the lognormal distribution

follow a lognormal law. This ensures the right sign, as the lognormal is always positive and the negative sign is enforced in the utility equation. The means and standard deviations correspond to the logarithm of the coefficients. The mean of each coefficient can be calculated with the appropriate expression for the lognormal law. Interestingly, the running cost coefficient turns out to be statistically significant in this case.

4.4.4 Kirchhoff Model and Box–Cox Model

The Kirchhoff model is a direct alternative to MNL. The probabilities are not related to the exponential of the systematic utilities, but to their power function. The probabilities will depend on their ratio, rather than on their difference:

$$p_k = \frac{\prod_{c\in C} \left(a_{kc}\right)^{\beta_c}}{\sum_{h\in K} \left(\prod_{c\in C} \left(a_{hc}\right)^{\beta_c}\right)}. \tag{4.43}$$

The Box–Cox model can be considered to be a generalization on the MNL model and Kirchhoff model. It proposes MNL where in the systematic utility, the Box–Cox transformation is applied to the explanatory variables:

$$v_k = \sum_{c\in C} \beta_c \cdot \lambda\left(a_{kc}, \alpha_c\right). \tag{4.44}$$

The Box–Cox transformation of a variable x is defined as $\lambda(x, \alpha)$ according to Eq. (4.45).

$$\lambda(x,\alpha) = \begin{cases} Log(x) & \text{if } \alpha = 0 \\ \frac{x^{\alpha}-1}{\alpha} & \text{otherwise} \end{cases} \tag{4.45}$$

The transformation works for positive variables, so that the attributes are here assumed to be positive. The transformation function for $\alpha \neq 0$ tends to $Log(x)$ for $\alpha \rightarrow 0$.

Assuming as usual that the parameter θ is already included in the parameters β_c, the resulting expression for the choice probabilities is as follows:

$$p_k = \frac{Exp\left(\sum_{c \in C} \beta_c \cdot \frac{(a_{kc})^{\alpha_c} - 1}{\alpha_c}\right)}{\sum_{h \in K} Exp\left(\sum_{c \in C} \beta_c \cdot \frac{(a_{hc})^{\alpha_c} - 1}{\alpha_c}\right)}. \tag{4.46}$$

If all $\alpha_c = 1$, the MNL model is obtained. If all $\alpha_c = 0$, the Kirchhoff model is obtained. This way, the transformation parameters α_c will be related to the sensibility to the absolute and relative difference between the systematic utilities of the alternatives.

The following numerical example illustrates how the different models respond to various choice situations. The parameter settings for the various models are given in Table 4.10 (Tables 4.11, 4.12 and 4.13).

Table 4.10 Model parameters

Model	Parameters
Kirchhoff	$\beta = -4$, $\alpha = 0$
Logit	$\beta = 0.25$, $\alpha = 1$
Box–Cox	$\beta = -1$, $\alpha = 0.5$

Table 4.11 Choice probabilities for two alternatives with cost $a_1 = 5$ and $a_2 = 10$

Alternative	Cost	Kirchhoff (%)	Logit (%)	Box–Cox (%)
1	5	94	78	86
2	10	6	22	14

Table 4.12 Choice probabilities for two alternatives with cost $a_1 = 105$ and $a_2 = 110$

Alternative	Cost	Kirchhoff (%)	Logit (%)	Box–Cox (%)
1	105	55	78	62
2	110	45	22	38

Table 4.13 Choice probabilities for two alternatives with cost $a_1 = 50$ and $a_2 = 100$

Alternative	Cost	Kirchhoff (%)	Logit (%)	Box–Cox (%)
1	50	94	100	100
2	100	6	0	0

4.4.5 Model Estimation and Test

The most important parameters of RUMs are those characterizing the systematic utility β_c (and in the case of the Box–Cox model, also the parameters α_c). Depending on the particular model, other parameters shaping the distribution of the random term are to be estimated (e.g., the θ_m parameters of the nested logit model). Let us include all the parameters of the model in a vector $\boldsymbol{\beta}$.

To estimate the parameters of a discrete choice model, it is necessary to characterize the observed decisions of the individuals. Let π_{uk} be the observed probability that user $u \in U$ chooses alternative $k \in K_u$ in a given choice context (where all attributes are known). The typical case is that of binary variables, where $\pi_{uk} = 1$ for the one alternative k chosen by the user and $\pi_{uh} = 0$ for all other alternatives $h \neq k$.

For the estimation of the parameters, a maximum-likelihood approach can be applied. The method consists in maximizing the joint probability (i.e., the likelihood L) of the individuals choosing their chosen alternatives according to the model. Under the assumption of independent observations, the likelihood can be expressed as the product of the probability that the model (for given parameter $\boldsymbol{\beta}$) associates with the choice of each user:

$$L(\boldsymbol{\beta}) = \prod_{u \in U} \prod_{k \in K_u} (p_{uk}(\boldsymbol{\beta}))^{\pi_{uk}}. \tag{4.47}$$

To simplify the estimation process, which consists in finding parameters $\boldsymbol{\beta}^*$ that maximize $L(\boldsymbol{\beta})$, a convenient transformation of Eq. (4.47) is the opposite of its logarithm. Because each probability is between 0 and 1, the optimization problem is reduced to finding the minimum of a positive additive objective function, whose lower bound is zero:

$$LL(\boldsymbol{\beta}) \equiv Log(L(\boldsymbol{\beta})) = \sum_{u \in U} \sum_{k \in K_u} \pi_{uk} \cdot Log(p_{uk}(\boldsymbol{\beta})), \tag{4.48}$$

$$\boldsymbol{\beta}* = ArgMin(-LL(\boldsymbol{\beta})). \tag{4.49}$$

This log-likelihood approach is shown in Eq. (4.48). It is important to notice that this estimation approach is valid for any discrete choice model. The selection of different models (such as the MNL model, the Kirchhoff model or the Box–Cox model) will change the expression of the probabilities $p_{uk}(\boldsymbol{\beta})$, but the estimation method will hold.

When selecting the most appropriate model and validating its performance, there are three different criteria that should be taken into account: its microeconomic consistency, its goodness of fit and its forecasting capability.

This microeconomic consistency is the most basic criterion for analysing models, but at the same time the most important. When the parameters $\boldsymbol{\beta}^*$ from the systematic utility are estimated, their sign and magnitude must be consistent with what one would expect based on the implicit microeconomic theory. For example,

in the case of public transportation route choice, most of the level-of-service attributes (such as fare, times and transfers) represent a disutility for the traveller; this way, their parameters β_c should be negative.

Additionally, based on the discrete choice model's structure and estimated parameters, it is possible to obtain marginal rates of substitution between attributes (particularly, monetary valuations for the different attributes) and elasticities. The magnitude of these variables should be compared with empirical applications for the given study context, in order to evaluate their correctness and consistency. For example, the ratio of time and monetary cost parameters represents the value of time (VOT) and such value can be compared with other calibration results for similar contexts and/or expectations about users' willingness to pay. For example, a VOT of 2 €/h could be appropriate for developing countries, but it might be too low for developed countries. Furthermore, for the same variable (e.g., time) split into several components (such as waiting and on-board time), the corresponding parameter absolute values are expected to be increasing with their disutility.

The goodness of fit of a model is the econometric fit between the observed data and the modelled data. Generally, several formal tests and statistics can be used to check the model capability to match data, particularly users' choices.

In the case of maximum-likelihood estimators, and for large samples, some of the most used tests help checking the significance of the parameters. Tests can be made on the single parameter or on the whole parameter vector. For the single parameter, the null hypothesis that it is statistically equal to zero and its estimate β_c^* differs from zero only for sampling errors is tested by the t-statistic:

$$t_c = \frac{\beta_c^*}{\sigma_c}, \tag{4.50}$$

where σ_c is the parameter standard deviation. Particularly, the standard deviations of the parameters are obtained as square roots of the diagonal terms of the opposite of the inverse of the Hessian matrix of second derivatives of the likelihood function with respect to the parameters, calculated in the solution $\boldsymbol{\beta}^*$.

The statistic in Eq. (4.50) has a Student's t-distribution. However, for the typical sample size, it is assumed that the t-statistic is distributed as a standard normal variable, which is the limit of the Student's t-distribution for large samples.

Then, the hypothesis that the parameter β_c^* is statistically equal to zero (and thus is useless for the model) is rejected at a 95 % confidence level if the value of t in Eq. (4.50) is greater than 1.96 or less than −1.96.

The likelihood ratio (*LR*) test checks the null hypothesis that the true value of the parameter vector $\boldsymbol{\beta}^*$ obtained by the maximum-likelihood estimation process is equal to zero through the statistic:

$$LR = 2 \cdot (LL(\boldsymbol{\beta}^*) - LL(\mathbf{0})), \tag{4.51}$$

which is asymptotically distributed as a chi-squared variable with a number of freedom degrees equal to the number of parameters. Accepting the hypothesis that

all the parameters are equal to zero means that all the alternatives have the same probability, and then, the calibrated model does not give any specific information on users' choices. A high value of *LR* ensures that the null hypothesis can be rejected with high probability.

A simple but useful index to check the model capability to match user choices is the so-called ρ^2, a normalized measure within the interval [0, 1] defined as the opposite of the relative difference between the maximum-log-likelihood estimation value $LL(\boldsymbol{\beta}^*)$ and the log-likelihood of a model with null parameters $LL(\mathbf{0})$:

$$\rho^2 = 1 - \frac{LL(\boldsymbol{\beta}^*)}{LL(\mathbf{0})}. \tag{4.52}$$

This statistic is strictly related to *LR*. When ρ^2 takes value zero, then $LL(\boldsymbol{\beta}^*) = LL(\mathbf{0})$ and the model has no explanatory capability. On the contrary, if $LL(\boldsymbol{\beta}^*) = 0$, the model matches perfectly the choices made by users. In fact, if all the probabilities estimated by the model are equal to 1 for the chosen alternative, the log-likelihood value is zero. For intermediate values, the higher the index (i.e., the closer to 1), the better the goodness of fit is. This index can be used to compare alternative models, independently of their nature (e.g., compare a MNL model with the Kirchhoff model). Particularly, an adjusted ρ^2 statistic can be used to compare models with different number of parameters:

$$\rho^2 = 1 - \frac{LL(\boldsymbol{\beta}^*) - |C|}{LL(\mathbf{0})}. \tag{4.53}$$

Where $|C|$ is the number of parameters estimated in the model.

Another way of assessing goodness of fit is to compare to the ASC model instead of the null one. The ASC model has ASCs for each alternative, except the reference, and all other parameters are null. This model predicts market share of each alternative exactly, and it is a more stringent benchmark.

It is also informative to investigate the percentage of users (of the same sample used for the calibration or—better—of another sample used only for the validation) for which the application of the calibrated model provides a good prediction. A prediction is said to be *clearly right* if the model assigns to the alternative actually chosen by the user, the maximum probability and this are higher than a given threshold α. Then, it is possible to plot the percentage of clearly right predictions for each α in the interval [50, 100 %]. The more this decreasing curve stays close to the upper bound of 100 % before dropping to 0 % for $\alpha = 100$ %, the better the model is.

As the main purpose of the models is their use, analysing their forecasting capabilities is essential. A traditional way to do this is to use a validation sample (a subset of the data that are not used when estimating), on which the models can be applied. The results can be compared in terms of forecasting log-likelihood or choice recovery. The first preference recovery (FPR) index is the aggregate of individuals that actually choose the maximum systematic utility alternative.

An additional approach to evaluate the forecasting capability of a model is to compare the observed level of services with the level of services according to the model. For example, in the case of public transportation networks, one could compare:

(i) the modelled and observed load profiles,
(ii) the modelled and observed number of transfers,
(iii) the modelled and observed travel times,
(iv) the modelled and observed modal usage.

Even more, instead of just comparing average values, it is convenient to compare the histograms of attribute values.

4.4.6 Reference Notes and Concluding Remarks

Discrete choice models based on random utility approach are the most used to estimate transport demand. Furthermore, theoretical hypotheses underlying RUMs represent an important advancement as regards descriptive or statistical approaches, because they make possible understanding the behavioural mechanisms generating user trip choices.

Starting from the reference works of the Nobel Prize Daniel McFadden, a wide body of literature exist where both new discrete choice models and hypotheses have been developed to simulate transport demand with greater and greater accuracy. Here, only some main works are described, which are among the most useful for the aims of the book.

For a general overview, Daganzo (1979) analyses both theoretical and practical features of the multinomial probit (MNP) model within the set of discrete choice models. Ben-Akiva and Lerman (1985) provide a comprehensive analysis of RUMs, starting from the general theory framework and then showing several model specifications leading to different discrete choice models. Louviere et al. (2000) propose a similar review with more emphasis on the use of stated preference (SP) methods and an update of econometric approaches to choice modelling. More recently, Train (2009a, b) describes the most recent discrete choice methods, proposing again an update review ranging from logit (included nested logit and cross-nested logit) to probit and mixed logit. Emphasis is also given to parameter estimation procedures, such as maximum likelihood, method of simulated moments and method of simulated scores. Concerning estimation techniques, the reference book by Judge et al. (1988) provides a comprehensive review of inferential approaches and theory and practice in econometrics, from the simpler to the more complex procedures to investigate and model the process that is thought to have generated the available data.

Concerning more in-depth studies and applications, Box–Cox discrete choice models are discussed by Gaudry and Dagenais (1979), while Postorino (1993) proposes a comparison of different discrete choice models, including dogit and

probit–logit models, and their elasticity. To solve the problem of dependency among public transport modes in Tel Aviv, Vovsha (1997a, b) studies a cross-nested model derived from the generalized extreme value class, which considers cross-similarities between several single and combined modes. Bhat (2001) uses a quasi-random maximum simulated likelihood method to estimate mixed MNL models. Papola (2003) discusses cross-nested logit (CNL) models reformulated as a generalization of two-level hierarchical logit models. To overcome the problem of compensatory behaviour embedded in several RUMs, Cantillo and Ortuzar (2005) formalize a semi-compensatory two-stage discrete choice model. Starting from the combined use of stated preference (SP) and revealed preference (RP) data, Hensher (2008) proposes a flexible mixed logit model considering alternative error structures. An approximate computation of choice probabilities in mixed logit models by using the Taylor expansion of the standard logit function is presented in Kalouptsidis and Psaraki (2010).

4.5 Mode and Route Choice

Sebastian Raveau, Odd Larsen and Kjell Jansson

This section applies the concepts of discrete choice modelling introduced in Sect. 4.4, to two specific choices that a traveller faces: which transport mode and, within the mode, which route to take. More sophisticated models are currently being developed, which jointly deal with mode and route choice at once. These models are particularly suited for modelling multimodal choices, where transfers across modes are considered within the routing strategy. These models are, however, out of the scope of this book, and only some discussion will be made in the case of long-distance mode and route choices.

4.5.1 Factors that Influence Mode and Route Choices

Usually, the attributes included in mode and route choice models are all quantifiable and mostly limited to travel time components, fare and transfers. Nevertheless, there is the intuition that many more factors influence behaviour and decisions. It is worth noticing that even though the decisions of the individuals are based on systematic utility levels v_{uk}, the level of service of modal alternatives tends to be measured as a generalized cost c_{uk}. In terms of transportation modelling, both definitions are equivalent when the generalized cost is considered to be a direct disutility, so $v_{uk} = -c_{uk}$. This is valid under the assumption that the positive component of the utility, which is normally related to the activity at destination, is invariant.

The most disaggregated level on analysis is given by a particular route of a particular mode. Therefore, we will centre our analysis on characterizing routes. On

a more aggregated level, the cost for each alternative mode for a given origin–destination pair can be obtained as an aggregation of the costs of its routes from origin to destination (for instance, using the logsum concept). Alternative approaches to obtain the modal costs could be through weighted averages of the route costs, or through a welfare measure such as the expected maximum utility (i.e., again using the satisfaction concept).

The most traditional and significant variables used to explain route choice behaviour are the fare and the travel time. Users tend to look for the fastest and cheapest way of getting from their origin to their destination. Regarding travel time, many components can be considered: in-vehicle time, waiting time at the origin station and all subsequent transfer stations, walking time when accessing or egressing the network and walking time when transferring. These different time components should be considered separately to address their different perception [captured through different parameters β_c in the utility definition of Eq. (4.25)] and importance in the travellers' decision-making process.

Regarding the transferring experience, the traditional approach is to consider the total number of transfers of each alternative route; as the actual transferring time is captured by the walking and waiting time variables, this variable solely captures the disutility of having to transfer. To further understand the transferring valuation, we can differentiate between possible types of transfers. In terms of stations' layout, the transfers can be made between ascending levels (i.e., going up), even levels (usually walking across the platform) and descending levels (i.e., going down). In terms of stations' infrastructure, the transfers can be assisted (made completely using escalator and/or lift), semi-assisted (made partially using escalator and/or lift and partially on foot) and non-assisted (made completely on foot). Additional services (such as toilets, ATMs or shopping stores) affect ambient quality and can influence the choice of transfer stop.

The level of comfort and crowding experienced by the public transport users during their trip is also an important factor. Sections 7.2 and 7.3 describe how models may capture the comfort perception.

Other intangible factors, mentioned in literature as variables, are the routes' topology and geometry, measurements of safety and routes' reliability (traditionally in terms of the different time components, but it can be extended to other attributes with uncertainty). Finally, socioeconomic characteristics that can influence the decisions of the individuals are as follows: the purpose of the trip, the income level (especially when related to fare), gender and age of the traveller, fare type (full fare or discount pass), time of the day the trip is made and the frequency of the journey (e.g., daily, weekly, monthly, first time ever).

4.5.2 Route Choice Set Generation Methods

When using models based on random utility theory to understand and predict the route choice decisions of the travellers, it is necessary to explicitly generate a set of

alternative routes (from which the individuals will choose their best alternative). Unlike other discrete choice contexts (particularly, mode choice), transport networks tend to offer a large number of routes between a given origin and destination. Enumerating all these alternatives cannot be done in practice and at the same time is unlikely that travellers are able to process and consider all the possible alternatives. This way, generation methods have been proposed to construct reasonable choice sets.

All proposed methods are heuristics, and it is good practice to validate the generated choice sets against observations of actually chosen routes, if these are available.

This section is intentionally very concise. The interested reader can find more details about this important subject in Cascetta (2009a, b, c) and Ortuzar and Willumsen (2011a, b).

4.5.2.1 K-Shortest Path Methods

A simple way of obtaining "good and reasonable" routes for the travellers to choose from is to obtain the best k paths between the origin and destination, based on a particular definition of link cost. The number of routes to be obtained (i.e., k) is to be defined by the modeller. A possible disadvantage of this method is the potential similarity between the routes generated, which could only differ in a link sometimes. A possible alternative solution is to obtain the shortest paths based on different cost definitions (also called "labels"), such as minimum travel time, less turns, and less transfers. This method, called "labelling approach", has a behavioural support as it guarantees that each route generated is optimal from a certain point of view and is more in line with the way humans tend to identify relevant choice characteristics.

4.5.2.2 Stochastic Methods

Assuming that the route choice model is based on a stochastic utility function, it seems reasonable that the alternative routes to be considered by the travellers are also defined using a stochastic utility approach. This way, a simple extension of the shortest path methods is to obtain shortest random paths. For this, a random term is added to the given (deterministic) definition of link costs, and then, shortest paths are computed for a certain number of simulated draws of these random terms. Note that this randomization is different from the random error term in RUMs, as here we perturb observable (systematic) utility attributes, whereas the random term captures the unobserved part of utility. Like in the deterministic shortest path methods, the links from the obtained shortest paths may be penalized (or removed) to reduce the overlapping of the routes in the choice set.

Instead of randomizing the arc costs, it is also possible to randomize the β_c parameters that determine the cost. This is practically equivalent to synthesizing

users with different preferences and typically leads to a more suitable variety of paths based on attributes.

4.5.2.3 Constrained Enumeration Methods

This approach is based on the assumption that travellers have certain behavioural rules to select their possible routes. Any route that complies with a set of restrictions imposed by the travellers is to be included in the choice set. These restrictions may be related to the cost of the route (e.g., alternative routes should not be more than 30 % longer than the shortest route), to the logic of the route (e.g., alternative routes should not possess loops) or to the topology of the route (e.g., alternative routes should not take the travellers farther away from the destination).

4.5.2.4 Constructive Methods

Multiple shortest path searches are carried out sequentially. To avoid obtaining routes that are too similar, at each iteration, the cost of the arcs of the currently shortest path is increased, so that it is less attractive in the next iteration. The iterations are terminated when either the desired number of paths has been found, or despite the cost increases, no new shortest paths are generated. This allows to identify a set of paths which do not overlap substantially.

4.5.3 Route Choice Models with Correlation

It is in principle straightforward to apply an MNL model to route choice, once the choice set is generated and utilities calculated for each alternative route. However, MNL models lack a structure for capturing correlation between routes, due to overlapping. In urban public transportation networks, the routes linking a given origin–destination pair will typically have many overlaps due to common arcs, so the independent error assumption of the MNL model is unrealistic. Different extensions of the MNL model have been proposed to explicitly capture correlation between alternative routes, some of which are presented in this section.

To further understand the problem of route overlapping, let us consider the simple network depicted in Fig. 4.23a, where we have three alternative routes with the same cost c, two of which share a common link with cost $c \cdot (1-\alpha)$. If MNL were to be applied, the probability of choosing any route (in particular the upper link, which has no overlapping) is 1/3, independent of the value of α. This is consistent with the case depicted in Fig. 4.23b, where $\alpha = 1$ and there is no overlapping (the three alternatives are independent and have a probability of being chosen of 1/3). Nevertheless, if $\alpha = 0$ as shown in Fig. 4.23c, the two overlapping routes collapse into a single alternative, and the probability of choosing any route

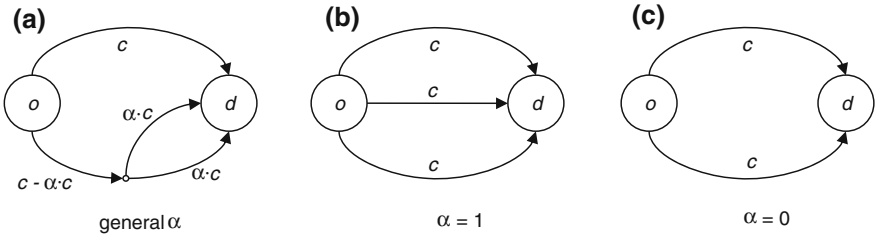

Fig. 4.23 Arc overlapping example

(in particular the upper link, which has no overlapping) is 1/2. This way, it is clear that the probability of choosing the upper link should vary between 1/3 and 1/2 depending on the value of α, something that the MNL model cannot do.

4.5.3.1 C–Logit Model

The simplest extension of the MNL model to address arc overlapping is the C–Logit model, which maintains the simplicity of the MNL model and adds a correction term to the systematic utility. The basic idea behind the model is to handle route interdependence via an attribute of each path k called the *commonality factor* CF_k. The probability of choosing route k is then given by Eq. (4.54):

$$p_k = \frac{Exp\left(\sum_{c\in C} \beta_c \cdot a_{kc} + \beta \cdot CF_k\right)}{\sum_{h\in K_u} Exp\left(\sum_{c\in C} \beta_c \cdot a_{hc} + \beta \cdot CF_h\right)}. \tag{4.54}$$

As before, c ranges over different attributes (cost, time); β is an additional (negative) parameter of the same kind of β_c to be estimated. The commonality factor can be defined in different ways; traditionally, Eq. (4.55) is considered:

$$CF_k = Log\left(\sum_{h\in K} SC_{kh}\right). \tag{4.55}$$

The so-called *similarity coefficient* SC_{kh} between two alternatives k and h is defined as follows:

$$SC_{kh} = \left(\frac{c_{kh}}{\sqrt{c_k \cdot c_h}}\right)^{\gamma}, \tag{4.56}$$

where c_k is the total cost of route k, c_h is the total cost of route h, c_{kh} is the common cost of routes k and h (due to overlapping) and γ is a positive parameter to be estimated.

4.5.3.2 Paired Combinatorial Logit

Another model based on the MNL structure, the aired combinatorial logit (PCL), exploits the fact that correlation between routes can be defined in pairs. The logic beneath the PCL is that routes are chosen among pairs of alternatives within the choice set. The choice probability is defined by Eq. (4.57), where p_{kh} is the marginal probability of choosing the pair of alternatives k and h among the total number of possible pairs and $p_{k|kh}$ is the conditional probability of choosing route k given the chosen pair of alternatives k and h. Conditional and marginal probabilities depend on the similarity between routes within the chosen pair:

$$p_k = \sum_{h \in K-k} p_{kh} \cdot p_{k|kh} \,. \tag{4.57}$$

Both probabilities p_{kh} and $p_{k|kh}$ will depend on the correlation between routes. The correlation between routes k and h can be expressed through the similarity coefficient SC_{kh}. The expressions for the choice probabilities are given by Eqs. (4.59) and (4.58):

$$p_{k|kh} = \frac{Exp\left(\frac{v_k}{1-SC_{kh}}\right)}{Exp\left(\frac{v_k}{1-SC_{kh}}\right) + Exp\left(\frac{v_h}{1-SC_{kh}}\right)} \,. \tag{4.58}$$

$$p_{kh} = \frac{(1 - SC_{kh}) \cdot \left(Exp\left(\frac{v_k}{1-SC_{kh}}\right) + Exp\left(\frac{v_h}{1-SC_{kh}}\right)\right)^{1-SC_{kh}}}{\sum_{i \in K} \sum_{\substack{j \in K \\ j > i}} (1 - SC_{ij}) \cdot \left(Exp\left(\frac{v_i}{1-SC_{ij}}\right) + Exp\left(\frac{v_j}{1-SC_{ij}}\right)\right)^{1-SC_{ij}}} \,. \tag{4.59}$$

Like with the C–Logit's commonality factor, there are many alternative ways to define the PCL's similarity coefficient. For example, the similarity coefficient can be defined by Eq. (4.56).

4.5.3.3 Cross-Nested Logit

The logic beneath applying the CNL model to route choice is that routes are chosen among nests, which correspond to the arcs of the network. In this particular case, we will denote the generic arc with m to recall that they play the role of nests. The nest for arc m comprises all routes that traverse m. Clearly, each route may belong to several nests. The choice probability is defined by Eq. (4.60), where p_m is the marginal probability of choosing nest m (there are as many nests as arcs in the network) and $p_{k|m}$ is the conditional probability of choosing route k given the chosen nest m.

$$p_k = \sum_{m \in A} p_m \cdot p_{k|m} \,. \tag{4.60}$$

The conditional and marginal probabilities depend on the degree of inclusion of alternative k in nest m. This degree of inclusion is given by variables α_{mk} that add up to 1 and can be defined by Eq. (4.61), where c_m is the cost of arc m, c_k is the total cost of route k and Δ_{mk} equals 1 if arc m belongs to route k and 0 otherwise:

$$\alpha_{mk} = \frac{c_m}{c_k} \cdot \Delta_{mk}. \tag{4.61}$$

The expressions for the choice probabilities are given by Eqs. (4.63) and (4.62), where $\delta_m = \theta_m/\theta_0$ (in the case where each arc is a nest, it is assumed that all δ_m are equal to δ):

$$p_{k|m} = \frac{\alpha_{mk}^{1/\delta_m} \cdot Exp\left(\frac{v_k}{\theta_m}\right)}{\sum_{h \in K} \alpha_{mh}^{1/\delta_m} \cdot Exp\left(\frac{v_h}{\theta_m}\right)}, \tag{4.62}$$

$$p_m = \frac{\left(\sum_{k \in K} \alpha_{mk}^{1/\delta_m} \cdot Exp\left(\frac{v_k}{\theta_m}\right)\right)^{\delta_m}}{\sum_{a \in A} \left(\sum_{k \in K} \alpha_{ak}^{1/\delta_a} \cdot Exp\left(\frac{v_k}{\theta_a}\right)\right)^{\delta_a}}. \tag{4.63}$$

To better understand the above formula, consider that for a nested logit model, based on Eq. (4.36), the following Eq. (4.64) holds and allows for an alternative expression of the probabilities (4.38) and (4.39) that coincide with the above Eqs. (4.62) and (4.63) for binary variables α_{mk} that partition the alternatives into separate nests:

$$Exp\left(\frac{w_m}{\theta_0}\right) = \left(\sum_{k \in K_m} Exp\left(\frac{v_k}{\theta_m}\right)\right)^{\delta_m}. \tag{4.64}$$

Note also that the main block of Eqs. (4.62) and (4.63) can be rewritten, so as to eliminate the parameters θ_m, leaving only the parameters δ_m and θ_0. Moreover, the latter can be absorbed as usual in the attribute coefficients. Thus, the parameters to actually calibrate are the δ_m and the β_c (this is also true for the NL model, given that the CNL is a generalization):

$$\alpha_{mk}^{1/\delta_m} \cdot Exp\left(\frac{v_k}{\theta_m}\right) = \left(\alpha_{mk} \cdot Exp\left(\frac{v_k}{\theta_0}\right)\right)^{1/\delta_m} = \left(\alpha_{mk} \cdot Exp\left(\sum_{c \in C} \beta_c \cdot a_{kc}\right)\right)^{1/\delta_m}. \tag{4.65}$$

4.5.4 Urban Case Study: Santiago de Chile Transit System

To illustrate the application of mode and route choice modelling techniques, two analyses of the public transport system of Santiago, Chile (called Transantiago), are presented. The first case study focuses on mode choice (among private and public modes), and the second study case focuses on route choice (within the public transport system).

4.5.4.1 Mode Choice

The data used in this study correspond to the fourth wave of the Santiago Panel, a five-day pseudo-diary considering information about work trips in the morning peak. The initial sample consisted of 303 Santiago residents working at one of the five campuses of the Pontificia Universidad Católica de Chile. There are up to five choices (i.e., one for each weekday) for each of the 258 respondents who answered correctly the survey, leading to 1107 observations after cleaning the data set of inconsistencies.

Ten different transport modes were considered, both single and combined:

(i) car driver,
(ii) car passenger,
(iii) shared taxi,
(iv) metro,
(v) bus,
(vi) car driver/metro,
(vii) car passenger/underground,
(viii) shared taxi/underground,
(ix) bus/underground,
(x) shared taxi/bus.

For each mode, information was available about walking, waiting and in-vehicle travel times, cost and number of transfers made. Regarding the users, the panel gathered information about socioeconomic variables, such as age, education and income, among others. A mixed logit model (Sect. 4.4.3) was estimated, where the only random coefficient was that for the cost (which interacted with the individual's wage rate). A full set of ASC was included, leaving the bus ASC as baseline. The results of the estimation are summarized in Table 4.14.

All parameters presented in Table 4.14 have the expected signs, and most of them are statistically significant at 95 % confidence level (waiting and walking times are kept in the model, as policy variables). Based on the obtained parameters, it is possible to calculate monetary valuations for all the variables (Table 4.15). Table 4.15 presents the subjective values of travel, waiting and walking times computed from the above results, as well as the valuations associated with the transfers. As the parameter related to cost is random, these valuations were

Table 4.14 Estimation results for mode choice in Santiago

Variable	Parameter	t-Value
Cost/wage rate (h)—mean	−0.031	−5.35
Cost/wage rate (h)—standard deviation	0.093	−8.58
In-vehicle time (h)	−0.012	−5.79
Waiting time (h)	−0.010	−1.75
Walking time (h)	−0.016	−1.38
Transfers	−1.047	−5.00
ASC car driver	1.600	3.12
ASC car passenger	−2.410	−4.48
ASC shared taxi	−1.247	−6.46
ASC metro	0.792	2.47
ASC car driver/metro	1.698	3.00
ASC car passenger/metro	−1.043	−2.73
ASC shared taxi/metro	−2.041	−4.14
ASC bus/metro	1.686	6.03
ASC shared taxi/bus	−1.369	−2.05
Sample size	1.107	
Corrected ρ^2	0.421	

Table 4.15 Monetary valuations for mode choice in Santiago

Variable	Valuation
In-vehicle time	1.55 €/h
Waiting time	1.17 €/h
Walking time	1.94 €/h
Transfers	2.41 €/transfer

computed using simulation. For all the valuations shown, an individual mean wage rate of €3.74 per work hour was considered; this value was obtained from the Santiago Panel's individual information.

It is important to stress out that the results are to be considered valid for the specific class of respondents (i.e., university workers), while different monetary values may result using a different sample. This also partly explains the difference in the values found in this example with the one described in the following section.

4.5.4.2 Route Choice

The data for this study come from a RP study conducted in 2011 for the entire public transport system, which was complemented with information from a 2008 survey on the metro system. The data from the surveys (demand information) were complemented with information regarding the level of services (supply information) provided by the public transport authorities. It is important to consider that

Transantiago is a fare-integrated system, on which the mode and route decisions can be made simultaneously.

The study considers ten explanatory variables:

(i) the fare,
(ii) the in-vehicle time,
(iii) the waiting time,
(iv) the walking time (when transferring),
(v) the number of transfers (distinguishing by the modes involved),
(vi) the mean occupancy of the route,
(vii) the probability of not boarding, when the occupancy is high,
(viii) the probability of finding a seat, when the occupancy is low,
(ix) a binary variable that captures if the route takes the traveller away from the destination,
(x) a binary variable that captures if the route takes the traveller back to the origin at some point.

With the available data (which consisted of 17,141 individuals), a C–Logit model (cf. Sect. 4.5.3.1) was estimated. The set of alternatives was generated with a link penalty approach (cf. Section 4.5.2.1), using different specifications of the links' costs (which consisted of different combinations of weights for the time components). The results of the estimation are summarized in Table 4.16.

All the parameters presented in Table 4.16 have the expected signs and are statistically significant at 95 % confidence level. The only positive parameter is the one related to getting a seat (which is the only variable that generates utility). It can be seen that travellers perceive the different time components differently. Transfers

Table 4.16 Estimation results for route choice in Santiago

Variable	Parameter	t-Value
Fare (€)	−0.105	−6.32
In-vehicle time (h)	−0.122	−12.12
Waiting time (h)	−0.168	−3.79
Walking time (h)	−0.217	−3.53
Transfers—metro to metro	−0.075	−3.32
Transfers—metro to bus/bus to metro	−0.098	−3.41
Transfers—bus to bus	−0.120	−4.98
Mean occupancy (pax/m^2)	−0.107	−4.31
Probability of not boarding	−0.197	−4.15
Probability of getting a seat	0.165	2.81
Going back to the origin	−0.145	−6.93
Going away from the destination	−0.157	−7.11
C–logit's commonality factor	−0.412	−3.55
Sample size	17.141	
Corrected ρ^2	0.382	

Table 4.17 Monetary valuations for route choice in Santiago

Variable	Valuation
In-vehicle time	1.16 €/h
Waiting time	1.60 €/h
Walking time	2.07 €/h
Transfers—metro to metro	0.71 €/transfer
Transfers—metro to bus/bus to metro	0.93 €/transfer
Transfers—bus to bus	1.14 €/transfer
Mean occupancy	1.02 €/(pax/m^2)
Probability of not boarding	1.88 €
Probability of getting a seat	1.57 €
Going back to the origin	1.38 €
Going away from the destination	1.50 €

involving metro are preferred. The parameter of the commonality factor, related to route correlation due to overlapping, is significant. This means that at some degree, travellers tend to perceive overlapping routes as non-independent alternatives. Based on the obtained parameters, it is possible to directly calculate monetary valuations for all the variables (Table 4.17).

4.5.5 Long-Distance Case Study: Stockholm Regional Buses

The purpose of this section is to discuss the interface between public transport route choice models and mode choice models for long-distance travel.

In models that handle short-distance travel—typically trips within an urban area—public transport assignment usually deals simultaneously with all PT submodes (i.e., bus, tram, subways, train). The choice of PT submode becomes a route choice decision, and the assignment model allows for transfer between submodes. The different submodes might have different weights for travel time and possibly waiting time that reflect differences in comfort and may also have different fares. Simultaneous treatment of all public transport modes in an assignment model is a natural choice if it can be assumed that the important factors that influence mode choice for PT are included as parameters or variables in the assignment model.

For long-distance travel, however, it is quite common to have separate assignments for each public transport submode (air, coach, train and possibly high-speed train defined as a separate mode) and to handle the choice between modes by a RUM. In this case, the choice model may also include car. This approach implies that a main mode has to be defined for multimodal trips, and a second or third mode used for a journey must be handled by a separate access/egress model that gives access to and egress from the main mode. A major reason for using RUMs for long-distance travel is that it may be difficult to capture all the factors relevant for

mode choice in long-distance travel by the parameters and variables in an assignment model.

On the other hand, the use of an assignment model that handles all PT modes (or subsets of PT modes) simultaneously also for long-distance travel has certain advantages. One obvious advantage is the ability to handle multimodal trips in a more realistic way. This is especially important in the absence of a comprehensive model for access/egress to different modes. Another important advantage is better treatment of schedule delay, the deviation between preferred departure and/or arrival times and the departure and arrival times offered by the different alternatives.

For long-distance travel, the frequency of public transport supply is usually much lower than that of local transport. As a consequence, travellers use timetables and decide on an itinerary for the trip. Actual departure and arrival times for different modes usually become more important than waiting times at terminals and stops due to the effect on deviations between preferred and actual departure/arrival times. This may also affect mode choice. A mode (or combination of modes) that may otherwise be preferred might not be chosen if departure or arrival time implies a large schedule delay compared to the alternatives.

Variability of travel conditions, as well as fares of public transport alternatives, is generally greater for long-distance travel. The use of yield management schemes by operators, where prices are set as a function of demand, poses a particular challenge as it implies that the passengers, even for the same departure, can have widely different fares, depending on when they booked the journey.

To deal adequately with the users' requirements, the modellers need to carefully observe the specific aspects of this particular type of travel. Some examples are given below:

- How to deal with heterogeneity of traveller preferences in many attributes. How to categorize them?
- How to deal with supply attributes other than timetable, such as capacity, comfort, fare, reliability, personal service and connections to other means of travel?
- How to deal with timetables? By using a full list of all timetables, which would be difficult for future situations as well as may be cumbersome for any existing reference situation, or by using a simplified representation such as only specifying the ride time and headway for each line per mode?
- What combinations of travel route alternatives should be considered for each travel decision and how should travel demand be distributed between various travel alternatives, i.e., what type of model should be used for this distribution?

4.5.5.1 Standard Model

In order to look at some implications related to modelling long-distance travel choices, we can use a simple example taken from Larsen et al. (2010). The supply of services for journeys on a hypothetical OD pair is shown in Table 4.18.

Table 4.18 Example for one OD pair

Alternative	On board (min)	Headway (min)	Acc + Egress (min)
Train 1	120	60	20
Train 2	130	90	20
Train 3	140	60	20
Bus	120	120	25
Car	160	0	10

For simplicity, we assume that the cost is the same for all alternatives and that time has the same weight irrespective of mode and time component. The combined mode and route choice model to be described below shall distribute travellers between the three modes (train, bus and car), but also between the 3 train routes.

The conventional approach is to make an assignment for each mode, often even for each submode of public transport separately and then feed each generalized cost per mode into a separate mode choice model. How will the conventional approach work for this example? Assume that frequency-based assignment (described in detail in Sect. 6.2) is used for route choice, although the long headways cast doubt on some of its prerequisites. Route choice results are given in Table 4.19.

In this simple case, explicit assignment is needed only for the train alternative, while bus only gets a schedule delay (waiting time) of half the headway. The three last lines in Table 4.19 are the times that will be used by a RUM for mode choice. The generalized cost for mode train is a weighted average for the train routes. The mode choice model must use the same coefficients on different generalized cost components as route choice model to avoid inconsistency. With the same coefficients, all we need for a simple multinomial logit model is the scale parameter that reflects the variance of the random term in the utility functions (see Sect. 4.4.1). If we use a nested logit model with a nest for PT (train and bus), we then need 2 parameters for the model. The second parameter shall in this case reflect that the random terms of train and bus are positively correlated, i.e., they share some of the unobserved mode-specific attributes.

Table 4.19 Route choice results for each mode separately

	Share	Minutes			
		On-board time	Mean schedule delay	Acc + Egress time	Generalized cost
Train 1	0.6368	120.00	22.05	20.00	
Train 2	0.1986	130.00	14.46	20.00	
Train 3	0.1646	140.00	12.00	20.00	
Train		125.28	18.89	20.00	164.17
Bus		120.00	60.00	25.00	205.00
Car		160.00	0.00	10.00	170.00

Table 4.20 Mode shares with a simple multinomial logit model treating each mode separately

Scale parameter	0.257	0.128	0.086	0.064	0.051	0.043	0.032
Train	0.817	0.676	0.611	0.568	0.536	0.513	0.476
Bus	0.000	0.004	0.018	0.042	0.067	0.089	0.129
Car	0.183	0.320	0.370	0.391	0.398	0.399	0.395

In Table 4.20, we give the mode shares for a simple multinomial logit model and different values of the scale parameters.

With the highest value for the scale parameter (small variance), the bus gets virtually no demand, mainly because of the high value for schedule delay, and car has a fairly low market share. As the scale parameter decreases, the mode shares become more equal and they approach 1/3 in the limit when the scale parameter is zero. For the train mode, the 3 lines will always have the shares in Table 4.20.

Now, let us assume that bus and train share all the unobserved attributes, i.e., the mode-specific random terms in the "utility functions" are identical for all travellers. In that case, all that matters for the choice between train and bus is included in Table 4.19, and we might as well use a combined assignment for train and bus.

With the same type of assignment, we now get the distribution on PT alternatives and generalized cost for input to a RUM as in Table 4.21.

Notice that the mean schedule delay for bus now drops from 60 to 14.45 min and the distribution on train routes is altered and so is the mean schedule delay for these alternatives. Note that this value reflects also a behavioural assumption: by using a single route choice model across bus and train, we express that the passenger waits for all options simultaneously and can choose bus or train at the last minute depending on the actual timetable. This is significantly different from the situation before where the decision between bus and train had to be made up front and a bus passenger was then stuck with the long headway. While the logit model for 3 modes with separate assignment for each mode has the IIA property, this is now only the case for PT and car, while the assignment on PT does not have this property. If we now use the generalized costs of PT and car as inputs to a binomial logit model (i.e., the specific case of MNL with only two alternatives) with the same set of scale parameters as in Table 4.20, we get the mode shares in Table 4.21.

The mode shares for train and bus separately are taken from the assignment in Table 4.22. We see that the mode shares are significantly different from Table 4.20 irrespective of the value of the scale parameter. In particular, for bus, the differences

Table 4.21 Mode shares with a binomial logit model for public transport (bus + train) and car

Scale parameter	0.257	0.128	0.086	0.064	0.051	0.043	0.032
PT	0.8909	0.7400	0.6688	0.6278	0.6027	0.5869	0.5650
Train	0.7470	0.6205	0.5608	0.5265	0.5054	0.4922	0.4738
Bus	0.1438	0.1195	0.1080	0.1014	0.0973	0.0948	0.0912
Car	0.1091	0.2600	0.3312	0.3722	0.3973	0.4131	0.4350

Table 4.22 Assignment and generalized cost with bus and train as a single mode

	Share	Minutes			
		On-board time	Mean schedule delay	Acc + Egress time	Generalized cost
Train 1	0.5446	120	19.91	20	
Train 2	0.1664	130	13.32	20	
Train 3	0.1276	140	11.14	20	
Bus	0.1615	120	14.45	25	
PT		124.22	16.81	20.81	161.83
Car		160.00	0.00	10.00	170.00

in mode share in the two tables are very pronounced, but also the mode shares for train and car are strongly affected by the choice of model.

Let us also look at the extreme case when "everything" that matters for the travellers is included in Table 4.18. This will be the case if the importance of unobserved attributes of different modes is exceedingly small or constant and equal across travellers and modes. In that case, there will be instances of timetables for the PT alternatives that will make car the best alternative for some travellers. To avoid numerical problems (division by zero), car is given a headway of 1 min in the assignment. Results are given in Table 4.23.

The generalized cost for all mode alternatives is now computed as a weighted average over all modes and routes. Again, we get a distribution on modes and among the PT lines that are different from what we have in Tables 4.22 and 4.23 and the mean schedule delay for the different PT alternatives is affected because travellers that experience high values for schedule delay with PT will choose car.

The example demonstrates that results depend very much on the model specification for mode and route choice. The effect is a combination of behavioural assumptions (which choices are made at which stage?) and the treatment of correlations between choice alternatives.

Table 4.23 Assignment model used for all alternatives combined

Alternative	Share	Minutes			
		On-board time	Mean schedule delay	Acc + Egress time	Generalized cost
Train 1	0.4160	120	13.44	20	
Train 2	0.1265	130	8.61	20	
Train 3	0.0680	140	4.57	20	
Bus	0.1331	120	10.49	25	
Car	0.2565	160	0.00	10	
All		132.88	8.39	18.10	159.37

4.5.5.2 Combined Model

As shown by Larsen et al. (2010), it is useful to obtain more realistic results by combining mode and route choice in a single RUM, at least for simple cases like the example used above. The correlation between alternatives within the same mode is captured by using identical error terms for them. The section proposes a solution approach based on simulation. In the simulation framework, it is also easy to use heteroscedastic error terms, which generally is a difficult problem to handle in RUMs. Combined models may also make it easier to avoid inconsistencies with respect to coefficients on generalized cost components in assignments and in RUMs.

There are, however, several problems that must be solved before such combined models can be applied on a large scale. Among these are:

- How to handle multimodal trips in terms of discrete choice error terms?
- How to estimate/calibrate the parameters of combined models? This includes both the relative weights of different travel time components and the scale parameters for discrete choice terms.

4.5.6 Reference Notes and Concluding Remarks

This chapter provides a summary of the state-of-the-art mode and route choice, with an emphasis in public transport networks. The contents of the chapter can be divided into three main topics: route choice generation methods, route choice models with correlation and travel behaviour study cases.

Regarding route choice generation, several methods have been proposed and applied in the literature. Most shortest path methods (such as k-shortest paths, labelling approach and stochastic methods) are based on single shortest paths (Bellman 1958; Dijkstra 1959). There are several extensions to multiple shortest paths (e.g., Hoffman and Pavley 1959; Yen 1971; Hadjiconstantinou and Christofides 1999), which differ in terms of considerations and computational efficiency. The labelling approach was proposed by Ben-Akiva et al. (1984) and has been applied since to private transport and public transport. The link elimination approach was first presented by Azevedo et al. (1993), and the link penalty approach was first presented by De la Barra et al. (1993); these methods have been adapted in later studies. The idea behind the stochastic methods originates from Sheffi and Powell (1982) and has been extended since. Finally, the constrained enumeration method was proposed by Prato and Bekhor (2006).

In terms of modelling approaches for dealing with correlation due to path overlapping, the first extension of the traditional MNL model was the C–Logit, proposed by Cascetta et al. (1996), for which different specifications of the commonality factor have been proposed (see Cascetta 2009a, b, c). The PCL was proposed by Chu (1989), developed by Koppelman and Wen (1998) and adapted to

route choice by Prashker and Bekhor (1998). The CNL was proposed by Vovsha (1997a, b) and adapted to route choice by Prashker and Bekhor (1998).

The mode choice study case on Santiago was conducted by Yañez et al. (2010), while the route choice study case on Santiago was conducted by Raveau et al. (2011) and later expanded by Raveau and Muñoz (2014). The long-distance study case on Stockholm was conducted by Larsen et al. (2010).

References

Ajzen I (1991) The theory of planned behavior. Organ Behav Hum Decis Process 50:179–211

Andriotti GK (2009) Prospect theory multi-agent based simulations for non-rational route choice decision making modelling. PhD thesis, Universität Würzburg, Würzburg, Germany

Avineri E (2004) A cumulative prospect theory approach to passengers behavior modeling: waiting time paradox revisited. J Intell Transp Syst 8:195–204

Avineri E (2009) Nudging travellers to make better choices. In: Proceedings of the international choice modelling conference, Leeds, UK

Azevedo JA, Santos Costa MEO, Silvestre Madera JJER, Vieira Martins EQ (1993) An algorithm for the ranking of shortest paths. Eur J Oper Res 69:97–106

Baron J (2008) Thinking and deciding. Cambridge University Press, New York

Bates J (2000) History of demand modelling. In: Henscher DA, Button KJ (eds) Hand-book of transport modelling. Elsevier's Handbooks in Transport

Bellman R (1958) On a routing problem. Q Appl Math 16:87–90

Ben-Akiva M, Lerman S (1985) Discrete choice analysis, theory and application to travel demand. MIT Press, Cambridge

Ben-Akiva ME, Bergman MJ, Daly AJ, Ramaswamy R (1984) Modeling inter-urban route choice behaviour. In: Volmuller J, Hamerslag R (eds) Proceedings of the 9th international symposium on transportation and traffic theory, VNU Science Press, Utrecht, The Netherlands, pp 299–330

Beria P, Maltese I, Mariotti I (2012) Multicriteria versus cost benefit analysis: a comparative perspective in the assessment of sustainable mobility. Eur Trans Res Rev 4:137–152

Bhat CR (2001) Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Transp Res B 35:677–693

Binetti M, Borri D, Circella G, Mascia M (2005) Does prospect theory improve understanding transit user behavior? In: Proceedings of computers in urban planning and urban management, London, UK

Cambridge Systematics (2010) Travel model validation and reasonableness checking manual, 2nd edn. Cambridge, Massachusetts

Cantillo V, de Ortuzar JD (2005) A semi-compensatory discrete choice model with explicit attribute thresholds of perception. Transp Res B 39:641–657

Cascetta E (2009a) Transportation systems engineering: theory and methods. Kluver Academic Publishers, Dordrecht

Cascetta E (2009b) Transportation systems engineering: theory and methods. Kluver Academic Publishers, Dordrecht

Cascetta E (2009c) Transportation systems engineering: models and applications. Springer, New York

Cascetta E, Nuzzolo A, Russo F, Vitetta A (1996) A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks. In: Lesort JB (ed) Proceedings of the thirteenth international symposium on transportation and traffic theory. Pergamon, Lyon, France, pp 697–711

Ceder A, Chowdhury S, Taghipouran N, Olsen J (2013) Modelling public transport users' behavior at connection points. Transp Policy 27:112–122
Chu C (1989) A paired combinatorial logit model for travel demand analysis. In: Proceedings of the 5th world conference on transportation research, Ventura, USA, pp 295–309
COM—The European Commission (2007) Greenbook 2007—towards a new culture for urban mobility. Commission of the European Communities, Brussels, Belgium
Daganzo CF (1979) Multinomial probit: the theory and its application to demand forecasting. Academic Press, New York
De la Barra T, Pérez B, Anez J (1993) Multidimensional path search and assignment. In: Proceedings of the 21st PTRC summer annual meeting, Manchester, England, pp 307–319
De Ramos GM, Daamen W, Hoogendoorn S (2010) Toward a better understanding of the reference point in a travel behavior context. In: Proceedings of the 11th TRAIL congress, The Netherlands
De Ramos GM, Daamen W, Hoogendoorn S (2013) Modelling travellers' heterogeneous route choice behavior as prospect maximizers. J Choice Modell 6:17–33
De Ramos GM, Daamen W, Hoogendoorn S (2014) A state-of-the-art review: developments in utility theory, prospect theory, and regret theory to investigate travellers' behaviour in situations involving travel time uncertainty. Transp Rev: A Transnational Transdisciplinary J 34:46–67
Department for Transport (2014) Transport analysis guidance: WebTAG. Department for transport, UK
Dijkstra EW (1959) A note on two problems in connection with graphs. Numer Math 1:269–271
EVA TREN (2008) Improved decision-aid methods and tools to support evaluation of investment for transport and energy networks in Europe. Deliverable 1. Evaluating the state-of-the-art in investment for transport and energy networks. 6th Framework Programme, European Union
Federal Highway Administration (2007) The transportation planning process: key issues. Transportation Planning Capacity Building Program, FHWA-HEP-07-039. FHWA, USA
FGSV—Forschungsgesellschaft für Straßen-und Verkehrswesen (2001) Leitfaden für Verkehrsplanungen, Köln. FGSV, Germany
Friedrich M. (1994) Rechnergestütztes Entwurfsverfahren für den ÖPNV im ländlichen Raum (Computer assisted design of public transport systems in rural areas, Dissertation). Schriftenreihe des Lehrstuhls für Verkehrs- und Stadtplanung, Heft 5, Technische Universität München, Germany
Friedrich M (2011) Wie viele? Wohin? Womit? Was können uns Verkehrsnachfragemodelle wirklich sagen? Tagungsbericht Heureka 11. FGSV Verlag, Köln, Germany
Friedrich M (2015) Multimodal transport planning. Lecture notes, Stuttgart University, Germany
Gao S, Frejinger E, Ben-Akiva M (2010) Adaptive route choices in risky traffic networks: a prospect theory approach. Transp Res C 18:727–740
Gaudry MJI (1980) Dogit and logit models of travel mode choice in Montreal. Canadian J Econ/Revenue Canadienne d'Economique 13:268–279
Gaudry MJI, Dagenais MG (1979) Heteroscedasticity and the use of Box-Cox transformations. Econ Lett 2:225–229
Hadjiconstantinou E, Christofides N (1999) An efficient implementation of an algorithm for finding k-shortest simple paths. Networks 34:88–101
HEATCO (2005) Developing harmonised European approaches for transport costing and project assessment. Deliverable 1: current practice in project appraisal in Europe. European 6th Framework Programme, EU
Hensher DA (2008) Empirical approaches to combining revealed and stated preference data: some recent developments with reference to urban mode choice. Res Transp Econ 23:23–29
Hinloopen E, Nijkamp P (1990) Qualitative multiple criteria choice analysis. Qual Quant 24:37–56
Hjorth K, Fosgerau M (2011) Using prospect theory to investigate the low value of travel time for small time changes. Transp Res B 46:917–932
HMT (2003) Green book: appraisal and evaluation in central government. HMSO, London, UK
Hoffman W, Pavley R (1959) A method for the solution of the n-th best path problem. J Assoc Comput Mach 6:506–514

Huang C. (2013) Examining decision making surrounding the use of managed lines by Katy freeway travellers: a prospect theory approach. PhD thesis, Texas A&M University, College Station, USA
Jou R, Chen K (2013) An application of cumulative prospect theory to freeway drivers' route choice behaviours. Transp Res A 49:123–131
Judge GG, Hill CR, Griffiths WE, Lütkepohl H, Lee T-C (1988) Introduction to the theory and practice of econometrics, 2nd edn. Wiley, New York
Kahneman D, Tversky A (2007) Choices, values and frames. Cambridge University Press, New York
Kalouptsidis N, Psaraki V (2010) Approximations of choice probabilities in mixed logit models. Eur J Oper Res 200:529–535
Koppelman F, Wen C (1998) Alternative nested logit models: structure, properties and estimation. Transp Res B 32:289–298
Larsen OI, Jansson K, Lang H (2010) On combining discrete choice and assignment models. In: Sumalee et al (eds) Proceedings of the 15th international conference of Hong Kong society for transportation studies 2010, transportation and urban sustainability
Litman T (1999) Evaluating public transit benefits and cost. Victoria Transport Policy Institute, Victoria, BC
Logan TK, Padgett DK, Thyer BA, Royce D (2006) Program evaluation: an introduction. Thomson Brooks/Cole, Belmont
Lohes D (2011) Grundlagen der Straßenverkehrstechnik und Verkehrsplanung, vol 2. Verlag für Bauwesen, Berlin, Germany, Verkehrsplanung
Louviere JJ, Hensher DA, Swait J (2000) Stated choice methods and analysis. Cambridge University Press, Cambridge
Lyk-Jensen VS (2007) Appraisal methods in the Nordic countries—infrastructure assessment. Report 3, Danish Transport Research Institute for Transport and Energy Ministry, Denmark
Manktelow K (2012) Thinking and reasoning. Psychology Press, East Sussex
Mc Fadden D, Train K (2000) Mixed MNL models for discrete response. J Appl Econometrics 15:447–470
Meyer MD, Miller EJ (2001) Urban transportation planning. The McGraw-Hill Companies, New York
Nijkamp P, Rietveld P, Voogd H (1990) Multi-criteria evaluation in physical planning. Elsevier Science, Amsterdam
OECD ECMT (2005) National systems of transport infrastructures planning. ECMT 2004 Round Table 128, Paris
Ortuzar J, Willumsen LG (2011a) Modelling transport, 4th edn. Wiley, West Sussex
Ortuzar J, Willumsen L (2011b) Modelling transport, 4th edn. Wiley, New York
Papola A (2003) Some developments on the cross-nested logit model. Transp Res B 38:833–851
PIARC (2004) Economic evaluation methods for road projects in PIARC member countries. PIARC
Postorino MN (1993) A comparative analysis of different specifications of modal choice models in an urban area. Eur J Oper Res 71:288–302
Prashker JN, Bekhor S (1998) Investigation of stochastic network loading procedures. Transp Res Rec 1645:94–102
Prato CG, Bekhor S (2006) Applying branch and bound technique to route choice set generation. Transp Res Rec 1985:19–28
Prochaska J, Di Clemente C (1986) Toward a comprehensive model of change. In: Miller W, Heather N (eds) Treating addictive behaviours. Plenum Press, New York
Raveau S, Muñoz JC (2014) Analyzing route choice strategies on transit networks. In: Proceedings of 93rd annual meeting of the transportation research board, Washington, DC
Raveau S, Muñoz JC, De Grange L (2011) A topological route choice model for metro. Transp Res A 45:138–147
Saaty TL (1990) Multi-Criteria decision making: the analytic hierarchy process. RWS Publications, Pittsburgh

Sheffi Y, Powell WB (1982) An algorithm for the equilibrium assignment problem with random link times. Networks 12:191–207

Timmermans H (2010) On the (Ir)relevance of prospect theory in modelling uncertainty in travel decisions. Eur J Transp Infrastruct Res 10:368–384

Train K (2009a) Discrete choice methods with simulation. Cambridge University Press, New York

Train KE (2009b) Discrete choice methods with simulation. Cambridge University Press, New York

Trochim WM (2006) Research methods knowledge base. Drake University, Des Moines

Tversky A, Kahneman D (1979) Prospect theory: an analysis of decision under risk. Econometrica 47:263–292

Tversky A, Kahneman D (1992) Advances in prospect theory: cumulative representation of uncertainty. J Risk Uncertainty 5:297–323

Van de Kaa EJ (2010) Prospect theory and choice behaviour strategies: review and synthesis of concepts from social and transport sciences. Eur J Transp Infrastruct Res 10:299–329

Vovsha P (1997a) Application of cross-nested logit model to mode choice in Tel Aviv, Israel, metropolitan area. Transp Res Rec 1607:6–15

Vovsha P (1997b) The cross-nested logit model: application to mode choice in the Tel Aviv metropolitan area. Transp Res Rec 1607:13–20

Wen CH (2010) Alternative tree structures for estimating nested logit models with mixed preference data. Transportmetrica 6:291–309

Wen CH, Koppelman FS (2001) The generalized nested logit model. Transp Res B 35:627–641

Wholey JS (2004) Handbook of practical program evaluation. Jossey-Bass, San Francisco, California

Wickens C, Lee J, Liu Y, Becker SEG (2004) An introduction to human factors engineering. Pearson/Prentice Hall, Upper Saddle River

World Bank (1996) Sustainable transport: priorities for policy reform. World Bank, Washington, DC

Xu H, Zhou J, Xu W (2011) A decision-making rule for modelling travelers' route choice behaviour based on cumulative prospect theory. Transp Res C 19:218–228

Yáñez MF, Raveau S, de Ortúzar JD (2010) Inclusion of latent variables in mixed logit models: modelling and forecasting. Transp Res A 44:744–753

Yen JY (1971) Finding the K shortest loopless paths in a network. Manage Sci 17:712–716

Yoon KP, Hwang CL (1995) Multiple attribute decision making: an introduction. Sage University Papers series on Quantitative Applications in the Social Sciences, 07–104. Thousands Oaks, Sage, California, USA

Chapter 5
From Transit Systems to Models: Data Representation and Collection

Klaus Noekel, Guido Gentile, Efthia Nathanail and Achille Fonzone

This chapter deals with the data that form input and output of passenger route choice models. All information about supply and demand that is relevant to passenger route choice must be captured in a formal way in order to be accessible to mathematical choice models. Over time standard conventions for this formalisation have emerged. In order to avoid repetition in Part III, they are presented once in Sect. 5.1. Similarly, Sect. 5.2 formalises the output of route choice models as quantitative indicators which are then fed back into higher level decision models and ultimately into the evaluation of different investment options. Finally, Sect. 5.3 reviews real-world data sources from which either model input can be extracted or against which model output can be compared to validation and calibration. Here, the impact of ITS is particularly visible, as availability, diversity and volume of data have increased rapidly.

K. Noekel (✉)
PTV AG, Haid-Und-Neu-Strasse 15, 76131 Karlsruhe, Germany
e-mail: klaus.noekel@ptvgroup.com

G. Gentile
DICEA—Dipartimento di Ingegneria Civile Edile e Ambientale, Sapienza University of Rome, Via Eudossiana, 18, 00184 Rome, Italy
e-mail: guido.gentile@uniroma1.it

E. Nathanail
University of Thessaly, Pedion Areos, 38334, Volos, Greece
e-mail: enath@civ.uth.gr

A. Fonzone
Transportation Research Institute, Edinburgh Napier University, 10 Colinton Road, EH10 5DT Edinburgh, UK
e-mail: A.Fonzone@napier.ac.uk

G. Gentile and K. Noekel (eds.), *Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems*, Springer Tracts on Transportation and Traffic 10, DOI 10.1007/978-3-319-25082-3_5

5.1 Input: Demand and Supply

Klaus Noekel and Guido Gentile

The many possible forms of public transport were described informally in Chap. 2. In this section, we develop a formal data model with a standard terminology which covers a large spectrum of conventional transit systems. The main exceptions are demand–responsive systems without pre-defined lines and/or timetables, where transport services are organised in flexible tours (e.g., personal rapid transit or minibus-taxis in developing countries).

To keep the data model concise and compact, we focus on the aspects that are relevant input to any transit assignment model, describing passenger route choice and service performance. We ignore instead features that are of interest only in other areas of public transport planning or in operations, as well as special cases that would complicate the presentation of models and algorithms without explaining novel phenomena.

Some basic definitions and notations were already introduced in Sect. 4.2.1; here, they will be further specified and developed.

5.1.1 Travel Demand and Its Segmentation

In this section, the segmentation of travel demand is addressed from different points of view: space (zones), time (intervals), users (classes) and means (modes).

5.1.1.1 Zones

People spend their time doing *activities* (such as staying at home, working, shopping) in different *locations*. These activities convey them some utility, but require *trips* from one location to another. In the following, we refer to *travellers*, or *passengers*, when focussing on public transport.

The locations where passengers engage in activities are aggregated into larger geographic entities called *zones*. Thus, the land is partitioned into a set of zones, denoted Z.

All socio-economic activities located in each zone $z \in Z$ are assumed to be concentrated in one single point, called *centroid*, where trips can start and end. In the transport network, each zone centroid z is associated with one (or more) *origin* node $O_z \subseteq N$ and with one (possibly different) *destination* node $D_z \in N$. Origins and destinations make up two node sets, which are denoted $O = \cup_{z \in Z} O_z$ and $D = \cup_{z \in Z} D_z$. Destinations have the same cardinality as zones. The same is true for origins, unless a space-time network is considered, in which case one origin is created for each zone and for each possible departure time, because the latter is embedded in the network topology. The concept of space-time network is described in detail in Sect. 6.3.

5.1.1.2 User Classes and Modes

If passengers are not homogeneous in their personal characteristics and trip purposes (and then in their preferences), the travel demand may be segmented into a set G of *user classes*.

Trips can be performed in a set M of different ways called *transport modes*, among which users choose (modal split). Each mode is a set of *transport systems* and of rules on how/where it is possible to access/egress them. A transport system is a specific technological solution for travelling, not only with its own infrastructures and vehicles (means of transport), but also with its specific operations and regulations. Most private modes (e.g., car, bicycle, motorcycle and feet) have only one transport system; although if parking is considered explicitly, also the pedestrian part of these journeys may become significant and shall then be explicitly represented. Public transport is a typical case of a mode where several transport systems may be used during one single trip (e.g., bus, tram, metro, train and feet). Actually, in metropolitan areas, these transport systems are so much interconnected into a transit network that the whole public transport can be thought as one transport system.

Generally, each arc of the network can be traversed using several transport systems, but at a different time and cost (the cost is infinite if an arc is denied to a specific transport system).

However, a different approach is adopted here: we introduce sub-networks, consisting of separate layers of arcs and nodes that are reserved for one specific transport system. This is mostly convenient to represent *combined modes* (such as park and ride, bike and ride, and their return; car and bike sharing) whose trips include relevant legs on different transport systems and usually require to access and egress parked vehicles. Transfers between *transport system networks* (e.g., from car to transit, from walking to car, from walking to lines, and vice versa) may occur only through *inter-modal arcs* (e.g., parking arcs and stop arcs).

Here, a (combined) transport mode is defined as a set of transport system networks through a connection matrix with Boolean entries among them to specify which inter-modal arcs can be used; a simple mode can be seen as a particular combined mode that can use only one transport system network (only the corresponding element of the matrix on the diagonal is true). Origin and destination nodes can be considered as two particular networks without arcs that are included in every mode, while the connector arcs exiting origins and entering destinations can be considered as particular inter-modal arcs. The key idea is that it should not be possible to reach the destination from the origin, unless using a feasible sequence of transport systems for that combined mode.

In this way it is possible to exploit the network topology to express constrains and rules, such as park and ride users have to switch from car to transit at some inter-modal arc (exchange parking) during their trip. This combined mode can be implemented through a connection matrix with only the following entries equal to one, as depicted in Fig. 5.1: from origins to car, from car to transit (including walking) and from transit to destinations.

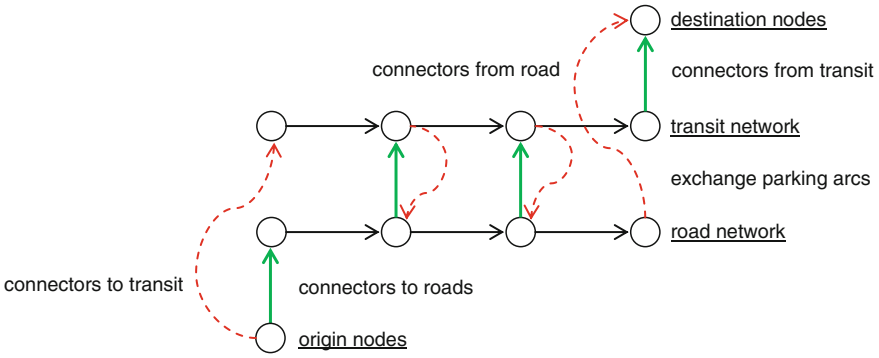

Fig. 5.1 Park and ride implemented through inter-modal (*vertical*) arcs among transport system networks (*horizontal arcs*); the *green arcs* are included in this transport mode, while the *dashed red arcs* are not

Another relevant example is car sharing, where users can freely switch during their trip from the pedestrian network to the road network (and vice versa) through inter-modal arcs; these can be dedicated parking slots or any node of the road network where parking is allowed (even if parking occurs along streets, it is usually represented at nodes), depending on the rules of the service. This combined mode can be implemented through a connection matrix with only the following entries equal to one, as depicted in Fig. 5.2: from origins to walking, from walking to cars and vice versa where parking the shared vehicle is permitted, from walking to destinations.

The proposed approach results in a multi-modal network where each mode $m \in M$ can use only a subset $A_m \subseteq A$ of arcs (and an arc is included in a subset of

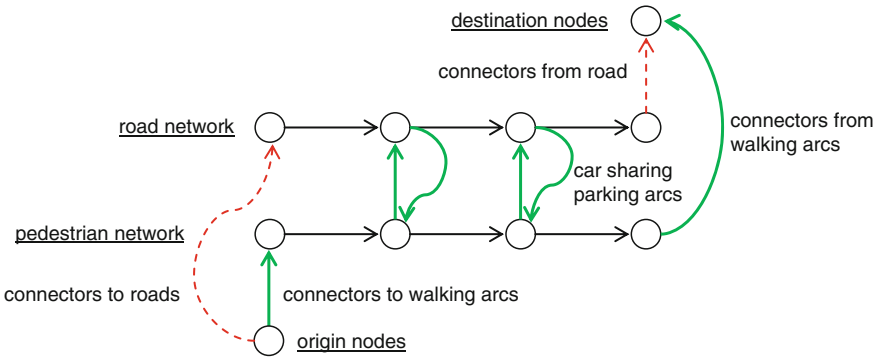

Fig. 5.2 Car sharing implemented through inter-modal (*vertical*) arcs among transport system networks (*horizontal arcs*); the *green arcs* are included in this transport mode, while the *dashed red arcs* are not

modes) while there is no need to formally introduce the transport system networks, which will remain implicit in our formulation.

In some applications, routes on combined modes must use at least once a given system (e.g., a trip on public transport shall use at least a transit line to distinguish it from a solely pedestrian trip; a train trip shall use at least a railway line to distinguish it from a generic trip on public transport) or must use the available transport systems in a given order (for example to comply with some fare rule). In some of these cases, travel cost will depend on the combinations/sequences of arcs traversed. To represent this through the graph topology, some of the transport system networks (for example the pedestrian network) are duplicated to separately reproduce access, transfer and egress. This leads to the concept of journey layers that is presented in Sect. 6.2.5 as a flexible paradigm to simulate complex fare structures.

Taxi, car-pooling and in general demand–responsive systems are the most difficult modes to code, because routes are flexible.

5.1.1.3 Scenarios and Time Intervals

Each simulation set-up (input and output) refers to a specific *scenario* with a particular day type (e.g., winter working day of the base year) and measures (intervention) being evaluated. The reference to the scenario will be omitted in the following.

If within-day variations of passenger demand and/or network performances are relevant (e.g., in schedule-based models), the *assignment period* (e.g., one day or the morning peak) is discretized into η subsequent *intervals* separated by an ordered set of *instants* whose generic integer index is $t \in [0, \eta] = T$ and whose generic clock *time* is $\tau_t \in \Re_+$. The set of clock times is denoted $T = \{\tau_t : t \in T\}$. By convention, we refer to an interval $t \in T$ as to its initial instant, so that its duration is $h_t = \tau_{t+1} - \tau_t$; therefore, the interval associated with the last instant η is dummy or may represent a possible discharging phase after the assignment period. An additional instant $\eta + 1$ with $\tau_{\eta+1} = \infty$ is also introduced for consistency, to represent events occurring after the assignment period $[\tau_0, \tau_\eta]$ or events not referred to a specific time.

In dynamic assignment models, time intervals are introduced not only for the description of travel demand and supply variations within the day, but are also used in the network computation of route choice, flow propagation and congestion models.

In practice, the assignment period is usually partitioned into (fewer) *demand time slices*, which result from the aggregation of several (shorter) time intervals. But in the following, for the sake of simplicity, we will not consider this opportunity and in our data model the demand time slices will coincide with the time intervals T.

If a space-time network is used, nodes have also an embedded time coordinate; thus, the system dynamic is represented through the topology of the so-called *diachronic graph*.

If a macroscopic dynamic model is considered, all network object attributes are represented as semi-continuous functions of time (sometimes called temporal profiles).

In micro-simulation models, time-based events are handled sequentially.

In static models, the time dimension is simply neglected.

5.1.1.4 Origin–Destination Matrices and Routes

Travel demand is expressed in the form of *Origin–Destination* (*O–D*) *matrices*. In the simplest case, a single matrix will suffice. Otherwise, a separate O–D matrix is associated with each class $g \in G$, mode $m \in M$ and time interval $t \in T$. The flow of class g users that travelling on mode m departs during time interval t from origin $o \in O \subseteq N$ directed towards destination $d \in D \subseteq N$ is denoted d_{odmgt}, which is the generic entry of the *O–D matrix* $\mathbf{d}_{ODmgt}$.

For every *Origin–Destination pair* $od \in O \times D$, we assume that there exists a set K_{odm} of *routes* connecting on the sub-graph (N, A_m) of mode $m \in M$ the origin o to the destination d. Typically, each route $k \in K_{odm}$ is represented by an acyclic path that is a concatenated sequence of arcs $A_k \subseteq A$ and nodes $N_k \subseteq N$ whose *origin node* k^- is o and *destination node* k^+ is d; however, more general topological representations of routes are possible (e.g., paths with cycles and hyperpaths, the latter introduced in Sect. 6.1.3).

Let K be the set of all routes: $K = \cup_{odm \in O \times D \times M} K_{odm}$. The origin, destination and mode of route $k \in K$ are denoted, respectively: $O_k \in O$, $D_k \in D$, $M_k \in M$.

The portion of the O–D demand of class $g \in G$ using path $k \in K_{odm}$ is called *route flow* and denoted q_{kgt}.

In static models, the reference to the time interval is omitted, while if a space-time network is used, the reference to time is embedded in the origin node. If there is only one user class, then the reference to it is omitted. If the focus is on one mode (e.g., public transport, which includes the transit lines and the pedestrian network), then the reference to it is omitted.

5.1.2 Transport Network and Transit Services

In this section, the input data describing the transit network and its all features are introduced.

5.1.2.1 Base Network

The topology and the main attributes of the infrastructures, such as roads and rail, as well as the abstract connections between locations, are described through the *base network* (B, E), where B is the set of *vertices* and $E \subseteq B \times B$ is the set of *edges* (note

that here the names vertex and edge are used instead of node and arc to avoid confusion between the input data and the model representation of the network). Each vertex has geographic coordinates, and edges are described by polylines with intermediate points.

A subset $E^{walk} \subseteq E$ of edges are walkable, and others are introduced just to describe the line routes but are not walkable. The generic edge $a \in E$ has a *length* l_a and a *walking speed* s_a^{walk}. We assume that walkable edges have by definition a positive speed: E^{walk}: $\{a \in E: s_a^{walk} > 0\}$.

The centroid of each zone $z \in Z$ is associated with an *origin node*, denoted $B_z^{orig} \in B$, and with a (possibly different) *destination node*, denoted $B_z^{dest} \in B$. The edges that exit from an origin vertex or enter into a destination vertex are called *connectors*, as they are aimed at representing, respectively, access and egress of the base network.

5.1.2.2 Stops and Lines

A transit network consists of a set S of stops between which services operate. A *stop* $s \in S$ is indeed a unique location (with geographic coordinates) where passengers can board and/or alight from transit services, e.g., a particular platform or curbside. The defining characteristic of a stop is that transfers within a single stop take zero walking time. Published timetables commonly group together several "stops" (in the above interpretation) under a single name, e.g., the two stops on opposite sides of the street served by the two directions of a bus line. By definition, these are two distinct stops rather than one, because any transfer between them requires non-zero walking time for crossing the street. There is no qualitative difference between such a transfer and one that involves walking to a differently named stop.

Often, several stops and platforms are functionally grouped in *stations*, sometimes called stop areas. These sets of stops may be perceived by passengers and managed by operators as one network element. The aggregation of stops is not explicitly formalised in the proposed data model, but nothing prevents from introducing it in practice.

While passengers are waiting at a stop, they are guaranteed to observe all services departing from there. In addition, they may, or may not, observe static information about scheduled service or dynamic information about actual service. Unless stated otherwise, we assume that this information only relates to services at the current stop.

Transit services are organised in a set L of lines. A *line* $\ell \in L$ serves in one direction an ordered set of stops, its *stop sequence* or *itinerary* denoted $S_\ell \subseteq S$, with no repetitions (Fig. 5.3). Thus, circular lines and side-trips within lines (i.e., a deviation from the main arteria to serve a stop located along an adjacent local street) are excluded to simplify the presentation. Stops which are passed without boarding and alighting are here omitted from the stop sequence. The *first stop* of a line, denoted $S_\ell^- \in S_\ell$, and its *last stop*, denoted $S_\ell^+ \in S_\ell$, are also called line *terminals*. The *successive stop* of stop $s \in S_\ell - S_\ell^+$ is denoted $s_\ell^+ \in S_\ell$; the *previous stop* of stop

$s \in S_\ell - S_\ell^-$ is denoted $s_\ell^- \in S_\ell$. The part of a line between one stop $s \in S_\ell - S_\ell^+$ and the successive one s_ℓ^+ is called a *line segment*; by convention, we refer to it as to its first stop, so that the line segment associated with the last stop is dummy. Clearly, more lines can share the same stop.

Movements of transit vehicles for purely operational purposes, e.g., pull-in, pull-out or dead-head runs (empty workings), are ignored. Circle lines may allow passengers to continue their trip without transferring at (arbitrary) terminals. This is ignored here for simplicity.

Similar to stops, published timetables commonly group together under the same name, the line serving a given stop sequence and the line serving the reverse sequence (and possibly still other variations of the stop sequence). In first approximation, the grouping has no consequence for passenger route choice and therefore we treat as distinct lines, the two directions as well as any service that has a different stop sequence. We also ignore special cases where boarding may be allowed at a stop, but not alighting, or vice versa.

During the interval $t \in T$, the generic line $\ell \in L$ is characterised by:

- a strictly positive *running time* $t_{\ell st}^{run}$, for each line segment $s \in S_\ell - S_\ell^+$;
- a non-negative *dwell time* $t_{\ell st}^{dwell}$, for each stop $s \in S_\ell - S_\ell^- - S_\ell^+$.

Unless stated otherwise running and dwell times are assumed to be deterministic.

The dwell time is required to allow alighting and boarding passengers, respectively, in and out the carrier, but it can also represent a longer stationing. In practical application, dwell times are often assumed null. Otherwise, the dwell time can depend on the alighting and boarding flows as a result of a model representing the related congestion phenomena.

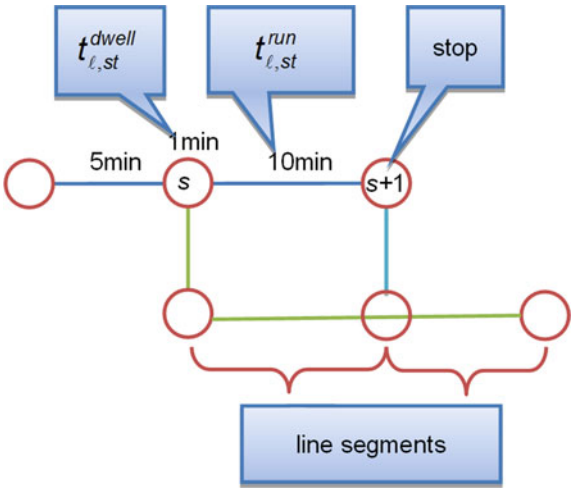

Fig. 5.3 Stops, lines and line segments; running and dwell times. All *circles* denote stops, differently *coloured lines* denote different transit lines

Often, the running time is not a direct input, but derives instead from more aggregated data sources. In practical applications, each line segment $s \in S_\ell - S_\ell^+$ is associated with an acyclic path on the base network, whose *support* edges are denoted as $B_{\ell s} \subseteq E$; this is essential to plot the line on a map. The length of the line segment $l_{\ell s}$ is then obtained as the sum of such edge lengths:

$$l_{\ell s} = \sum_{a \in B_{\ell s}} l_a. \tag{5.1}$$

The running time of line segment $s \in S_\ell - S_\ell^+$ is the sum of the travel time of its support edges plus a *stop time* t_ℓ^{stop} characteristic of the line:

$$t_{\ell st}^{run} = t_\ell^{stop} + \sum_{a \in B_{\ell s}} \frac{l_a}{s_{at}}, \tag{5.2}$$

where s_{at} is the *commercial speed* along edge $a \in B_{\ell s}$ of segment s during interval $t \in T$, under the assumption that all lines using that edge are characterised by the same transport system.

Commercial speeds of edges are defined by public transport companies (or by mobility agencies as an input for competitive service offers) based on:

- the performances of the mean of transport;
- the road speeds and the intersection regulations; and
- the level of road traffic congestion.

The stop time of the line takes the following into account:

- the accelerations and decelerations manoeuvres that are necessary to make stops;
- the time for opening and closing doors.

Automatic vehicle location (AVL) systems are very useful to properly set these values based on day type, hour and road section.

The generic line $\ell \in L$ is also characterised by the following:

- a strictly positive *alighting time* t_ℓ^{alight} that besides the time for getting off the carrier may represent accessory operations, such as baggage claim and passing custom;
- a strictly positive *boarding time* t_ℓ^{board} that besides the time for getting on the carrier, which is usually negligible, may represent the a priori anticipation of passengers in reaching the stop which is actually necessary for schedule-based services, also as a safety margin for coincidences against potential delays.

A route may contain transfers between different lines. Each portion between two transfers (or before the first/after the last transfer) is called a *leg*.

5.1.2.3 Runs and Timetables

Each line $\ell \in L$ is served by an ordered set of runs, called its *run sequence* and denoted R_ℓ. The set of all runs $R = \cup_{\ell \in L} R_\ell$ defines the whole transit service, while the line of run $r \in R$ is denoted $L_r \in L$.

A *run* $r \in R_\ell$ is constituted by one vehicle serving all stops of its line in order. The *first run* and the *last run* of line $\ell \in L$ are denoted, respectively, $R_\ell^- \in R_\ell$ and $R_\ell^+ \in R_\ell$. The *successive run* of run $r \in R_\ell - R_\ell^+$ is denoted $r^+ \in R_\ell$; the *previous run* of run $r \in R_\ell - R_\ell^-$ is denoted $r^- \in R_\ell$. As before, the part of a run between one stop $s \in S_\ell - S_\ell^+$ and the successive one $s_\ell{}^+$ is called a *run segment*. Where in reality some runs of a line may be confined to only a sub-sequence of the stops, this can be represented as serving all stops of a separate, shorter, line.

We assume that each run $r \in R_\ell$ has a well-defined schedule with an *arrival time* τ_{rs} for each stop, $s \in S_\ell - S_\ell^-$, and a *departure times* θ_{rs} for each stop $s \in S_\ell - S_\ell^+$ but the last one, at least in the form of a working timetable defined by the operator.

Optionally, only the departure time $\theta_r = \theta_{rs}$ from the first stop $s = S_\ell^-$ is provided as a specific run property, because it implicitly determines all other arrival and departure times downstream, for each other stop, through line running and dwell times as follows, $\forall s \in S_\ell - S_\ell^+$:

$$\theta_{rs} = \begin{cases} \theta_r, & \text{if } s = S_\ell^- \\ \tau_{rs} + t_{\ell st}^{dwell}, & t \in T : \tau_t \le \tau_{rs} < \tau_{t+1}, \quad \text{otherwise} \end{cases}, \tag{5.3a}$$

$$\tau_{rs^+{}^\ell} = \theta_{rs} + t_{\ell st}^{run}, \quad t \in T : \tau_t \le \theta_{rs} < \tau_{t+1}. \tag{5.3b}$$

For a given timetable, it is possible to calculate the following variables:

- the running time t_{rs}^{run} of run $r \in R_\ell$ on segment $s \in S_\ell - S_\ell^+$,

$$t_{rs}^{run} = \tau_{rs_\ell^+} - \theta_{rs}; \tag{5.4}$$

- the dwell time t_{rs}^{dwell} of run $r \in R_\ell$ at stop $s \in S_\ell - S_\ell^- - S_\ell^+$,

$$t_{rs}^{dwell} = \theta_{rs} - \tau_{rs}. \tag{5.5}$$

Then, if the schedule is provided as a direct input, the line running times at any stop $s \in S_\ell - S_\ell^+$ and dwell times at any stop $s \in S_\ell - S_\ell^- - S_\ell^+$ can be determined for each interval $t \in T$ as an average of the running and dwell times of all runs departing and arriving during the interval, respectively:

$$t_{\ell st}^{run} = \sum_{r \in R_\ell : \tau_t \le \theta_{rs} < \tau_{t+1}} t_{rs}^{run}, \tag{5.6}$$

$$t_{\ell st}^{dwell} = \sum_{r \in R_\ell : \tau_t \le \tau_{rs} < \tau_{t+1}} t_{rs}^{dwell}. \tag{5.7}$$

5.1.2.4 From Schedule to Headways and Frequencies

The frequency of a line is the number of run departures from a particular stop in a given time interval; i.e., it is the flow of carriers departing from the stop. Only under the assumption of constant (i.e., not time-varying) running and dwell times, the frequency of a line is for all stops identical to that of the first stop, if the latter is constant as well. Otherwise, the frequency at different stops follows the dynamic propagation law for FIFO flows of vehicles departed from the first stop (see Sect. 6.4.4). The time interval between two successive departures of a line from a given stop is called the headway.

Scheduled times refer to a priori planned operations, i.e., without disturbances, and may differ from actual arrival and departure times that occur in practice. If services are punctual, these two coincide; otherwise, the difference between actual and scheduled times is called *delay*.

The *headway* $h_{\ell st}$ of line $\ell \in L$ at a given stop $s \in S_\ell - S_\ell^+$ during interval $t \in T$ is then represented here as an independent (this is an often implicit assumption) random variable with a given distribution $\varphi_{\ell st}^h(h)$. The two most common assumptions are given as:

- the exponential distribution, which entails the lowest possible regularity;
- the deterministic distribution, which entails the highest possible regularity.

The *frequency* $f_{\ell st}$ of line $\ell \in L$ at a given stop $s \in S_\ell - S_\ell^+$ during interval $t \in T$ is defined as the inverse of the headway expected value:

$$f_{\ell st} = \frac{1}{E\left(h_{\ell st}\right)}. \tag{5.8}$$

The *irregularity* of the service, which is another fundamental feature of the line, is related also to the other parameters of the headway distribution, as illustrated later on (see Sect. 6.2.1); in particular, it is here defined as the ratio between the standard deviation of the headway and its expected value, which is also called *variation coefficient*:

$$\sigma_{\ell st} = \frac{SD\left(h_{\ell st}\right)}{E\left(h_{\ell st}\right)}. \tag{5.9}$$

As mentioned earlier, the frequency is often assumed to be constant along the whole line; in this case, the reference to the stop and time interval is omitted.

Headways and frequencies can also be retrieved from the schedule. For a given timetable, it is possible to calculate the following variables:

- the *departure headway* h_{rs}^{dep} of run $r \in R_\ell - R_\ell^-$ from stop $s \in S_\ell - S_\ell^+$;
- the *arrival headway* h_{rs}^{arr} of run $r \in R_\ell - R_\ell^-$ to stop $s \in S_\ell - S_\ell^-$;

$$h_{rs}^{dep} = \theta_{rs} - \theta_{r^-s}, \tag{5.10}$$

$$h_{rs}^{arr} = \tau_{rs} - \tau_{r^-s}. \tag{5.11}$$

Then, if the schedule is provided as a direct input, with the same approach used to compute (5.6) and (5.7), the line frequency at any stop $s \in S_\ell - S_\ell^+$ can be determined for each interval $t \in T$ as the inverse of the average headway of all runs departing during the interval:

$$f_{\ell st} = \frac{1}{E\left(h_{rs}^{dep}, \quad \forall r \in R_\ell - R_\ell^- : \tau_t \le \theta_{rs} < \tau_{t+1}\right)}. \tag{5.12}$$

The above-defined irregularity can be estimated from the schedule as follows:

$$\sigma_{\ell st} = f_{\ell st} \cdot SD\left(h_{rs}^{dep}, \quad \forall r \in R_\ell - R_\ell^- : \tau_t \le \theta_{rs} < \tau_{t+1}\right). \tag{5.13}$$

5.1.2.5 Capacity and Comfort

All transit services have a limited capacity which is determined by the size of vehicles operating the line as well as by the service frequency. Assuming that the service is operating regularly, the *line capacity*, that is the maximum flow of passengers served at a given stop, can be determined simply as a product of the vehicle individual capacity and the line frequency.

The *vehicle capacity*, that is the maximum number of on-board passengers, is usually made up of the number of available seats, also called *seating capacity*, plus an assumption on how many passengers can find a standing space. Hence, the *standing capacity* is obtained by multiplying the available space for standing with a factor representing the maximum number of passenger per square metre, also called *crush capacity*.

Operators may assume different crush capacities under different conditions. Transport for London, for example, assumes a "physical" crash capacity of 7 pax/m^2 and a "practical" crash capacity of 5 pax/m^2, to take into account the uneven load of passengers on the same train.

Each vehicle has doors to allow passengers in and out the carrier. The maximum flow of passengers that the vehicle can exchange with the stop when doors are open is called the *door capacity*. When doors are dedicated to entry flows and exit flows, the door capacity is given by the sum of a *boarding capacity* and an *alighting capacity*.

Bus stops and train platforms have a maximum number of passengers that can be safely hosted; this is the so-called *stop capacity*. Operators may regulate the access to the whole station with the aim of preventing the number of waiting passengers exceeding the capacity of a single stop.

In transit networks, we can then distinguish among several relevant types of capacity:

- vehicle capacity κ_ℓ^{veh} determines the maximum number of passengers which can physically be (stand and seat) on-board of the typical carrier that serves line $\ell \in L$; this is the sum of
- seating capacity κ_ℓ^{seat}, the number of available seats;
- standing capacity κ_ℓ^{stand}, the number of available standing places;
- crush capacity κ_ℓ^{crush}, the maximum passengers per square metre;
- door capacity κ_ℓ^{door} determines the maximum number of passengers which can board and alight in a unit of time the typical carrier that serves line $\ell \in L$; this is the sum of
- boarding capacity κ_ℓ^{board}, the maximum flow of passenger that can board when (dedicated) doors are open; and
- alighting capacity κ_ℓ^{alight}, the maximum flow of passenger that can alight when (dedicated) doors are open.
- stop capacity κ_s^{stop} determines the maximum number of passengers which can safely be hosted at once by the platform of stop $s \in S$.

The line capacity is thus a derived feature rather than a direct input.

While speed mainly influences travel times and frequency mainly influences waiting times, capacity is the main service characteristic that influences transit congestion phenomena related in primis to discomfort (seat unavailability and vehicle overcrowding), but also to queuing and irregularity; it can thus have indirectly a profound effect on route choice.

Stops possess characteristics that can significantly reduce the high psychological burden of waiting, due to the fact that passengers must pay attention and continuously check for carrier arrival whose instant is unknown. The provision of proper information (e.g., the arrival times of line vehicles) can play a crucial role, as well as the ergonomy of the stop in general (e.g., shelter, seating, air conditioning, entertainment, safety and security, protection from road noise and pollution). For longer waits, when the passenger can spend some time in the station before boarding a specific run without concerns about the departure time, the presence of other activities (e.g., shops) can have a relevant value.

Lines possess additional characteristics that may heavily impact on on-board comfort: seat ergonomy, vehicle style, air conditioning, cleanness and silence inside carriers, safety and security on-board, additional services such as (Wi-Fi) telecommunications, refreshments and entertainments. Another relevant (bundle) characteristic is the physical *mean* of transport used to provide the service (e.g., bus, tram, metro, regional train, coach, high-speed train, plane) on which passengers may have preferences.

Finally, line segments are characterised with passengers flows, which can be put in relation to capacities to obtain seat unavailability and vehicle overcrowding.

Stops $s \in S$ and lines $\ell \in L$ are then also characterised by other (more qualitative than quantitative) attributes that are however relevant to the route choice of passengers, which make up two vectors $\mathbf{a}_s$ and $\mathbf{a}_\ell$, respectively; among these, the

Boolean variables identify the mean of transport and the comfort features, for example:

- the stop has a shelter;
- the stop has dedicated shops;
- the line is a bus;
- the line is a train; and
- the line has air conditioning.

These (often very important) features may represent attributes of the generalised cost, differently perceived by each user class $g \in G$, which are connected with the travel or waiting time that gives the exposure to discomfort. They can then be synthesised in a *stop discomfort coefficient*, denoted γ_{sg}^{stop}, and a *line discomfort coefficient*, denoted $\gamma_{\ell g}^{line}$. Usually, the following linear form is assumed:

$$\gamma_{sg}^{stop} = 1 + \sum_{c \in C^{cs}} \beta_{cg}^{stop} \cdot a_{sc}, \tag{5.14}$$

$$\gamma_{\ell g}^{line} = 1 + \sum_{c \in C^{cl}} \beta_{cg}^{line} \cdot a_{\ell c}, \tag{5.15}$$

where β_{cg}^{stop} and β_{cg}^{line} are the utility coefficients relative to stop and line attributes a_{sc} and $a_{\ell c}$, while C^{cs} and C^{cl} are the set of such attributes.

However, discomfort at stops and on-board is typically the result of a more complex model which represents congestion phenomena such as overcrowding (see Sect. 7.2.1).

Thus, discomfort is not only caused by the crowding on-board but also at stops, as well as on some pedestrian links inside stations. For example, crowded platforms, stairs, escalators and lifts may be perceived not only as inconvenient but also as a safety hazard, up to the point that station access control is sometimes applied by some operators during peak hours.

5.1.2.6 Fares

The major non-temporal element of generalised cost is *fare*, the monetary fee of using public transport. A vast variety of fare schemes is in use throughout the world, and a complete taxonomy is outside the scope of these notes. Here, we focus on fare schemes which are widely applied and, at the same time, can be easily modelled in route choice.

Note first that passenger route choice refers here to the decision for a single trip. However, some fare options with lower cost per trip are conditional upon an up-front payment, such as the purchase of a monthly pass. Such options are always taken in relation to a larger number of expected trips in a certain time frame. These buying decisions, such as other long-term mobility choices, e.g., the decision to own a car, are part of an earlier stage in demand modelling. One way to abstract

from this earlier choice step is to assume that travel demand is segmented into user classes for the different multi-trip fare available options, e.g., occasional passengers, who buy a trip ticket, and commuters, who hold a flat subscription. Moreover, fares influence passenger route choice only if alternative routes for the same trip have different monetary cost associated with them. If the fare is constant or does not depend on the particular route, then fare may be relevant to mode choice, but drops out of route choice.

Among the fare schemes that do influence route choice, we find the following typical forms (Fig. 5.4).

- *Distance-based fare*. The fare is a function of the distance covered. This function is usually assumed to be monotonically increasing, but is otherwise of arbitrary shape, possibly with discontinuities, possibly degressive rather than linear. Moreover, this may apply to either the complete passenger trip or to each trip leg between two consecutive transfers.
- *Zone-based fare*. The fare is a function of the number of traversed *fare zones* that are usually an aggregation of traffic zones (typically, the fare zones are concentric). The function has the same properties as in the previous case.
- *Short-distance fare*. The fare is a constant amount, but only valid on trips which do not exceed a specified threshold on distance, travel time and/or number of stops.

The most general way of reproducing the above schemes is to associate a *route fare* c_{kg}^{fare} to each route $k \in K$ and class $g \in G$. But this is a rather unpractical data structure; we can obviate it by introducing a specific procedure capable of computing on the fly the fare of a given route based on the specific pricing rules. However, this approach requires a model with explicit path enumeration and some scripting in the assignment algorithm.

An alternative approach that is particularly convenient for the representation of fares is based on the explicit representation of legs through specific arcs. A leg is a direct connection between two stops of a same transit line, which represents a sequence of line segments. This way it is possible to reproduce non-additive costs in the route choice.

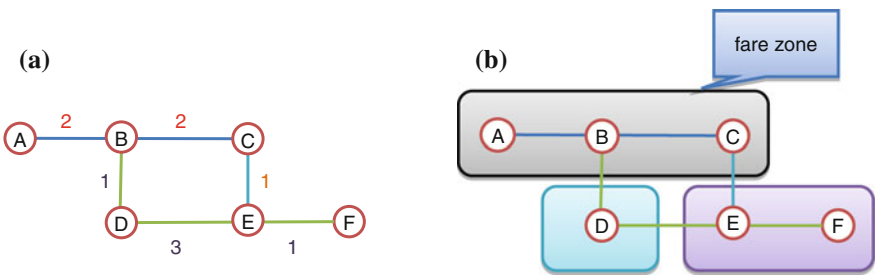

Fig. 5.4 **a** Example of an additive distance-based fare with line segment fares. **b** Example of a zone-based fare with stops grouped into fare zones (fare amounts not shown)

A much easier way of coping with monetary costs is to associate a *kilometric fee* to each line segment $c_{\ell s}^{kfee}$ for each $s \in S_\ell - S_\ell^+$ (which applies also to the corresponding run segments), and a one-time *boarding fee* $c_{\ell s}^{bfee}$ paid at accessing the service of line $\ell \in L$ at stop s. In this case, fares are said to be *additive*; implicit path enumeration is then possible, since the monetary cost for a given route can be obtained by summing over the fare of traversed line segments and boarded lines. A *fee multiplier* γ_g^{mfee} can be introduced to represent possible discounts for users of class $g \in G$. This simplified representation of monetary costs can result to be inadequate, depending on the particular application, especially if the objective is to design the fare structure.

5.1.2.7 Data Model Simplification

To simplify the data model (i.e., reduce the number of fields) and to make it more flexible for possible extensions, the attributes of the supply and of the demand objects can be obtained from a set of parameter bundles F, where each bundle $f \in F$ is a *dictionary of parameters* with values associated with keys. This approach is particularly convenient where the characteristics of network elements assume a reduced range of combined values, e.g., where link types are considered. This way of memorising attributes can be effectively applied to all real valued variables introduced so far, with three noticeable exceptions:

- the length of edges,
- the departure times of runs and
- the demand flows.

To this end, we introduce the following associations:

- $F_a \in F$ for each edge $a \in E$;
- $F_g \in F$ for each class $g \in G$;
- $F_s \in F$ for each stop $s \in S$;
- $F_\ell \in F$ for each line $\ell \in L$; and
- $F_{\ell s} \in F$ for each stop $s \in S_\ell$ of line $\ell \in L$.

5.1.2.8 Network Topology for Transit Assignment

The network used in transit assignment is not a direct input, but it is rather built-up from base data. Each different model/method constructs in general a specific topology with possibly different arc and node types, as detailed in Part III (see, for example, Sect. 6.2.2 for static networks and Sect. 6.3.1 for space-time networks).

The multimodal *network* of transport infrastructures and services is represented by means of a directed *graph* (N, A), where N is the set of *nodes*, each characterised by geographic coordinates (as well as by a temporal coordinate in space-time networks) and $A \subseteq N \times N$ is the set of *arcs* (ordered couples of nodes), each

representing an atomic *trip segment* of a specific *type* (e.g., walking from one point to another, waiting for a given interval or for a given event, riding on-board a line from a stop to the subsequent one, driving from one intersection to the next) on a specific transport system (e.g., public transport, car, bike). The sequence of trip segments of the same type is called *trip phase* or *trip leg*. Different models may disarticulate trips in different ways and identify different arc types.

The initial node of the generic arc $a \in A$ is referred to as *tail* and denoted $a^- \in N$, while the final node is referred to as *head* and denoted $a^+ \in N$. The set of arcs exiting the generic node $i \in N$ is referred to as its *forward star* and denoted $i^+ = \{a \in A: a^- = i\}$. Symmetrically, the set of arcs entering node $i \in N$ is referred to as its *backward star* and denoted $i^- = \{a \in A: a^+ = i\}$. The same notation for tail, head, forward and backward star can be applied to arcs and nodes of the base network.

Let $k = (N_k \subseteq N, A_k \subseteq (A \cap N_k \times N_k))$ be a *sub-graph* of (N, A). It is useful to introduce the following concepts:

- the *successor arcs* of $i \in N_k$, denoted $i_k^+ = i^+ \cap A_k$;
- the *predecessor arcs* of $i \in N_k$, denoted $i_k^- = i^- \cap A_k$;
- the *origin nodes* of k, denoted $k^- = \{i \in N_k: i_k^- = \varnothing\}$ (with no predecessors);
- the *destination nodes* of k, denoted $k^+ = \{i \in N_k: i_k^+ = \varnothing\}$ (with no successors).

An *acyclic path* k is a concatenated sequence of arcs that connects its origin to its destination, i.e., an acyclic sub-graph (N_k, A_k) with:

- $|k^-| = 1$, i.e., one origin node;
- $|k^+| = 1$, i.e., one destination node;
- $|i_k^+| = 1, \forall i \in N_k - k^+$, i.e., one successor, except for the destination node which has none;
- $|i_k^-| = 1, \forall i \in N_k - k^-$, i.e., one predecessor, except for the origin node which has none.

5.1.3 The Example Network

This section introduces a small network, depicted in Fig. 5.5, which will be used from now on in the text as a reference case, particularly in Part III to explain frequency-based and scheduled-based assignment models through numerical examples. The network consists of four stops (nodes contained in the oval enclosures) where passengers may transfer board and alight, and four lines with different headways, commercial speeds and vehicle capacities. There is also a fast pedestrian (e.g., bike sharing) connection between Stops 1 and 2.

The red and green lines are regular busses with high frequency. The maroon line is a metropolitan train with lower frequency but higher speed. The black line is a minibus with very high frequency but slower speed. The red line serves a separate itinerary, while the other 3 lines operate along the same corridor.

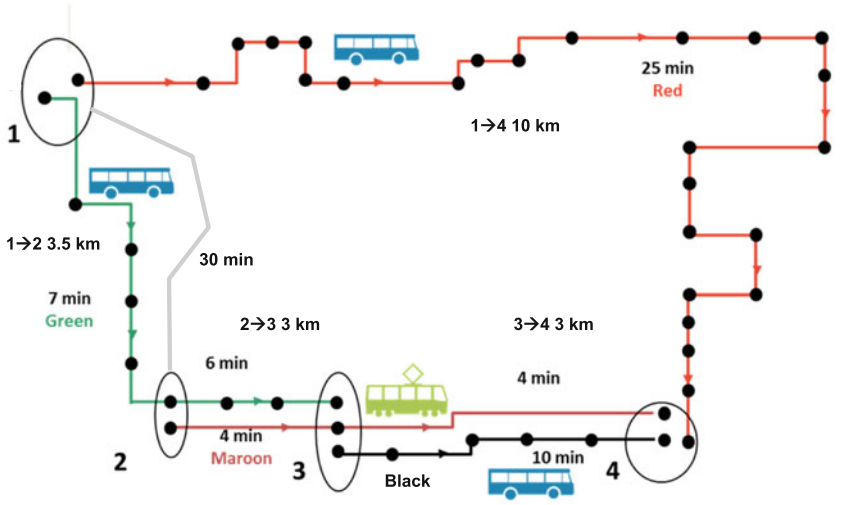

Fig. 5.5 A simple transit network with 4 stops and 4 lines (from Spiess and Florian 1989)

The main line features (frequency, commercial speed and vehicle capacity) are provided in Table 5.2, while Table 5.3 shows the running times of line segment computed on the basis of the supporting edges lengths that are reported in Table 5.1. The timetable with the passage of each run at each stop, reported in Table 5.4, is used only for schedule-based models, where services are assumed to be considerably less frequent, while the resulting travel times of the run sections are consistent with the commercial speeds of the lines. We can imagine that Stop 4 is downtown and all the demand is directed there in the morning (see Table 5.5).

5.1.4 Reference Notes and Concluding Remarks

It should be noted that the example network described above only includes the demand of public transport, while it completely ignores the interaction between public and private transport vehicles. This is a major simplification and may imply a

Table 5.1 Base edges

Tail	Head	Length (km)	Walking speed (km/h)
1	2	3.5	7
2	1	3.5	7
2	3	3	0
3	4	3	0
1	4	10	0

Table 5.2 Line data

Line	Frequency (veh/h)	Expected headway (min)	Commercial speed (km/h)	Vehicle capacity (pax)
Line 1—Red	10	6 = 60/10	24	80
Line 2—Green	10	6 = 60/10	30	80
Line 3—Maroon	4	15 = 60/4	45	80
Line 4—Black	20	3 = 60/20	18	80

Table 5.3 Line segment data

Line	Line segment	Running time (min)
Line 1—Red	(1, 4)	25 = 60 × 10/24
Line 2—Green	(1, 2)	7 = 60 × 3.5/30
Line 2—Green	(2, 3)	6 = 60 × 3/30
Line 3—Maroon	(2, 3)	4 = 60 × 3/45
Line 3—Maroon	(3, 4)	4 = 60 × 3/45
Line 4—Black	(3, 4)	10 = 60 × 3/18

Table 5.4 Timetable

	Line 1		Line 2			Line 3			Line 4	
	Stop 1	Stop 4	Stop 1	Stop 2	Stop 3	Stop 2	Stop 3	Stop 4	Stop 3	Stop 4
Run1	7.38	8.03	7.30	7.37	7.43	7.56	8.00	8.04	7.45	7.55
Run2	8.08	8.33	8.00	8.07	8.13	8.26	8.30	8.34	8.00	8.10
Run3	8.38	9.03	8.30	8.37	8.43	–	–	–	8.30	8.40
Run4	–	–	9.00	9.07	9.13	–	–	–	9.00	9.10

Table 5.5 O–D pairs and travel demand

Origin	Destination	Flow (pax/h)
Stop 1	Stop 4	300 = 5 pax/min
Stop 2	Stop 4	360 = 6 pax/min
Stop 3	Stop 4	240 = 4 pax/min

reduction of the scope of the models that are tested on it. On the other hand, this allows getting a closer comprehension of the models in their applications to transit networks.

In the following chapters, this network will be used to discuss some worked examples that will help to explain important phenomena and modelling issues such as:

- *Run* choice in schedule-based models versus *line* choice in frequency-based models.
- Shortest hyperpath versus shortest path in the modelling framework of route choice.
- The effect on route choice of no information; wayside information by means of countdown displays at the stops; and "continuous information" by means of held-held devices such as smart phones.
- In-vehicle overcrowding, discomfort and seat availability.
- Passenger overcrowding in the form of mingled queues or FIFO queues.
- Development of analytical versus simulation-based assignment models.

5.2 Output: Indicators

Efthia Nathanail, Irina Jackiva and Klaus Noekel

5.2.1 Introduction

This chapter introduces the output indicators most often used to assess the transit network performance.

Output indicators aim to evaluate a transit operation in quantitative or qualitative terms of performance, level of service, environment, economy or safety focusing on issues such as generalised cost, waiting/travel time, safety aspects, environmental impacts and perceived, desired or expected quality of service. While the primary output of the route choice model is determined by the specific method (e.g., analytic vs. simulation-based, macroscopic vs. microscopic), further indicators may be derived depending on the objective of the evaluation.

An *indicator* can be defined as "a parameter, or a value derived from parameters, which points to, provides information about, describe the state of a phenomenon/environment/area, with a significance extending beyond that directly associated with a parameter value" (OECD 2003).

Alternatively, indicators comprise measurements towards meeting an objective, estimating the effect of resource mobilisation, assessing the quality of a system or comparing the performance of alternative solutions, and in any case, they provide quantified information that facilitates stakeholders to communicate and make individual or coordinated decisions.

We will discuss what output indicators are, which properties they should possess, how they are derived and finally, which are the most commonly used.

5.2.2 Purpose of Indicators and Selection Criteria

Indicators provide the measurable input to the evaluation techniques described in Sect. 4.1.4. More specifically, indicators provide answers to the following questions:

- How well does a transit system perform under different scenarios, including ITS?
- Which scenarios are most effective in enhancing the performance of the transit system, and from the perspective of which stakeholder?
- How does performance of one transit system compare to other systems (benchmarking)?

Various users are involved in the transit system evaluation, each with his own objectives and indicators for quantifying them. The *operators* (*transit agencies*) seek to analyse indicators that are related to the performance of the system, that mainly affect its cost effectiveness, such as vehicle kilometres and load factor which consequently lead to the estimation of operational costs (in fuel consumption or in €) and revenues (in €). *Passengers* (*end users*) consider travel time (wait and in-vehicle time) and delay, accessibility to the transit network (in walking time), cost (fare) and other quality indicators, such as crowding on-board or at stations/stops, reliability and accuracy and safety. *Community and local authorities* look more into the quality of life indicators, such as environmental impacts (pollutant emissions, energy consumption and noise). *Technological stakeholders* measure the system performance (especially when ITS is involved) in terms of success and failure rates of data transmission for fleet monitoring and asset management (smart stops, passenger information systems, etc.). *Policy-makers* consider the multiple and sometimes conflicting impact of ITS on transit, and the impact of transit on the overall transport system.

Different output indicators may be used for different levels of analysis. For example, when short-term analysis of localised interventions is conducted, as in the case of a specific line rerouting, the actual travel time (and delay) and passenger volumes along the line may be computed per trip or per hour (e.g., passengers per hour per line). Also, specific operational modifications in the line (e.g., information provision to passenger about vehicle arrival time at stop) may be studied at a microscopic level and result in the estimation of boardings/alightings on certain routes. Macroscopic level analysis, or analysis which refers to a longer term, may require that shifts of passenger volumes and travel times are computed for all origin–destination pairs in the study area. In the case of a new metro service being introduced in a city, the total ridership per year may be the required indicator for the evaluation.

Output indicators are applied in a wide variety of analyses, thus many different opinions may arise as to which are good indicators. For the definition of the indicator set, the analyst has to take into account the purpose of the analysis. The

selection of the proper indicators usually considers the following criteria (TRANSFORUM 2006; TOOLQIT 2007):

- Relevance—the objectives of the analysis should be covered. Thus, different indicators may be used for benchmarking (relative figures) and assessment (absolute figures).
- Representativeness—policy (and stakeholders') objectives and measures have to be covered, and particular characteristics of the system analysed should be taken into account.
- Completeness—all relevant policy objectives, measures, external factors and involved parties should be covered. As the technology options (ITS) evolve, new indicators may need to be devised in order to capture all effects of a proposed investment.
- Level and Stage of analysis—Strategic versus tactical and versus operating level define the level of detail of the indicators. Also, planning, development or implementation may affect the selection set.
- Clarity and comprehensiveness—the indicators should be well defined and the involved stakeholders should be considered, as their needs may differ in terms of information requirements and level of expertise.
- Reliability and consistency.
- Data feasibility and measurability—this is particularly important where modelled and actual data comparison is foreseen. However, this criterion should not restrict the possibility of selecting indicators that are not measurable at the time of analysis, especially when developing an evaluation framework, because data availability may change over time and is case dependent.
- Interdependence—(set of indicators when you check the validity of a model).

Ideally, output indicators should be given as:

- possible to measure;
- possible to forecast; and
- easily observed or estimated.

In some cases, it may be justified to select unique indicators even though they are not available or accessible at the time of analysis, as data availability may change over time.

5.2.3 Definition of Indicators

The remainder of the chapter focuses on quantitative indicators based on model results. See Sect. 5.3 for aspects of data collection from the field.

The data model for the output of passenger route choice forms an extension of the data model for input and specifically of its canonical graph representation. In basic terms, the primary output consists of route alternatives with passenger volumes.

More specifically, every route choice model defines the choice set available to travellers. In the graph representation of the network, the choice set will be a set K_{od} of *routes* (paths or hyperpaths) from each origin $o \in O$ to each destination $d \in D$. If the model is dynamic, then each alternative is also fixed in time. For space-time networks, the time instant is implicitly embedded in the origin node $o = (i \in B, t \in T)$; otherwise, this is achieved by explicitly ascribing the time interval of departure. A route is space and time is also called a *connection*. For example, two different sets of nodes in the space-time network for the same sequence of edges constitute two separate choice alternatives, e.g., two connections using successive runs of the same line. To be feasible, the times along a connection must be non-decreasing and time differences between the tail and head of each arc must conform to the travel times specified in the supply model, e.g., the running times of runs and the walking times.

Recall that demand is specified as the flow d_{odgt} of passenger trips made by group $g \in G$ from origin $o \in O$ to destination $d \in D$ departing during interval $t \in T$. Since each passenger must choose one alternative from the choice set K_{od}, the main result of the route choice model (or assignment) is given by the flow q_{kgt} for each $k \in K_{od}$, such that:

$$\sum_{k \in K_{od}} q_{kgt} = d_{odgt}. \tag{5.16}$$

All direct indicators are defined on the basis of q_{kgt} using a small number of *combination principles* described below. Thus, a base indicator is the VOLume of a route $k \in K_{od}$, defined for a given time interval $t \in T$ or for the whole assignment period as:

$$VOL_{kt} = h_t \cdot \sum_{g \in G} q_{kgt}, \tag{5.17}$$

$$VOL_k = \sum_{t \in T} VOL_{kt}. \tag{5.18}$$

Note that, if the route flows are measured in terms of passenger loads or numbers, then there is no need of multiplying the instantaneous flow q_{kgt} by the corresponding interval duration h_t.

The principle of *aggregation* derives an indicator from a source indicator by summing the source indicator over a subset of objects. For the route flows, this can be done in space, time or along any structure of the transit supply. As an example of spatial aggregation, consider the subset $K_s \subseteq K$ of all routes which include boarding at some given stop $s \in S$, then:

$$PBS_s = \sum_{k \in K_s} VOL_k, \tag{5.19}$$

is the total number (or volume) of passengers boarding at stop s at some point during their trip. Likewise, if subset $K_\ell \subseteq K$ denotes the subset of all routes which include running on given line $\ell \in L$, then:

$$PRL_\ell = \sum_{k \in K_\ell} VOL_k. \tag{5.20}$$

is the total number of passengers riding line ℓ at some point during their trip.

In macroscopic models of dynamic assignment, any indicator may be aggregated by time by summing over only those connections that visit certain nodes within a given time interval. In this way, temporal profiles of volumes can be extracted at any desired temporal resolution, e.g., quarter-hourly boardings at a given stop.

Indicators may also be aggregated along any hierarchy in the supply model. If the lines belong to different sub-modes (e.g., bus, metro, commuter rail), then the total ridership of each sub-mode may be found by summing volume over its constituent lines.

Although summation is by far the most prevalent aggregation function, other operations may be used instead, to define the minimum, maximum, average or arbitrary percentiles of a source indicator over a given subset.

Aggregation is a powerful tool for benchmarking where often primary indicators are aggregated all the way up to network-wide averages, to be compared between cities.

By using the principle of *composition,* new indicators can be derived from two or more source indicators. These indicators may originate either directly from the route choice model, or from the demand or supply model input. As a typical example, the indicator passenger-kilometres for Route k may be defined as:

$$PKR_k = VOL_k \cdot DIS_k. \tag{5.21}$$

Multiplying the flow indicator by the distance indicator that is given by the sum of supply side indicator length over all arcs along the route:

$$DIS_k = \sum_{a \in A_k} l_a. \tag{5.22}$$

Similarly, the passenger-kilometres for line $\ell \in L$ can be defined by a combination of aggregation (considering the running arcs A_ℓ^{run} of the line) and composition:

$$PKL_\ell = \sum_{k \in K} VOL_k \cdot \left(\sum_{a \in A_k \cap A_\ell^{run}} l_a \right). \tag{5.23}$$

Indicators may even be composed exclusively from supply or demand indicators, to serve as the basis of further composition. Example: the Service-Kilometres of line

$\ell \in L$, the total distance travelled by all of its runs, can be stated as (if a space-time network is considered, then the cardinality $|R_\ell|$ of the runs is omitted because the running arcs refer explicitly to runs):

$$SKL_\ell = |R_\ell| \cdot \left(\sum_{a \in A_\ell^{run}} l_a \right). \tag{5.24}$$

Composition of passenger-kilometres and service-kilometres yields the (distance-weighted) average loading of line $\ell \in L$:

$$ALL_\ell = \frac{PKL_\ell}{SKL_\ell}. \tag{5.25}$$

The last combination principle is an important special case of composition, i.e., the *normalisation* by calculation of relative or weighted average indicators, which allows indicators to be compared between otherwise different base objects.

As an example of normalisation, we present here an indicator which plays a central role in mode choice models built on level of service indicators from a route choice model, i.e., the (volume-weighted) average travel time for a given origin–destination pair $od \in O \times D$ and interval $t \in T$:

$$ATT_{odt} = \frac{\sum_{k \in K_{od}} t_{kt} \cdot VOL_{kt}}{\sum_{k \in K_{od}} VOL_{kt}}, \tag{5.26}$$

where t_{kt} is the travel time of route k for passengers departing at instant t. The latter is to be calculated, not simply as a sum of the travel times t_{at} for users entering at instant t each arc $a \in A_k$ composing path k, but fully taking into consideration the concatenation of travel times, as it will be clarified in Part III, for example in Sect. 6.1.13. For demand modelling, this indicator is further aggregated by time for the whole assignment period.

Taking into consideration other characteristics of the path, such as the distance and the generalised cost, one can define similar indicators at a lower level, such as the average generalised cost, or at an upper level, such as the average travelled distance:

$$AGC_{odgt} = \frac{\sum_{k \in K_{od}} c_{kgt} \cdot q_{kgt}}{\sum_{k \in K_{od}} q_{kgt}}, \tag{5.27}$$

$$ATD_{od} = \frac{\sum_{k \in K_{od}} DIS_k \cdot VOL_k}{\sum_{k \in K_{od}} VOL_k}. \tag{5.28}$$

These kinds of indicators are also called *skim matrices*.

Table 5.6 Output indicators of transit assignment (€ denotes arbitrary currency unit)

Category	Indicator	Unit
Demand	Total passenger trips	1
	Traffic intensity = Total passenger trips/population of service area	1
	Traffic density = Total passenger trips/size of service area	1/km^2
	Passenger km (per line or total)	km
	Passenger hours (per line or total)	h
Utilisation of supply	Passenger volume per line or line segment	1
	Boardings/alightings/transfers per stop	1
Level of service	Travel time, broken down into in-vehicle, wait, access/egress	h
	Travel distance	km
	Commercial speed = Travel time/travel distance	km/h
	Air-line distance speed = Travel time/air-line distance	km/h
	Density = passenger volume/capacity	1
	Boarding failures	1
	Delay = actual travel time − scheduled travel time	h
Operating effort	Fleet size = Number of vehicles	1
	Vehicle km (per line or total)	km
	Vehicle hours	h
	Timetable efficiency = Total service km/total vehicle km	1
Economy	Total operating cost	€
	Total operating cost/total passenger trips	€
	Total operating cost/total vehicle km or total vehicle hours	€/km €/h
	Total fare revenue	€
	Total fare revenue/total revenue	1
	Total revenue/total passenger trips	€
	Cost coverage = Total revenue/total operating cost	1
Environment	Pollutant emissions, including green-house gases	t
	Energy consumption	kW
	Noise	db
Safety	Vehicle accident rate = Vehicle accidents/vehicle km	1/km
	Passenger accident rate = Passenger accidents/passenger km	1/km

Finally, indicators from the route choice model may themselves feed into external models where they are input to equations for deriving, e.g., economic or environmental indicators.

The basic and most commonly used direct and indirect indicators from passenger route choice models are shown in Table 5.6. For simplicity reasons, these indicators refer only to one unit. Most indicators can also be extracted per time interval, or per user class.

5.2.4 *Displaying the Output*

Depending on the interests, familiarity, expertise of the stakeholders outputs can be presented in different formats. The most universal format is a tabulation of indicators per object (e.g., boardings and alightings per stop, passenger volumes per line segment, revenue and operating cost by line). In many cases, presentation of numerical indicators as business charts will improve legibility. If the objects to which the indicators relate are geo-referenced, then indicators can alternatively be displayed on a thematic map. Multiple indicators can be combined in a compact display through combination of symbols, scaling and colour-coding. Figures 5.6 and 5.7 show passenger volumes along a line in all three forms, as a list, a business chart and a thematic map.

Travel demand and level of service indicators (travel time, travel cost, number of transfers, etc.) among zones may be expressed in matrix form, where the matrix dimension (number of rows and columns) is determined by the number of zones in the network. Matrices of indicators are commonly called *skim matrices*. If the zones are geo-referenced, then matrix values can be displayed on the map as well, as straight lines from origin to destination. In case of travel demand, these lines are called *desire lines*.

Other graphical formats include choropleth maps for indicators representing densities by area and isochrone curves which colour-code destinations by travel time from a given origin, thus visualising a measure of accessibility (Fig. 5.8).

Line	Stop	Boardings	Alightings	Volume
003	Karlsruhe Hauptbahnhof	451	0	451
003	Poststrasse	197	2	646
003	Tivoli	399	184	862
003	Werderstrasse	1136	174	1823
003	Baumeisterstrasse	775	129	2469
003	Mendelssohnplatz	372	61	2781
003	Kronenplatz/Universität	16	1569	1228
003	Kronenplatz/Universität	442	7	1663
003	Marktplatz	209	212	1659
003	Herrenstrasse	392	292	1759
003	Europaplatz	511	391	1879
003	Mühlburger Tor	396	365	1910
003	Kunstakademie	18	606	1322
003	Synagoge	5	204	1123
003	Lilienthalstrasse	2	110	1014
003	Berufsakademie	0	58	956
003	Heidehof	1	804	154
003	Neureut-Heide	0	154	0

Fig. 5.6 Database of passenger volumes along a line

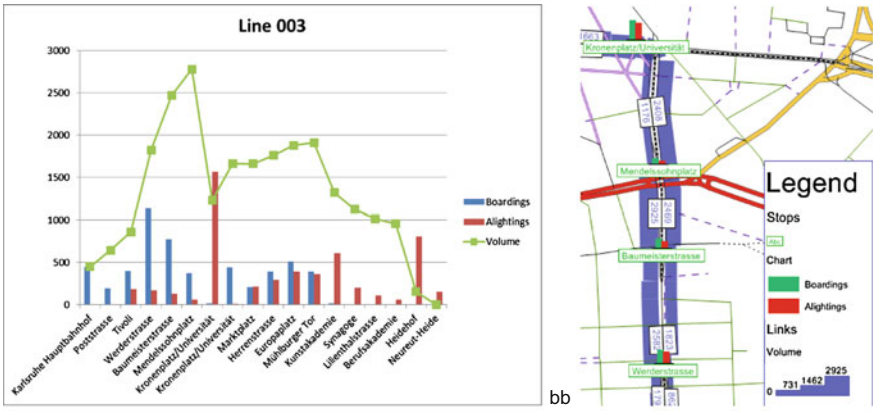

Fig. 5.7 Different presentations of boardings, alightings and passenger volumes along a line

Fig. 5.8 Isochrones visualising travel times from one zone to all others

5.2.5 Reference Notes and Concluding Remarks

In OECD (2003), a fundamental definition of indicators is given, as well as guidelines on the definition of meaningful indicators. Several EU projects dealt with the definition of indicators for characterising environmentally sustainable transport. Among these, TOOLQIT (2007) and TRANSFORUM (2006) formulated desirable properties of indicators.

5.3 ITS Data for Transit Assignment

Achille Fonzone, Maria Bordagaray, Alicia Rodriguez and Maria Nadia Postorino

Reliable data are fundamental for transport modelling. In the past, gathering information to design, run and evaluate transport systems was a burdensome activity normally disjoint from day-to-day operations. Often accurate data collection was not affordable for transport agencies, operators and consultancies. Also when considerable resources could be employed, technical constraints limited the quantity and quality of the data which could be gathered. The situation has radically changed with the advent of intelligent transport systems (ITS). ITS can be the source of huge amounts of reliable, detailed and cheap data with the potential for substantially improving the results of transit models.

At the beginning of their deployment, the possibility of using ITS for planning data collection was overlooked. ITS were (and still are in many cases) conceived by transit agencies exclusively as tools for network management (e.g., for fare collection and management) and operations (e.g., to track vehicles or to regulate priority at signalised traffic lights), to enhance security (e.g., through video surveillance) and safety (e.g., for automatic train breaking) and to disseminate static and dynamic information to customers. Not only the use of ITS as source of information for planning purposes was disregarded, but collecting and storing (possibly extra) data to be used in assignment models and other planning applications was seen as harmful to the efficient implementation of ITS. The only initial applications of ITS data for research were studies of incidents and special events.

Nowadays, the stress generated by the growing demand on systems of increasing complexity, with limited resources and subject to tight environmental regulations, calls for very thorough planning and therefore for much more accurate models than before. The development and the implementation of sophisticated models is allowed and fostered by the increasing capacity of collecting, elaborating, transmitting and storing information. Hence, the use of ITS as planning data source has boosted and technology developers have been encouraged to update software and hardware consequently.

The wide range of functions of ITS data applications can be categorised into two groups: real-time monitoring and offline operations. Archived data are particularly interesting for the models described in this book although assignment models are more and more widely playing an important role in real-time applications. The present chapter aims to provide an illustration of the implementations of ITS recorded data which feed into transit assignment models, with the hope that acknowledging achievements and shortcomings of existing approaches may inspire future research on the applications of this type of data to transport modelling and transit assignment. The chapter is organised as follows: Firstly, data which can be collected by transit ITS and its potential in public transport planning are described. Then, the advantages of using this kind of data are discussed in comparison with

traditional data collection techniques. Finally, some of the main approaches to transit ITS data analysis are described which have been put forward to derive information for assignment models from raw data.

5.3.1 Data from Transit ITS

The provision of public transport services is a complex process, entailing several activities which form a loop. A simplified model of the process is shown by the blue blocks at the top of Fig. 5.9. The objectives are established at political level to pursue the larger interests of the community. Standards are enforced through regulations. Networks and service characteristics are designed to achieve the objectives set by the policy-makers complying with the regulations. Transit operations require control and monitoring systems and procedures, and possibly imply communication of information on the services to travellers. Effectiveness and impacts of the service are appraised. The outcome is used to update policies and plans. In this process, the results of assignment models are used to evaluate scenarios and to support decision-making before, during and after service provision. ITS are available to assist with planning, operations and evaluation. In this chapter, we show how data collected while running the ITS service can be used to improve the quality of the assignment models.

A complete description of data collected by ITS exceeds the scope of this book and in any case it would become obsolete rapidly because ITS are continuously evolving both forms the technological and the functional point of view. Hence, in this section, we only present the systems and the data which so far have been used or identified as source of information for transit assignment models.

5.3.1.1 Automatic Vehicle Location (AVL)

AVL systems have been introduced to track vehicles' position mainly for two purposes: security and assessment of real-time performance against schedule. Consequently, they associate information concerning a specific run such as the vehicle ID, the line and the service number to information regarding the time and the location of an event (e.g., door opening or presence at certain locations). Events can be polled according to two procedures: location-at-time (time-point level) and time-at-location (stop level). The latter kind of polling geo-referenced information allows a better assessment of a wide range of indicators concerning the actual service performance, given that the presence of vehicles at stops or stations can be monitored in a more accurate way. AVL systems can be elaborated to derive important input data for assignment models such as dwell and running times, headways and regularity of service, reliability indicators.

As an example, the London iBus AVL system (in its 2013 version) logs a record containing location, speed and odometer values of the bus every second. In

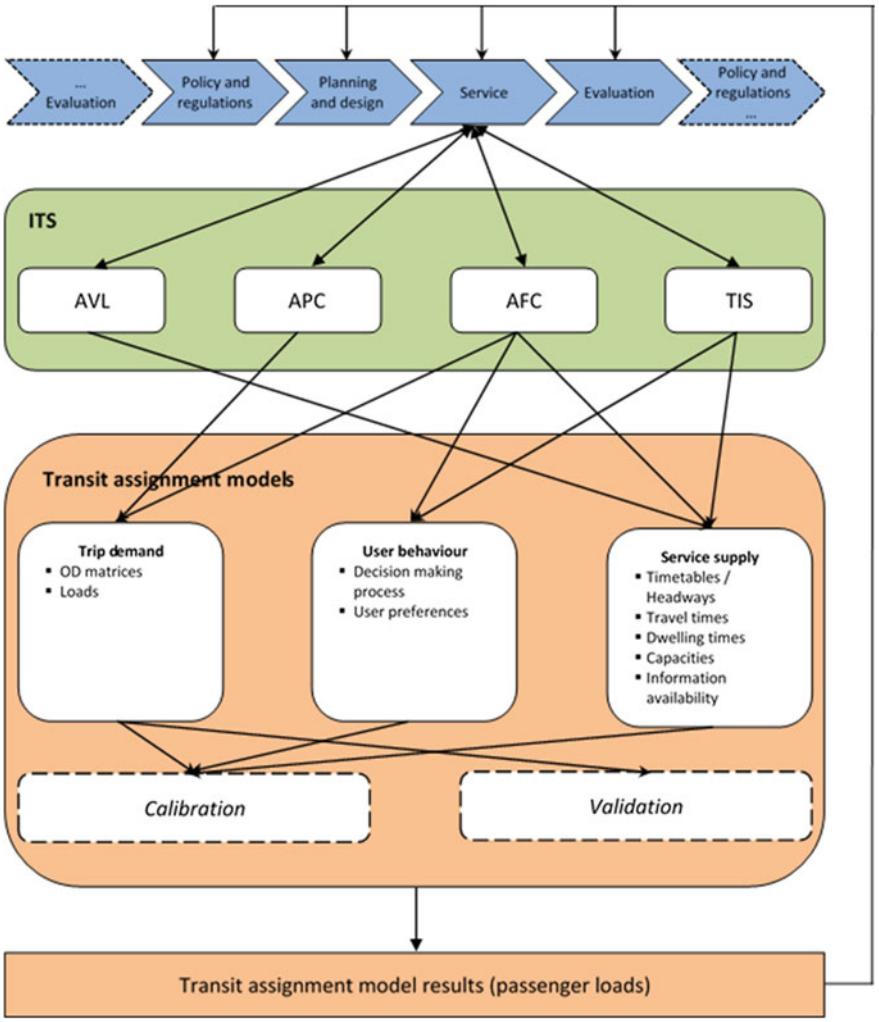

Fig. 5.9 ITS and assignment models in the process of transit service provision

addition, the system keeps track of the times at which the doors are opened and closed. This information allows calculating several bus performance measures which can be used for assignment, such as drive time between two stops and the dwell time at stops. The data from iBus can be aggregated at a higher level and further elaborated to obtain system performance indicators, like those in Table 5.7 which are regularly published by Transport for London. However, the extraction of such indexes requires ad hoc parsers. The standardisation of AVL system outputs could increase the quality of the information available for planning applications and

Table 5.7 TfL bus system performance indicators calculated using the iBus data set

Type of service	Indicator
All buses	% Vehicle kms operated
	% kms lost for staff reasons
	% kms lost for mechanical reasons
	% kms lost for traffic reasons
High-frequency services	Average excess wait (minutes)
	Average actual wait (minutes)
	% Chance of waiting <10 min
	% Chance of waiting 10–20 min
	% Chance of waiting 20–30 min
	% Chance of waiting >30 min
Low frequency services	% Departing on time
	% Departing early
	% Departing 5–15 min late
	% Non arrival

reduce the time to obtain them. Note that standards are already available and widely used to feed data into journey planners and AVL systems.

While the iBus implementation in London is one example of AVL, transit agencies with multiple operators have a need to link AVL systems of different vendors. Real-time vehicle locations of different operators have to be accessed either for operational purposes (e.g., rescheduling, short turnarounds) or for dynamic information of passengers. Different AVL systems can exchange data according to the CEN/TS 15531 standard (SIRI Service Interface for Real-time Information). The Association of German Transport Companies (VDV) publishes additional standards which are applied in Germany and several other countries to link AVL's of different public transport operators such as railway and local operators. The most common standard is VDV 453 (VDV 2008).

Some AVL implementations are embedded within a whole set of applications with real-time transit vehicle location as one of the primary services. These comprehensive AVL implementations are called Intermodal Transport Control Systems (ITCS).

5.3.1.2 Automatic Passenger Counter (APC)

APC systems provide counts of boarding and alighting passengers. Each on and off movement is recorded, and then the data is converted into counts. The information is used by transit agencies and operators mainly to derive ridership rates and to deal with revenue management. Passenger counts can be used in assignment model calibration and validation. Note that sometimes APC systems do not store location information. In this case, they can be used for assignment purposes only if they can be matched with AVL data.

5.3.1.3 Automated Fare Collection (AFC)

Besides keeping track of passenger ticket payments, AFC systems may store valuable information concerning used stops and lines, time of the trip and personal information. These data are valuable to observe traveller behaviour, to estimate O–D matrices and to infer journey time. However, where flat fare schemes are in place, usually passengers are required to validate their electronic tickets/cards (several forms of electronic payment methods are available: magnetic strip cards, contact less cards also in combination with bank cards, Near Field Communication (NFC) devices) only at the beginning of their travel. This makes the use of AFC data sets more complicated.

The two most accredited standard bodies in the field of transit smart card are the Integrated Transport System Organisation (ITSO) and the Calypso network. Both organisations promote standardisation to foster interoperability. In particular, the latter proposes also a proprietary data model. It is not easy to access the complete structure of the data sets provided by the AFC systems because they are usually covered by confidentiality agreements. Some indications can be inferred by published papers or by the limited data sets made available by operators and agencies (Table 5.8). Data collected through AFC systems have applications: besides those related to network planning and so to assignment (strategic level studies), they are analysed to derive information for schedule adjustment (tactical level), network monitoring and detection of irregularities in the smart card system itself (operational level).

5.3.1.4 Traveller Information Systems (TIS)

Traveller Information Systems include journey planners and real-time information systems. In some publications, they are also called Advanced Traveller Information Systems (ATIS). These tools can be important sources of information about time-dependent travel times. It has to be noted that TIS data may derive from AVL systems. However, TIS data are publicly available, whereas AVL data might be more difficult to access. Furthermore, in case of journey planners, often TIS information is the result of the statistical elaboration of raw AVL data which can be directly inputted in assignment models. Recently, the possibility is under evaluation to extract information on demand dynamics from the logs of the queries to Internet-based TIS.

5.3.2 ITS and Traditional Data Collection Techniques

A transit assignment model is a procedure that uses a mathematical representation of a public transport system to make predictions about passenger loads of transit vehicles and/or lines. In an assignment procedure, the demand of trips between

Table 5.8 Data from selected AFC systems

Agency	Transit system	Some of the data collected for each transactions	Reference
Société de transport se l'Outaouais Gatineau, Quebec	Bus	Date and time of the validation, status of the transaction (boarding acceptance, boarding refusal, and transfer), card ID, fare type, route ID, route direction, stop ID, bus ID, driver ID, run ID, and internal database ID	Pelletier M-P, Trépanier M, Morency C (2011) Smart card data use in public transit: a literature review. Transp Res Part C: Emerg Technol 19 (4):557–568
Transport for London (*Oyster card*) London, United Kingdom	Bus, tube, tram, DLR, London overground and most national rail services in London	Day of the week, sub-system, Station the trip started at, station the trip ended at, entry time of the trip in minutes after midnight, exit time of the trip in minutes after midnight[a], zones of product types involved in the trip (pay-as-you-go, season ticket, mixed), application of the daily capping[b] (for pay-as-you-go trips), fare before any discounts fare after usage-based discounts, route number of the bus, if a bus has been boarded	"Oyster card journey information" at http://www.tfl.gov.uk/info-for/open-data-users/our-feeds . Accessed on 16 Sept 2014

[a]Not available for transactions concerning bus use

[b]Once a passenger has paid a certain amount of money for his travels in the TfL network in a certain day, he/she is not charged any longer for travelling in that day

specific locations (e.g., cities, neighbourhoods or stops) is distributed among different services according to the characteristics of the services and on the basis of the presumed behaviour of passengers. Good-quality input data about trip demand, public transport supply and user behaviour are fundamental to obtain valid and reliable results from assignment models (orange box in Fig. 5.9). In Fig. 5.9, some of the information can be provided by ITS and used to specify, calibrate and validate transit assignment models. Geographic maps are not mentioned in the figure because they are not supplied by transit ITS. Maps are clearly essential though for a realistic representation of both demand and supply. Nowadays, many

maps are freely available on the Internet, although it is often necessary to pay services to access the most precise and up-to-date maps.

If ITS cannot be used for data collection, traditional survey techniques have to be applied. These are normally based on sampling and include questionnaires to study passenger behaviour and estimate O–D matrices (through revealed preference or stated preference approaches), passenger counts (which gives flows for calibration), and direct observations of headways and travel times. The quality of the data sets is determined by the size and the representativeness of the samples. In particular, these should be chosen to reproduce the spatial and temporal distribution of the characteristics of demand and supply.

Data collection methods not using ITS suffer from several drawbacks:

- High collection costs: Since manual surveys and observations are very labour-intensive activities, they require large expenditure for staff.
- Incomplete data sets: Because of the high costs involved, data can be collected only for short periods. Often observations are made exclusively in peak periods and the observation span is limited to few days. Finally, respondents may not answer all questions.
- Uncertain validity: Most information concerning traveller behaviour can only be collected through stated preference surveys, which sometimes do not provide a truthful description of reality.
- Inaccuracy and inconsistency of data: The quality of data can be impaired by the lack of training, distraction or fatigue of surveyors. Inter-observer reliability issues (i.e., different observers can approach measurement in different ways) can affect the consistency of collected data sets.

These problems can be solved by the use of ITS data:

- Data sets needed for assignment models are a by-product of ITS installed for system management. The additional cost of collecting data for planning purposes is practically nil, whereas generally there are extra expenditures to store and manage the very large data sets used for offline elaborations.
- ITS is normally in operation throughout the provision of service. Hence, they make available data sets which cover the whole range of operational conditions of the transit systems.
- Measurement does not rely on human interventions. This reduces the risk of inaccurate and inconsistent data sets. However, it has to be noted that if a system is not calibrated, all the data will be affected and it might be difficult to pinpoint the problem.

Sometimes ITS in operation in the same transit system have been implemented by different suppliers, in different periods and without the need for integration, because in general each system can achieve the goal for which it has been implemented as a stand-alone application. This implies that relating different data sets to use them in the same models may not be straightforward. For instance, in London, the AFC (Oyster card) and the AVL (iBus) systems do not share the same time stamp clock. Therefore, linking information on user behaviour which can be

extracted from the former to the performance of the system recorded by the latter is not straightforward but requires complex data processing.

Major transportation agencies (for instance, TriMet in Portland, Chicago Transit Authority, Transport for London) apply ITS records for offline analysis. Interestingly, manual data are still needed to complete the information not provided by the ITS as well as to validate the results yielded by algorithms or models calibrated with automatic records. Pre-processing of raw data relies on assumptions which need validation, see for instance the problems related to trip reconstruction described in the following. The availability of ITS data does not eliminate the need for traditional in-depth studies on traveller behaviour, for instance to investigate trip purposes and to estimate activity models. In particular for transit systems, household surveys may still be needed to assess walking distances or to gather information on the available mode alternatives.

5.3.3 ITS Data Applications

The data collected by ITS normally need elaboration before they can be used in assignment models. This section discusses the shortcomings of raw ITS data and presents the logic underpinning the methods to derive useful information concerning trip demand, user behaviour and service supply from the unprocessed data sets.

5.3.3.1 Boarding Stop

Several techniques have been developed to infer the boarding location of a trip matching AFC and AVL data sets when the AFC system does not keep track of this kind of information. In the easiest case, the AVL system stores the door opening and closing times, and the AFC system the line number and time stamp of each transaction. With this information, the AFC transactions that occurred in a certain period can be directly associated with the corresponding stop. When the synchronisation of AFC and AVL is not perfect or the location-at-time polling procedure is adopted (therefore the opening and closing times might be unknown), a tolerance time of 3–5 min is introduced to match the AFC transactions for which a perfectly corresponding "open doors" period does not exist. For instance: If the opening and closing times are available but the synchronisation of the clocks of the AFC and AVL systems is not reliable, non-assigned transactions recorded 3–5 min before the opening or after the closing events of a given "open doors" period are allotted to that period. If interchanges are recorded by the fare payment system, the procedure can be repeated for every leg of a trip. A more complex methodology has been proposed in which firstly consecutive events recorded within a certain time interval are grouped into the same cluster. Then stops are assigned to clusters by probabilistic models, which assume the travel speeds between stops are randomly

distributed. Both these procedures have success rates of about 90 % in terms of AFC transactions that are eventually associated with a stop and a line. However, there is no available data on the correctness of the inference.

5.3.3.2 Alighting Stop

To identify transfer points and trip destination, a trip-chaining technique is commonly applied based on the hypothesis that the alighting point of a trip leg coincides with the starting point of the next one. In this chapter, "trip leg" is defined as the portion of the public transport service route between the point in which the passenger boards and that in which he alights.

In particular, it is assumed that the last trip of a day finishes where the first of the following one begins. This assumption implies that passengers do not use private transport and the existence of symmetry in travel patterns. Trip-chaining methods can consider also the geography of transit networks: Candidate alighting points for a trip leg are those stops of the line used by the traveller whose the distance from the boarding stop of the next trip does not exceed a certain threshold. Euclidean distances ranging from 400 m to 2 km have been recommended. The limit is defined on the basis of experience or intuition and no experimental evidence can be found in the literature. The stop closest to the boarding point of the next trip is usually selected as the alighting stop of a given trip. If trip segments are missing in the data (i.e., when symmetry is not present in the data of a smart card because the owner has travelled some legs using a different mode), one or more alighting stops may not be identified and thereby the trip sequence along a day or across successive days cannot be reconstructed. The methods available in the literature allow inferring the destination of more than 70 % of the recorded trips. Some approaches have been validated by means of small traditional O–D surveys and groups of volunteers. The predictions are correct in more than 85 % of cases.

An alighting point in a chain can be the destination of a trip or a transfer point within a trip. The two kinds of alighting points are distinguished applying mainly spatial and/or temporal criteria. The most common condition to detect transfers concerns the time spent at a certain location: If the time distance between two consecutive transactions of the same user that were not recorded on the same line (if the transactions concern the same line, then it is more likely that the alighting point is an intermediate destination than transfer point) is smaller than a given threshold, then the event is classified as a transfer. Different time thresholds may apply to changes between different transport modes or involving non-travelling activities such as buying a new ticket. A more sophisticated detection method entails the comparison of the period of the change with the schedule of the lines that the travellers might have used: if they do not board one of the first alternatives arriving at the origin of the next segment, the alighting event is considered a destination. Land use can also be considered to evaluate whether the passenger might have carried out an activity between two consecutive boarding transactions, in which case the alighting point is considered as a trip destination. Sensitivity analyses of

the results provided by the above-described methods can be performed to study the parameters characterising trip chaining (e.g., the variability of walking speed, the maximum distance between two stops which can be considered part of the same trip and the time lag between the transactions defining a transfer).

5.3.3.3 Passenger Arrival

Service schedule and reliability affect the arrival of travellers at stops. Headways greater than 10–15 min lead to schedule-based decision-making and so to arrivals concentrated near the departure times of the different runs, whereas higher frequencies make consideration of exact departure times less important and the arrivals are uniformly distributed. AFC data can help studying the passenger arrival patterns, including their consistency in time. However, the possibility of using AFC data sets to study arrivals depends on the moment in which passengers validate their ticket: In fact, if the validation occurs on-board, the AFC records only show the boarding time and there is no information as to the arrival at the stop. The arrival patterns can be visualised by the cumulative distributions of the arrivals recorded by the AFC system. Cumulative distributions referring to different times can be compared to examine the variation of the arrival behaviour within the day and day by day. Matching AFC and AVL data, the relation between arrival patterns and service reliability can be investigated as well. Depending on the location of the validating device, this measure may be directly related to the crowding rate in the station or the platform and also might be connected to the willingness to find a seat, which is sometimes different depending on the type of the day and the purpose of the trip. The crowding may also be related to the delays that the service has experienced. The heterogeneity in the arrivals might be tested by assessing each user's arrival spread.

5.3.3.4 Route Choice

Any assignment procedure has an underpinning choice model. It is usually assumed that travellers aim to minimise a generalised cost function which can include journey time, expenditure for fuel or tickets, transfer penalties and other kinds of disutility (Golledge and Gärling 2001). Research has found evidence that travelling is not only a means to an end, so that it is solely a cost to be minimised, but it is associated with a positive utility (Mokhtarian and Salomon 2001). However, in most applications, journey time is the only factor accounted for in the cost function. In case of public transport systems, the journey time includes at least the time spent on-board and the waiting time. In networks with high-frequency services and common lines, the travellers can reduce their expected journey time by adopting a strategic behaviour like that described in Spiess and Florian (1989) (see also Sect. 7.1 Strategies and Information). There is not much behavioural research on route choice in transit networks; in particular, observations of actual behaviour are very

difficult to collect with traditional methods. Implementation of AFC data sets allows observing the existence of the hyperpath-based route choice approach and calibrating the related cost function (Schmöcker et al. 2013). In complex networks, a trip with a single entry–exit transaction may include transfers which are not recorded because they do not involve further ticket validation. In this case the assignment of the demand to specific routes/trip legs and the inference of transfer points can be carried out through discrete choice models. Alternatively, fixed route selection percentages can be applied if they are known. GPS trajectories extracted, e.g., from mobile phone data and map-matched to the transit network may help disambiguate chosen routes. Detailed information on trip legs and routes loads can be used to estimate random utility models providing insights into user preferences.

5.3.3.5 Passenger Journey Time

Journey time is a crucial cost element in all assignment procedures. AVL systems are conceived to track vehicles; therefore, the data that they collect can be used to assess travel and dwell times in a straightforward way. In several studies, APC data have been used to obtain travel time for each O/D pair once all possible destinations have been estimated. The travel time can also be derived from the trip queries made to a Customer Information System (CIS) based on geo-referenced AFC records. If accessing and leaving a particular transit network requires crossing controlled gates as it is common in rail networks, the portion of journey time inside the system can be easily calculated as the elapsed time between the entry and the exit transactions. Other measures—such as access and egress time between gates and platforms, wait time, in-vehicle time—can be derived when it is possible to assign each trip to a particular service. From individual (portions of) journey times, network indicators can be computed such as ranges, percentiles and probability distribution functions. For a detailed analysis, such measures should ideally be available for each O–D.

5.3.3.6 Service Reliability

Vehicle running times are affected by network regulations and operations (including signal timing and maintenance), transit operator behaviour, congestion also due to the interactions with other transport systems (for instance, the performance of bus services is much influenced by the conditions of private car mode), accidents and incidents, and the demand for boarding and alighting. Each of these factors and their interaction can cause deviations from schedules that generate additional costs both to passengers (longer journey time) and to operators (service unreliability can reduce the attractiveness of the service and so lead to a loss of revenues, and it can break vehicle deployment or cause knock-on effects on the following runs covered by a delayed vehicle).

Service performance can be evaluated comparing the actual running times provided by the AVL system with the scheduled times. The Transit Capacity and

Quality of Service Manual measures the level of service related to on-time performance in terms of percentage of arrivals with less than 5-min delay. However, it has to be noted that unreliability is associated not only with late arrivals but also with early departures, which is an even worse cause of inconvenience to passengers.

The reliability of transit services can be described having as unit of analysis stops rather than lines, i.e., looking at the variations of headways instead of running times. Schedule deviations can be derived not only by the analysis of AVL data but also from AFC data in case the ticket system includes a location system or even when the ticket transaction does not contain any geographical information but it is possible to process data to assign boarding to stops. Deviations calculated from ITS data can be usefully related to information concerning traffic and dwell times, lift usage, passenger counts and weather conditions.

5.3.3.7 Crowding

The disutility attached to on-board time increases if passengers have to stand while travelling especially if they are squeezed in an overcrowded vehicle. Therefore, seat availability and crowding definitely affect route choice. In dealing with the effects of crowding extreme values need to be assessed. Their computation requires large samples that can be obtained in a much easier way from ITS. The combined use of information collected by AVL, APC and AFC systems allows determining the relation between loads and headway deviation, for instance through regression techniques. Visualisation of time–distance diagrams and 3D load profiles often provides helpful insights into the system dynamics.

5.3.3.8 Capacity Constraints

The irregularity of headways may lead to the so-called bus bunching, i.e., by the fact that different buses serving the same line or common lines arrive at a stop at the same moment. Bus bunching generates a reduction of the system capacity experienced by travellers. It can be caused by several factors, some internal to the service such as schedule, stop spacing and behaviour of the bus operator, other external such as congestion, weather conditions, fluctuating demand or signal timing. Occurrence and features of transit vehicle bunching can be easily recognised by plotting spatial–temporal graphs of AVL data. A better understanding of the phenomenon—and a basis for prediction—can be achieved by including information on boarding (provided by AFC systems) in space-time diagrams.

5.3.4 O–D Matrix Estimation by Traffic Counts

Without ITS, O–D matrices can only be obtained through expensive surveys in which passengers are asked information about the origin and the destination of their current journey. Because of their cost, this kind of surveys cannot but be limited to short periods and restricted portions of the networks. Therefore, the possibility of using data from AFC systems is appealing. However, seldom the stored AFC records contain all the necessary information, for instance:

- They might not keep track of used stops and runs/lines.
- Normally, information at the starting point of the trip is collected because it is fundamental to determine the ticket price. But where flat fare schemes are in force, travellers are not required to validate again their ticket at the end of the trip or when they transfer, so alighting stops are not known.

Note that in any case the usage of AFC (and AVL) data can only provide information on boarding and alighting points and not on the actual origin and destination of the trip, which has to be investigated with other means instead. Inferences regarding the origin of trips are possible when personal details of the holders (e.g., their postcode) are associated with the AFC cards. Also data collected by APC can be used to infer O–D matrices, as explained in the box below.

The most used approaches to estimate the origin/destination transport demand can be broadly grouped into three classes: direct estimation, model estimation and estimation from traffic counts. Direct estimation methods expand sample surveys using sampling theory results. Estimation by demand models allows the explicit simulation of the demand level and/or modal split variations mainly in the framework of multi-stage models. The estimation from traffic counts requires an initial, known origin/destination travel demand and a set of traffic counts measured on the links of the considered transportation network. The common aim of all the approaches proposed to solve this problem was to reproduce some measured link traffic flows starting from the current estimate of the origin/destination demand and by means of an assignment model.

Consider a transit network described by a graph made by nodes and links (as explained below, links can represent specific runs or line segments). The most general formulation of the count-based travel demand estimation problem is given as follows:

$$Min\left(\psi_1 \cdot z_1\left(\mathbf{d}, \mathbf{d}^0\right) + \psi_2 \cdot z_2(\mathbf{q} = q(\mathbf{d}), \mathbf{q}^m), \quad \mathbf{d}^{LB} \le \mathbf{d} \le \mathbf{d}^{UB}\right), \tag{5.29}$$

where

- $\mathbf{d}$ is the demand vector to be estimated, whose element d_{od} is the trip demand between the origin–destination pair od, leaving during a given reference time period;
- $\mathbf{d}^0$ is the target initial, known estimate of the trip demand;

- $\mathbf{q}$ is the arc flow vector obtained by assigning the current demand estimate $\mathbf{d}$ to the transportation network and whose element q_a is the traffic or passenger flow on the arc a;
- $\mathbf{q}^m$ is the surveyed arc traffic or passenger flow vector;
- $q(\mathbf{d})$ is the assignment function that yields arc flows for given demand flows, and this can be a simple flow propagation on given path shares or a full equilibrium problem, static or dynamic;
- z_1 is a function evaluating the distance between the current estimate of the trip demand and the target;
- z_2 is a function evaluating the distance between the arc flows obtained by assigning the current trip demand estimate and the observed ones (note that only the elements of vector $\mathbf{q}$ which have a corresponding element in $\mathbf{q}^m$ are relevant);
- ψ_1 and ψ_2 are weights that measure, respectively, the confidence in the target demand and in the counted link traffic flows;
- $\mathbf{d}^{\mathrm{LB}}$ and $\mathbf{d}^{\mathrm{UB}}$ are the lower and upper bound of the demand vector to be estimated.

Passenger counts provided by APC systems, in a more or less aggregate form, represent the amount of traffic $\mathbf{q}^0$ moving on a given arc in a given reference time period.

A few considerations can be made regarding this minimisation problem. First of all, the demand can be assumed constant in a given reference time period or time-dependent according to whether its estimate is used for long-term or short-term planning purposes. Generally, in reality the demand is time-dependent, but for long-term planning purposes it can be assumed constant in a given reference time period and equal to its average value in it. In this case, the demand can be estimated by using mean traffic flows measured in the same time interval on some links of the transportation network. This approach is also referred to as "static". In the second case, for short-term planning purposes, traffic flows measured in different reference time periods (e.g., every 15 min, or every hour) are used to improve the estimate of the corresponding time-dependent travel demand. The results are several demand vectors each one corresponding to a given time period. The information contained in the traffic flows measured in shorter time periods can also be used to estimate the demand vector for longer time periods. In this case, the travel demand vector for a given reference time period is the average value on several sub-periods. In both cases, the approach is also referred to as "dynamic".

The arc flow vector $\mathbf{q}$ is obtained by assigning the current demand estimate $\mathbf{d}$ to the transportation network; therefore, the transport supply and the transport assignment models have to be specified in the initial hypotheses. The supply models used to represent public transport services can be grouped into two main classes: schedule-based and frequency-based (see Sect. 6.1). In the schedule-based approach, transit services are represented as single runs by using a time-space graph. In the frequency-based approach, transit services are represented by using lines and frequencies rather than single runs. Another characteristic of the supply

model is related to travel costs. Two approaches can be identified according to whether the travel costs are considered constant (uncongested networks) or flow-dependent (congested networks). However, for transit services, the link travel costs are considered constant in most cases.

Let introduce the *assignment matrix* **M**, whose elements $m_{a\ od}$ are the fractions of demand d_{od} using each arc a. The number of rows of **M** is equal to the number of count links and the number of columns to the number of *od* pairs. The elements of the assignment matrix can be obtained from the path shares p_k of each route $k \in K_{od}$ and the incidence matrix **Δ** whose elements Δ_{ak} are 1 if arc a belongs to path k and 0 otherwise:

$$m_{a\,od} = \sum_{k \in K_{ok}} p_k \cdot \Delta_{ak}. \tag{5.30}$$

The assignment matrix can be used to simplify Problem (5.29) by assuming:

$$q(\mathbf{d}) = \mathbf{M} \cdot \mathbf{d}\,. \tag{5.31}$$

From an operational point of view, the estimation of the travel demand starting from an initial, known estimate is obtained by using counted link flows on a subset of real links of the transport network. These values are combined with the initial estimate in order to improve it and obtain new values of the travel demand that better fit the measured traffic flow values. In other words, the current travel demand estimate is assigned to the network and the obtained traffic flow values are compared with the counted ones. If the differences between the assigned and the counted values and between the target and the current origin/destination travel demand values are greater than a prefixed threshold, the current travel demand estimate is modified until assigned and counted traffic flow values are comparable.

An important aspect of the above problem is the location of the links where the counts are made. Although underestimated in several cases, the choice of the Count Link Set (CLS) is crucial because it affects significantly the quality of the trip demand final estimation. The most common criteria to choose such links can be grouped as follows:

- *O–D covering rule*. Let α be a threshold value of link use percentage. A given *od* pair is said to be "covered" by a link a if the element $m_{a\ od}$ of the assignment matrix is greater than α. The links of the CLS should then be such that all *od*-s are covered. Often it is assumed α equal to zero. From a practical point of view, if a given origin–destination pair is not covered by any of the chosen count links, i.e., no chosen count links belong to one of the paths serving the origin–destination pair, then the corresponding *od* demand value cannot be updated by using the information contained in the CLS.
- *Maximal flow fraction rule*. For a given O–D pair, the links of the CLS should be chosen in order to guarantee that the maximum flow fraction Max($m_{a\ od}$, $\forall a \in$ CLS) is as high as possible. Such rule can solve conflicts among O–D pairs

when choosing the count links. In fact, although low-flow links are generally avoided, they may have a high $m_{a\ od}$ value for a given *od* pair. In other words, a link used by most of the trip demand for a given *od* pair contains more information than other links for that *od* pair. Then, it should be chosen as part of the CLS although the total link traffic flow on it is rather small.

- *Maximal flow-intercepting rule*. Practical experiences show that the origin–destination trip demand final estimation can vary greatly according to the CLS, although the number of links in the CLS is the same. Particularly, the more the chosen count links "collect" traffic flow on the network, the better the origin–destination trip demand estimation is. In fact, more collected traffic flows mean more available information to solve the origin–destination trip demand estimation problem. For a given maximum number of count links for the CLS, the chosen links should intercept as much traffic flow as possible.
- *Link independence rule*. Not all the count links may contain "new" or "original" information. It is well known that the algebraic sum of traffic flows at a transport node is equal to zero—in other words, incoming traffic flows are equal to outgoing traffic flows. Then, consider the simple case where only one link converges at a node and only two links diverge from the same node. If the incoming link and the two outgoing links are chosen as part of the CLS, the information they contain is not independent—in other words, there is no added information because there is a linear dependence among such links. As a general rule, the count links should be chosen in order to avoid linear dependence among the traffic flows measured on them.

When the vehicles of a public transport system are equipped with APC, selecting a CLS means to identify the routes on which the equipped vehicles have to be used. In the ideal situation, APC is present in all the vehicles and therefore counts are available for all links.

The availability of massive traffic data (bib and open) regarding not only flows but also travel times (e.g., FCD speeds, passenger tracks) and the development of modern techniques for derivative-free optimisation allow today to calibrate the assignment model against all sort of measurements. This is true not only in terms of (possibly aggregated) demand parameters (emission and attraction rates; departure temporal profiles) but also in terms of the main supply parameters (e.g., speeds and capacities, for types of line vehicles, for example). Problem (5.29) becomes:

$$Min\left(\begin{array}{c}\psi_1 \cdot z_1\left(\mathbf{d}, \mathbf{d}^0\right) + \psi_2 \cdot z_2(\mathbf{q} = q(\mathbf{d}, \boldsymbol{\delta}), \mathbf{q}^m) + \\ \psi_3 \cdot z_3\left(\boldsymbol{\delta}, \boldsymbol{\delta}^0\right) + \psi_4 \cdot z_4(\mathbf{t} = t(\mathbf{d}, \boldsymbol{\delta}), \mathbf{t}^m)\end{array}, \quad \begin{array}{c}\mathbf{d}^{LB} \le \mathbf{d} \le \mathbf{d}^{UB}, \\ \boldsymbol{\delta}^{LB} \le \boldsymbol{\delta} \le \boldsymbol{\delta}^{UB}\end{array}\right), \tag{5.32}$$

where

- $\mathbf{d}$, $\mathbf{d}^0$, $\mathbf{d}^{LB}$, $\mathbf{d}^{UB}$ are the vectors of demand parameters: calibrated, initial, lower bound, upper bound;

- $\boldsymbol{\delta}$, $\boldsymbol{\delta}^0$, $\boldsymbol{\delta}^{LB}$, $\boldsymbol{\delta}^{UB}$ are the vectors of supply parameters: calibrated, initial, lower bound, upper bound;
- $\mathbf{q}$, $\mathbf{q}^m$ are the vectors of traffic flows: resulting from the assignment model, measured on the field;
- $q(\mathbf{d}, \boldsymbol{\delta})$ and $t(\mathbf{d}, \boldsymbol{\delta})$ are the functionals of the assignment model in terms of traffic flows and travel times;
- z_1, z_2, z_3, z_4 are distance functions;
- ψ_1, ψ_2, ψ_3, ψ_4 are weights of the distance functions in the objective function.

5.3.5 Perspectives

The data collected by transit ITS cannot fulfil completely the demand for input data required by transport planning. Some of the shortcomings have been discussed above. Another important limitation is that transit ITS do not provide information on other kind of transport systems. This is a problem because contemporary travel patterns are increasingly complex and multimodal, and what happens at the boundaries between different transport systems can largely affect travellers' decisions, and so the results of assignment procedures. For instance, a long-distance train passenger who, at the end of her rail leg, has to take a bus to complete her journey may decide to travel to a further station if from that station a bus service with higher frequency is available. Furthermore, subjects interested in using ITS data for planning applications may not have access to it because the data owners are not willing to or cannot share it, because of commercial reasons (e.g., when a transport operator should provide data which could generate advantages for his competitors), because the applications for which the data are required are not of interest for the owner (e.g., when data are required for research purposes) or for privacy issues.

In some cases, the limitation of the scope and availability of transit ITS data can be overcome by collecting information through mobile crowdsensing (i.e., using mobile devices not only for communication between users, but also to collect data to be used to measure a phenomenon of common interest) and/or from social media. The potential of these information sources remains to be explored. They can be very useful also beyond the transportation domain, for instance in urban planning and in other areas affected by mobility patterns. Research in these areas has concerned mainly information gathered using wireless telecommunication technologies, such as cellular telephony, Wi-Fi and Bluetooth. Crowdsensing shows some of the same benefits associated with the use of transit ITS data, for instance the low cost, since there is no need to install new infrastructure, and the sample size, which can be remarkable due to the huge penetration of mobile phones providing access to such technologies. In addition, crowdsensing allows tracking travellers from the very beginning to the very end of their journeys and can provide data in real time. In the

following, for illustrative purposes, some ideas are presented on the use of wireless telecommunication technologies to collect data for transport models.

Data from crowdsources can be obtained from the communication network or from the same users. Normally, network is able to locate users as they move because it knows the base station or router to which the phone is attached to or the devices connected to certain Bluetooth receiver. Therefore, user's paths and speeds can be estimated and then matched with the data concerning the itineraries of train and bus line to determine service loads. Retrieving data from the networks gives access to the most complete data sets, concerning all the users using the network. However, this approach relies on the availability of network operators to share their data. Mobile phones can retrieve information about the network they are attached to in every moment, as well as additional information about their current context (speed, orientation…) which can be of great help to infer the input needed in transport models. The main drawback of using mobile devices as direct source of information is that the number of people involved might not be as extensive as in the previous case. Some mobile operating systems allow to record these data in a file that can be processed later on. Although the applications to retrieve the information and transmit it are transparent for the user, there is still the need for the user to install the applications on her mobile phone, which leads to a reduction in the population participating. Moreover, users may be reluctant to disseminate personal and sensitive information.

Mobile telephony systems have been widely used as location and tracking technology for several applications, such as deriving O–D matrices, mapping geographical mobile phone usage at different times of the day for urban analysis and planning, and estimating general traffic data or studying human mobility patterns. This method relies on the following idea. Each mobile phone is attached to a Base Station (BS) whose geographical location is known. Therefore, the mobile phone user's location can be approximated by the BS location. This process can be extended to track bus journeys. The bus stop sequence of a line is usually described in terms of the geographical location of each stop. This can be translated to the mobile network domain by mapping each bus stop into the BS (cell) from which the user receives signal, and adding fictitious stops when two consecutive real ones are far apart, so as to ease tracking. Then, when a user boards a bus, the sequence of BSs he/she is being attached to as the bus moves can be matched with an AVL database to see which bus line the user has taken. Post-processing is needed in order to detect when the user actually takes and leaves the bus, and to improve the matching between user and AVL information. The main advantage of this method is the complete independence from third parties, both transportation and network operators. Wi-Fi technology can also be used for data collection purposes. The idea is similar to the one behind the use of mobile networks, replacing BSs by access points (APs). If a user connects to an AP from her phone, she is within a radius of 80 metres from the AP. This information can reveal when a traveller takes or leaves a bus, or if a trip is made by train or subway, taking advantage of the Wi-Fi access provided by some transportation operators in these modes (e.g., Madrid bus lines or New York subway). The raw data (from Wi-Fi routers or phones) have to be

processed in order to filter events such as users connected to the bus Wi-Fi network but not travelling, or to take into account users who do not connect to the AP right when they take the bus, among others. Two main drawbacks are associated with this solution. The collaboration of transport operators is required to obtain router logs and data on the bus carrying each router. The sample size would also be smaller than when using the mobile phone network, as not every mobile phone is Wi-Fi enabled and not every person with a Wi-Fi-enabled phone actually connects to the transportation Internet services. The last example of wireless technology that can be leveraged for information collection purposes is Bluetooth. It is a short-range communication method, which requires the two connected devices to be less than 10 m apart. This feature allows more accurate tracking of travellers, very useful for example when the coverage area of a BS or AP is too wide to be sure of when a user arrives to a bus stop and when she leaves. If Bluetooth devices are installed at each bus stop, as in Madrid, passenger arrival at or departure from a stop can be inferred from the connection and disconnection of the portable device to the hot spot. The main drawback, shared with the Wi-Fi technology, concerns the sample size, as Bluetooth is neither included in every phone nor used by all people.

5.3.6 *Reference Notes and Concluding Remarks*

The increasing awareness of the potential of ITS data has encouraged the scientific community to develop research focused on the use of the available data sets for modelling purposes. As a result, nowadays many transport agencies and operators can benefit from the opportunities offered by this new source of information, overcoming the limitations of the conventional manual data collection methods. This is the case, for instance, of New York City Transit (Reddy et al. 2009) and Transport for London (Gordon 2012; Wang et al. 2011).

The very first applications of archived data are reviewed by Furth et al. (2003), who also suggest further ambitious analyses that have been addressed by the scientific community in the latest years. Also Utsunomiya et al. (2006) contribute to the identification of the potential of ITS data on transit planning.

A comprehensive and detailed description of the ITS and their role in data collection with a specific focus on travel demand forecast is provided by Enei (2012) and Shibayama and Lemmerer (2013). Obtaining the O–D matrix of a transit system has been one of the first objectives of the application of ITS data. In this field, Barry et al. (2002) put forward a trip-chaining approach to infer the destinations when only entries are recorded. The idea has been the basis of the much research concerning the estimation of O–D matrices. More recently, the identification of transfers has become a major research topic, for instance Chu and Chapleau (2008), Bagchi and White (2005), Munizaga and Palma (2012) and Seaborn et al. (2009). Another application of automatic data to the study of demand concerns the passenger arrival time distribution (Csikos and Currie 2008).

Numerous models and algorithms have been proposed to evaluate the performance of the supply side of a transit system. A comparison of actual versus scheduled running times is carried out by Golani (2007), Lin et al. (2008), Matias et al. (2010), Salicrú et al. (2011), and Strathman et al. (2002). Service reliability in terms of compliance with schedules is studied by Camus et al. (2005) and Chan (2007). El-Geneidy et al. (2006) identify the factors that have an impact on running times and service reliability. A headway regularity metric is also developed and calibrated by Lin and Ruan (2009) with one-month AVL and APC data of Chicago Transit Authority (CTA). AVL and AFC data are analysed also by Strathman and Kimpel (2003) to determine the relation between headway deviation and loads in the case of the TriMet system. Hammerle et al. (2005) deal with the quality of bus service, whereas bus bunching is the topic of Moreira-Matias et al. (2012).

The third component (together with demand and supply models) of assignment procedure, the route choice model, is studied by Wilson et al. (2009) in case of CTA. Path choice has been widely modelled from data regarding entry and exit points of transit journey (Frumin and Zhao 2012; Kusakabe et al. 2010). Nielsen et al. (2009) propose a stochastic route choice model for the Copenhagen suburban rail system in the presence of delays. Finally, an interesting contribution is given by Sun and Xu (2012) who estimate the variability of each stage of a trip to assess the expected value and the variance of travel times.

Graphics and other visualisation techniques are developed and applied in the literature to understand data, visualise demand patterns, evaluate indicators such as the loads of each run the performance of the service versus the schedule, locate accidents and design alternative routes or schedules (see for example, Chu et al. 2009; Feng et al. 2011; Rahbee 2008).

Finally, a few studies exist using data sources such as wireless networks that will probably integrate the information provided by transit ITS data in the near future. Cellular telephony and Wi-Fi networks have been used as location and tracking technologies for applications such as deriving O–D matrices (Caceres et al. 2007, 2008; Sohn and Kim 2008), localising cell phone usage at different times of the day for urban analysis and planning (Ratti et al. 2006; Reades et al. 2009; Sevtsuk and Ratti 2010), and studying human mobility patterns (González et al. 2008).

References

Alfred Chu KK, Chapleau R (2008) Enriching archived smart card transaction data for transit demand modeling. Transp Res Board 2063:63–72

Bagchi M, White PR (2005) The potential of public transport smart card data. Transp Policy 12:464–474

Barry JJ, Newhouser R, Rahbee A, Sayeda S (2002) Origin and destination estimation in New York City with automated fare system data. Transp Res Board 1817:183–187

Caceres N, Wideberg JP, Benitez FG (2007) Deriving origin–destination data from a mobile phone network. IET Intel Transport Syst 1:15–26

Caceres N, Wideberg JP, Benitez FG (2008) Review of traffic data estimations extracted from cellular networks. IET Intel Transport Syst 2:179–192

Camus R, Longo G, Macorini C (2005) Estimation of transit reliability level-of-service based on automatic vehicle location data. Transp Res Rec 1927:277–286

Chan J (2007) Rail transit OD matrix estimation and journey time reliability metrics using automated fare data. Thesis, Massachusetts Institute of Technology, USA

Chu KKA, Chapleau R, Trépanier M (2009) Driver-assisted bus interview. Transp Res Board 2105:1–10

Csikos D, Currie G (2008) Investigating consistency in transit passenger arrivals: insights from longitudinal automated fare collection data. Transp Res Board 2042:12–19

El-Geneidy AM, Strathman JG, Kimpel TJ, Crout D (2006) Effects of bus stop consolidation on passenger activity and transit operations. Transp Res Board 1971:32–41

Enei R (2012) The potential role of ICT in favouring a seamless co-modal transport system. Deliverable 3.1 of COMPASS. 7th Framework Program, European Union

Feng W, Figliozzi M, Price S, Feng W, Hostetler K (2011) Techniques to visualize and monitor transit fleet operations performance in urban areas. In: Proceedings of the 90th annual meeting of transportation research board. Washington, D.C., USA

Frumin M, Zhao J (2012) Analyzing passenger incidence behavior in heterogeneous transit services using Smartcard data and schedule-based assignment. Transp Res Board 2274:52–60

Furth PG, Hemily B, Muller THJ, Strathman JG (2003) Uses of archived AVL-APC data to improve transit performance and management : review and potential. Transp Res Board 113

Golani H (2007) Use of archived bus location, dispatch, and ridership data for transit analysis. Transp Res 1992:101–112

Golledge RG, Gärling T (2001) Spatial behavior in transportation modeling and planning. In: Goulias KG (ed) Transportation systems planning: methods and applications, CRC Press, New York

González MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. Nature 453:779–782

Gordon JB (2012) Intermodal passenger flows on London's public transport network. Massachusetts Institute of Technology, USA

Hammerle M, Haynes M, Mcneil S (2005) Use of automatic vehicle location and passenger count data to evaluate bus operations: experience of the Chicago Transit Authority, Illinois. Transp Res Rec 1903:27–34

Kusakabe T, Iryo T, Asakura Y (2010) Estimation method for railway passengers' train choice behavior with smart card transaction data. Transportation 37:731–749

Lin J, Ruan M (2009) Probability-based bus headway regularity measure. IET Intel Transport Syst 3:400–408

Lin J, Wang P, Barnum DT (2008) A quality control framework for bus schedule reliability. Transp Res E 44:1086–1098

Mokhtarian PL, Salomon I (2001) How derived is the demand for travel? Some conceptual and measurement considerations. Transp Res A 35:695–719

Moreira-Matias L, Gama J, Mendes-Moreira J, Sousa JF (2010) Validation of both number and coverage of bus Schedules using AVL data. In: Proceedings of the 13th international IEEE conference on intelligent transportation systems (ITSC). Madeira, Portugal

Moreira-Matias L, Ferreira C, Gama J, Sousa JF (2012) Bus bunching detection by mining sequences. In: Advances in data mining. Applications and theoretical aspects, Springer, Berlin

Munizaga MA, Palma C (2012) Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile. Transp Res C 24:9–18

Nielsen OA, Landex A, Frederiksen RD (2009) Passenger delay models for rail networks. In: Wilson NHM, Nuzzolo A (eds) Schedule-based modeling of transportation networks. Springer, New York

OECD (2003) OECD environmental indicators. Development, measurement and use. OECD Environmental Directorate, Paris

Rahbee AB (2008) Farecard passenger flow model at Chicago transit authority, Illinois. Transp Res Board 2072:3–9
Ratti C, Frenchman D, Pulselli RM, Williams S (2006) Mobile landscapes: using location data from cell phones for urban analysis. Environ Plann B 33:727–748
Reades J, Calabrese F, Ratti C (2009) Eigenplaces: analysing cities using the space—time structure of the mobile phone network. Environ Plann B 36:824–836
Reddy A, Lu A, Kumar S, Bashmakov V, Rudenko S (2009) Entry-only automated fare-collection system data used to infer ridership, rider destinations, unlinked trips, and passenger miles. Transp Res Board 2110:128–136
Salicrú M, Fleurent C, Armengol JM (2011) Timetable-based operation in urban transport: Run-time optimisation and improvements in the operating process. Transp Res A 45:721–740
Schmöcker JD, Shimamoto H, Kurauchi F (2013) Generation and calibration of transit hyperpaths. Transp Res C 36:406–418
Seaborn C, Attanucci J, Wilson NHM (2009) Analyzing multimodal public transport journeys in London with smart card fare payment data. Transp Res Board 2121:55–62
Sevtsuk A, Ratti C (2010) Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. J Urban Technol 17:41–60
Shibayama T, Lemmerer H (2013). The role of ICT in travel data collection. Deliverable D4.2 of COMPASS. 7th Framework Program, European Union
Sohn K, Kim D (2008) Dynamic origin–destination flow estimation using cellular communication system. IEEE Trans Veh Technol 57:2703–2713
Spiess H, Florian M (1989) Optimal strategies: a new assignment model for transit networks. Transp Res B 23:83–102
Strathman JG, Kimpel TJ, Kenneth J, Gerhart RL, Callas S (2002) Evaluation of transit operations: data applications of Tri-Met's automated bus dispatching system. Transportation 29:321–345
Strathman JG, Kimpel TJ, Callas S (2003) Headway deviation effects on bus passenger loads : analysis of Tri-Met's archived AVL-APC data. Report PR126. Portland State University Centre for Urban Studies, Oregon
Sun Y, Xu R (2012) Rail transit travel time reliability and estimation of passenger route choice behavior. Transp Res Board 2275:58–67
TOOLQIT (2007) Project website. 6th Framework Program, European Union
TRANSFORUM (2006) Project website. 6th Framework Program, European Union
Utsunomiya M, Attanucci J, Wilson N (2006) Marketing and fare policy potential uses of transit smart card registration and transaction data to improve transit planning. Transp Res Rec 1971:119–126
VDV (2008) Integration interface for automatic vehicle management systems – VDV 453, Version 2.4, Schrift des Verbands Deutscher Verkehrsunternehmen. https://www.vdv.de/service/downloads.aspx?id=100844&forced=true, Accessed 20 Oct 2015
Wang W, Attanucci JP, Wilson NHM (2011) Bus passenger origin-destination estimation and related analyses using automated data collection systems. J Public Transp 14:131–150
Wilson NHM, Zhao J, Rahbee A (2009) The potential impact of automated data collection systems on urban public transport planning. In: Schedule-based modeling of transportation networks, (eds) Wilson A. Nuzzolo, Springer, New York, USA

Part III
The Theory of Transit Assignment - Guido Gentile

Chapter 6
The Theory of Transit Assignment: Basic Modelling Frameworks

Guido Gentile, Michael Florian, Younes Hamdouch, Oded Cats and Agostino Nuzzolo

In this chapter, the different basic assumptions for the development of assignment models to transit networks (frequency-based, schedule-based) are presented together with the possible approaches to the simulation of the dynamic system (steady state, macroscopic flows, agent-based). The main functional components of uncongested assignment and user equilibrium (route choice, flow propagation, arc performances) are also illustrated here in their general form, while the various demand and supply phenomena emerging in transit systems (regularity, congestion, information) are dealt with in the following Chap. 7.

G. Gentile (✉)
DICEA - Dipartimento di Ingegneria Civile, Edile e Ambientale,
Sapienza University of Rome, Via Eudossiana, 18, 00184 Rome, Italy
e-mail: guido.gentile@uniroma1.it

M. Florian
CIRRELT, University of Montreal, Pavillon André-Aisenstadt, CIRRELT CP 6128,
Succursale Centre-ville Montréal, QC H3C 3J7, Canada
e-mail: mike.florian@cirrelt.ca

O. Cats
Department of Transport and Planning, Delft University of Technology, P.O. Box 5048
2600 GA Delft, The Netherlands
e-mail: o.cats@tudelft.nl

Y. Hamdouch
College of Business & Economics, United Arab Emirates University, P.O.Box 15551
Al Ain, United Arab Emirates
e-mail: younes.hamdouch@uaeu.ac.ae

A. Nuzzolo
Department of Enterprise Engineering, University of Rome Tor Vergata,
Via Del Politecnico 1, 00133 Rome, Italy
e-mail: nuzzolo@ing.uniroma2.it

G. Gentile and K. Noekel (eds.), *Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems*, Springer
Tracts on Transportation and Traffic 10, DOI 10.1007/978-3-319-25082-3_6

6.1 Formulating and Solving Transit Assignment

Guido Gentile

In this section, a general mathematical framework for the formulation and solution of transit assignment is presented, which allows for different models, ranging from uncongested assignment to user equilibrium, from static to dynamic. The main functional components of assignment models (route choice, flow propagation, arc performances) are illustrated here with some specific reference to transit networks, but the simulation of public transport services is analysed with more proper detail in the sections that follow. The behavioural concept of strategy is introduced, together with its formulation through hyperarcs and hyperpaths.

6.1.1 Schedule-Based Versus Frequency-Based Services and Models

6.1.1.1 Information Provision and Passenger Decision-Making

A fundamental dichotomy in modelling transit services arises from the question whether or not passengers know or care about timetables. If service is so irregular or so frequent that passengers find no convenience in timing their arrival at a stop with that of a specific run, or simply the schedule is unavailable to them for whatever reason, then users perceive the line in terms of headways between subsequent run departures from that stop (often we refer to carrier arrivals instead of departures, ignoring dwell times).

Actually, while a timetable usually exists for management reasons (most transport companies do program the service in terms of runs in order to allocate vehicles and drivers), it is a specific choice of the operator to determine how much and which schedule information shall be provided to the public. Indeed, there might be issues of reliability and/or usefulness for such timetables. Due to road congestion (if transit carriers share the infrastructure with private vehicles), driver random behaviour, traffic signals, as well as passenger congestion (if dwell times depend on boarding and alighting loads), service regularity may be so poor that it turns out misleading to publish the programmed schedule. Moreover, when a service lags, some runs can be delayed or cancelled by the operator without the need of informing the public. On the other hand, it is not interesting to learn a published schedule by passengers when regularity is so poor that it is not really possible to identify which run of a same line is going to be served by an arriving carrier. Finally, it may be not useful nor possible to memorize the timetable if lines are very frequent (e.g., a metro passing every 3 min).

On the contrary, if actual arrivals are fairly regular with respect to the schedule (passengers are still able to associate a delayed carrier arrival with a specific run) and carrier arrivals are fairly infrequent (e.g., a regional bus passing every 30 min), then passengers perceive the service in terms of runs. This is particularly true for transit

systems that require a seat reservation by users before boarding, as this clearly refers to a specific run.

In the following, this model dichotomy in passenger behaviour on the demand side is solved as an operator decision on the supply side.

In practice, we assume that if the operator publishes full information about the timetable, then the scheduled arrival and departure times of all runs at all stops are regular, and passengers are (at least in principle) able (and thus willing) to plan their complete trip before departure. This form of information provision/perception and consequent decision-making is called *schedule-based*.

Alternatively, scheduled times at stops remain unpublished and refer to a priori planned operations, i.e., without disturbances, but they may differ from actual arrival and departure times that occur in practice. Headways are then represented as random variables with a given distribution, while the frequency is equal to the inverse of the expected headway. The operator may publish only the stop sequence of each line (and possibly their frequency). Based on their travel experience, passengers figure out the (expected) running and dwell times, as well as the frequency and regularity of transit lines (but not the exact scheduled times). This form of information provision/perception and consequent decision-making is called headway-based or *frequency-based* services.

Clearly, in the same transit network, there can coexist services that are frequency-based and schedule-based. This requires non-trivial treatment from the modelling point of view; otherwise, we have to accept the limitations connected to one of the two main approaches.

6.1.1.2 Model Results for Design and Operation

In the above section, the differentiation between schedule-based and frequency-based services has been explained from the user point of view. On the other hand, the purpose of modelling travel behaviour in transit assignment is functional to obtaining passengers' loads and service performances that are used for design and operation.

To this end, we can distinguish as follows:

- schedule-based models, which aim at determining passenger loads on each single run of the service, as well as the actual run trajectories (diagram in time and space along the line stops), since due to delays these may differ from the planned timetables;
- frequency-based models, which aim at determining the average loads on the lines and the possibly emerging phenomena of macroscopic congestion.

The first approach is more suitable for management in real time, because services are daily operated in terms of runs by public transport companies, and the second one is for planning offline, because services are usually yearly designed in terms of lines by mobility agencies. However, in the future, more attention shall be probably devoted to design the requirements of transit operations as an interconnected

network of services, by also optimizing transfers in terms of total passenger delays; this objective clearly requires schedule-based assignment models.

Moreover, schedule-based models are in general richer than frequency-based models. Indeed, it is always possible to aggregate the results obtained for each run into results for each line. Clearly, more detailed output is obtained with a more detailed input, which might be unavailable or irrelevant during the preliminary phases of service planning, and with more complex models, which may require many parameters and high computing times. Therefore, the choice of the modelling approach shall be strictly linked to the actual need of the design task.

In the following, we will often refer to schedule-based and frequency-based assignment considering the above point of view of supply (model), rather than that of demand (service).

6.1.2 *Multiclass Flows and Performances on Multimodal Networks*

In this section, the topological (structural) relations among flows and among performances (separately) at the two different levels of arcs and routes are presented; no functional component is introduced here. We refer in general to 'routes' and not simply to 'paths' to later include (next section) the concept of 'hyperpaths' that is used in transit assignment to represent passenger strategic behaviour.

The topology of the transport network (supply) is represented through a directed graph (N, A), with nodes N and arcs A, on which a set of routes K (paths, for the moment) is defined to connect the different O–D pairs of trips made by users of various classes G (demand) with some modes M (see Sect. 5.1.2.8). In general, but even more notably in transit networks, each arc represents an atomic trip segment of a specific type (e.g., walking from one point to another, waiting for a given interval or for a given event, riding on-board a line from a stop to the subsequent one, driving from one intersection to the next) on a specific transport system (e.g., public transport, car, bike). The sequence of trip segments of the same type is called trip phase or trip leg. Different models may disarticulate trips in different ways and identify different arc types.

Arcs and routes are characterized with variables to quantify flows and performances for each class of users; arcs belong implicitly to one transport system network (one road link is represented by different arcs for pedestrians, cars and to support transit services), with the exception of those used for inter-modal changes (e.g., the stop arcs that connect the pedestrian network and the line network introduced in Sect. 6.2.2); routes belong explicitly to one (simple or combined) mode (for details see Sect. 5.1.1.2).

In static models and in space-time network models (such as in schedule-based models where a diachronic graph is used to represent the temporal dimension within the network topology, as in Sect. 6.3), the reference to time is usually omitted; this is the assumption adopted in the following, while extensions to other kind of dynamic models are presented in Sects. 6.4 and 6.5.

At the network level, flows and performances of arc $a \in A$ for users of class $g \in G$ are defined as follows:

q_{ag} class specific flow;
q_a volume (aggregation of all class flows);
t_a travel time (the same for all classes);
γ_{ag} value of time;
c_{ag}^{nt} non-temporal cost;
c_{ag} generalized cost.

Flows and volumes express in general the number of users passing through a given section in a given time interval. But in space-time networks, where the arc topology embeds natively the simulation time, flows represent actually a number of users (loads); for example, the passengers travelling on a given run section.

The volume of arc $a \in A$ is obtained by summing up the flows of each class $g \in G$, possibly multiplied by a specific equivalency coefficient ω_{ag}, which may differ by arc type, plus a base volume q_a^0, which represents flow components that are not modelled directly:

$$q_a = q_a^0 + \sum_{g \in G} q_{ag} \cdot \omega_{ag}. \tag{6.1}$$

In case of passenger flows, the typical assumption is given as $\omega_{ag} = 1$ and $q_a^0 = 0$.

The generalized cost of arc $a \in A$ for users of class $g \in G$ is obtained multiplying the travel time by the value of time plus the non-temporal cost:

$$c_{ag} = c_{ag}^{nt} + \gamma_{ag} \cdot t_a. \tag{6.2}$$

The value of time of each class differs by arc type and may depend on volumes (discomfort) like the travel time itself (congestion); these phenomena are the subjects of later Sects. 7.2, 7.3 and 7.4 and are essential in transit equilibrium models. The non-temporal cost is in turn the sum of several disutility components, including monetary costs and user preferences with respect to a large variety of arc attributes (e.g. , length, steepness, tortuosity, landscape, pollution, presence of economic activities).

At the trip level, flow and costs of route $k \in K$ for users of class $g \in G$ are defined as follows (recall that the notion of route k embeds its origin $O_k \in O$, destination $D_k \in D$ and mode $M_k \in M$):

c_{kg}^{na} non-additive cost;
c_{kg} generalized cost;
q_{kg} class specific flow.

The generalized cost of route $k \in K$ for users of class $g \in G$ can be obtained by summing up the costs of the corresponding arcs plus a non-additive term, which may represent fares or any nonlinear component of disutility perceived by users (e.g., walking time):

$$c_{kg} = c_{kg}^{na} + \sum_{a \in A} c_{ag} \cdot \Delta_{ak}, \tag{6.3}$$

where Δ_{ak} is the number of times that a user travelling on route $k \in K$ passes through arc $a \in A$. For acyclic paths, it is given as follows:

$$\Delta_{ak} = \begin{cases} 1, & \text{if } \ a \in A_k \\ 0, & \text{otherwise} \end{cases}. \tag{6.4}$$

In case where the disutility associated with users to each route can be represented as a linear combination of its network element costs, then the supply model is said to be *additive*, i.e., the terms c_{kg}^{na} are all null.

The flow on arc $a \in A$ of class $g \in G$ users is the sum of each route flow (of that same class) multiplied by the number of time it passes through that arc:

$$q_{ag} = \sum_{k \in K} q_{kg} \cdot \Delta_{ak}. \tag{6.5}$$

The flow q_{kg} on route $k \in K_{odm}$ of class $g \in G$ users results from the choice among all routes connecting origin $o \in O$ to destination $d \in D$ on mode $m \in M$, and is thus obtained as:

$$q_{kg} = d_{odmg} \cdot p_{kg}, \tag{6.6}$$

i.e., by multiplying:
d_{odmg} is the demand flow of class g users travelling from o to d on mode m;
p_{kg} is the probability that user of class g choose route k.

6.1.3 Strategies and Hyperpaths

A strategy is in general a plan to achieve a goal under conditions of uncertainty. In game theory, a strategy refers to the rules that a player will use to choose among the available options. A strategy may recursively look ahead and consider what can happen in each contingent state of the game depending on the previous possible actions.

Applying this concept to route choice, the goal of the traveller was to reach the destination of his/her trip at a minimum expected (perceived and generalized) cost.

Travel strategies include diversion points (nodes), where users may exploit information acquired along the trip, about variables that are preventively seen as random unknowns, and on this base can make en-route decisions on how to proceed towards the destination. In most cases, the information is actually acquired at the diversion node, but modern information systems may change this circumstance.

This is for example the case of a passenger waiting at a stop for a subset of attractive lines (among all those available at the stop) that he/she wishes to board for reaching his/her destination. When he/she realizes which line is served by the vehicle that is approaching the stop, then he/she can decide whether to board it or keep waiting, depending on whether the line is attractive or not (line probabilities in Sect. 7.1).

In other cases, the outcome of the random variable that becomes known to the user at the diversion point directly determines the action undertaken without an actual decision made by the user. This is for example the case of a passenger boarding a line vehicle who may get, or not, a seat depending on the availability on-board and on how lucky he/she is with respect to the other boarding passengers (fail-to-sit probabilities in Sect. 7.2.2). A similar example is that of a passenger waiting for a line on a crowded platform who may get, or not, on the arriving vehicle depending on the available space on-board and on how lucky he/she is with respect to the other waiting passengers (fail-to-board probabilities in Sect. 7.3.3).

Strategic behaviour is thus connected with the presence of random variables which determine a probability for each one of the considered options among those available at a diversion point and the corresponding expected cost. A travel strategy is then described by a 'complete' iterative sequence of route diversions, starting from the origin, until the destination is reached for each possible combination of events, given the considered options (in this sense, complete).

From a topological point of view, a convenient way of formalizing this kind of strategies on a transit network (but not only) is to introduce hyperarcs and hyperpaths.

A *hyperarc* is a non-empty set of diversion arcs (also called its *branches*) exiting from a *diversion node* $i \in N^{div} \subseteq N$; i.e., a subset of its forward star i^+. The set of *diversion arcs* is $A^{div} = \{i^+ : i \in N^{div}\}$. Note that the number of hyperarcs that can be defined on a network may be very large, because each of them identifies a different combination of arcs exiting from a diversion node (the considered options among those available).

The generic hyperarc $\breve{a} \subseteq i^+$, with $i \in N^{div}$, has a singleton *tail*, denoted $\breve{a}^- = i$, while a set of nodes constitutes its *head*, denoted $\breve{a}^+ = \{a^+ : a \in \breve{a}\}$. Let H be the set of hyperarcs defined on the transport network (not necessarily all combinations of diversion arcs exiting from a same diversion node make up a hyperarc of H). Each branch $a \in \breve{a}$ of a hyperarc $\breve{a} \in H$ is characterized by the following variables:

$p_{a|\breve{a}}$ the *diversion probability* of using branch a among all branches $\breve{a}$ of the hyperarc;

$t_{a|\breve{a}}$ the *conditional travel time* connected using branch a as part of the hyperarc $\breve{a}$.

The (*combined*) *travel time* $t_{\breve{a}}$ of the hyperarc is then given by:

$$t_{\breve{a}} = \sum_{a \in \breve{a}} t_{a|\breve{a}} \cdot p_{a|\breve{a}}. \tag{6.7}$$

The *conditional cost* $c_{a|\breve{a}g}$ connected using branch $a \in \breve{a}$ as part of the hyperarc $\breve{a} \in H$ for users of class $g \in G$ is proportional to its travel time through the value of time γ_{ag}:

$$c_{a|\breve{a}g} = \gamma_{ag} \cdot t_{a|\breve{a}} + c^{nt}_{ag}. \tag{6.8}$$

The (*combined*) *cost* $c_{\breve{a}g}$ of the hyperarc is then given by:

$$c_{\breve{a}g} = \sum_{a \in \breve{a}} c_{a|\breve{a}\,g} \cdot p_{a|\breve{a}} = \gamma_{\breve{a}^- g} \cdot t_{\breve{a}} + \sum_{a \in \breve{a}} c^{nt}_{ag} \cdot p_{a|\breve{a}}; \tag{6.9}$$

the latter assumes that the value of time γ_{ig} is equal to all diversion arcs exiting from the same tail node $i = \breve{a}^-$. This expression is useful because models often provide directly the combined travel time $t_{\breve{a}}$ instead of the conditional travel time $t_{a|\breve{a}}$.

In the following, for notation consistency, it is intended that if $a \notin \breve{a}$ then $p_{a|\breve{a}} = 0$, $t_{a|\breve{a}} = 0$, $c_{a|\breve{a}\,g} = 0$.

The generic *hyperpath* k is a 'bush' of arcs that connects its origin to its destination, i.e., an acyclic sub-graph (N_k, A_k) with:

- $|k^-| = 1$, i.e., one origin node;
- $|k^+| = 1$, i.e., one destination node;
- $|i_k^+| = 1, \forall i \in N_k - k^+ - N^{div}$, i.e., one successor arc, except for the destination node which has none, and for diversion nodes which may have more than one;
- $|i_k^+| \geq 1, \forall i \in N_k \cap N^{div}$, i.e., one or more successor arcs at diversion nodes, which make up one hyperarc, i.e., $i_k^+ \in H$;
- $|i_k^-| \geq 1, \forall i \in N_k - k^-$, i.e., one or more predecessors, except for the origin node which has none;
- $i_k = \varnothing, \forall i \notin N_k$, just for notation consistency.

In the example of Fig. 6.1, there are 7 possible hyperarcs exiting from the diversion node $i \in N^{div}$, i.e., all the possible combinations of diversion arcs a, b and c: $\{a\}$, $\{b\}$, $\{c\}$, $\{a, b\}$, $\{a, c\}$, $\{b, c\}$, $\{a, b, c\}$; but among them only one hyperarc, i.e., $i_k^+ = \breve{a} = \{a, b\}$, can belong to a given hyperpath k.

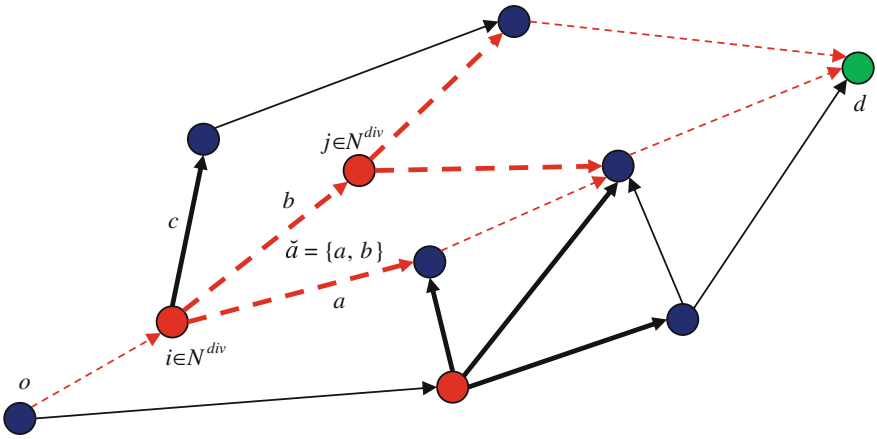

Fig. 6.1 Example of a hyperpath k from origin $o = k^-$ to destination $d = k^+$. The hyperpath is depicted in *dashed red lines*. The diversion nodes are in *red*. The *bold lines* are diversion arcs

It is intended that exiting from a diversion node, no diversion arc can be used per se in a hyperpath but only hyperarcs can; clearly, it is possible to define a singleton hyperarc made of only one diversion arc.

A path can be seen as a hyperpath that does not include diversions. In the following, the term 'route' will then denote indifferently paths or hyperpaths; the proposed formulation is valid for both cases, unless otherwise specified.

In particular, a strategy can be formalized, from a topological point of view, as a hyperpath that connects the origin–destination pair of the trip. Each strategy has an expected cost which is considered by users to make their route choice before starting the trip.

The cost of a hyperpath (i.e., the cost of the underlined strategy) is defined as the sum of its arc costs and of its hyperarc branch costs, multiplied by the probability of using these arcs when following that route; in this sense, it may be additive (if the non-additive cost is null). Equation (6.3) becomes:

$$c_{kg} = c_{kg}^{na} + \sum_{a \in A - A^{div}} c_{ag} \cdot \Delta_{ak} + \sum_{a \in A^{div}} c_{a|(a^-)_k^+ g} \cdot \Delta_{ak}, \tag{6.10}$$

where $\mathbf{\Delta}_{ak}$ denotes now the probability of using arc a (possibly as a branch of a hyperarc) when travelling on route k, and $(a^-)_k^+ \in H$ is the one hyperarc made up by the successor arcs of the diversion node $a^- \in N^{div}$ on hyperpath k.

Note that the conditional cost $c_{a|\check{a}g}$ may differ substantially from the cost c_{ag}; it is usually lower, and from this derives the convenience of considering a hyperpath instead of a simple path (this is the case of attractive lines). In other cases (e.g., fail-to-sit or fail-to-board), there is no cost difference, but a hyperpath is actually the only available route.

The *arc-route probabilities* depend on the hyperarc diversion probabilities through the following recursive equation, which can be solved in topological order (from the origin to the destination of the route):

$$\Delta_{ak} = \begin{cases} 1, & \text{if } a \in A_k - A^{div} \\ p_{a|(a^-)_k^+}, & \text{if } a \in A_k \cap A^{div} \\ 0, & \text{otherwise} \end{cases} \cdot \begin{cases} 1, & \text{if } a^- = k^- \\ \sum_{b \in (a^-)^-} \Delta_{bk}, & \text{otherwise;} \end{cases} \tag{6.11}$$

the first term is the conditional probability of using arc a from its initial node a^- along hyperpath k, and the second term is the absolute probability of using its initial node.

The proper extension to hyperpaths of the structural cost Eq. (6.3) requires to formally change the network model from a graph to a *hypergraph* (N, $\check{A} = A \cup H$), where hyperarcs are native elements:

$$c_{kg} = c_{kg}^{na} + \sum_{a \in A - A^{div}} c_{ag} \cdot \Delta_{ak} + \sum_{a \in H} c_{\check{a}g} \cdot \Delta_{\check{a}k}, \tag{6.12}$$

where $\mathbf{\Delta}_{\check{a}k}$ denotes the probability of using hyperarc $\check{a}$ when travelling on route k.

The structural flow equation, given by Eq. (6.5), can instead be extended immediately to hyperpaths under the new interpretation of Δ_{ak} as a arc-route probability.

However, in hyperpath-based models, the structural Eqs. (6.10) and (6.5) are not merely related to the network topology, but are rather the result of a functional model which describes en-route decisions and/or events connected with random variables yielding the diversion probabilities.

In more advanced models (see the case of information about next arrival for each line provided at stops presented in Sect. 7.1), the en-route diversions of a hyperarc reproduce indeed a strategic rerouting which depends on the destination, mode and class of the traveller; in this case, the diversion probabilities and the conditional travel times are denoted $p_{a|\check{a}\ dmg}$ and $t_{a|\check{a}\ dmg}$, respectively. As a consequence, based on Eq. (6.11), the arc-route probabilities would depend also on the class (while destination and mode are intrinsic in the route). Clearly, $p_{a|\check{a}\ dmg} = 0$ and $t_{a|\check{a}\ dmg} = \infty$ if $a \notin A_m$.

Finally, the number of hyperpaths defined on the network can be huge, although finite, because of the many possible combinations of diversion arcs each one represented by a different hyperarc.

Based on these considerations, although strategies can be formally represented by hyperpaths, their explicit enumeration is prohibitive. Thus, an implicit enumeration approach is usually adopted, as explained in the following Sect. 6.1.5.

6.1.4 Sequential Route Choice and Flow Propagation

The route probabilities of Eq. (6.6) depend in turn on the route costs, for example, through a random utility model (see Sects. 4.4 and 4.5):

$$p_{kg} = p_{kg}\left(c_{hg}, \quad \forall h \in K_{odm}\right). \tag{6.13}$$

Route probabilities must satisfy the following consistency and non-negativity constraints:

$$\sum_{k \in K_{odm}} p_{kg} = 1, \quad p_{kg} \geq 0. \tag{6.14}$$

Equation (6.13) defines the route choice model in case of *explicit enumeration* of routes, while Eqs. (6.6) and (6.5) define the corresponding flow propagation model.

This basic model, where routes are chosen jointly, may be inadequate to describe passenger behaviour; as the number and complexity of paths increases, users can become unable to memorize and compare the available alternatives, as this would require too high cognitive faculties. Moreover, explicit path enumeration may be heavy from a computational point of view.

Decision-makers tend to simplify choice contexts that are too complex. A path, after all, is not an elementary concept, because it is constituted by a sequence of arcs.

In case of additive supply models, we can then assume that users reach their destination through a sequence of (more simple) choices at nodes, where the local alternatives are the arcs of the forward start. This approach is based on *implicit enumeration* of routes and requires to introduce the following variables, referred to users of class $g \in G$ directed towards destination $d \in D$ on mode $m \in M$:

- p_{admg} *probability* that users take arc $a \in A$ *conditional* on being at its tail node;
- w_{idmg} *expected cost* perceived by users to reach the destination from node $i \in N$;
- q_{idmg} *flow* of users traversing *node* $i \in N$.

Sequential route choice models are generally referred to destinations, as this is the most natural way to address the problem from user's perspective, and it is also the only possible way to proceed if one wants to introduce the concept of strategies (see Sect. 6.1.3). Travelling passengers aim to reach their destination and can take en-route decisions only in reaction to future events based on incoming information.

Consider the local choice at node $i \in N - \{d\}$, while for $i = d$ it is: $w_{idmg} = 0$, $p_{admg} = 0\ \forall a \in i^+$.

The cost of each alternative, also called *remaining cost* and denoted as w_{bdmg}, is obtained as the sum of the arc cost $b \in i^+ \cap A_m$ and the expected cost perceived by the user to reach the destination from its final node b^+:

$$w_{bdmg} = c_{bg} + w_{b^+dmg}. \tag{6.15}$$

These costs jointly determine the *conditional probabilities* of each arc $a \in i^+$ through a discrete choice model:

$$p_{admg} = \begin{cases} p_{admg}\left(w_{bdmg}, \forall b \in i^+ \cap A_m\right), & \text{if } a \in A_m \\ 0, & \text{otherwise} \end{cases}. \tag{6.16}$$

Note that users of mode m can take only arcs of this mode.

Arc conditional probabilities must satisfy the following consistency and non-negativity constraints:

$$\sum_{a \in i^+} p_{admg} = 1, \quad p_{admg} \geq 0. \tag{6.17}$$

Any discrete choice model provides together with the probability of each alternative the so-called *satisfaction*, i.e., the expected value of the maximum utility resulting from the choice. We assume that the expected cost perceived by users coincides with the opposite of the satisfaction in the local choice at the node:

$$w_{idmg} = w_{idmg}\left(w_{bdmg}, \quad \forall b \in i^+ \cap A_m\right). \tag{6.18}$$

Equation (6.18) for all nodes $i \in N - \{d\}$ (given a triplet dmg) can be seen as a system of nonlinear equations, where unknowns are the node costs w_{idmg}. Under the

assumption that only *efficient routes* are considered, i.e., paths getting closer to the destination with respect to some fixed cost or distance metric, which is typically acceptable in transit networks, the above system is triangular and can be easily solved by substitution, processing nodes in reversed topological order with respect to the chosen metric. Then, Eq. (6.16) can be computed in no particular order. It is interesting to recall that in case of Logit model, by introducing the concept of 'weights' as the negative exponential of costs scaled by the distribution parameter, the above system can be transformed in a system of linear equations (the first step of Dial's algorithm).

The case of deterministic choices deserves particular attention. Equations (6.18) and (6.16) for each $i \in N - \{d\}$ and $a \in i^+ \cap A_m$, respectively, become the following:

$$w_{idmg} = Min\left(w_{admg}, \quad \forall a \in i^+ \cap A_m\right), \tag{6.19}$$

$$p_{admg} \cdot \left(w_{admg} - w_{idmg}\right) = 0. \tag{6.20}$$

The complementarity condition represented by Eq. (6.20) is the formulation of Wardrop's First Principle for the local choice. The result is a one-to-many mapping where multiple flow patterns may correspond to one cost pattern if there are alternatives of equal cost.

The probability of each path $k \in K_{odm}$ from origin $o \in O$ to destination $d \in D$ on mode $m \in M$ can be determined a posteriori as the product of all the arc conditional probabilities making up the route (this result does not apply to hyperpaths):

$$p_{kg} = \prod_{a \in A} \left(p_{admg}\right)^{\Delta_{ak}}. \tag{6.21}$$

This equation is not required in the assignment model itself; however, path information is necessary to undergo post-evaluation (see Sect. 5.2.3), because many result indicators are calculated on the basis of path flows, regardless the fact that a sequential or strategic approach (both yielding arc probabilities) has been used in the route choice model.

Indeed, in sequential models, the typical way of performing flow propagation avoids the need to introduce paths, by solving a system of linear equations for all nodes $i \in N$ (given a triplet dmg), where unknowns are the node (exit) flows q_{idmg}. Each equation represents the following conservation of flows at the node.

$$q_{idmg} = d_{idmg} + \sum_{a \in i^-} q_{a^- dmg} \cdot p_{admg}, \tag{6.22}$$

where the exit flow is equal to the demand flow plus the entry flow. The latter is in turn given by the sum over the node backward star of each arc tail flow multiplied by the corresponding arc conditional probabilities. The demand flow d_{idmg} is null if

i is not an origin. Under the assumptions of efficient routes, the above system is triangular and can be easily solved by substitution, processing nodes in direct topological order with respect to the chosen metric (such as in the second step of Dial's algorithm). In the general (non-triangular) case, the coefficient matrix of system (6.22) is highly sparse, given that each equation involves only the adjacent arcs entering a node; this feature can be exploited by solution algorithms such as BiCGstab. Preconditioning by a triangularized solution (i.e., solving the problem without taking into account non-efficient arcs) has great advantages.

The arc flows of a specific user class can then be obtained as an aggregation of all contributions for each destination and mode:

$$q_{ag} = \sum_{d \in D} \sum_{m \in M} q_{admg}, \tag{6.23}$$

where q_{admg} is the product of the node flow and the arc conditional probability:

$$q_{admg} = q_{a^- dmg} \cdot p_{admg}. \tag{6.24}$$

It is worth warning that sequential models provide the same results (flows) of the corresponding route choice models only for some elementary case (e.g., deterministic, logit).

6.1.5 Sequential Model and Strategies

The proposed sequential model for route choice can be immediately extended to represent a strategy-based behaviour. In this case, the conditional probability p_{admg} of a diversion arc $a \in A^{div}$ is the result of two models:

- the local choice $p_{\check{a}dmg}$ among the hyperarcs exiting from the diversion node a^-, and
- the hyperarc diversion probabilities $p_{a|\check{a}\ dmg}$ depending on random events.

$$p_{admg} = \sum_{\check{a} \subseteq \left((a^-)^+ \cap A_m\right) : \check{a} \in H} p_{\check{a}dmg} \cdot p_{a|\check{a}dmg}. \tag{6.25}$$

The local choice probabilities require to compute the remaining cost $w_{b\check{a}dmg}$ for reaching the destination using each hyperarc $b^{\check{}}$available at node $i = a^-$. This is equal to the average, weighted by the diversion probabilities $p_{b|b^{\check{}}\,dmg}$, among its branches $b \in b^{\check{}}$, of the sum between the arc conditional cost $c_{b|b^{\check{}}\,dmg}$ and the expected cost w_{b^+dmg} from its final node b^+. Based on (6.9) and (6.15), it is given as follows:

$$w_{\breve{b}dmg} = \frac{c_{\breve{b}dmg} + \sum_{b \in \breve{b}} p_{b|\breve{b}\,dmg} \cdot w_{b^+ dmg}}{\sum_{b \in \breve{b}} p_{b|\breve{b}\,dmg}} = \frac{\gamma_{\breve{b}^- g} \cdot t_{\breve{b}dmg} + \sum_{b \in \breve{b}} p_{b|\breve{b}\,dmg} \cdot w_{bdmg}}{\sum_{b \in \breve{b}} p_{b|\breve{b}\,dmg}}; \tag{6.26}$$

the latter assumes that the travel time of the arc branches per se is null as it is already included in that of the hyperarc.

The reason for rescaling the probabilities in Eq. (6.26) and not on Eq. (6.25) is to allow models, such as the fail-to-board probabilities of Sect. 7.3.3, where the sum of the hyperarc diversion probabilities is less than one, i.e., where some flow is eliminated from the network during the flow propagation.

The hyperarc diversion probabilities result from an adaptation strategy to circumstances rather than a choice among alternatives. They are strictly related to the particular stochastic process under consideration; on transit networks, en-route random events may depend on line frequencies and on remaining capacities, as well as on expected costs to destination (see Sects. 7.1, 7.2.2 and 7.3.3). Equations (6.16) and (6.18) become, respectively:

$$p_{\breve{a}dmg} = p_{\breve{a}dmg}\left(w_{\breve{b}dmg}, \forall \breve{b} \subseteq \left((a^-)^+ \cap A_m\right) : \breve{b} \in H\right), \tag{6.27}$$

$$w_{idmg} = w_{idmg}\left(w_{\breve{b}dmg}, \forall \breve{b} \subseteq (i^+ \cap A_m) : \breve{b} \in H\right). \tag{6.28}$$

It is worth noting again that there is a noticeable difference between the hyperarc choice probabilities $p_{\breve{a}dmg}$ and the arc diversion probabilities $p_{a|\breve{a}dmg}$. The former are choice shares among possible route alternatives, the latter are the outcome percentages from random events. The arc conditional probabilities p_{admg} resulting from the route choice model are a combination of both, as evident from Eq. (6.25). Therefore, in the presence of hyperarcs and consequent strategy-based behaviour (in case of fail-to-sit and fail-to-board probabilities, there is no other option than strategies), the transit assignment model shall be extended to include the representation of physical phenomena providing, possibly congested, diversion probabilities.

6.1.6 *Shortest Paths and All-or-Nothing Assignment*

The computation of shortest trees rooted at zone centroids is at the base of most assignment algorithms, even when the route choice model is not deterministic but stochastic, and even when a sequential (arc-based) model is adopted instead of a path-based one. Therefore, in the following, we give some basic information about this problem.

In case of transit networks, the root of the tree is typically a destination and not an origin; this is a natural choice for strategic models.

Let us consider the problem for users of class $g \in G$ directed towards destination $d \in D$ on mode $m \in M$. Most shortest tree algorithms solve actually the dual problem of finding the minimum cost to reach the destination from each node $i \in N$ by repeatedly applying to every arc $a \in A_m$ the following *Bellman update*, until no further cost improvement is possible (Bellman 1958):

$$w_{a^- dmg} \leftarrow Min\left(w_{a^- dmg}, w_{admg}\right). \tag{6.29}$$

The above minimization checks whether using arc a at a cost of $w_{admg} = c_{ag} + w_{a+dmg}$ can improve the current cost w_{a-dmg} to reach the destination d from its initial node a^-.

The (expected) cost of each node (also called label) is initially set to infinity, except for the destination, whose cost is obviously zero.

Whenever a cost label is updated, that node is inserted in a list of nodes to be visited. The algorithm starts by initializing this list with the destination. Nodes are iteratively extracted from the list and Eq. (6.29) is applied to each arc of its backward star. If at each iteration a node with the least cost is extracted, then no node will be extracted twice, provided that all arc costs are non-negative. An effective way of (pseudo) ordering the nodes is by introducing a *bucket list*, where the space of expected costs is partitioned in (many small) n^b buckets of equal span δ^b; identifying the proper bucket for the insertion of a node i with cost w_{idmg} in the list requires just an integer division: $w_{idmg} \backslash \delta^b$. The resulting algorithm of Dijkstra (1959) is particularly suited for transit networks, which are characterized by anisotropic costs and non-planar graphs, and provides also a topological order of the nodes given by the inverse order of their extraction from the list.

In case of acyclic graphs, the Bellman update can be applied in inverse topological order, without the need of handling a list of nodes to be visited.

At each successful (convenient) update, the algorithm records also a, as the successor arc of its tail node a^- (or equivalently a^+ as its successor node); in other terms, p_{admg} is set to one, while for the other arcs of the node forward star the probability is set to zero. This information can be exploited in a so-called *All-Or-Nothing assignment* to shortest paths, where the travel demand is propagated by solving (6.22) in topological order. Then (6.24) is applied to obtain arc flows for each destination.

This yields one of the possible (extremal) outcomes of the deterministic model for route choice (6.19) and (6.20).

6.1.7 Extension to Shortest Hyperpaths

In principle, the computation of shortest hypertrees requires applying to every hyperarc $\breve{a} \in H$ *and* the following revised version of the Bellman update, in addition of applying (6.29) to every arc $a \in A_m$, until no further cost improvement is possible:

$$w_{\breve{a}^- dmg} \leftarrow Min\left(w_{\breve{a}^- dmg}, w_{\breve{a}dmg}\right). \tag{6.30}$$

However, the extension of the proposed Dijkstra algorithm to strategies is not trivial and some issues arise:

- to calculate in (6.30) the value of $w_{\breve{a}dmg}$ through (6.26), the algorithm has to wait for all the heads of hyperarc $\breve{a}$ to be extracted (indeed, all such nodes must have a cost value and the head cost of the branch included in the backward star of the node currently visited is not enough);
- the resulting node cost of the hyperarc tail can be lower than those of (some of) its heads (such as when arcs with negative cost are considered), which prejudices the *label setting* approach of the Dijkstra algorithm (although nodes are extracted from the list in order of cost, a node with a lower cost will be extracted after a node with a higher cost, so that a node already extracted can be further optimized);
- this implies that the optimal strategy can involve so-called *absorbing cycles* (e.g., an unlucky boarding passenger unable to seat, who then alights at next stop and walks back to wait again for the line at the previous stop, thus gaining another chance of seating on-board);
- each further cycle would have a smaller probability to happen, but a *label correcting* approach (i.e., the node cost can be modified even if the node has been already extracted form the list, thus requiring its insertion again—so, nodes can be extracted more than once) would induce infinite updates; shortest hyperpath would then require to solve the problem as a system of nonlinear (the minimum function) equations.

However, a hyperpath is by definition an acyclic sub-graph; to avoid this kind of paradoxes requires some additional rules in the search. For example, a label setting approach (i.e., the cost is not updated if the node has already been extracted) can be forced, unless the node is a diversion (to allow waiting for hyperarcs to be processed), or unless the correction derives from the successor of the node. This allows to eliminate absorbing cycles (if no cycle of diversion nodes exists), which can be justified with a risk-adverse behaviour: passengers never take twice chances, even if on average this maybe convenient, because it can result sometime in a higher cost. A complete analysis of this heuristic goes beyond the scope of this short note, whose aim was rather to raise some concern on the implementation.

6.1.8 Uncongested Assignment Versus User Equilibrium

If no congestion phenomena are considered to be relevant, then transit assignment reduces to a simple chain of sub-models: a flow-independent performance model, a route choice model, a flow propagation model. This can be solved by computing the following sequence of equations that for given arc performances yield arc flows:

with explicit path enumeration: (6.2) → (6.3) → (6.13) → (6.6) → (6.5), or (6.31)

with implicit path enumeration: (6.2) → (6.15) → (6.18) → (6.16) → (6.22) → (6.24) → (6.23). (6.32)

In presence of congestion or discomfort, we have to replace (6.2) with proper arc performance functions:

$$c_{ag} = c_{ag}\left(q_{bg'}, \quad \forall b \in A, \ \ \forall g' \in G\right). \tag{6.33}$$

Using (6.33) and (6.5) in (6.3) as follows:

$$c_{kg} = c_{kg}^{na} + \sum_{a \in A} c_{ag}\left(\sum_{h \in K} q_{hg'} \cdot \Delta_{bh}, \quad \forall b \in A, \ \ \forall g' \in G\right) \cdot \Delta_{ak}, \tag{6.34}$$

yields the so-called *supply function*:

$$c_{kg} = c_{kg}\left(q_{hg'}, \quad \forall h \in K, \ \ \forall g' \in G\right). \tag{6.35}$$

The relation represented by (6.33) closes an 'internal' loop in the model, because the arc flows provided by the uncongested assignment will change the arc costs, thus requiring to update route choice, and so on.

In case of recurrent congestion phenomena (discomfort and delay occurring every day at the same time), the most common paradigm adopted in the simulation of transit networks is the well-known User Equilibrium.

By definition, a User Equilibrium on a transit network is achieved when no passenger finds convenient to change route (as mentioned earlier, a route can be a single path connecting the O–D pair defining the trip of the passenger, or a hyperpath, in case of strategy modelling). This implies assuming that passengers are rational decision-makers, i.e., they minimize their (perceived) cost.

The introduction of arc performance functions that are able to reproduce the relevant congestion phenomena on transit networks makes the assignment problem more complex than the case of road networks. This is due to the *non-separability* of these phenomena: the cost of an arc depends on the flows of other adjacent arcs, and not only on the flow of the arc itself. Moreover, this dependency is in general not symmetric nor monotonic. The only noticeable exception is the case of over-crowding on-board discomfort.

In essence, the existence of an equilibrium is guaranteed (sufficient condition) by the continuity of the arc cost function, while the uniqueness of the equilibrium is guaranteed (sufficient condition) by the positive definiteness of the arc cost function Jacobian (in strict form, for deterministic choice models). As mentioned above, the

latter does not hold in general; however, in standard situations, the non-uniqueness does not typically occur but counterexamples can be made.

In the particular case of separable (and monotone) arc cost functions, the equilibrium assignment can be formalized and solved through an (convex) optimization model where the objective function is the sum of cost integrals (see Sect. 7.2.3). Otherwise, more complex formulations are required, such as variational inequalities or fixed-point problems. The framework that follows is based on the latter paradigm.

In the two figures below, white boxes indicate variables, grey boxes indicate functions, green boxes indicate input, and red boxes denote post-evaluation.

In the case of transit assignment, the cost functions will also provide the hyperarc diversion probabilities, which are essential in strategic route choice models, through the computation of line probabilities and fail-to-board or fail-to-sit probabilities (see Chap. 7).

The above schemes describe how the outlined variables and their structural relations can be organized in a concatenation of models to yield different kinds of fixed-point problems that can be introduced to formulate transit equilibrium assignment.

In general, a *fixed-point problem* finds a point $\mathbf{x}$ in a given subset X of a multidimensional space. This point $\mathbf{x}$ is mapped by the fixed-point function $f(\mathbf{x}) \in X$ on the point itself:

$$\text{find } \mathbf{x} \in X: \mathbf{x} = f(\mathbf{x}). \tag{6.36}$$

In case of a one-to-many mapping $f(\mathbf{x}) \subseteq X$, we shall substitute in the problem defined by Eq. (6.36) the equality symbol '=' with the belonging symbol '$\in$'; deterministic choice models are the examples of this modified instance.

In transport assignment, the space of search can be that of arc flows, arc costs, route flows or route costs, while the mapping results from the chain of models in the above schema that starting from the chosen fixed-point variable with a full round brings back to it. Figures 6.2 and 6.3 present in particular the case where $\mathbf{x}$ is the vector of arc flows, which is the typical modelling choice.

Fixed-point problems constitute thus a natural framework for equilibrium assignment. However, they present a drawback with respect to more classical optimization models: the lack of rapidly convergent algorithms prevents precise calculations of the equilibrium solutions which may be required when comparing scenarios.

In assignment models, the simple iteration of the fixed-point function does not converge in general (it is not a so-called *contraction*). Therefore, to solve the fixed-point problem, we typically use the method of successive averages (MSA), where at each iteration $n = 1, 2, \ldots$ the new equilibrium iterate is obtained as a convex combination between the current equilibrium flows and the application (to them) of the fixed-point function; in case of arc flows, the latter is also called *Network Loading Map* and the resulting flows are denoted q_{ag}^{nlm}:

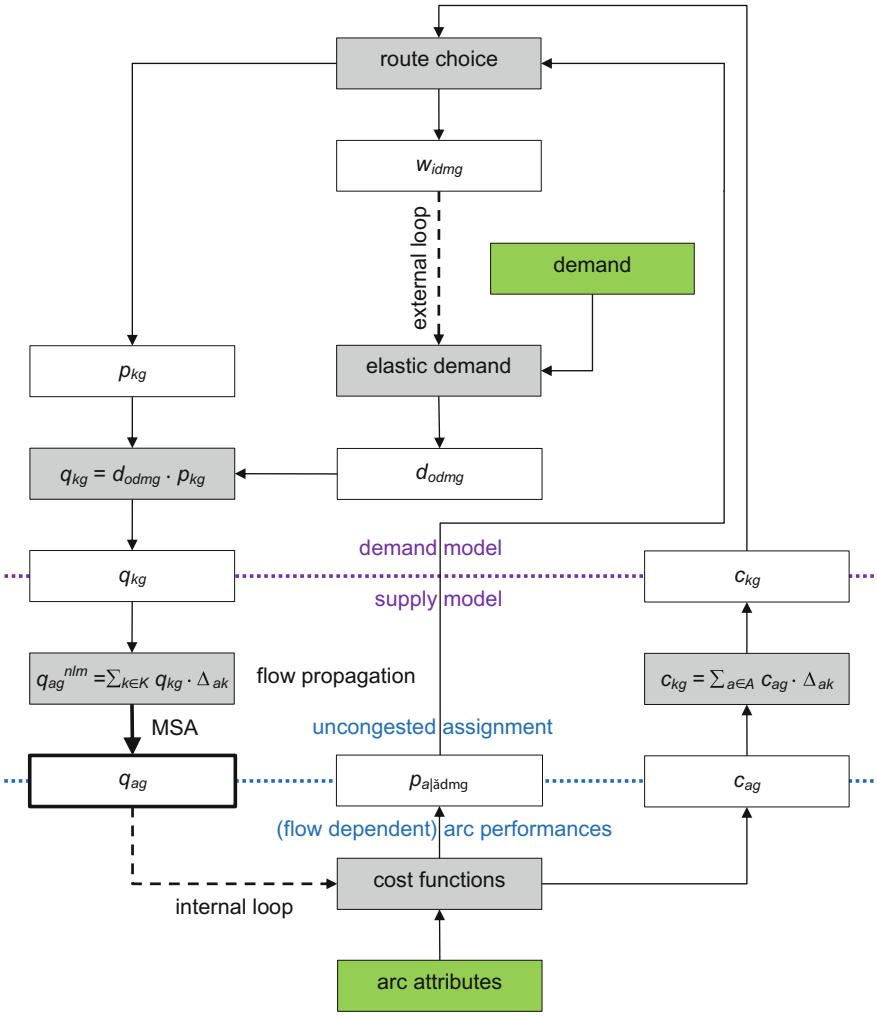

Fig. 6.2 Fixed-point formulation of equilibrium models on multimodal networks with explicit path enumeration

$$q_{ag} \leftarrow q_{ag} + \frac{1}{n} \cdot \left(q_{ag}^{nlm} - q_{ag} \right). \tag{6.37}$$

The MSA (see Sect. 4.2) is presented above in its simpler form, where the coefficient of the convex combination is the inverse $1/n$ of the iteration number. This actually provides the average of all the flows resulting in the network loading maps obtained so far; thus, slow convergence is somehow intrinsic.

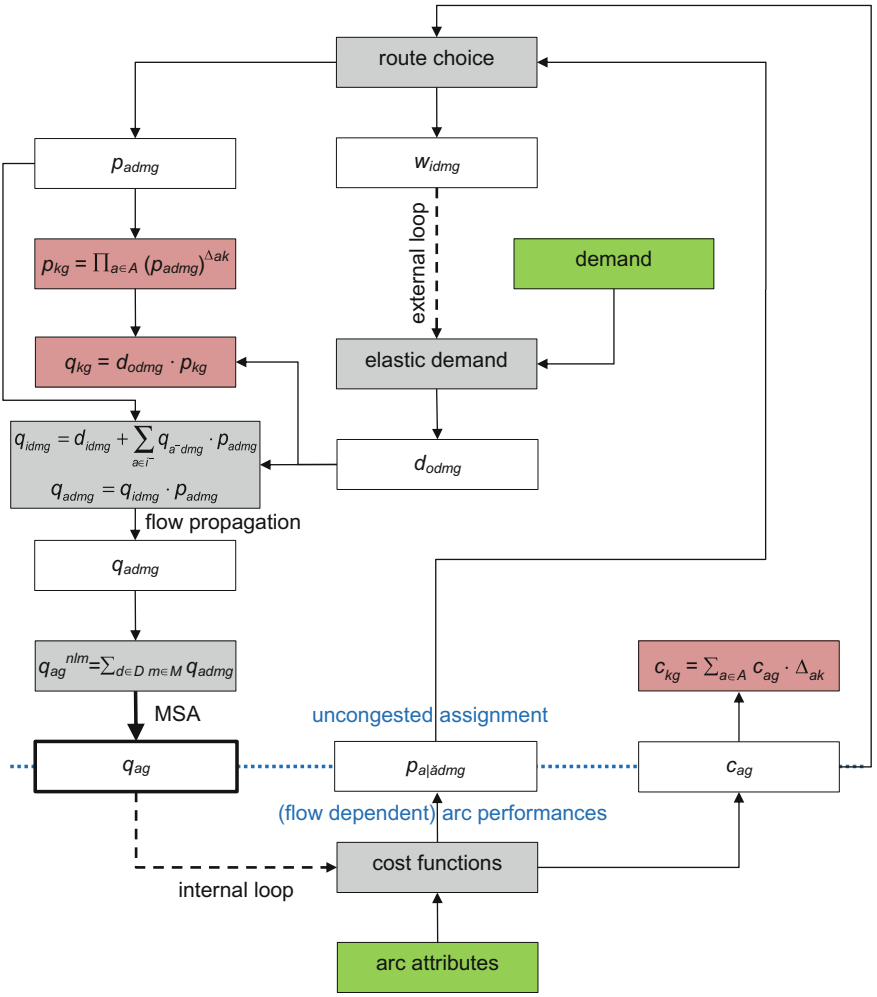

Fig. 6.3 Fixed-point formulation of equilibrium models on multimodal networks with implicit path enumeration. Path variables are obtained a posteriori for evaluation purposes

6.1.9 Fixed Versus Elastic Demand

Elastic demand is in general the dependence of O–D demand (flow) matrices from O–D skim (cost) matrices. This may involve different levels (stages) of the demand model (see Sect. 4.2.3.2), from generation rate, to distribution pattern and/or modal split (including departure time choice in case of dynamic models).

Elastic demand introduces a second 'external' loop in the model scheme of Figs. 6.2 and 6.3. But this can also be regarded as a *fork and join*, without the need of formulating a bi-level problem. Nevertheless, for traditional reasons, the

(few) commercial software that allows for elastic demand modelling adopt a two-step iterative approach, solving the internal loop before updating the external loop; typically, the external loop is not solved with a high precision and no averaging process such as the MSA is applied to it.

6.1.10 User Equilibrium Versus Day-to-Day Evolution

As shown in the previous section, from an algorithmic point of view the computation of a user equilibrium consists of an iterative process. Each iteration corresponds to a single assignment on the transport network, which represents the interaction of supply and demand through route choice, flow propagation and performance functions. This process resembles also the chronological evolution of the system from non-equilibrium to a possible equilibrium state, with the single assignment time frame being one day; indeed, a day is the typical temporal horizon considered in cyclic travel decisions, as well as in most human activities. Hence, one fixed-point iteration can be regarded as a day and everything that happens in this time frame as *within-day*. The internal loop of the user equilibrium model therefore corresponds to a *day-to-day* dynamic process of route choice, while the external loop corresponds to longer term travel choices (e.g., mode, destination and trip frequency); however, if the fork and joint approach is considered in the analysis of elastic demand, then also these choices are seen as part of day-to-day dynamics.

While user equilibrium models define a priori the relevant state of the system as that in which average flows and costs (demand and supply) are mutually consistent, inter-period (or day-to-day) dynamic assignment models simulate the evolution of the system over a sequence of similar periods (days), and its possible convergence over time to a stable condition. Under some rather mild assumptions, the equilibrium configurations can be interpreted as *attractors* of the dynamic system. This allows us to analyse the stability of equilibria and provides a statistical description of transient states. Although the mathematical analysis of dynamic systems is out of the scope of this book, it must be clear that the existence of a unique equilibrium is just one of the possible cases; the day-to-day process does not necessarily lead to a steady state (static or within-day dynamic) and may oscillate among different equilibria or even show a chaotic pattern.

Each user may update the route choice made for the current day based on the information gathered on the route costs during the previous trips. Possibly, all previous experiences contribute to the knowledge of the network developed by the user, although the learning process typically privileges the relevance of latest trips. In this evolutionary interpretation of equilibrium, in general users would experience every day a different cost on the same route, because congestion may induce other users to change their route or because random events may affect loads and performances; whether the experienced costs or other information sources induce a considerable change in the expectations that motivated the current choice, then a user will consider changing route.

Thus, day-to-day dynamic assignment models require the explicit representation of two phenomena:

- users' learning and forecasting mechanisms for utility updating; that is, how present route choices are influenced by experience on previous transport costs (memory);
- users' choice updating behaviour; that is, how present route choices are influenced by the choices made on previous days (habit).

The *utility updating* model describes in which way expected (or predicted) utilities on day n are influenced by experienced utilities on previous days (and possibly by other sources of information). In principle, a disaggregate approach can describe the updating of the individual perceived utilities of each single user (agent); otherwise, the utility updating can be applied to their averages (systematic utilities) considered by several users (demand component), or directly to the generalized costs, which are the main drivers of route choice.

In the following, it is assumed that referring to the generic path $k \in K_{odm}$ utilized by the travellers of class $g \in G$ in day n, the expected costs $\tilde{c}_{kg}^{n+1}$ of next day $n + 1$ are a convex combination (exponential filter) of the actual costs c_{kg}^{n} incurred in day n resulting from the supply function given in Eq. (6.35) based on the actual flows q_{kg}^{n} and the current expected costs $\tilde{c}_{kg}^{n}$:

$$c_{kg}^{n} = c_{kg}\left(q_{hg'}^{n}, \quad \forall h \in K, \ \ \forall g' \in G\right), \tag{6.38}$$

$$\tilde{c}_{kg}^{n+1} = \alpha_{g}^{learn} \cdot c_{kg}^{n} + \left(1 - \alpha_{g}^{learn}\right) \cdot \tilde{c}_{kg}^{n}, \tag{6.39}$$

where the average weight $\alpha_{g}^{learn} \in (0, 1]$ attributed by the users of class g to the actual costs is usually assumed to be independent of the day. Note that given the structural linear relationship given by Eq. (6.3) between arc and (additive) path costs, the exponential filter can also be applied to arc costs; this would also have a physical interpretation, since during each trip on the network a traveller gathers experience on arc costs that are part of several paths.

The *choice updating* model describes in which way route choices on day $n + 1$ are influenced by choices made on previous days. In the following, it is assumed that each day some users repeat the choices made in the previous day, and others reconsider (although do not necessarily change) their choices. Then, the flows q_{kg}^{n+1} of next day $n + 1$ are a convex combination (exponential filter) of the flows $\tilde{q}_{kg}^{n+1}$ that would result from the route choice model (6.13) based on the expected costs $\tilde{c}_{kg}^{n+1}$ of next day $n + 1$ and the current flows q_{kg}^{n}:

$$\tilde{q}_{kg}^{n+1} = p_{kg}\left(\tilde{c}_{hg}^{n+1}, \quad \forall h \in K_{odm}\right) \cdot d_{odm}, \tag{6.40}$$

$$q_{kg}^{n+1} = \alpha_g^{choup} \cdot \tilde{q}_{kg}^{n+1} + \left(1 - \alpha_g^{choup}\right) \cdot q_{kg}^{n}, \tag{6.41}$$

where the probability $\alpha_g^{choup} \in (0, 1]$ that a user of glass g reconsiders the choice made on the previous day is usually assumed to be independent of the day, while the complement $1 - \alpha_g$ is the probability that the choice of the previous day is repeated. In some models, the choice updating is neglected assuming $\alpha_g^{choup} = 1$.

Under this evolutionary interpretation of the equilibrium model, the main system variables are both the costs (utilities) and the flows (choice probabilities), which can be summarized in a state vector $\mathbf{x} = (\mathbf{c}_{KG}, \mathbf{q}_{KG})$, as in Fig. 6.4.

The within-day dynamic consists of a flow propagation procedure plus a new computation of performances; these may be possibly calculated at once (see 6. Dynamic Network Loading in Sect. 6.4). During the day, travellers execute their trip and accumulate experience concerning generalized costs.

Then, day-to-day dynamic takes place. The learning process filters the latest information about the network cost pattern gathered during the last day with the

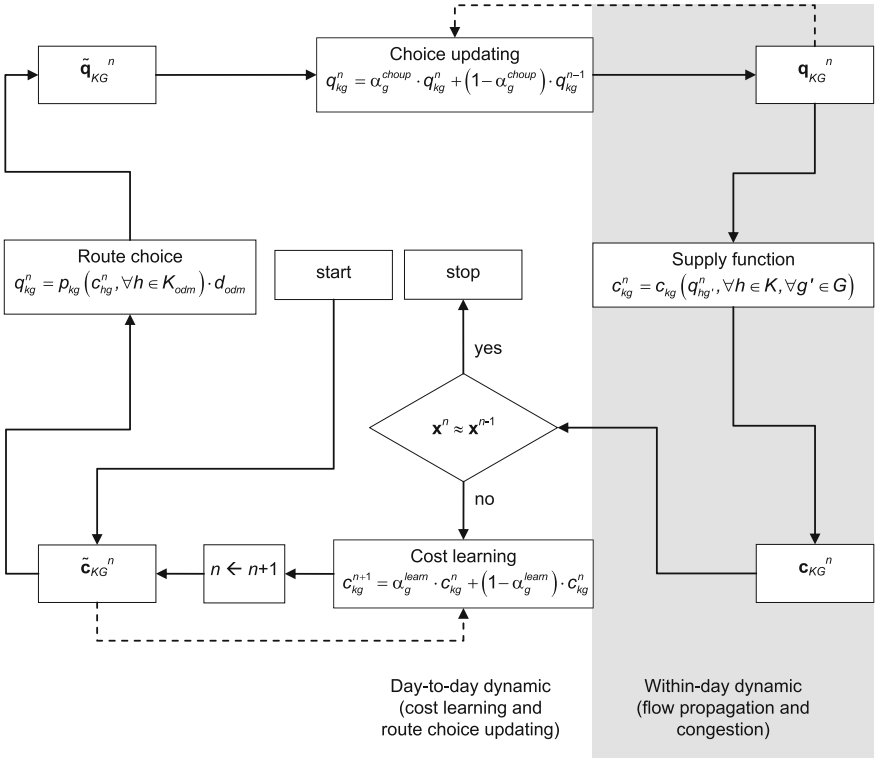

Fig. 6.4 Iterative flow propagation/congestion (within-day-dynamic) and cost learning/route choice updating process (day-to-day dynamic)

experience accumulated during all previous trips, updating the latter. The next day, the travellers can update their route choice on this basis in order to improve their objectives. But only a portion of them will actually reconsider the previous choice, probabilistically; the actual path flows that will load the network follow accordingly.

As travellers increase their experience with the system, their mental map extends and their expectations reflect more closely to the actual network performance. But a major issue in the application of the cost learning filter regards the update of path costs that have not been utilized by travellers in the previous day. In theory, only the cost of the utilized path should be actually revised by each single traveller. Instead, it is common practice to update the cost of all paths, independently from the fact that they have been used or not. To justify this approach, we can assume some form of collective awareness where information is shared among users; this is not far from reality in a changing world of social networks and travel information based on crowd sourcing. This is more credible in the context of probabilistic models where each path available to a demand component is travelled in a given day by at least a small proportion of users.

Clearly, this assumption accelerates the day-to-day process towards a possible equilibrium. If instead travel demand is represented through individual agents (see Sect. 6.5) with their own memory (in contrast to the collective memory of the above schema), the proposed process (possibly) leading towards equilibrium (which involves learning, choosing and congestion) is slower and must be guided necessarily by random perturbation of expected costs for each simulated day, as otherwise there is the risk of having individuals trapped on bad paths because of wrong estimation of their available alternatives.

6.1.11 Path-Based Versus Arc-Based

Similar to route choice and flow propagation, from a topological point of view, there are two main kinds of assignment models: *path-based* and *arc-based*.

In the first case, the relevant routes are explicitly enumerated; they can be identified in advance or generated during the assignment process (column generation). The route cost can include non-additive terms.

In the second case, the arc conditional probabilities result from a sequential model with implicit enumeration of routes, where users directed towards a given destination are recursively split among the arcs of the node forward star. Only additive cost structures are allowed.

Looking at route choices, path-based models are the most natural approach and are also richer in terms of modelling opportunities. In this case, for example, sophisticated stochastic models can be easily formulated using random utility theory, including correlation among alternatives (e.g., Probit, Cross Nested Logit, C-Logit). Moreover, route costs do not have to be necessarily additive with respect

to arc cost, thus allowing to evaluate fancy fares structures and nonlinear disutilities (see Sect. 4.5.2).

However, path-based models usually require to preliminary identify and explicitly enumerate all the relevant route alternatives. Although on transit networks the number of good alternatives is definitely less than those emerging (due to congestion and grid topology) on an urban road network, this task may be cumbersome in terms of computation and hard in terms of modelling. Actually, explicit route enumeration requires a specific selection model, since the number of acyclic paths (and even more, hyperpaths) on a transport network is finite but can be extremely large, so much to make the problem with exhaustive enumeration practically unsolvable.

Column generation during equilibrium assignment (i.e., build up and store new paths at each iteration) is actually available only for deterministic models (in a stochastic framework the process would hardly converge) and provides a reduced set of used paths with respect to the whole set of possible equilibrium paths. Indeed, it is well known that the solution of deterministic equilibrium may be unique (under monotonicity conditions on the arc performance function) in terms of arc flows, but it is not in general unique in terms of path flows. Because equilibrium solutions in terms of paths obtained through column generation are rather poor, they are not suitable for post-processing procedures, such as O–D matrix estimation from traffic counts and critical link analysis (i.e., to identify all path flows using a given link).

Arc-based models are more robust with respect to these issues and are therefore often chosen for the implementation of commercial software. Moreover, it is always possible to retrieve a practical set of used paths starting from the arc conditional probabilities, using Eqs. (6.21) and (6.6), for example not considering paths whose probability is below a certain threshold. Finally, when considering strategies, arc-based models are almost a necessity.

6.1.12 *Deterministic Versus Stochastic Route Choice*

From a behavioural point of view, there are two main kinds of route choice models: *deterministic and stochastic* (or probabilistic). More details about route choice models are provided in Sect. 4.5; the purpose of this section is then to highlight some issues related to the assignment model.

Deterministic models assume homogeneity of attribute preferences for users of the same class and perfect information, i.e., passengers have a good knowledge of the network performance pattern (travel costs and speeds) in space and time for the current-day type. In this case, the rationality of the decision maker brings to the choice of minimum-cost alternatives.

The alternatives are routes connecting an O–D pair, in case of path-based models, or arcs exiting from a node, in case of arc-based models.

The most commonly applied paradigm for stochastic models is *random utility* theory, where it is assumed that users are rational decision-makers who associate a

utility to each *travel alternative* of a (finite) *choice set* and choose the best among them. The modeller is not able to evaluate exactly these utilities for each user, due to several factors, among which:

- heterogeneity of preferences among users of a same class with respect to the same attributes of alternatives;
- subjective errors in the perception of objective attributes by users (incomplete information);
- measure errors in the evaluation of real attributes by the modeller.

Then, the modeller can represent the utility of these alternatives as a multivariate random variable with a joint distribution (if correlations among alternatives are relevant). As a result, it is only possible to calculate the probability that each alternative has to be chosen, i.e., to have the highest utility. If the variance of random utilities is null, the model reduces to the case of *deterministic* behaviour with perfectly informed users who choose (the) best alternatives.

Despite many years of research about stochastic assignment models, also for transit assignment, the fact is that still most of the methods implemented in commercial software and actually used in practice consider a deterministic behaviour. Clearly, stochastic models are much more flexible and realistic in reproducing passenger flow patterns. Nonetheless, advantages of deterministic model are given as follows:

- easier to understand from a theoretical point of view (not from the mathematical one);
- their results are easier to interpret and analyse;
- have no behavioural parameter to be calibrated; and
- are more reliable and robust from a computational point of view.

For these reasons, if the actual aim of the modeller is to analyse the sensitivity to design variables in a project and not to reproduce reality, deterministic models can represent a valid opportunity.

Another reason for opting to deterministic models is that the stochastic models which are able to suitably reproduce the correlations (e.g., due to path overlapping) among alternatives are not yet robust enough for scenario comparisons; in particular, the Probit model requires too many Monte carlo iterations of the main assignment loop to achieve a reasonable stability.

However, we shall be aware that deterministic models tend to transfer the motivation for a plurality of used paths serving a same O–D pair from behaviour heterogeneity in route choice to congestion.

6.1.13 Static Versus Dynamic Assignment

Static models are based on the following assumptions of *stationarity*: the network can be described with constant flows and performances during the assignment period. This requires that travel demand as well as all supply features is constant for

a sufficient period of time and that the network works in under-saturated conditions, i.e., no permanent queue is observed. Thus, queues at transit stops can be suitably modelled in a static framework only if each waiting passenger is able to board the next-arriving vehicle.

In dynamic models, the fact that travelling takes time and that network elements have a capacity is explicitly considered, not merely as a component of disutility. The following phenomena can be modelled:

- the route costs and the corresponding choices refer to specific departure times and shall be computed considering the concatenation of travel times, i.e., each arc cost shall be evaluated at the instant when a passenger following that route enters it (dynamic route choice);
- passenger flows move on the network consistently with travel times (dynamic propagation);
- exit flows on network elements satisfy the presence of capacity constraints (queues);
- entry flows on network elements satisfy the presence of occupancy constraints (spillback).

Five ways of incorporating dynamic aspects in transit assignment can be identified:

- *space-time network* models, where daytime is built-in the topological structure of a diachronic graph (see Sect. 6.3);
- *quasi-dynamic* models, where a layer sequence of static models is defined, each referred to a time interval, to reproduce some dynamic phenomena, such as queuing;
- *macroscopic* models, where passengers and vehicles are represented as a (semi) continuous fluid characterized by temporal profiles (see Sect. 6.4);
- *microscopic* models, where individual passengers and vehicles are represented as discrete particles;
- *mesoscopic* models, where in terms of travel behaviour passengers and vehicles are represented as individual agents or packets of agents, and moved accordingly on the network, while their interaction (congestion and travel times) is reproduced through aggregated traffic models (see Sect. 6.5.4).

Space-time network models consider the concatenation of dynamic route choice but adopt a graph-based representation of flow propagation, with no possibility of reproducing the effects of passenger congestion on run delays (dwell times) in a consistent way. Moreover, some limitations arise in the simulation of passenger queues at stops and their effects on travel times; for example, FIFO queues cannot be reproduced and only mingling is possible.

Quasi-dynamic models introduce a chronological sequence of static layers each representing a fairly long-time interval. Usually, this time discretization is such that passengers complete a relevant portion of their trip within a same interval (e.g., 15 min). The concatenation of times is neglected in the route choice, by assuming *instantaneous* route costs that are computed separately for each static layer; this

holds true also for the dynamic flow propagation on the network as travel demand is loaded from origins to destinations during the same layer, without considering that the movement of passengers takes time. However, a proper congestion model can be adopted which allows for explicit reproduction of queues at stops, and the extra passengers who are not able to board during the current time interval due to capacity constraints can be shifted to the next temporal layer as additional demand components which behave according to current costs of the new layer.

Macro-, micro- and meso-models allow the simulation of all dynamic phenomena (dynamic route choice, dynamic propagation, queues and spillback).

6.1.14 Simulation-Based Versus Analytical Models

Public transport systems involve lots of complex relations among variables, many of which can be suitably described as random outcomes of erratic events that may change significantly from day-to-day (e.g., the actual number of passengers waiting at the stop, the actual arrival time of a vehicle and the actual travel time of a run). Most of the aspects regarding passenger information and service congestion that affect route choice on transit networks (see Chap. 7) are highly dependent on these unpredictable phenomena.

Random events involve both demand and supply. On the demand side, individual trip decisions are taken each day regarding the actual departure time or route choice. On the supply side, actual travel times of line vehicles are affected by road traffic and driver behaviour. Moreover, dwell times of vehicles at stops and queuing times of passengers at stops depend on the loads of passengers boarding, alighting and riding each line run (congestion), while the propagation of flows along passenger routes depend on the travel times on the transit network. The strategic behaviour of passengers at stops may amplify the effect of random events because in reaction en-route decisions are taken which further divert flows. On this basis, travel times and passenger flows become all correlated random variables.

A major distinction among the available approaches to transit assignment can then be made between:

- *analytical formulations*, where model results yield directly the expected values of the output variables (loads and performances),
- *simulation tools*, where model results yield one possible outcome of the output variables, so that (in theory) several repetitions of the model are necessary to obtain a stable average of each variable and (more interestingly) the shape of its distribution.

Simulation-based models for within-day transit assignment are highly flexible and more suitable to reproduce all such complex-correlated phenomena. This comes at the price of unstable results, which can be a serious drawback when the final aim is that of comparing design scenarios. However, this disadvantage may be alleviated if each within-day simulation is considered in the context of a day-to-day

evolution framework (see Sect. 6.1.10), as this gives some justifications to the lack of (enough) repetitions.

However, also the design based on precise results in terms of expected values presents some limitations. For example, a robust project should be taken into consideration the random distribution of the output, rather than only the average values of the outcomes, so as to guarantee good performances in the majority of cases. In this respect, simulation-based models can effectively support robust design with the calculation of results in terms of percentiles based on the a priori definition of safety margins against unfavourable cases.

There are two main contexts of application for simulation-based models:

- in *real-time* applications, many of the variables can be retrieved directly from the field (e.g., the current estimation and forecast of vehicles arrival provided by an AVL system);
- in *offline* applications, the same information must instead be elaborated on the basis of synthetic values extracted from random variables with known distributions.

Different levels of aggregation are possible in simulation-based transit assignment and some models actually make use of relations among average variables, such as travel times and flows, instead of looking at individual passengers and vehicles (mesoscopic models). If individual passengers are simulated then also their preferences can be synthesized and the user classes are substituted by distribution of parameters.

Simulation models can really make a difference in reproducing the effect of information about random events and the reaction of travellers. We can define and distinguish the following types of events:

- *minor events* are perturbations of the cost pattern in which passengers incur while travelling without anticipated knowledge, whose relevance or frequency is not sufficient to induce a strategic or rerouting behaviour;
- *recurrent events* are outcomes of systematic phenomena on which passengers have expectations and may be informed at some point en-route, thus allowing a strategic behaviour;
- *major events* are relatively rare but serious accidents on which passengers do not have expectations because their frequency is low, but whose relevant impact may induce rerouting.

Minor events affect the distribution of the corresponding arc costs which are perceived by users. But without prior information, users will choose the best path on average, possibly associating an additional risk-adverse cost to variances. Analytical formulations, which are based directly on the averages, are still appropriate because the expected value of a sum of random variables (path cost) is the sum of the expected values (arc costs), and the same is true for variance.

Recurrent events induce a strategic behaviour where the cost and the probability of local alternatives depend recursively on the expected costs of the diversions possibly encountered later during the trip towards the destination (see Sect. 7.1).

Only if the events are independent and are informed locally, then analytical formulations through the introduction of hyperarcs are actually suitable to reproduce average phenomena.

If the information is anticipated (which today is possible through mobile communication) and/or the random events are strongly correlated, then the simulation approach becomes unavoidable to reproduce the reaction of passenger in terms of en-trip route choices. Decision points are not anymore fixed (e.g., stops) as in the classical strategy representation based on hyperarcs, because the information can reach the passenger virtually anywhere and at any time. Upon each further injection of information, the passenger will reconsider all available alternatives to reach his/her destination and possibly update his/her route choice.

This usually requires the recomputation of attributes (in primis, travel and wait times) for a predetermined choice set of paths from that point to the destination. However, this practical approach (paths can be stored in computer memory) is not fully satisfactory because it does not take into account that the alternatives should be strategies with recursive diversions and not simple paths: this way only the first (current) diversion is properly considered. On the other hand, the explicit selection and storage of hyperpaths is prohibitive.

A possible alternative to path storage is the sequential route choice (see Sects. 6.1.4 and 6.1.5), where decisions are reconsidered locally by hypothesis; hyperpaths do not have to be explicitly enumerated but instead the expected cost of optimal strategies from nodes to destination is constantly updated. In this case, also the knowledge possibly acquired in a day-to-day learning process is stratified on node variables (expected cost to destination) and not on paths, by considering the cost actually suffered in the last within-day simulation from that node to the destination.

Major events and rerouting can be reproduced, not only with simulation models, but also through analytical models with a rolling horizon approach. This means that the analytical model is restarted every say 5 min to provide a prediction for the next say 60 min, by considering as a 'warm' initial state the result of the previous simulation; each iteration yields possibly different results from the previous one because new information and events are included in the simulation, affecting both supply characteristics and passenger behaviour.

6.1.15 Reference Notes and Concluding Remarks

The introduction of hyperarcs and hyperpaths for the representation of strategies on transit networks is due to work of Gallo et al. (1993).

A detailed presentation of stochastic (and deterministic) equilibrium models based on fixed-point problems for multiclass assignment on multimodal networks with elastic demand is provided in Cantarella (1997). With particular reference to transit networks, Nielsen (2000) uses a type of probit model to represent stochastic route choice.

Sequential route choice models have been proposed by many authors, among which Gentile and Papola (2006), who provide a general theoretical framework and several solution algorithms. Its consistent formalization with respect to multimodal transport networks and strategies, with the specific role of hyperarc diversion probabilities, can be considered an original contribution of this book.

Day-to-day dynamic processes in transport modelling were first proposed by Cascetta and Cantarella (1995) and by Watling (1999) in the framework of car assignment to road networks.

6.2 Frequency-Based Assignment on Transit Static Networks

Guido Gentile and Michael Florian

In this section, frequency-based (or headway based) models for static assignment on transit networks are presented in their basic version, without involving strategic behaviour of passengers with respect to common lines and information or congestion phenomena on the supply side, which will be analysed in Chap. 7.

Although the public transport service is organized with runs for each transit line and is thus actually available at stops only at discrete times, in frequency-based models the basic representation of supply is continuous (like that of cars on a road network) and the flow of vehicles can be seen as a moving walkway. The main issue is then the representation of the passenger wait times required at stops to access the available transit lines, which depend on the vehicle headways.

In this framework, the service is perceived by passengers in terms of probabilistic departure events of lines from the stops, because the timetable is not relevant in the route choice due to high frequency or low regularity. The line headway at any stop can be then represented as a random variable with given statistical distribution, and the frequency is defined as the inverse of its expected value.

The main characteristic of frequency-based models is thus their capability of reproducing discontinuous transit services by means of a continuous network representation. This implies to identify waiting as a separate trip phase through specific arcs. To this end it is necessary, on one side to calculate the expected wait time corresponding to a given headway distribution, on the other side to build up a proper topological representation of the transit graph.

6.2.1 Headway Distributions and Wait Times

Frequency-based models were originally based on the assumption that passengers arrive randomly at stops and service headways are *deterministic* (regular) and independent. In this case, the wait time has a uniform distribution equal to the

frequency from zero to the inverse of the frequency (i.e., the given headway, which is also the maximum wait time); the expected wait time is simply equal to one half of the frequency inverse. However, these assumptions are inconsistent with statistical analysis of real-world data since constant headways can be obtained only under perfect service regularity (see Sect. 7.4).

On the other hand, instead of evenly spaced headways, one can consider the case where transit service is completely unpredictable (irregular) and can thus be described as a Poisson arrival process of rare events. This assumption results in a (negative) *exponential* distribution of the headways and of the wait times, which implies the 'memory less' property: the elapsed wait time gives no further indication about the remaining wait time (the conditional distribution of an exponential function is indeed that same exponential function). The expected wait time is equal to the inverse of the frequency; it is thus twice as long than the case with deterministic headways. This shows the relevance of the assumption regarding headway distributions.

In frequency-based assignment, the headway distribution is typically a characteristic of the whole line; but to represent service perturbation along the line (e.g., bouncing), it should be modelled as stop specific; the AVL systems today allow for such a more detailed input (see Sect. 5.1.2). In the following, we refer in general to a *service headway h* of a given line at a given stop during a given interval (thus the indices ℓst are omitted).

The headway is modelled as a random variable with an independent probabilistic distribution, i.e., a density function $\varphi^h(h)$. As mentioned earlier, the inverse of its expected value $f = 1/E(h)$ is called the *frequency*, which is the main parameter of the headway distribution.

A flexible representation of service regularity can be obtained under the assumption that headways adhere to an *Erlang* distribution, which describes the sum of n independent Poisson processes:

$$\varphi^h(h) = \begin{cases} \frac{Exp(-n \cdot f \cdot h) \cdot (n \cdot f)^n \cdot h^{n-1}}{(n-1)!}, & \text{if } h \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{6.42}$$

This distribution (see Fig. 6.5) can bridge the gap between the two above extreme cases, by letting the parameter n vary from 1 (exponential—perfect uncertainty) to ∞ (deterministic—perfect regularity). Note that in the above formula $f \cdot n$ is the frequency of the n independent Poisson processes, while f is the frequency of vehicle departures from the stop.

The minimum of independent exponential random variables is also distributed exponentially with a frequency parameter that equals the sum of the random variable parameters. Therefore, the expected wait time for the first vehicle serving a set of attractive lines equals the inverse of the cumulative frequency.

To obtain the analogous result in case of common lines with deterministic headways, i.e., to obtain an expected wait time which is half the inverse of the cumulative frequency, would require that departures from the same stop of different

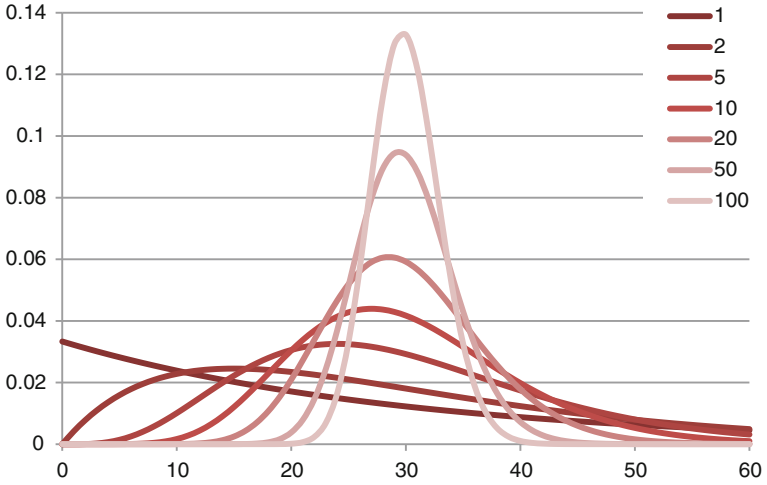

Fig. 6.5 Probability density function of the Erlang headway distribution for different values of n ranging from 1 to 100 and $f = 1/30$. For $n \to \infty$ the impulse function at $h = 1/f = 30$ is obtained

lines are equally spaced and in that sense perfectly coordinated, which is in contrast to the assumption of independent headways. But this would be theoretically possible only in case of identical line headways. Perfect correlation is thus practically impossible and assuming the above wait time expression for common lines with deterministic headways is just an optimistic approximation.

This justifies a more detailed analysis of independent headways and resulting wait times for the case of common lines that is developed in Sect. 7.1. Instead, in the following, we address the case of just one line for general headway distributions.

6.2.1.1 Mathematical Derivation

Because the headway h is random, then also the wait time (for a given line at a given stop) is random. Assuming that passengers arrive uniformly distributed at the stop, the probability density function $\varphi^w(t)$ of the wait time is related to the headway distribution through the formula:

$$\varphi^w(t) = f \cdot \bar{\Phi}^h(t), \tag{6.43}$$

where, by definition, it is:

$$\bar{\Phi}^h(h) = 1 - \Phi^h(h) = \int_h^{h^{max}} \varphi^h(h) \cdot dh, \tag{6.44}$$

and h^{max} is the maximum headway (it can be $h^{max} = \infty$).

Proof We now prove the validity of Eq. (6.43).

To this end, we shall first show that:

$$f = \frac{1}{\int_0^{h^{max}} \bar{\Phi}^h(h) \cdot dh}. \tag{6.45}$$

Differentiating by parts, it is:

$$\bar{\Phi}^h(h) \cdot dh = d\left(\bar{\Phi}^h(h) \cdot h\right) - h \cdot d\bar{\Phi}^h(h). \tag{6.46}$$

The following 3 statements hold true:

$$\begin{aligned} &\left[\bar{\Phi}^h(h) \cdot h\right]_0^{h^{max}} = \bar{\Phi}^h\left(h^{max}\right) \cdot h^{max} - \bar{\Phi}^h(0) \cdot 0 = 0 \cdot h^{max} - 1 \cdot 0 = 0 \\ &- h \cdot d\bar{\Phi}^h(h) = -h \cdot d\left(1 - \Phi^h(h)\right) = h \cdot d\Phi^h(h) = h \cdot \varphi^h(h) \cdot dh \\ &\int_0^{h^{max}} h \cdot \varphi^h(h) \cdot dh = E[h]\,. \end{aligned} \tag{6.47}$$

Based on (6.47), taking the integral of (6.46) on both sides between $h = 0$ and $h = h^{max}$ yields:

$$\int_0^{h^{max}} \bar{\Phi}^h(h) \cdot dh = E[h]. \tag{6.48}$$

Because by definition it is: $f = 1/E(h)$, then Eq. (6.48) is equivalent to (6.45), which shows the relation between the frequency and the integral of the distribution function; this is actually a general property of non-negative random variables.

Now, the fact that the wait time is exactly equal to t occurs for some value of headway h not lower than t (otherwise the passenger cannot have waited that long), if the passenger arrives at the stop $h - t$ time units (e.g., minutes) after the previous vehicle departure. Given that passenger arrivals at stops are uniformly distributed, each one of these possible events has a probability that is proportional to $\varphi^h(h)$, through a constant, say α. Therefore, summing up these probabilities yields the (density) of probability that the wait time is equal to t:

$$\varphi^w(t) = \alpha \cdot \int_t^{h^{max}} \varphi^h(h) \cdot dh = \alpha \cdot \bar{\Phi}^h(t). \tag{6.49}$$

Like any probability density function, the integral of $\varphi^w(t)$ over all possible wait times is 1:

$$\int_0^{h^{max}} \varphi^w(t) \cdot dt = 1. \tag{6.50}$$

Then, substituting the right-hand side of Eq. (6.49) into to the integrand of Eq. (6.50), based on Eq. (6.45) gives:

$$\alpha = \frac{1}{\int_0^{h^{max}} \bar{\Phi}^h(h) \cdot dh} = f. \tag{6.51}$$

Finally, based on Eqs. (6.51), (6.49) gives Eq. (6.43), which proves the assertion. ♦

In the case of Erlang headway distributions, based on formula (6.43), the probability density function of the wait time is given as:

$$\varphi^w(t) = \begin{cases} f \cdot Exp(-n \cdot f \cdot t) \cdot \sum_{i=0}^{n-1} \frac{(n \cdot f \cdot t)^i}{i!}, & \text{if } t \geq 0 \\ 0, & \text{otherwise,} \end{cases} \tag{6.52}$$

while the probability of waiting for more than t is:

$$\bar{\Phi}^w(t) = \begin{cases} Exp(-n \cdot f \cdot t) \cdot \sum_{i=0}^{n-1} \left(1 - \frac{i}{n}\right) \cdot \frac{(n \cdot f \cdot t)^i}{i!}, & \text{if } t \geq 0 \\ 1, & \text{otherwise.} \end{cases} \tag{6.53}$$

In case of deterministic headways, we obtain a uniform distribution of wait times:

$$\varphi^w(t) = \begin{cases} f, & \text{if } 0 \leq t \leq \frac{1}{f} \\ 0, & \text{otherwise} \end{cases}, \tag{6.54}$$

while the probability of waiting for more than t is:

$$\bar{\Phi}^w(t) = \begin{cases} 1 - f \cdot t, & \text{if } 0 \leq t \leq \frac{1}{f} \\ 0, & \text{if } t > \frac{1}{f} \\ 1, & \text{otherwise} \end{cases}. \tag{6.55}$$

Based on Eq. (6.52) with $n = 1$, for exponential headways (irregular service), the expected wait time is given as:

$$t^{wait} = \int_0^{\infty} \varphi^w(t) \cdot t \cdot dt = \frac{1}{f}. \tag{6.56}$$

Based on Eq. (6.54), for deterministic headways (regular service), the expected wait time is given as:

$$t^{wait} = \int_0^{\frac{1}{f}} \varphi^w(t) \cdot t \cdot dt = \frac{0.5}{f}. \tag{6.57}$$

The general formula for the expected wait time, under the assumption of uniform passengers' arrivals at stops depends only on the first and second moments of the headway distribution and not on its pdf:

$$t^{wait} = 0.5 \cdot \frac{E(h^2)}{E(h)}. \tag{6.58}$$

Proof We now prove the validity of Eq. (6.58).

Assume that a sequence of n independent random headways is given, which represent time intervals between two consecutive carrier arrivals at a stop, each of duration h_j, with $j = 1, \ldots, n$. Let passengers' arrivals at the stop be uniformly distributed.

The probability p_j that a passenger arrives during interval j is proportional to the headway h_j:

$$p_j = \frac{h_j}{\sum_{i=1}^{n} h_i}. \tag{6.59}$$

The expected value of the wait time t, conditional to the above event is given as:

$$E(t|j) = 0.5 \cdot h_j. \tag{6.60}$$

Based on the law of total probability, the expected wait time is obtained from the conditional expectations as follows:

$$E(t) = \sum_{j=1}^{n} E(t|j) \cdot p_j. \tag{6.61}$$

Using Eqs. (6.59) and (6.60) in Eq. (6.61), we obtain:

$$E(t) = \sum_{j=1}^{n} 0.5 \cdot h_j \cdot \left(\frac{h_j}{\sum_{i=1}^{n} h_i} \right) = 0.5 \cdot \frac{\sum_{j=1}^{n} h_j^2}{\sum_{j=1}^{n} h_j} = 0.5 \cdot \frac{\frac{\sum_{j=1}^{n} h_j^2}{n}}{\frac{\sum_{j=1}^{n} h_j}{n}} = 0.5 \cdot \frac{E(h^2)}{E(h)}, \tag{6.62}$$

which proves the assertion. ♦

6.2.1.2 Service Irregularity and Variation Coefficient

From an engineering point of view, the effect of service irregularity on the expected wait time can be reproduced through the variation coefficient $\sigma \geq 0$, introduced in Sect. 5.1.2.4. Because in general it is:

$$Var(h) = E\Big((h - E(h))^2\Big) = E\big(h^2\big) + E(h)^2 - 2 \cdot E(h) \cdot E(h) = E\big(h^2\big) - E(h)^2, \tag{6.63}$$

and considering the definition $\sigma = SD(h)/E(h)$, it is:

$$1 + \sigma^2 = 1 + \left(\frac{SD(h)}{E(h)} \right)^2 = 1 + \frac{Var(h)}{E(h)^2} = \frac{E(h^2)}{E(h)^2}. \tag{6.64}$$

Then, Eq. (6.58) can be rewritten as:

$$t^{wait} = \frac{0.5}{f} \cdot \big(1 + \sigma^2\big) = \frac{1}{f} \cdot \sigma^2 + \frac{0.5}{f} \cdot \big(1 - \sigma^2\big). \tag{6.65}$$

Again, for deterministic headway, it is: $\sigma = 0$; while, for exponential headway it is: $\sigma = 1$. Therefore, the above equation can be seen as the convex combination between the exponential wait time and the deterministic wait time through the square of the variation coefficient.

In case of Erlang headway distribution, where it is $E(h) = 1/f$ and $E(h^2) = (1 + 1/n)/f^2$, based on Eq. (6.58), the expected wait time is given as:

$$t^{wait} = \frac{0.5}{f} \cdot \left(1 + \frac{1}{n}\right). \tag{6.66}$$

A simple comparison between Eqs. (6.65) and (6.66) shows how the second parameter of the Erlang distribution is related to the headway variation coefficient: $n = 1/\sigma^2$.

The variation coefficient of the headway is the most common measure of service (ir)regularity and can be easily obtained from Automated Vehicle Location (AVL) data. Common values for different levels of transit right-of-way are available from transit planning guides (e.g., TCQSM, TRB 2013).

6.2.2 The Static Transit Network

In this section, the network topology for the static transit assignment with frequency-based services is derived starting from the input data (see Sect. 5.1).

A transit trip consists in general of several phases:

- accessing a transit stop from the origin, usually by walking;
- waiting at that stop for a transit vehicle;
- boarding a dwelling vehicle;
- travelling (or running, or riding) in the vehicle (on board) through a sequence of stops;
- alighting the vehicle at another stop;
- (possibly) transferring between two transit stops, usually by walking;
- (possibly) repeat the phases from waiting to transferring a certain number of times;
- and finally, egress from a transit stop to the destination, usually by walking.

Each trip phase is (possibly) represented by a sequence of arcs with a same type (which specifies the nature of the trip phase) on the *transit network*; the latter is composed by:

- the *pedestrian network*, including centroids and connectors, as well as access, egress, walking and transfer links;
- the *line network*, with a sub-network for each transit line articulated in boarding, running, dwelling and alighting arcs plus the stops shared by several lines;
- intermodal arcs at each stop to connect the pedestrian network with the line network.

In the frequency-based approach, to represent the topology of the transit network, several layers of nodes are then introduced, among which we can distinguish:

- the *base nodes* $N^{base} = B$, coinciding with the vertices B, including
- the *origin nodes* $O = \{B_z^{orig} : \forall z \in Z\} \subseteq N^{base}$ (each zone $z \in Z$ is associated with an *origin vertex*, denoted $B_z^{orig} \in B$), and
- the *destination nodes* $D = \{B_z^{dest} : \forall z \in Z\} \subseteq N^{base}$ (each zone $z \in Z$ is associated with a *destination vertex*, denoted $B_z^{dest} \in B$);
- the *stop nodes* $N^{stop} = S$, coinciding with the stops S (each stop $s \in S$ is associated with a *stop vertex*, denoted $B_s^{stop} \in B$);
- the *line nodes* N_ℓ, with one layer for each line $\ell \in L$.

A further specialization of line nodes is required by different models to represent specific phenomena. The key feature of frequency-based models is the representation of waiting as a separate trip phase. To this aim, when building-up the graph supporting the transit assignment model, the stop must be exploded into a set of arcs and nodes. There are several ways to do so; the scheme depicted in Fig. 6.6 allows to track most passenger flows and to reproduce (later on) the relevant congestion phenomena. Two nodes for each stop of line $\ell \in L$ are then introduced, so as to represent consistently dwelling and running:

- the *arrival node* $N_{\ell s}^{arr} \in N_\ell, \forall s \in S_\ell - S_\ell^-$;
- the *departure node* $N_{\ell s}^{dep} \in N_\ell, \forall s \in S_\ell - S_\ell^+$.

A typical way of building-up the transport network is to introduce the following types of arcs:

- the *pedestrian arcs* $A^{walk} = E^{walk}$;
- the *stop arcs* $A^{stop} = \left\{\left(B_s^{stop}, s\right) : \forall s \in S\right\} \cup \left\{\left(s, B_s^{stop}\right) : \forall s \in S\right\}$;
- the *running arcs* $A^{run} = \left\{\left(N_{\ell s}^{dep}, N_{\ell s[+\ell]}^{arr}\right) : \forall S \in S_\ell - S_\ell^+, \forall \ell \in L\right\}$;
- the *dwelling arcs* $A^{dwell} = \left\{\left(N_{\ell s}^{arr}, N_{\ell s}^{dep}\right) : \forall s \in S_\ell - S_\ell^- - S_\ell^+, \forall \ell \in L\right\}$;
- the *waiting arcs* $A^{wait} = \left\{\left(s, N_{\ell s}^{dep}\right) : \forall s \in S_\ell - S_\ell^+, \forall \ell \in L\right\}$;
- the *alighting arcs* $A^{alight} = \left\{\left(N_{\ell s}^{arr}, s\right) : \forall s \in S_\ell - S_\ell^-, \forall \ell \in L\right\}$.

It is useful to denote $L_a \in L$ the line associated with arc $a \in A$, if any. It is also useful to distinguish stops from base nodes, as this allows us to separate the line network (to which the stop node belongs) from the base network, which includes pedestrian and support arcs; the two may even derive from two separate data source.

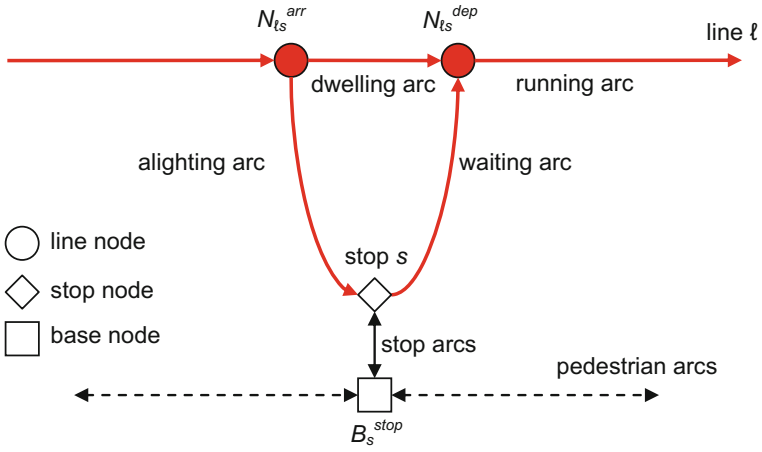

Fig. 6.6 Typical topology of the public transport network for frequency-based models. Arcs and nodes of the sub-network for this specific line are depicted in *red*

Two dummy stop (inter-modal) arcs are, in this case, introduced to connect each stop $s \in S$ to the associated base node $B_s^{stop} \in B$ (e.g., the closest one), from the latter to the former and vice versa. Note that the arcs describing the walking paths internal to a station are pedestrian arcs, and not stop arcs.

In some model, the stops S are directly a subset of vertices B, i.e., $B_s^{stop} = s$, in which case it is: $N^{stop} \subseteq N^{base}$ and no stop arc is required.

In some model, the pedestrian network is not explicitly introduced. The stops are directly part of the base nodes, while two stops are possibly connected by a *connection arc* that represents the shortest path on the hidden/implicit pedestrian network between the two stops.

Each zone centroid is connected to one or several base nodes (and vice versa) via particular pedestrian arcs that are called *connectors*, which represent the average access (egress) time/cost from a location in the zone to that node (or stop). But connectors should not be used to cross a zone. This rule can be enforced on the network by splitting the centroid of each zone $z \in Z$ into two different nodes, the *origin vertex* $B_z^{orig} \in B$ and the *destination vertex* $B_z^{dest} \in B$.

Support arcs represent the transport infrastructures used by line vehicles (e.g., road, reserved lanes, rail and tram tracks) and are not used by passengers directly. They are introduced for aggregating inputs and outputs, for plotting on maps the itineraries of the lines, and possibly to reproduce the mixed traffic congestion deriving from the concomitant use of roads between public and private transport means.

The proposed configuration has one major limitation: it does not allow us to identify the flow of passengers transferring from one line to another line within the same stop, based solely on the arc volumes. This can be obviated by introducing, as in Fig. 6.7, the following additional arc type:

- the *transfer arcs* $A^{tran} = \{(N_{\ell s}^{arr}, N_{\ell' s}^{dep}) : \forall s \in S_\ell, \forall \ell \in L, \forall \ell' \in L - \ell : s \in S_{\ell'}\}$.

Transfer arcs have the same time/cost of the corresponding arcs for waiting the same line at that stop. A small cost is associated with the stop arcs, so that direct transfers are convenient when alighting and boarding at the same stop. More articulated topologies of stops can be defined to consolidate alighting and boarding flows including transfers on one single arc, which may be useful (but not necessary) to reproduce congestion phenomena.

In the following, we will refer to the more classical set-up of Fig. 6.6.

6.2.3 Arcs Travel Times and Costs

Once the topology of the transport network is defined, the relevant performance attributes for each arc are to be specified, with particular reference to the factors yielding the generalized costs of Eq. (6.2) that are travel time, value of time and non-temporal cost. The time associated with the different types of trip phases is typically perceived differently by passengers of a given user class $g \in G$; it is then

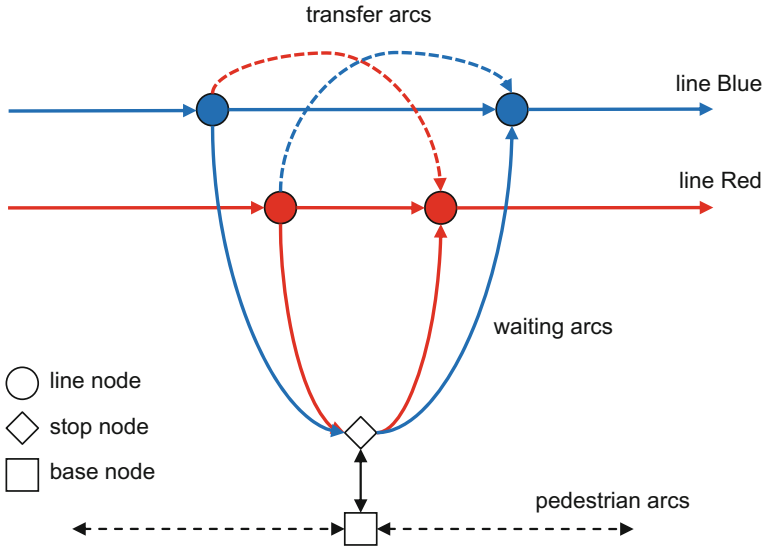

Fig. 6.7 Alternative topology with transfer arcs. Arcs and nodes of the sub-network for each one of the two lines are depicted in *red* and *blue*, respectively

transformed into costs multiplying a *base value of time* γ_g^{vot} by different weights; for example, walking and waiting are usually perceived as considerably more costly than riding.

For pedestrian arcs, the walking time, obtained as the ratio between the length of the arc l_a and the walking speed s_a^{walk} (introduced in Sect. 5.1.2), is multiplied by:

- the base value of time γ_g^{vot};
- a *walking discomfort coefficient* γ_g^{walk} which differs for each user class $g \in G$ (for example, elderly people suffer a higher discomfort for walking with respect to young people).

Monetary costs are assumed null. Thus, we have the following:

$$t_a = \frac{l_a}{s_a^{walk}}, \quad \gamma_{ag} = \gamma_g^{vot} \cdot \gamma_g^{walk}, \quad c_{ag}^{nt} = 0, \quad \forall a \in A^{walk}. \tag{6.67a}$$

Pedestrian arcs is one of the few noticeable cases in transit assignment where it would make some sense distinguishing the travel time per user class; but this would complicate the data structure unworthily.

Stop arcs are dummy; therefore, we assume a null cost and time:

$$t_a = 0, \quad \gamma_{ag} = \gamma_g^{vot}, \quad c_{ag}^{nt} = 0, \quad \forall a \in A^{stop}. \tag{6.67b}$$

For alighting arcs, the alighting time t_{ℓ}^{alight} (introduced in Sect. 5.1.2) is multiplied by the base value of time; monetary costs are assumed null, but a positive non-temporal cost associated with transfers is usually introduced:

$$t_a = t_{\ell}^{alight}, \quad \gamma_{ag} = \gamma_g^{vot}, \quad c_{ag}^{nt} = c_g^{tran}, \quad \forall a = \left(N_{\ell s}^{arr}, s\right) \in A^{alight}. \tag{6.67c}$$

The *transfer cost* for each user class $g \in G$ may represents a bundle of disutility components related to alighting and transferring, not necessarily connected with a measurable delay:

- the psychological stress of alighting (e.g., being aware of the current station);
- the psychological stress of possibly changing line;
- the additional travel time variance induced by the transfer.

For running arcs, the travel time $t_{\ell s}^{tran}$ of the line segment (introduced in Sect. 5.1.2) is multiplied by:

- the base value of time γ_g^{vot};
- the line discomfort coefficient γ_g^{vot} (introduced in Sect. 5.1.2);
- the *crowding discomfort coefficient* $\gamma_{\ell g}^{crowd}$ of the line segment $s \in S$ of line $\ell \in L$ for user class $g \in G$ that possibly depends on (separable) congestion through the (same) arc volume (as detailed in Sect. 7.2.1).

Monetary costs of the line segment are given by the kilometric fee $c_{\ell s}^{kfee}$ (introduced in Sect. 5.1.2) that is multiplied by a possible *fee multiplier* γ_g^{mfee} for user class $g \in G$. Thus, we have the following:

$$\begin{aligned} t_a &= t_{\ell s}^{run}, \quad \gamma_{ag} = \gamma_g^{vot} \cdot \gamma_{\ell g}^{line} \cdot \gamma_{\ell s g}^{crowd}(q_a), \\ c_{ag}^{nt} &= c_{\ell s}^{kfee} \cdot l_{\ell s} \cdot \gamma_g^{mfee}, \forall a = \left(N_{\ell s}^{dep}, N_{\ell s+\ell}^{arr}\right) \in A^{run}. \end{aligned} \tag{6.67d}$$

For dwelling arcs, the travel time $t_{\ell s}^{dwell}$ (introduced in Sect. 5.1.2) that possibly depends on (non-separable) congestion through the volumes of the alighting and waiting arcs (as detailed in Sect. 7.4.4) is multiplied by the base value of time γ_g^{vot}. Monetary costs are null. Thus, we have the following:

$$t_a = t_{\ell s}^{dwell}(\mathbf{q}_A), \quad \gamma_{ag} = \gamma_g^{vot}, \quad c_{ag}^{nt} = 0, \quad \forall a = \left(N_{\ell s}^{arr}, N_{\ell s}^{dep}\right) \in A^{dwell}. \tag{6.67e}$$

The cost of waiting arcs and transfer arcs derives mainly from the service discontinuity in time. The expected wait time $t_{\ell s}^{wait}$ of line $\ell \in L$ at stop $s \in S$ depends on the headway distribution through (6.58) or (6.65) and possibly depends on (non-separable) congestion (effective frequency) through the volume of the next running arc (as detailed in Sect. 7.3.2). This time is multiplied by:

- the base value of time γ_g^{vot};
- the *waiting discomfort coefficient* γ_g^{wait} which differs for each user class $g \in G$;
- the stop discomfort coefficient γ_{sg}^{stop} (introduced in Sect. 5.1.2);
- the *crowding discomfort coefficient* γ_{sg}^{crowd} of user class $g \in G$ at stop $s \in S$ that possibly depends on (non-separable) congestion through the volume of the waiting arcs (as detailed in Sect. 7.2.1).

The fixed fare of the line is applied here as a boarding fee $c_{\ell s}^{kfee}$ (introduced in Sect. 5.1.2). Thus, we have the following:

$$\begin{aligned} t_a &= t_{\ell s}^{wait}(\mathbf{q}_A), & \gamma_{ag} &= \gamma_g^{vot} \cdot \gamma_g^{wait} \cdot \gamma_{sg}^{stop} \cdot \gamma_{sg}^{crowd}(\mathbf{q}_A), \\ c_{ag}^{nt} &= c_{\ell s}^{bfee} \cdot \gamma_g^{mfee}, & \forall a &= \left(s, N_{\ell s}^{dep}\right) \in A^{wait} \\ & & \forall a &= \left(N_{\ell' s}^{arr}, N_{\ell s}^{dep}\right) \in A^{trans}. \end{aligned} \tag{6.67f}$$

The Eqs. (6.67)–(6.67f) allow to compute Eq. (6.2) and to obtain arc costs for the whole transit network.

As detailed in the following chapters, cost functions shall be associated with specific arc types to reproduce the effect of congestion on: crowding coefficients (Sect. 7.2.1), dwell times (Sect. 7.4.4) and wait times (Sect. 7.3).

The arc performance model proposed in this section includes many coefficients expressing the attitudes and preferences of the different user classes. The most effective way of determining their values is to calibrate a random utility model for route choice, based on an ad hoc survey with interviews to passengers of the study area including both revealed and stated preference questions (see Sect. 4.4.6).

6.2.4 *Waiting Costs in the Case of Known Timetable and Regular Service*

In case of lines with low frequency, the waiting cost resulting from Eqs. (6.67f) and (6.2) may be overestimated. Indeed, if the service is regular (i.e., $\sigma_{\ell s} = 0$), we can assume that passengers have the possibility of knowing the timetable; then, they will adopt a schedule-based behaviour (see Sect. 6.1.1). At least for the first line used in their journey, passengers can stay at home (or at the office, in the case of a return trip), where (see Sect. 6.3.7) the value (disutility) of time $(\gamma_g^{del} \cdot \gamma_g^{vot})$ is lower than that of waiting at the stop [γ_{ag} in Eq. (6.67f)], as some more useful activity can be done there, until it is time to walk towards the stop (the user delays his/her desired departure time). In general, we can then assume that the passenger faces the following alternative:

- stay at home until possible for a time equal on average to half of the headway minus the boarding time t_ℓ^{board} (introduced in Sect. 5.1.2 as a safety margin), which implies a disutility related to the delay with respect to the desired departure time (see Sect. 6.3.7), and then wait at the stop only for time t_ℓ^{board};
- go directly at the stop and wait on average for half of the headway.

Of course, a rational passenger will choose the most convenient in terms of generalized costs of the above two options; then, the wait time at the stop and the additional cost due to departure delay are, respectively, as follows:

$$t_a = Min\left(t_\ell^{board},\ \frac{0.5}{f_{\ell s}}\right), \qquad \begin{array}{c} \forall a = \left(s, N_{\ell s}^{dep}\right) \in A^{wait} \\ \forall a = \left(N_{\ell' s}^{arr}, N_{\ell s}^{dep}\right) \in A^{trans}, \end{array} \tag{6.68}$$

$$c_{ag}^{nt} = Min\left(\left(\frac{0.5}{f_{\ell s}} - t_\ell^{board}\right) \cdot \gamma_g^{del} \cdot \gamma_g^{vot},\ 0\right) + c_{\ell s}^{bfee} \cdot \gamma_g^{mfee} \qquad \begin{array}{c} \forall a = \left(s, N_{\ell s}^{dep}\right) \in A^{wait} \\ \forall a = \left(N_{\ell' s}^{arr}, N_{\ell s}^{dep}\right) \in A^{trans}. \end{array} \tag{6.69}$$

This model is typically applied to given lines on the whole network, without distinguishing between the first stop and additional transfers that are relative to a specific passenger trip. Thus, it can lead to some cost underestimation for transferring to regular lines with low frequency.

6.2.5 Route Choice and Uncongested Assignment

Any of the static methods for uncongested assignment presented in Sect. 6.1 can be used to analyse the transit network with a frequency-based model. In particular, we can adopt the path-based model of Eq. (6.31) or the arc-based model of Eq. (6.32), where arc performances given by Eq. (6.2) are specified by Eq. (6.67); route choice can be stochastic or deterministic, or a mixture for the different user classes.

In the following, an all-or-nothing assignment to shortest paths is illustrated for the example of Sect. 5.1.3. With respect to the network construction described in Sect. 6.2.2, the assignment graphs depicted in Fig. 6.8 simplify dwelling arcs and stop arcs. For the sake of simplicity, the only disutilities considered are the running times and the wait times (also depicted in the Figure). The latter are equal to the expected value of the headways assuming their exponential distribution. Demand flows to the destination stop 4 are reported below the origin stops. The colour of arcs is red for line 1, green for line 2, maroon for line 3 and black for line 4. The grey arc is a pedestrian connection.

The numerical computation presented in Table 6.1 results from the Dijkstra algorithm described in Sect. 6.1.6. The figures in brackets denote the Bellman

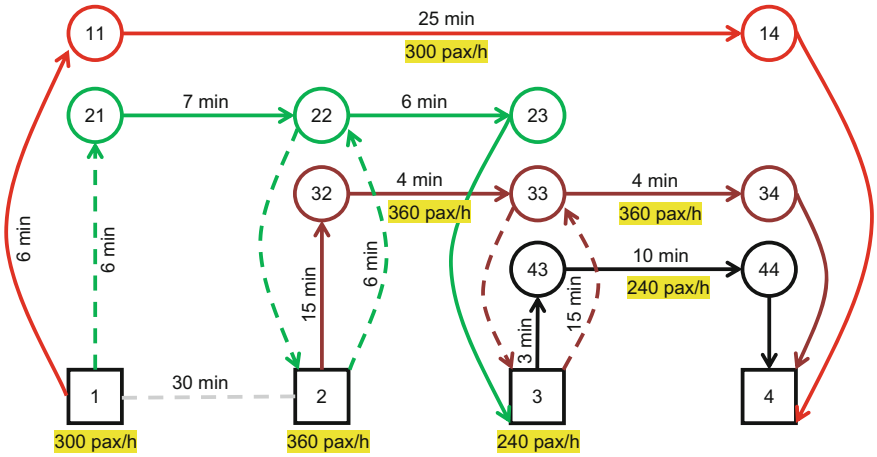

Fig. 6.8 Input data and results of an AoN assignment to shortest paths are applied to the example network

Table 6.1 Shortest tree computation for destination node 4 following the Dijkstra algorithm

Node	Expected cost (min)	Successor	Insertion order	Extraction order
1	(53 = 30 + 23) 31 = 6 + 25 (32 = 6 + 26)	(2) 11 (21)	14	14
2	23 = 15 + 8 (25 = 6 + 19) (61 = 30 + 31)	32 (22)	10	11
3	(19 = 15 + 4) 13 = 3 + 10	(33) 43	8	8
4	0		1	1
11	25 = 25 + 0	14	5	12
14	0 = 0 + 0	4	2	2
21	26 = 7 + 19	22	13	13
22	19 = 6 + 13 (23 = 0 + 23)	23 (2)	12	10
23	13 = 0 + 13	3	11	9
32	8 = 4 + 4	33 22	9	6
33	4 = 4 + 0 (13 = 0 + 13)	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

updates of node costs and successors which are not convenient and/or are later replaced by a better solution.

The shortest tree is identified recursively by following the successor nodes. The results of Table 6.1 show that all passengers take a direct route with no transfer. In particular, the shortest path from 1 to 4 is to use the red line; the shortest path from 2 to 4 is to use the maroon line, and the shortest path from 3 to 4 is to use the black line. The dashed arcs in Fig. 6.8 are not included in the shortest tree. The arc flows can be easily determined by propagating the demand flows along these paths; all flows are depicted in yellow in Fig. 6.8, where for simplicity only the running arcs are valorised.

6.2.6 Criticism of the Non-strategic Approach

The frequency-based model presented in this section ignores the possibility of combining transit lines that serve the same stop. At node 1, passengers have the choice between the red and the green line; at node 2 between the green line and the maroon line; and at node 3 between the maroon and the black line. But, a rational transit passenger could choose to board either lines, since he can actually reach the destination with both alternatives at a similar cost (see Table 6.1), while the wait time at the stop would decrease considerably since the resulting service frequency would combine that of two lines.

This leads to the notion of *attractive lines* at a stop, which is the set of transit lines at a given stop that a passenger may willing to board. These may be *common lines* that operate on the same corridor (a sequence of streets and stops) or transit lines that operate on different routes, but provide service with transfers to the same destination, which form a whole *strategy* formalized by a hyperpath (see Sect. 6.1.3).

More in general, the formulation of frequency-based model for choice route on transit networks depends on the assumptions made on the information that is available to passenger during the trip. If no information is available then the best choice would be the shortest (costliest) path.

The additional information that may become available during the trip is given as follows:

- departure of vehicles from the stop (visually obtained),
- some knowledge of transit timetables,
- estimated arrival time of vehicles at stops (also from remote via apps or vms),
- elapsed wait time at a stop,
- information on other transit lines by looking out the window once on board,
- vehicle occupancies.

On this basis, models of increasing complexity are formulated and solved in Sects. 6.3 and 7.1.

6.2.7 Reference Notes and Concluding Remarks

6.2.7.1 The Impact of ITS in Frequency-Based Models

The impact of ITS in frequency-based models emerges mainly through two indirect effects on variables:

- the reduction of headway variation coefficient that can be obtained by implementing a fleet control policy (e.g., holding vehicles at stops); this can lead up to halving the wait time (say from exponential headways with $\sigma_{\ell s} = 1$ to deterministic headways $\sigma_{\ell s} = 0$) obtaining the same effect of doubling the service frequency, especially for stops towards the end of the line;
- the reduction of waiting discomfort coefficient that can be obtained through better information to passengers at stops (e.g., displaying vehicle arrival times); also in combination with other measures (more comfortable stops), this can lead up to halving the cost of waiting, from say $\gamma_g^{wait} = 2$ to $\gamma_g^{wait} = 1$.

These interventions can produce a relevant impact on the quality of public transport, acting specifically on its peculiar cost components due to service discontinuity (wait times) and thus greatly helping to bridge the performance gap with private transport.

6.2.7.2 The Evolution of Frequency-Based Models in the Literature

The development of transit assignment models can be traced to a contribution by Dial (1967), who proposed a variant on the shortest path algorithm, originally used for routing private vehicles on road networks that takes into account the wait time of passengers at stops, which is the main phenomenon characterizing public transport networks. The wait times at the transfer stops were computed as half of the inverse of the combined frequency of all the lines serving the stop, thus already embedding an embryonal idea of the common line dilemma (Chriqui and Robillard 1975). This concept was later developed into an efficient algorithm for large networks by De Cea and Fernandez (1989).

Other early contributions to the transit route choice literature are those of Fearnside and Draper (1971), Le Clercq (1972), Andreasson (1976) and Last and Leak (1976). Most of these initial methods employed heuristic approaches, where a behavioural assumption leads directly to an algorithm without stating a formal model that would be solved by the computational procedure. These were largely inspired from assignment algorithms used on car networks, such as the all-or-nothing assignment to shortest paths and the stochastic (logit) multipath assignment, modified to reflect the wait times at stops which are inherent to transit networks.

Since the 1980s, a significant body of research (references in Sect. 7.1.8) was contributed to the study of transit route choice models where passengers are

assumed to know the frequency of the offered services but not the exact timetable. This assumption is reasonable for transit services in urban areas that operate with high line frequency; passengers arrive at a stop, either to start the trip or by transferring from another lines, and their wait time is related to the distribution of time intervals between successive vehicle arrivals, which is commonly referred to as the line headway. The typical hypothesis is that headways are independent with exponential (random) distribution (minimum regularity), or uniform (deterministic) distribution (maximum regularity), while the arrival of passenger at stops has a uniform distribution.

Another important stream of research is the theoretical analysis of headway distributions and their calibration with respect to real data. The proof of Eq. (6.58) is due to the seminal contribution of Osuna and Newell (1972). Further contributions were provided in the late 1970s and early 1980s (Jolliffe and Hutchinson 1975; Larson and Odoni 1981; Bowman and Turnquist 1981). A more general proof is given in Amin-Naseri and Baradaran (2014), who take also into account the correlation among subsequent arrivals. The formal derivation of wait time from the headway variation coefficient for Erlang distributions is an original contribution of this book.

6.3 Scheduled-Based Assignment on Transit Space-Time Networks

Guido Gentile, Younes Hamdouch and Markus Friedrich

In this section, schedule-based (or timetable based) models for transit assignment are presented in their basic version, without involving strategic behaviour and/or congestion phenomena.

The representation of supply shall take explicitly into account the fact that the public transport service is organized with runs for each transit line and is thus actually available not only at discrete places (the stops) but also at discrete times (the schedule). The main issue is then the representation of a discrete service, which can be accomplished through a suitable topological description of the transit network that incorporates timetables and other dynamic aspects of supply by introducing a diachronic graph.

There exist other approaches for the representation of schedule-based supply. One is to introduce a specific agent for the vehicle of each run in the context of a simulation model, as explained in Sect. 6.5.2. Another one can be achieved through a standard graph by macroscopic flow modelling with queuing; 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, respectively, as explained in Sect. 7.3.5. Considerations about such alternative frameworks will be provided in the referred sections, while this section is devoted to diachronic graphs.

In schedule-based assignment, the typical assumption is that passengers do consider the service timetable in their route choice, because this is available and reliable; they therefore select a path on the diachronic graph, which by construction embeds the departure time choice.

However, as already explained in Sect. 6.1.1, the schedule-based approach can be also confined to the description of the dynamic network loading, while a frequency-based perception of services, possibly including strategies, is considered for route choice. This can be simply achieved through a proper definition of arc costs on the diachronic graph, as shown in Sect. 6.3.3 (the cost of waiting arcs is null, while the cost of boarding will include the expected waiting).

Hyperpaths can be, explicitly or implicitly, defined on the diachronic graph to represent mingling queues of passengers at stops, as shown in Sect. 7.3.3 (the fail-to-board probability is associated with a hyperarc and a sequential route choice is considered, while iteration is required to reach equilibrium), or their strategic behaviour with respect to line arrivals at stop, as shown in Sect. 7.1.7 (the attractive set is built-up at stops in reverse chronological order and transmitted backward in time through waiting arcs on the diachronic graph).

6.3.1 *The Diachronic Graph*

The key feature of schedule-based transit services, from a modelling point of view, is that they can be easily represented through a *space-time network*, also called *diachronic graph*, where each single run has its own layer of topology.

In general, in a space–time network, each node has a specific time coordinate, beside space geo-coordinates. For the sake of simplicity, in the graphical representation, the *x*–*y* space is often reduced to one dimension, as depicted in Fig. 6.9;

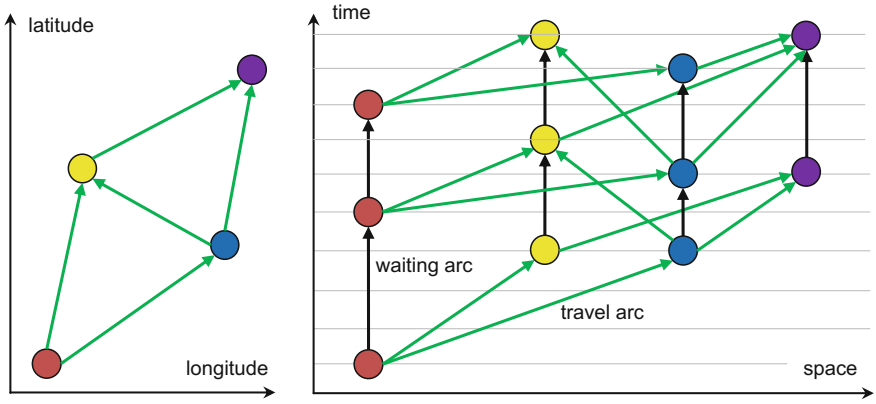

Fig. 6.9 Generic space-time network, or diachronic graph, and the corresponding base network. Waiting arcs are depicted in *black*, travel arcs are depicted in *green*

only in case of one transit line, it is easy to keep the metric of space consistent with the progressives of stops. Note that the same edge of the base network can have different travel times, for different entry times; but the FIFO rule is typically satisfied at the link (and path) level.

In particular, each node is defined here as a ordered couple of a vertex $i \in B$, which identifies its point in space, and a time index $t \in T \cup \eta + 1$, which identifies its instant in time (see Sect. 5.1.1), thus adopting a discrete representation of both dimensions (space and time). Recall that the additional instant $\eta + 1$, with $\tau_{\eta+1} = \infty$ is introduced here to represent events occurring after the assignment period $[\tau_0, \tau_\eta]$ or events not referred to a specific time.

To guarantee the consistency of the time-space network, we introduce the following index functions, which are used to shift any given instant $\tau \geq \tau_0$ to an instant of the predefined time discretization:

- $t^+(\tau)$ identifies the (next) time index $t \in T \cup \eta + 1$ such that $\tau_{t-1} < \tau \leq \tau_t$; $t^+(\tau_0) = 0$;
- $t^-(\tau)$ identifies the (previous) time index $t \in T$ such that $\tau_t \leq \tau < \tau_{t+1}$.

In the following, the network topology of the diachronic graph for transit assignment with schedule-based services is derived starting from the input data (see Sect. 5.1).

Like in the frequency-based approach, each trip phase (refer to the list in Sect. 6.2.2) is (possibly) represented by a sequence of arcs with a same type on the *transit network*; the latter is composed by:

- the *pedestrian network*, including centroids and connectors, as well as access, egress, walking and transfer links;
- the *line network*, with a sub-network for each transit run (and not for each line, like it is in frequency-based models), plus the stops and the waiting arcs shared by different lines;
- intermodal arcs at each stop to connect the pedestrian network with the line network.

To represent the topology of public transport services through a space-time network, several layers of nodes are introduced, among which we can distinguish:

- the *base nodes* $N^{base} = \{(i, t): \forall i \in B, \forall t \in T \cup \eta + 1\}$, including
- the *origin nodes* $O = \left\{\left(B_z^{orig}, t\right) : \forall z \in Z, \forall t \in T\right\} \subseteq N^{base}$, and
- the *destination nodes* $D = \left\{\left(B_z^{dest}, \eta + 1\right) : \forall z \in Z\right\} \subseteq N^{base}$, without a specific time coordinate;
- the *stop nodes* $N^{stop} = \{(s, t) : \forall s \in S, \forall t \in T \cup \eta + 1\}$;
- the *run nodes* N_r, with one layer for each run $r \in R_\ell$ of line $\ell \in L$.

Figure 6.10 shows a typical structure of the diachronic graph with the different node layers.

A further specialization of run nodes is required by different models to represent specific phenomena. Like in frequency-based models, there are several ways to

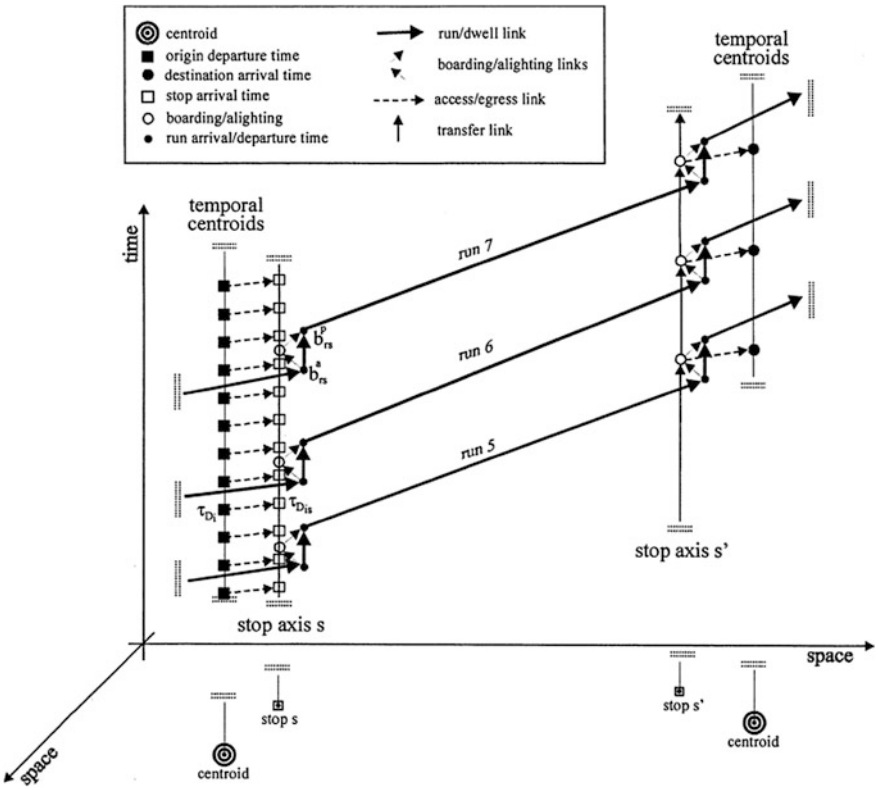

Fig. 6.10 Typical topology of the diachronic graph for schedule-based models

explode stops; the scheme depicted in Fig. 6.11 allows to track most passenger flows and to reproduce (later on) the relevant congestion phenomena. Two nodes for each stop of run $r \in R_\ell$ are introduced, so as to represent consistently dwelling (the idle vehicle at one stop) and running (the moving vehicle between two consecutive stops):

- the *arrival node* $N_{rs}^{arr} \in N_r, \forall s \in S_\ell - S_\ell^-$, with time coordinate $t^+(\tau_{rs})$ for a given scheduled time τ_{rs};
- the *departure node* $N_{rs}^{dep} \in N_r, \forall s \in S_\ell - S_\ell^+$, with time coordinate $t^-(\theta_{rs})$ for a given scheduled time θ_{rs}.

A typical way of building-up the diachronic graph is to introduce the following types of arcs:

- the *pedestrian arcs* $A^{walk} = \left\{\left((a^-, t), \left(a^+, t^+(\tau_t + l_a/s_a^{walk})\right)\right) : \forall a \in E^{walk}, \forall t \in T\right\}$;
- the *destination arcs* $A^{dest} = \left\{\left(\left(B_z^{dest}, t\right), \left(B_z^{dest}, \eta + 1\right)\right) : \forall z \in Z, \forall t \in T\right\}$;

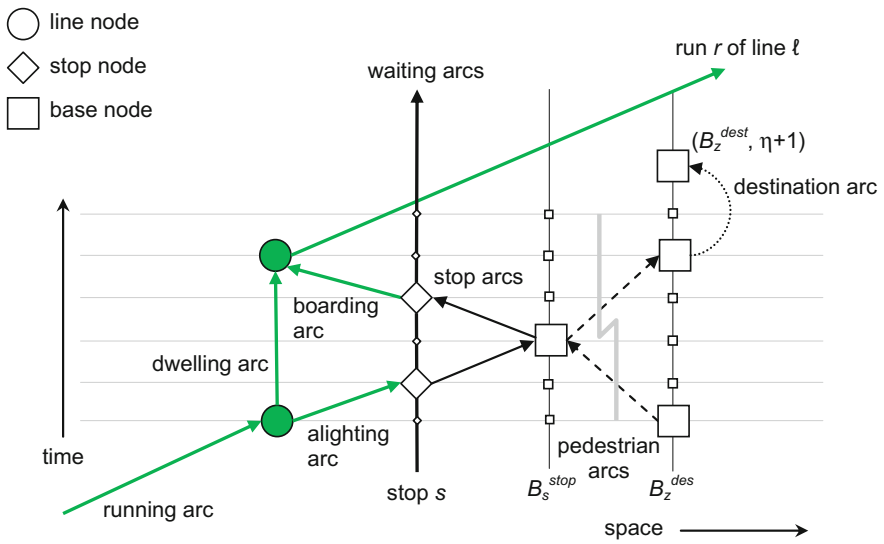

Fig. 6.11 Detail of the stop topology and of the connection with the pedestrian network

- the *running arcs* $A^{run} = \left\{ \left(N_{rs}^{dep}, N_{rs[+\ell]}^{arr} \right) : \forall s \in S_\ell - S_\ell^+, \forall r \in R_\ell, \forall \ell \in L \right\}$;
- the *stop arcs* $A^{stop} = \{ ((B_s^{stop}, t), (s, t+1)) : \forall s \in S, \forall t \in T \} \cup \{ ((s,t), (B_s^{stop}, t+1)) : \forall s \in S, \forall t \in T \}$
- the *waiting arcs* $A^{wait} = \{ ((s,t), (s,t+1)) : \forall s \in S, \forall t \in T$;
- the *dwelling arcs* $A^{dwell} = \{ (N_{rs}^{arr}, N_{rs}^{dep}) : \forall s \in S_\ell - S_\ell^- - S_\ell^+, \forall r \in R_\ell, \forall \ell \in L \}$;
- the *boarding arcs* $A^{board} = \{ ((s, t^-(\theta_{rs} - t_\ell^{\text{board}})), N_{rs}^{\text{dep}}) : \forall s \in S_\ell - S_\ell^+, \forall r \in R_\ell, \forall \ell \in L \}$;
- the *alighting arcs* $A^{alight} = \{ (N_{rs}^{\text{arr}}, (s, t^+(\tau_{rs} + t_\ell^{alight}))) : \forall s \in S_\ell - S_\ell^-, \forall r \in R_\ell, \forall \ell \in L \}$.

In this configuration, the intermodal arcs are only the stop arcs, while the boarding and alighting arcs are part of the line network. Note that running arcs connect two consecutive stops; they do not represent a *trip leg* that in some other models is introduced to jointly identify a sequence of stops between possible boarding and alighting.

Base and stop nodes are replicated for each instant of the predefined time discretization, while any arc shall connect two existing nodes by construction. For this purpose, the index function has been introduced, so that the instants which are used in the dynamic computation of route choice (concatenation) and flow propagation are fictitiously shifted to keep the time-space network consistent and acyclic. However, the travel time associated with each arc which is used in the computation of costs can be evaluated precisely. The typical time discretization in

schedule-based model for transit assignment has one-minute intervals; but shorter intervals shall be adopted to properly describe pedestrian networks with short edges, such as the connections inside a station (e.g., $l_a < 80$ m). Clearly, larger time intervals imply bigger approximations in the concatenation of costs and propagation of flows, while shortest time intervals imply more precise calculations at the price of higher computational costs due to the presence of many pedestrian and waiting arcs (ram and run-time are proportional to the number of arcs).

Acyclicity is a key feature of the diachronic graph topology, which turns very useful in the computation of shortest paths and flow propagation; in particular, the chronological order is in this case a topological order.

Arcs and nodes of the sub-network for this specific run are depicted in green; the bold arcs (including waiting at stops) make up the line network. Horizontal lines represent the instants of the predefined time discretization. The grey vertical line entails that a complex pedestrian network may be introduced to ensure the connection between origin, destinations and the base nodes of stops (i.e., access, egress and transfer).

6.3.2 Travel Costs in the Case of Run Choices

Once the space-time network is defined, the arcs can be characterized with exactly the same performance variables introduced for the frequency-based static model, since the representation of dynamics is here intrinsic in the topology of the diachronic graph. However, the travel times are to be here interpreted as a cost component used for route choice, while the speed of movements used in flow propagation and cost concatenation, as well as for the construction of paths/trajectories on the space-time network, are those connected with the temporal dimension of the diachronic graph.

If passengers make their route choice using the run schedule, the arc performance model developed in Sect. 6.2.3 for frequency-based services is still valid, with two noticeable differences:

- the travel time of boarding arcs includes only the constant value t_ℓ^{board} that is related to a safety margin compared to the departure, while the waiting phase is represented with a specific arc type;
- the running times and the dwell times are provided by the timetable, respectively, as $\tau_{rs+\ell} - \theta_{rs}$ and $\theta_{rs} - \tau_{rs}$.

The Eqs. (6.70a)–(6.70h) presented below allow us to compute (6.2) and obtain arc costs for the whole space-time network.

$$t_a = \frac{l_a}{s_a^{walk}}, \quad \gamma_{ag} = \gamma_g^{vot} \cdot \gamma_g^{walk}, \quad c_{ag}^{nt} = 0, \quad \forall a \in A^{walk}; \tag{6.70a}$$

$$t_a = 0, \quad \gamma_{ag} = \gamma_g^{vot}, \quad c_{ag}^{nt} = 0, \quad \forall a \in A^{stop}; \tag{6.70b}$$

$$\begin{aligned} t_a = t_\ell^{alight}, \quad \gamma_{ag} = \gamma_g^{vot}, \quad c_{ag}^{nt} = c_g^{tran}, \\ \forall a = \left(N_{rs}^{arr}, \left(s, t^+\left(\tau_{rs} + t_\ell^{alight}\right)\right)\right) \in A^{alight}; \end{aligned} \tag{6.70c}$$

$$\begin{aligned} t_a = \tau_{rs^+\ell} - \theta_{rs}, \quad \gamma_{ag} = \gamma_g^{vot} \cdot \gamma_{\ell g}^{line} \cdot \gamma_{rsg}^{crowd}(q_a), \\ c_{ag}^{nt} = c_{\ell s}^{kfee} \cdot l_{\ell s} \cdot \gamma_g^{mfee}, \quad \forall a = \left(N_{rs}^{dep}, N_{rs^+\ell}^{arr}\right) \in A^{run}; \end{aligned} \tag{6.70d}$$

$$t_a = \theta_{rs} - \tau_{rs}, \quad \gamma_{ag} = \gamma_g^{vot} \cdot \gamma_{\ell g}^{line}, \quad c_{ag}^{nt} = 0, \quad \forall a = \left(N_{rs}^{arr}, N_{rs}^{dep}\right) \in A^{dwell}; \tag{6.70e}$$

$$\begin{aligned} t_a = t_\ell^{board}, \quad \gamma_{ag} = \gamma_g^{vot} \cdot \gamma_g^{wait}, \quad c_{ag}^{nt} = c_{\ell s}^{bfee} \cdot \gamma_g^{mfee}, \\ \forall a = \left(\left(s, t^-\left(\theta_{rs} - t_\ell^{board}\right)\right), N_{rs}^{dep}\right) \in A^{board}. \end{aligned} \tag{6.70f}$$

The crowding discomfort coefficient $\gamma_{sgt}^{crowd} > 1$ of the segment $s \in S$ of run $r \in R_\ell$ of line $\ell \in L$ for user class $g \in G$ possibly depends on (separable) congestion through the number of passengers on-board given by the volume on the (same) arc (as detailed in Sect. 7.2.1).

For waiting arcs, the duration of the interval $\tau_{t+1} - \tau_t$ is multiplied by:

- the base value of time γ_g^{vot};
- the waiting discomfort coefficient γ_g^{wait};
- the stop discomfort coefficient γ_{sg}^{stop};
- the crowding discomfort coefficient γ_{sgt}^{crowd} of stop $s \in S$ for class $g \in G$ users entering the waiting arc at instant $t \in T$ that possibly depends on (separable) congestion through the (load) number of waiting passengers given by the volume on the (same) arc (as detailed in Sect. 7.2.1).

Monetary costs are assumed null. Thus, we have the following:

$$\begin{aligned} t_a = \tau_{t+1} - \tau_t, \quad \gamma_{ag} = \gamma_g^{vot} \cdot \gamma_g^{wait} \cdot \gamma_{sg}^{stop} \cdot \gamma_{sgt}^{crowd}(q_a), \quad c_{ag}^{nt} = 0, \\ \forall a = ((s,t),(s,t+1)) \in A^{wait}. \end{aligned} \tag{6.70g}$$

Note that given the possibility offered by longer waits, for scheduled services the stop discomfort coefficient γ_{sg}^{stop} (introduced in Sect. 5.1.2), besides ergonomy, depends also on the activities (e.g., shopping) that can be developed by the passenger at the specific stop.

Destination arcs are dummy; therefore, we assume a null cost and time:

$$t_a = 0, \quad \gamma_{ag} = \gamma_g^{vot}, \quad c_{ag}^{nt} = 0, \quad \forall a \in A^{dest}. \tag{6.70h}$$

6.3.3 Travel Costs in the Case of Line Choices

The cost model presented in the previous section is to be considered the reference one for schedule-based assignment. However, there are cases where the assignment of the diachronic graph is aimed to determine the load on each run, while passenger behaviour is still connected with the perception of service in terms of line frequencies.

In this case, the arc performance model developed in Sect. 6.2.3 for frequency-based services can be still applied, with few exceptions:

- the cost of boarding arcs shall include the expected waiting time and its disutility, but not the cost due to the crowding discomfort, as each coefficient γ_{sgt}^{crowd} is arc specific and depends on the load of waiting passengers at the stop entering at that specific instant;
- the cost of waiting arcs (that would in this case be null in principle) shall thus include only the crowding discomfort.

In this case, Eqs. (6.71a)–(6.71e) and (6.71h) are identical to Eqs. (6.70a)–(6.70e) and (6.70h), while for boarding and waiting arcs it is, respectively:

$$\begin{aligned} t_a &= t_\ell^{board} + t_{\ell st}^{wait}(\mathbf{q}_A), \quad \gamma_{ag} = \gamma_g^{vot} \cdot \gamma_g^{wait} \cdot \gamma_{sg}^{stop}, \\ c_{ag}^{nt} &= c_{\ell s}^{bfee} \cdot \gamma_g^{mfee}, \quad \begin{matrix} \forall a = \left((s,t), N_{rs}^{dep}\right) \in A^{board} \\ t = t^- \left(\theta_{rs} - t_\ell^{board}\right) \end{matrix} ; \end{aligned} \tag{6.71f}$$

$$\begin{aligned} &t_a = \tau_{t+1} - \tau_t, \quad \gamma_{ag} = \gamma_g^{vot} \cdot \gamma_g^{wait} \cdot \gamma_{sg}^{stop} \cdot \left(\gamma_{sgt}^{crowd}(q_a) - 1\right), \\ &c_{ag}^{nt} = 0, \quad \forall a = ((s,t),(s,t+1)) \in A^{wait}. \end{aligned} \tag{6.71g}$$

This way, the cost of waiting is arbitrarily separated in two discomfort components that are associated with two different arc types, i.e., boarding and waiting, respectively. This modelling choice is justified by several facts:

- the perceived wait time (6.71f) preventively estimated by passengers to make their route choice is linked with the headway distribution and primarily with the line frequency;
- the actual wait time (6.71g) suffered by passengers is that accumulated during the wait, under the assumption that the timetable embedded in the diachronic graph is a possible instance of what will actually occur in reality;
- by perceiving some cost also on the waiting arcs, passengers are induced boarding the first available run(s) of each line (under the typical assumption that $\gamma_{sgt}^{crowd} > 1$);
- the crowding discomfort coefficient at the stop depends on the number of waiting passengers, which may grow during the wait and is thus well represented by the load on the waiting arc.

Once again, be aware that the travel times evaluated by the above arc performance model are functional to the route choice model through perceived costs. The correct computation of an output indicator such as the total travel time is properly accomplished by taking into consideration the temporal coordinates embedded into the diachronic graph; indeed, by considering (6.71f) and (6.71g) wait times would be counted twice.

The expected wait time $t_{\ell st}^{wait}$ at stop $s \in S$ for line $\ell \in L$ at instant $t \in T$ depends on the headway distribution through (6.65), where the frequency $f_{\ell st}$ and the irregularity $\sigma_{\ell st}$ shall be evaluated through Eqs. (5.12) and (5.13), respectively, starting from the schedule. This can be done by considering fairly long-time intervals (say 1 h), which would be feasible with the memory ability of passengers in recalling temporal profiles of attributes. As an alternative, the frequency can be calculated as the inverse of the (departure) headway between the current run and the previous one. In case of queuing, the perceived wait time can also possibly depend on (non-separable) congestion (effective frequency) through the load of the next running arc (as detailed in Sect. 7.3.2).

As already mentioned, the crowding discomfort coefficient γ_{sgt}^{crowd} of stop $s \in S$ for user class $g \in G$ at instant $t \in T$ possibly depends on (separable) congestion through the number of waiting passengers given by the load on the corresponding arc (as detailed in Sect. 7.2.1).

The disutility connected with the possible difference between the desired departure time and the actual one is discussed in Sect. 6.3.7.

6.3.4 Route Choice and Uncongested Assignment

Any of the static methods for uncongested assignment presented in Sect. 6.1 can be used to analyse the transit network with a schedule-based model. In particular, we can adopt the path-based model given in Eq. (6.31) or the arc-based model given in Eq. (6.32), where arc performances in Eq. (6.2) are specified by Eq. (6.70) or by Eq. (6.71); route choice can be stochastic or deterministic, or a mixture for the different user classes.

Although there is no technical drawback in adopting an arc-based model, the traditional approach to the schedule-based assignment on the diachronic graph is path-based, because this allows to better represent some important attributes, such as fares and walking distance, that may be modelled as nonlinear components in the passengers disutility, as well as other behavioural aspects of route choice, such as the correlation among alternatives.

The different preferences of users on the many relevant attributes (e.g., walking time, wait time, on-board time, transfers, monetary cost, stop ergonomy and comfort) are not easy to synthesize in the deterministic coefficients of a given class segmentation. Hence, a random utility model can be used which incorporates passenger heterogeneity in a stochastic framework.

The correlation among route alternatives is a relevant aspect to take into account in stochastic assignment when the model is conceived to distinguish the service provided by each run of a same line. On the other hand, the number of non-dominated routes available on the space-time network and practical to consider for each O–D pair is usually limited.

In Sect. 4.5.2, some methods are presented to generate a 'good' set of paths, which is a critical step in this kind of assignment procedure; moreover, in the next section, the multi-path algorithm is presented as a route generation method specifically conceived for schedule-based models of long distance trips.

Under these considerations, path-based models allow for more opportunities than arc-based models:

- a proper selection of usable routes, which comes at the price of introducing rules for their identification;
- a proper representation of their correlation, which comes at the price of a more complex route choice model;
- the possibility of introducing non-additive costs, such as fares, which comes at the price of a more complex supply model.

A connector exiting from the origin node represents the access to the pedestrian network, which allows us to reach the first stop. The passenger waits at the platform a specific run for a certain number of time intervals and then boards the vehicle at the end of its dwell time, which instead occurs on the track (thinking of a railway example). The run departs from the stop (with the passenger on-board) at the scheduled time and arrives at the next stop again on schedule. The passenger travels along a sequence of line segments on such run until he/she alights upon vehicle arrival at the planned stop. The passenger may transfer to a new stop using the pedestrian network, or stay in the same stop, and the waiting–boarding–alighting sequence is repeated, until the last stop is reached. Finally, the passenger walks towards the destination and egresses there from the transit network via a connector; the trip on the diachronic graph actually ends with a dummy destination arc.

Each one of the above trip phases is represented through a sequence of arcs on the diachronic graph. Thus, the route costs c_{kg} can be obtained by summing up all the arc costs of the path, plus possibly a non-additive term, as in Eq. (6.3). Route probabilities p_{kg} can then be reproduced through any discrete choice model (e.g., based on random utility), as in Eq. (6.13).

In the schedule-based approach with space-time network, a trip starting at instant $t \in T$ from origin zone $z \in Z$ to destination zone $z' \in Z$ is represented as an (acyclic) path $k \in K_{od}$ from origin node $o = (B_z^{orig}, t) \in O$ to destination node $d = (B_{z'}^{dest}, \eta + 1) \in D$ on the diachronic graph; thus, the notion of departure time is embedded in the origin.

Travel demand d_{odg} represents here trips from origin node $o = (B_z^{orig}, t)$ to destination d, i.e., the number of passengers of class g who (wish to) depart from origin zone z during a time interval $(\tau_{t\text{-}1}, \tau_t]$ to reach d at a later time. It is assumed that all such passengers will behave like the one departing in the final instant of the

interval, who will consider the costs c_{kg} of the paths $k \in K_{od}$. The resulting path flows q_{kg} are consistent with the route probabilities p_{kg} as in (6.6). Finally, the arc flows and volumes are computed as in Eqs. (6.5) and (6.1).

Most of the existing models for schedule-based assignment adopt the above path-based approach, which can support also the simulation of real-time information about vehicle arrivals and the consequent en-route adaptation of the path choice. In that case, random utility models are though forced to represent also the events occurring at stop relative to the service departure time, in a context of imperfect regularity, while actually the two phenomena (random utility and random headways) follow in general quite different statistical laws. Moreover, service irregularity leads passengers to a strategic behaviour, which cannot be represented satisfactorily by stratifying the network knowledge, possibly acquired through a day-to-day learning process, in terms of path costs. Instead, strategies are well formalized through hyperpaths, which given their awkward explicit representation require de facto an implicit modelling of sequential arc choice towards the destination. For this reason, as already explained in Sect. 6.1.3, despite the advantages of path-based models in reproducing nonlinear attributes, arc-based models can better provide a suitable support for the future development of schedule-based algorithms, where:

- the assumption of perfectly reliable timetables is abandoned,
- the representation of supply variability becomes a key aspect of the simulation, and
- the diachronic graph reduces to a technical tool for the analysis of the loads resulting from the assignment on each run and is not anymore meant to reflect the mental map of the passenger.

6.3.5 *Branch and Bound Algorithm for Choice-Set Generation*

The multi-path algorithm (Friedrich et al. 2001) is here presented as a method for generating the set K_{od} of all potential routes from origin node $o = (B_z^{orig}, t) \in O$ to any destination node $d = (B_{z'}^{dest}, \eta + 1) \in D$ on the diachronic graph that are compliant with a set of given rules.

The construction algorithm builds-up iteratively a connection tree which may provide several paths from an origin (at a given time) to possibly every destination, as depicted in Fig. 6.12. The root of the tree is the origin node; a walk leg (i.e., a sequence of pedestrian arcs) is added to the tree to reach each stop of public transport in the vicinity (a proper distance threshold is to be defined) and the corresponding node (stop and arrival time) is added to a list of nodes to examine.

Then, for each one of the reached stops (contained in the list), a branch and bound approach is applied to visit and possibly add to the connection tree all transit legs (i.e., a sequence of line network arcs, such as board, run, dwell, run, dwell,

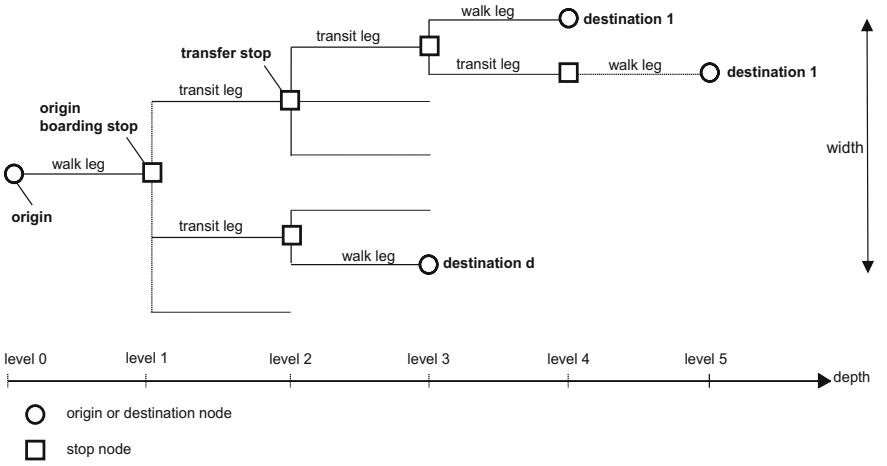

Fig. 6.12 Structure of the connection tree

alight) that bring from the current stop at the current time (current node) to another stop and satisfy a set of rules (to be specified in the following). When this happens, the final stop of the visited leg (at the arrival time) is added to the list of nodes to be further examined. Once all stops of one level are visited, the depth of the tree is increased to the next level. Moreover, additional walk legs are added to the tree if destinations are reachable in the vicinity.

The use of entire connection legs as tree edges simplifies and accelerates the search for new routes on the diachronic graph to a great extent; the combinatorial explosion of connections is primarily limited by the maximum number of transfers.

The construction of the connection tree includes the definition of a *search impedance* c_k^{imp} for each route k:

$$c_k^{imp} = \beta^{time} \cdot t_k^{tot} + \beta^{trans} \cdot n_k^{trans} + \beta^{fare} \cdot c_k^{fare}, \tag{6.72}$$

where:

- t_k^{tot} is the total travel time of the path (undistinguished for trip phase),
- n_k^{trans} is the number of transfers and
- c_k^{fare} is the fare, while
- β^{time}, β^{trans} and β^{fare} are global search parameters.

The functional form of this search impedance is thus similar but usually simpler than that of the systematic utility. Indeed, while the utility should reflect at best the perception of the travellers in the route choice, the impedance is used only to generate an appropriate choice set of paths. This can justify some somewhat different parameter values.

Now, we present the set of rules that can be applied in the branch and bound constructive search to determine if a given transit leg $(i, u) \rightarrow (j, v)$ from the stop of the currently examined node (i, t) to another stop j reachable at a later time with no transfer from i should be added to the connection or not:

- Rule 1. Temporal suitability. The run of the transit leg under consideration shall depart after the arrival time at stop i: $u > t$.
- Rule 2. Dominance. No other connection $k \in K_{oj}$ should already exist on the tree from origin o to stop j that dominates in all relevant aspects the one h created by adding the transit leg under consideration, i.e., such that $t_k^{tot} < t_h^{tot}$ and $n_k^{trans} < n_h^{trans}$ and $c_k^{fare} < c_h^{fare}$.
- Rule 3. Tolerance. The following constraints are satisfied: $c_h^{imp} \leq \alpha^{imp} \cdot Min(c_k^{imp} : k \in K_{oj}) + \chi^{imp}$ *and* $n_h^{trans} \leq \chi^{trans} n$, where $\alpha^{imp} > 1$ is the relative tolerance with respect to the least impedance path, χ^{imp} is an additive absolute tolerance with respect to the least impedance path, and χ^{trans} is the maximum number of transfers allowed within a connection.

When a connection is added, then its final stop is added to the list of nodes with the resulting arrival time. A tree level is explored completely before connection legs of the next level are considered. The procedure terminates when the list of nodes is empty or the maximum number of transfers (levels) is reached. Finally, for each destination d, one additional connection is added directly to the origin o, which contains a walking leg using the shortest path on the pedestrian network.

6.3.6 Computation of Shortest Tree on the Space-Time Network

A practical alternative to the preliminary explicit generation of all relevant paths is their iterative construction through the computation (and storage, if needed) of minimum-cost trees on the diachronic graph.

In space-time networks, each node has a specific time coordinate, with the exception of destination nodes, which have none. This way, with one shortest tree rooted at the destination node the minimum-cost path starting from every origin node (each one representing an origin centroid and a departure time) is obtained; let us see why this is convenient.

Usually, in a dynamic assignment problem, we are faced with the computation of the costless path to reach the destination centroid $B_d^{dest} \in B$ of zone $d \in Z$ (at any time) starting from the origin centroid $B_o^{orig} \in B$ of zone $o \in Z$ at a given instant $t \in T$, because the demand is specified and stratified for departure time. This can be achieved at once for all possible origin zones $o \in Z$ and departure times $t \in T$ with one single visit in reverse chronological order of the diachronic graph by initializing to zero the labels of the base nodes $(B_d^{dest} \in B, \forall e \in T)$ corresponding to the

destination centroid at all possible arrival times (these are not the destination node); the result of the algorithm would be therefore a forest and not a tree. However, the introduction of dummy destination arcs allows to initialize only the label of the destination node $(B_d^{dest}, \eta + 1)$ to zero.

Moreover, the computation of a shortest tree on the diachronic graph is trivial, since the graph is acyclic and has a natural topological order that is the chronological order (we can assume destination nodes have an infinite time). Under such conditions, a shortest tree can be easily computed by processing all nodes with Eq. (6.19), i.e., by applying the Bellman relation given in Eq. (6.29) to each arc of the forward star, in reverse chronological order starting from the destination, without the need of introducing a list of nodes to be visited. This approach is here referred to as the Pallottino algorithm (1998), who proposed and analysed several variants of this problem.

The example below presents the computation of the shortest tree to destination 4 by considering (Fig. 6.13) a simplified version of the diachronic graph topology with respect to that proposed in Sect. 6.3.1 applied to our test case of Sect. 5.1.3. In particular, for the sake of simplicity, the stop is here represented as a single node; thus, while the feasible connections among possible runs are correctly represented, the resulting arc loads are not capable of explaining boarding, alighting and transfer flows.

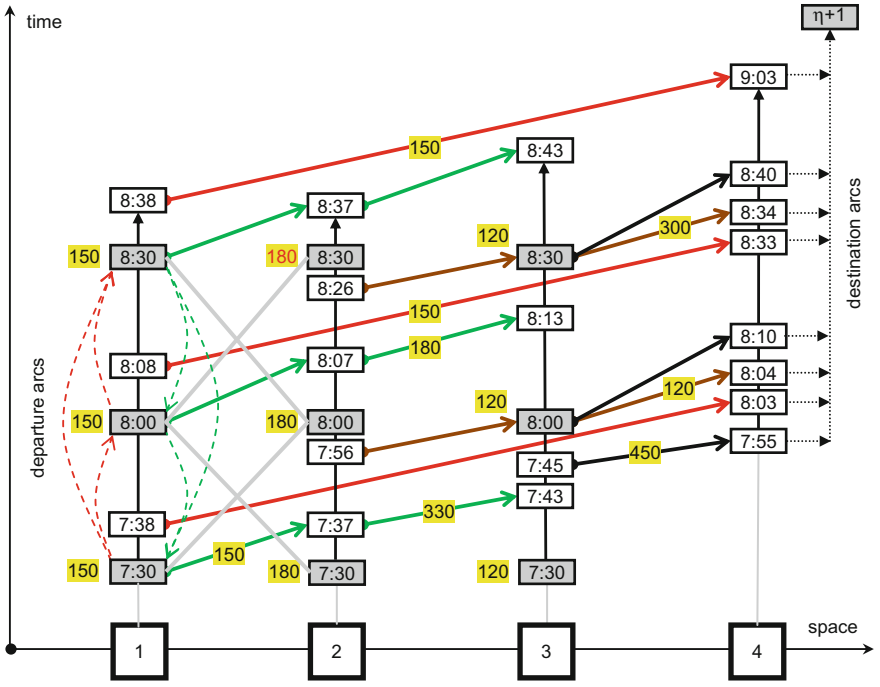

Fig. 6.13 Results of an AoN assignment to shortest paths on the diachronic graph applied to the example network

The colour of running arc is that associated with the line: red for line 1, green for line 2, maroon for line 3 and black for line 4. The grey time boxes represent the base nodes replicated for each instant of the time discretization, while the white time boxes represent departure and arrivals of runs. The grey lines between stops 1 and 2 represent the pedestrian arcs. The dashed arrows at stop 1 represent the departure options (red for delay and green for anticipation) as explained in Sect. 6.3.7.

In Table 6.2, each line shows the solution for a node of the diachronic graph. Nodes are visited in reverse chronological order from the destination and the best

Table 6.2 Shortest tree computation for destination 4 following the Pallottino algorithm

Stop	Time	Expected cost (min)	Successor stop	Successor time
4	$\eta + 1$	0		
4	9:03	0 = 0 + 0	4	$\eta + 1$
3	8:43	∞		
4	8:40	0 = Min(0 + 0, 23 + 0)	4	$\eta + 1$
1	8:38	25 = 25 + 0	4	9:03
2	8:37	∞ = 6 + ∞	3	8:43
4	8:34	0 = Min(0 + 0, 6 + 0)	4	$\eta + 1$
4	8:33	0 = Min(0 + 0, 1 + 0)	4	$\eta + 1$
1	8:30	33 = Min(7 + ∞, 8 + 25)	1	8:38
2	8:30	∞ = 7 + ∞	2	8:37
3	8:30	4 = Min(4 + 0, 10 + 0, 13 + ∞)	4	8:34
2	8:26	8 = Min(4 + 4, 11 + ∞)	3	8:30
3	8:13	21 = 17 + 4	3	8:30
4	8:10	0 = Min(0 + 0, 23 + 0)	4	$\eta + 1$
1	8:08	25 = Min(25 + 0, 22 + 33)	4	8:33
2	8:07	27 = Min(6 + 21, 19 + 8)	3	8:13
4	8:04	0 = Min(0 + 0, 6 + 0)	4	$\eta + 1$
4	8:03	0 = Min(0 + 0, 1 + 0)	4	$\eta + 1$
1	8:00	33 = Min(7 + 27, 8 + 25, 30 + ∞)	1	8:08
2	8:00	34 = Min(7 + 27, 30 + 33)	2	8:07
3	8:00	4 = Min(4 + 0, 10 + 0, 13 + 21)	4	8:04
2	7:56	8 = Min(4 + 4, 7 + 27)	3	8:00
4	7:55	0 = Min(0 + 0, 8 + 0)	4	$\eta + 1$
3	7:45	10 = Min(10 + 0, 15 + 4)	4	7:55
3	7:43	12 = 2 + 10	3	7:45
1	7:38	25 = Min(25 + 0, 22 + 33)	4	8:03
2	7:37	18 = Min(6 + 12, 19 + 8)	3	7:43
1	7:30	25 = Min(7 + 18, 8 + 25, 30 + 34)	2	7:37
2	7:30	25 = Min(7 + 18, 30 + 33)	2	7:37
3	7:30	25 = 13 + 12	3	7:43

local alternative of the forward star (such that the arc cost plus its head cost is minimum) is identified, thus providing expected cost and successor stop for the node under analyses.

At the end of the process, it is possible to reconstruct the shortest path starting from any origin node by following on the table the sequence of successor nodes. For example (see dark cells of the table), starting from stop 1 at 7:30, the sequence is: (board and ride the green line) stop 2 at 7:37, (ride the green line) stop 3 at 7:43, (alight from the green line and wait) stop 3 at 7:45, (board and ride the black line) stop 4 at 7:55, (reach the destination node) stop 4 at time $\eta + 1$.

The numbers at the left of the node and on the arcs in yellow depicted in Fig. 6.13 identify the passenger loads resulting from an AoN assignment to shortest paths on the diachronic graph; in red (stop 2 at 8:30) is depicted a demand load that is unable to reach the destination.

6.3.7 Departure Time Choice

The results obtained with the simulation of route choice only are not particularly satisfactory in the case of schedule-based models, because the departure time choice is not properly taken into account in a context where the availability of service is scarce in time.

For example, passengers who desire to depart at 8:30 from stop 2 do not have an available travel alternative if they must necessarily start their trip at 8:30; but, they may be willing to anticipate their trip, and this allows to have more travel alternatives. Passengers who desire to depart at 7:30 from stop 3 have a cost of 25 min by starting their trip exactly at 7:30; but they may be willing to postpone their trip at 8:00 when the cost to destination is only 4 min. Passenger who desire to depart at 8:00 from stop 1 have a cost of 33 min by starting their trip exactly at 8:00; but they may be willing to anticipate their trip at 7:30 when the cost to destination is only 25 min.

However, the shift from the desired departure time to the actual departure time conveys a cost (disutility) for anticipation or delay; therefore, the final trip decision will result from the combination of route opportunities available at different departure times and the above shift disutility.

In the framework of diachronic graphs, it is fairly easy to couple the route choice with the departure time choice. To this end, after the computation of the shortest tree to destination $d \in D$ and before performing the flow propagation, the demand d_{odg} of class $g \in G$ users directed to d that desire to depart form origin zone $z \in Z$ at instant $t \in T$ shall not be (necessarily) loaded on origin node $o = (B_z^{orig}, t) \in O$, but instead on the origin node $i = (B_z^{orig}, e) \in O$, with actual departure instant $e \in T$, which shows the best utility. The latter is given by the combination of the route (expected) cost w_{idg} from node i to destination d and of the cost (disutility) for anticipation $\tau_t - \tau_e$ (if $e \leq t$) or delay $\tau_e - \tau_t$ (if $e \geq t$) due to the shift of the actual

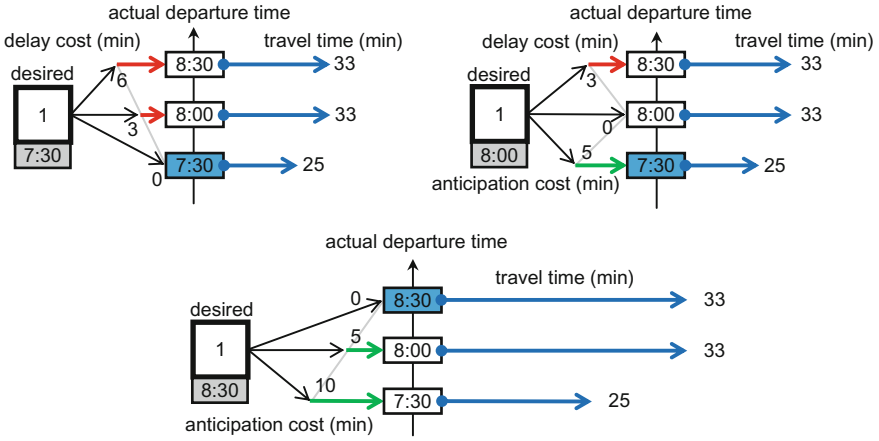

Fig. 6.14 Departure time choice for passengers leaving from origin stop 1 destination stop 4

departure time τ_e with respect to the desired departure time τ_t. Usually, the choice set of actual departure times that users of class $g \in G$ take into consideration is identified by considering a maximum anticipation t_g^{ant} and a maximum delay t_g^{del}.

Typically, a linear expression of the above disutilities for users of class $g \in G$ is assumed with different coefficients for anticipation t_g^{ant} and delay t_g^{del}; because the demand is specified for desired departure times, then usually $\gamma_g^{ant} > \gamma_g^{del}$; indeed, in this case users are doing some activity at the origin which ends at a given time. The total cost w_{odge} to reach destination d for users of class g that desire to depart at instant t from node $o = (B_z^{orig}, t)$ but instead depart at instant e from node $i = (B_z^{orig}, e)$ is then given by:

$$w_{odge} = w_{idg} + \begin{cases} \gamma_g^{ant} \cdot \gamma_g^{vot} \cdot (\tau_t - \tau_e), & \text{if } \tau_t - t_g^{ant} \le \tau_e \le \tau_t \\ \gamma_g^{del} \cdot \gamma_g^{vot} \cdot (\tau_e - \tau_t), & \text{if } \tau_t < \tau_e \le \tau_t + t_g^{del} \\ \infty, & \text{otherwise} \end{cases} . \tag{6.73}$$

On the contrary, if the demand is specified for desired arrival times, then the user is going to undertake an activity at the destination which will start at a given time. However, in this case, the shortest paths shall be computed from the origin, basically inverting the described approach.

A probabilistic model based on random utility (see Sect. 4.4) can also be used to split the demand on the alternative departure times, where the opposite of the above cost combination works as a systematic utility.

In an equilibrium assignment including departure time choice, some commuters will travel before the desired departure time and some after, so as to avoid the congestion of the peak period.

Below an example of departure time choice from origin 1 is presented, under the assumption of a disutility for delay equal to 6 min/h and for anticipation equal to

10 min/h. Passengers who desire to depart at 8:00 find more convenient to depart at 7:30 (Fig. 6.14).

In the case where the departure time choice can be considered as part of the route choice like if anticipation or departure delay are an additional phase of the trip, then the departure time choice can be simulated within the assignment model simply by extending the network into a *super-network* with departure arcs that connect the origin node (with its desired departure time) and any other feasible departure time. The cost associated with these arcs is the disutility of anticipation or delay which can take any form (e.g., linear or quadratic) as a function of the above-mentioned difference; interestingly in this particular case, the anticipation arcs will travel back in time.

6.3.8 Networks with Mixed Schedule-Based and Frequency-Based Services

This section shortly addresses the problem of modelling in transit assignment the case of networks where both schedule-based (SB) and frequency-based (FB) services are present.

When do passengers refer to timetables or not? This depends mainly on headways and on their regularity. Typical threshold for regular headways is around 10–20 min. But what if the network contains lines with both high and low headways? The structures of FB models with static network and SB models with space-time network are quite different and do not fit well together, so they cannot be combined simply in a same model. Alternative solutions are then:

- use the FB approach for the whole model and approximate the passenger behaviour for lines with low frequency, as proposed in Sect. 6.2.4;
- use the SB approach for the whole model and approximate the passenger behaviour for lines with high frequency, as proposed in Sect. 6.3.3.

One way is then to introduce in a static transit network a proper limit to the maximum wait time, so as to represent the convenience expected by passengers of timing their arrival at the stop with that of vehicles, and wait at home instead of at the stop, as in Eq. (6.69).

Another way of dealing with networks with mixed services is to apply the two cost functions (6.70) and (6.71) on the diachronic graph of the same model where appropriate. In this case, we shall assume that stops are dedicated either to SB or to FB services.

Finally, note that dynamic models (macroscopic assignment and simulation) can instead support natively the presence of both high and low-frequency services, as explained in Sects. 6.4 and 6.5.

6.3.9 *Reference Notes and Concluding Remarks*

Schedule-based model has been widely developed from the beginning of the new millennium exploiting the conceptual framework of space-time networks, which allow for the explicit representation of each single run, in contrast to the aggregated representation of service in terms of lines used in frequency-based modes.

The most natural and well-established approach to model run-based assignment involves the representation of transit supply, which is intrinsically discrete in time, as a diachronic graph (Nuzzolo and Russo 1998; Nguyen et al. 2001), where each run is modelled through a specific sub-graph whose nodes have space and time coordinates according to the timetable. As an alternative, it is possible to define a dual graph (Nielsen and Jovicic 1999; Moller-Pedersen 1999), where each run section is a node, while the arcs represent the connections at stops satisfying temporal consistency. A third approach is to explicitly generate a number of alternative paths that constitute the passenger choice set and then assign on them the demand by means of a random utility model (Tong and Wong 1999; Friedrich et al. 2001). In general, stochastic models are often considered to simulate route choice on dynamic transit networks with timetables (Hickman and Bernstein 1997; Nuzzolo et al. 2001; Nielsen 2004).

The use of super networks to explicitly represent the departure time choice in the assignment model is an original contribution of this book, as Sheffi (1984) exploited this approach to reproduce different forms of elastic demand (mode choice, destination choice) in the context of static assignment algorithms.

The presentation of a consistent approach to reproduce a frequency-based behaviour on a schedule-based supply and a schedule-based behaviour on a frequency-based supply (see Sect. 6.2.4), both obtained by introducing proper arc costs, which paves the way to simulate networks with mixed services, is an original contribution of this book.

6.4 Macroscopic Models for Dynamic Transit Assignment

Guido Gentile

Macroscopic models for dynamic assignment have been developed in the last 30 years, mostly for private traffic. Their aim was to reproduce road congestion and more specifically how travel times are affected by the forming and vanishing of vehicle queues. This gives rise to the so-called *Dynamic Network Loading* (DNL) problem, where the flow propagation is performed using fixed route choices (not to be confused with the Network Loading Map of Sect. 6.1.8, which includes elastic route choice). A second use case of DTA involves, indeed, elastic route choice and the focus is on how the flow pattern is affected by congestion, giving rise to the so-called *Dynamic User Equilibrium* (DUE) problem.

In transit assignment, the interest for macroscopic dynamic models derives essentially from the possibility to describe FIFO queues at bus stops formed by passengers that are not able to board the first arriving carrier due to a lack of remaining capacity on the vehicle; these oversaturation queues are not to be confused with the under saturation queues that are due to the discontinuity of the service. Another relevant phenomenon that can be well captured by macroscopic models is the variation of service frequencies along the line due to the impact of boarding and alighting flows on dwelling times, which can even lead to bouncing (see Sect. 7.4). On the contrary, the scheduled-based models based on space-time networks presented in Sect. 6.3, which are the most common form of dynamic models for transit networks, are not suited to represent congestion phenomena that affect travel times.

In macroscopic models, vehicles are represented as a partially compressible fluid, whose physical law in stationary flow states is fully described by the so-called *fundamental diagram*. It is an experimental relation between flow density and its speed; the assumption that this holds true also in transition states results in the *kinematic wave theory*, which supports a number of (first order) traffic flow models.

In dynamic macroscopic models for transit networks, private vehicles are replaced by passengers. In this case, though, the progression of the user fluid is not affected by its density, as the speed of carriers is practically independent of on-board passenger loads, if the vehicle has sufficient engine power. However, this is not true for pedestrian arcs, where the congestion among walking passengers may resemble vehicle congestion, as well as for boarding and alighting arcs, whose flows influence the dwell times as explained in Sect. 7.4.4. In some model, indeed, transit line vehicles are represented as another flow component which follows a fixed path; in this case, the progression of the two flow components is strongly interdependent, thus adding a degree of complexity (non-separability) to the assignment problem.

Congestion is the focus of dynamic models. On the supply side of the DUE problem, the attention is thus essentially devoted to the passenger queuing at stops and to the seating mechanism, with emphasis on the node models with capacities and priorities, rather than on the arc model. On the demand side of the DUE problem, however, other congestion phenomena play a relevant role affecting costs through the value of time, rather than the travel time; these are primarily connected through discomfort for overcrowding, as explained in Sect. 7.2.1.

In the following, we focus on the mathematical framework of the approach and on the demand side of the DUE problem, leaving the description of congestion phenomena to Chap. 7. Section 6.4.1 extends the equilibrium formulation for transit assignment as a fixed-point problem to the dynamic case. Sections 6.4.2 and 6.4.3 illustrate how the concatenation of travel times influences the flow propagation and route choice, respectively, in a dynamic model. Section 6.4.5 shows how dynamic flow propagation applies to service frequency. Finally, Sect. 6.4.6 presents some references.

While the general ideas embedded in macroscopic modelling for network dynamics (presented in Sect. 6.4.1) are relevant for transit assignment as soon as there are congestion phenomena affecting travel times (e.g., queues of passengers at

stops), the specific formulation of each component (presented in Sects. 6.4.2 and 6.4.3) can be considered an advanced material that is not necessary to understand the reminder of the book, with the exception of Sects. 7.3.4 and 7.3.5.

6.4.1 *Fixed-Point Formulations of Arc-Based Dynamic Assignment*

In this section, the fixed-point scheme of Fig. 6.3, introduced to formulate User Equilibrium problems on static transit networks, is extended to macroscopic models for dynamic assignment, yielding Fig. 6.15; moreover, some details on the resulting functional components are provided.

An arc-based (see Sect. 6.1.11) macroscopic model for dynamic transit assignment consists of the four sub-models presented below, where the variables are *temporal profiles*, i.e., (semi) continuous functions of the daytime τ. Here, in contrast to the case of diachronic graph, the network is just a spatial and functional representation of the transport system with no embedded time dimension. The reference network element is the generic arc $a \in A$; variables are referred to users of class $g \in G$ directed towards destination $d \in D$ on mode $m \in M$.

Network congestion model (NCM) takes as input the arc volumes $q_a(\tau)$ that are typically aggregated from destination-specific flows $q_{admg}(\tau)$, and the arc characteristics $\boldsymbol{\delta}_a(\tau)$. It yields as output the arc exit times $\theta_a(\tau)$ for a given entry time τ and the corresponding value of time $\gamma_{ag}(\tau)$. This sub-model aims at reproducing various congestion phenomena, from discomfort to queuing, which introduce an increasing

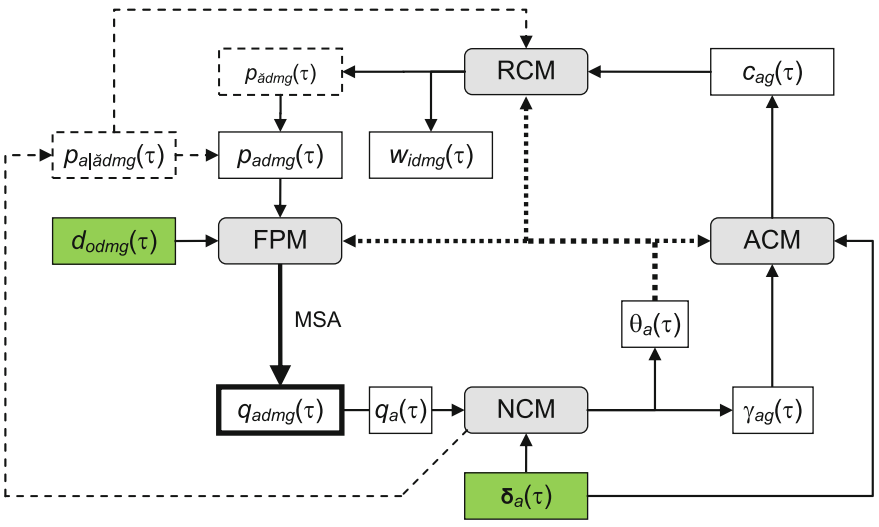

Fig. 6.15 Fixed-point formulation of DUE and DNL for arc-based macroscopic models

level of complexity; discomfort affects only costs, whereas queuing affects also times. Moreover, in transit assignment, some conditional probabilities derive from hyperarc diversion probabilities $p_{a|\breve{a}dmg}(\tau)$ (see Sect. 6.1.5); they are not the result of route choices, but are rather related to random events on the supply side, such as the attractive line probabilities (see Sect. 7.1) and the fail-to-board probabilities (see Sect. 7.3.3). The NCM shall also represent these physical phenomena.

Arc cost model (ACM) takes as input the arc travel times $\theta_a(\tau) - \tau$, the value of times $\gamma_{ag}(\tau)$ and the arc characteristics $\boldsymbol{\delta}_a(\tau)$. It yields as output the arc costs $c_{ag}(\tau)$ perceived by each user class, considering their different values of time and preferences. The need of handling the value of time as a separate variable from travel times and not directly in the ACM (as usual for traffic models) derives from the relevance of comfort in transit assignment, which can be heavily affected by overcrowding congestion, on-board and at stops (see Sect. 7.2.1).

Route choice model (RCM) takes as input the arc costs, as well as the arc travel times that allow for the dynamic concatenation of perceived utilities (see Sect. 6.1.13). It yields as output the expected costs (un-satisfactions, if route choice is based on a random utility model) to reach the destination from each node $w_{idmg}(\tau)$ that are then used to compute the arc conditional probabilities $p_{admg}(\tau)$ (the node costs can be seen as the dual variables of the arc probabilities). Interestingly, from an algorithmic point of view, it can be convenient to perform the computation of the latter directly in the FPM.

Flow propagation model (FPM) takes as input the travel demand $d_{odmg}(\tau)$ and the local choices, as well as the travel times that allow for the dynamic propagation of flows (see Sect. 6.1.13). It yields as output the arc flows of each class directed towards each destination, which are then aggregated into arc volumes. Arc flows by destination are needed as such by the more advanced NCM based on macro-, micro- or meso-simulations, as well as to apply gradient projection algorithms instead of MSA.

In the figure above, the rounded grey boxes are functionals; sharp white boxes are variables; and sharp green boxes are input. The bold box denotes the pivot variable of the fixed-point problem. The bold arrow closes the equilibrium loop and recalls that an algorithmic transformation of the pivot (e.g., through MSA or Gradient Projection) is required to ensure convergence. The dotted bold arrows highlight the crucial role of travel times in dynamic models. The dashed elements represent the extension to the case of strategic behaviour.

Two main cycles can be identified in the scheme of Fig. 6.15: inner and outer. The whole outer cycle is the DUE problem, while the inner cycle between FPM and NCM in the DNL problem. More specifically, the DUE can be then formalized as a fixed-point problem in terms of the arc flows:

$$\text{DUE} = \text{NCM} \rightarrow \text{ACM} \rightarrow \text{RCM} \rightarrow \text{FPM} \rightarrow [\text{MSA}] \rightarrow \text{NCM}. \tag{6.74}$$

The DNL is a sub-problem of DUE, which consists of seeking, for given route choices, an arc flow pattern consistent with the travel times through the arc performance model. DNL can be seen as a simplified DUE, without route choice. However, it still has a circular dependency to be solved iteratively in order to

guarantee temporal consistency (not more than few iterations in practice). Arc flows can be again considered as pivot variables of this fixed-point problem:

$$\text{DNL} = \text{NCM} \rightarrow \text{FPM} \rightarrow [\text{MSA}] \rightarrow \text{NCM} \ . \tag{6.75}$$

Both fixed-point problems, DUE and DNL, can be solved through the method of successive averages (MSA) considering as pivot variable the arc flows by destination $q_{admg}(\tau)$.

As an alternative, the DNL can also be solved in chronological order as a one-shot procedure (such as the link transmission models) without iteration by exploiting the acyclicity of causalities in time (i.e., an event occurring on an arc during a given time interval may have an effect on other arcs only in future time intervals), although this requires in practice a fine time discretization for short arcs. In general, the choice probabilities that are the input of DNL can be given in different forms: path probabilities (which requires their explicit enumeration), arc conditional probabilities per destination (which implies a sequential route choice model, as in the proposed framework), or non-destination-specific arc splitting rates (which does not guarantee the consistency of the loading with a given O–D matrix).

Note that in case of strategies, the DNL problem includes the update of the hyperarc probabilities through the NCM and their consequent use in the FPM. To correctly express this circumstance, in the scheme of Fig. 6.15 with respect to that of Figs. 6.2 and 6.3, the arc conditional probabilities $p_{admg}(\tau)$ derive explicitly from the hyperarc probabilities $p_{\check{a}dmg}(\tau)$ and the diversion probabilities $p_{a|\check{a}dmg}(\tau)$ as illustrated in Eq. (6.25). Clearly, if no strategic behaviour is considered, then all the dashed components of the scheme are discarded.

6.4.2 Propagation of Continuous Flows

To cope with the complexity of dynamic assignment, the concept of travel time is to be extended accordingly, as shown in Fig. 6.16. Let $\theta_a(\tau)$ be the exit time from arc a of a passenger who enters it at time τ. The inverse $\theta_a^{-1}(\tau)$ of the exit time profile yields the entry time of a passenger who exits it at time τ. The travel time $t_a(\tau)$ for a given entry time τ is then:

$$t_a(\tau) = \theta_a(\tau) - \tau. \tag{6.76}$$

As the travel time of any arc of non-null length is positive, the exit time profile $\theta_a(\tau)$ is always above the bisection with derivative 1. When travel times are increasing, its derivative is higher than 1; when travel times are decreasing, its derivative is lower than 1. If (strict) FIFO rule holds true, i.e., no overtaking is possible among passengers, then the derivative is always non-negative (positive).

In macroscopic models, passengers are represented as a partially compressible fluid. The flow is the amount of fluid traversing a given section at a given instant. It

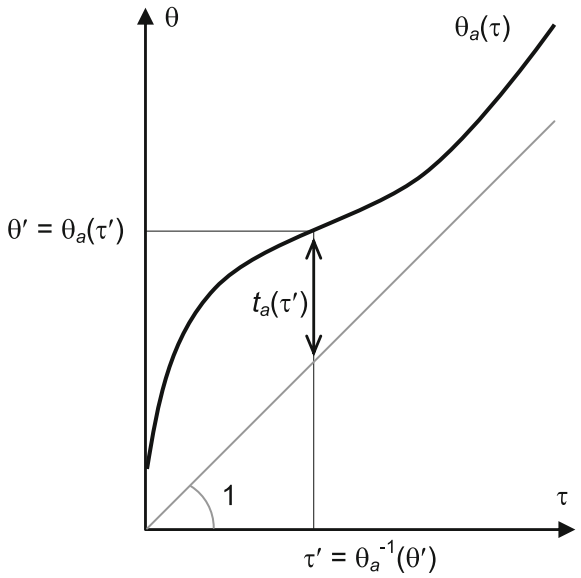

Fig. 6.16 The dynamic extension of the travel time variable: exit and entry time

is then not possible to talk generically of the passenger flow on a given network element, path or arc; instead, instantaneous inflows and outflows are to be defined and analysed, as well as entry and exit capacities. Indeed, at a solution of a DNL problem, the flow shall be consistent with the available capacities.

If the FIFO rule holds true, the cumulative outflow q_{ag}^{count} at time $\theta = \theta_a(\tau)$ when the passenger exits arc $a \in A$ is equal to the cumulative inflow q_{ag}^{cin} at the time $\tau = \theta_a^{-1}(\theta)$ when the passenger entered it:

$$q_{ag}^{cout}(\theta) = q_{ag}^{cin}(\tau). \tag{6.77}$$

Figure 6.17 shows how these dynamic variables are intrinsically connected: the horizontal distance between the cumulative outflow and inflow temporal profiles is the travel time, while their vertical distance is the number of passengers on the arc.

By taking the derivative of Eq. (6.77) with respect to τ while considering $\theta = \theta_a(\tau)$, the following result for instantaneous flows is obtained (cumulative flows are the integral in time of instantaneous flows):

$$q_{ag}^{out}(\theta) = \frac{q_{ag}^{in}(\tau)}{\frac{\partial \theta_a(\tau)}{\partial \tau}}, \tag{6.78}$$

showing that the outflow q_{ag}^{out} at time θ when the passenger exits arc $a \in A$ is equal to the inflow q_{ag}^{in} at time τ when the passenger entered it, divided by the derivative of the exit time at τ.

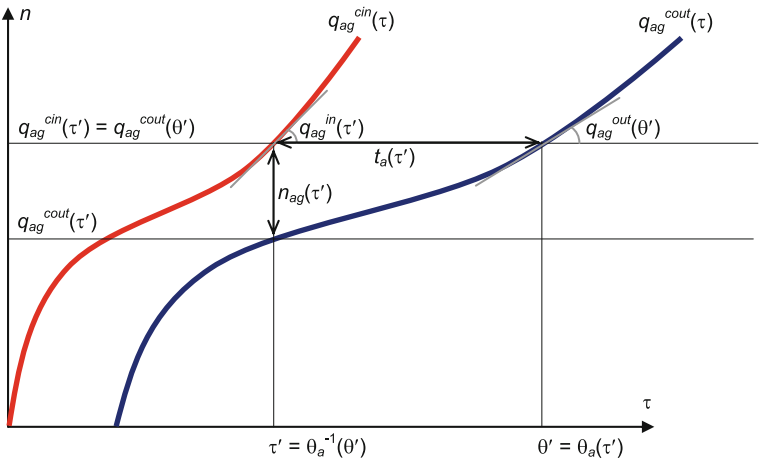

Fig. 6.17 Relation among the profiles of travel time, entry flow and exit flow, according to FIFO rule, in the space of cumulative flows

When the travel time is increasing (e.g., due to a growing queue), the outflow is smaller than the corresponding inflow; the opposite is true when the travel time is decreasing. Figure 6.19 shows how an arc speed v_a decreasing in time implies an arc flow q_a decreasing in space x, as the area of the two rectangles is equal.

Based on Eq. (6.78), the dynamic propagation of flows can be obtained by extending Eq. (6.22) expressing the flow balance of the node (see Fig. 6.18), as follows:

$$q_{idmg}(\theta) = d_{idmg}(\theta) + \sum_{a \in i^-} \frac{q_{a^-dmg}(\tau_a)}{\frac{\partial \theta_a(\tau_a)}{\partial \tau}} \cdot p_{admg}(\tau_a), \quad \tau_a = \theta_a^{-1}(\theta), \forall a \in i^-. \quad (6.79)$$

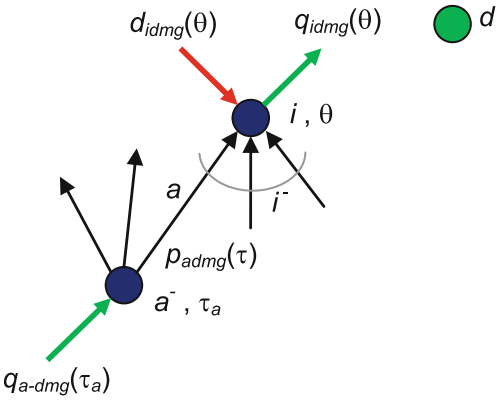

Fig. 6.18 Flow balance of node i at time θ for passengers directed towards destination d

The above time continuous model for network flow propagation can be transformed into a time discrete model by introducing the entry–exit map m_{aet} which denotes the share of users that enter arc a during interval e and exit it during interval t. Figure 6.19 shows how this share can be obtained from the exit time functional $\theta_a(\tau)$, as follows:

$$m_{aet} = \frac{Min\left(Min\left(\theta_a^{-1}(\tau_{t+1}), \tau_{e+1}\right) - Max\left(\theta_a^{-1}(\tau_t), \tau_e\right), 0\right)}{\tau_{e+1} - \tau_e}. \tag{6.80}$$

In this case, the dynamic flow propagation, given in Eq. (6.79), becomes a series of systems, one for each interval t of duration h_t, which can be solved in chronological order, so that the node flows of previous time intervals are always known:

$$q_{idmgt} = d_{idmgt} + \sum_{a \in i^-} \sum_{e<t} q_{a^- dmge} \cdot \frac{h_e}{h_t} \cdot m_{aet} \cdot p_{admge} + \sum_{a \in i^-} q_{a^- dmgt} \cdot m_{att} \cdot p_{admgt} \tag{6.81}$$

For sufficiently short-time intervals such that no user enters and exists any arc during the same interval (i.e., the aciclicity of causalities holds) it is: $m_{att} = 0$; the above systems become diagonal and their solution is trivial (nodes can be processed in any order). Otherwise, the solution algorithms proposed in Sect. 6.1.4 can be applied.

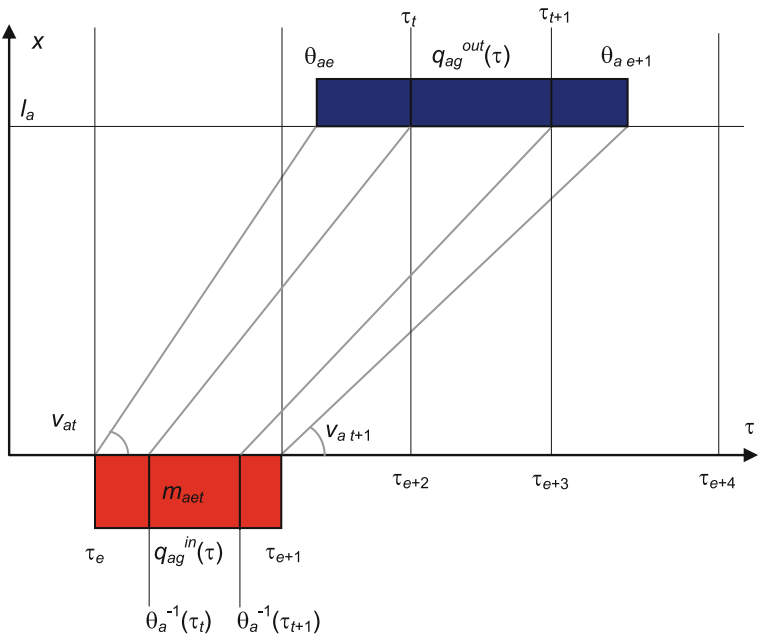

Fig. 6.19 Relation among the profiles of travel time, entry flow and exit flow, according to FIFO rule, in the case of discrete time intervals

6.4.3 Temporal Layer Formulation of Route Choice

The concatenation of time in dynamic route choices (see Fig. 6.20) can be ensured for arc-based models by substituting in Eqs. (6.18) and (6.16) the cost of each local alternative $b \in i^+$, denoted w_{bdmg}, with the cost of arc b for users entering it at time τ plus the expected cost to reach the destination from its final node evaluated at exit time $\theta_b(\tau)$:

$$w_{bdmg}(\tau) = c_{bg}(\tau) + w_{b^+dmg}(\theta_b(\tau)). \tag{6.82}$$

When time is discretized, Eq. (6.82) becomes the following:

$$w_{bdmgt} = c_{bgt} + w_{b^+dmge} + (\theta_{bt} - \tau_e) \cdot \frac{w_{b^+dmg\,e+1} - w_{b^+dmge}}{h_e}; \tag{6.83}$$

the temporal profile of the head expected cost is interpolated at the exit time $\tau_t + t_{bt}$ as a piecewise linear function between times τ_e and τ_{e+1}, where time index e is such that the corresponding interval of duration $h_e = \tau_{e+1} - \tau_e$ contains the exit time: $\tau_e \le \tau_t + t_{bt} \le \tau_{e+1}$, as shown in Fig. 6.21.

For what concerns the computation of local probabilities it is worth mentioning that to ensure the stability of equilibrium it is assumed that everybody behaves like the last passenger of the interval, otherwise some passenger would not suffer the effect of the congestion he/she generates.

For a suitable extension of the shortest paths problem to macroscopic dynamic models, *temporal layers* can be solved in reverse chronological order by setting the cost labels for passengers directed towards the current destination and leaving the node at the current time. If the largest interval of the adopted time discretization is smaller than the smallest arc travel time, then the dynamic shortest tree can be obtained by applying the Bellman update to all arcs in no particular order, otherwise the shortest path shall be processed in reverse topological order from destination to the furthest node, possibly adopting a Dijkstra algorithm as explained in Sect. 6.1.6.

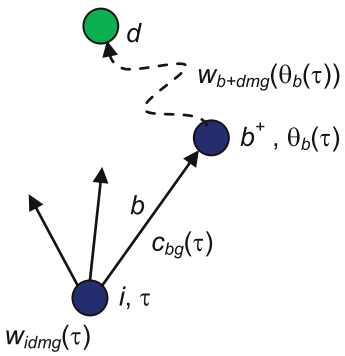

Fig. 6.20 Concatenation of travel times and local alternative cost

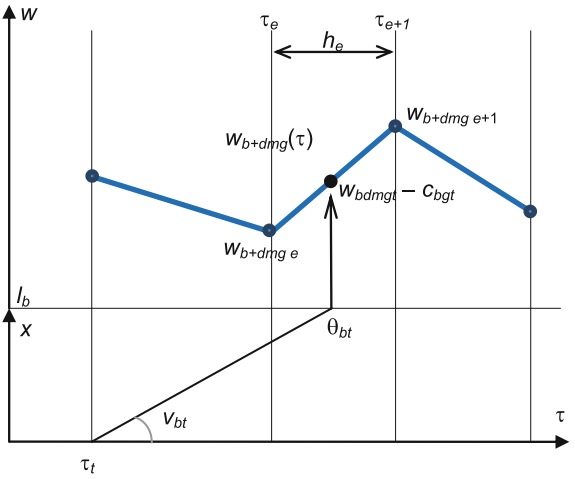

Fig. 6.21 Linear interpolation of cost label profile

6.4.4 Extension to Dynamic Hyperarcs

The extension of (6.82) and (6.26) to dynamic hyperarcs yields the following:

$$w_{\check{b}dmg}(\tau) = \frac{\gamma_{\check{b}^- g} \cdot t_{\check{b}dmg}(\tau) + \sum_{b \in \check{b}} p_{b|\check{b}\ dmg}(\tau) \cdot \left(c^{nt}_{bg}(\tau) + w_{b^+ dmg}\left(\theta_{b|\check{b}\ dmg}(\tau)\right)\right)}{\sum_{b \in \check{b}} p_{b|\check{b}\ dmg}(\tau)}, \tag{6.84}$$

where the combined exit time $\theta_{b|b^{\check{}}\ dmg}(\tau) = \tau + t_{b|b^{\check{}}\ dmg}(\tau)$ conditional to taking branch $b \in b^{\check{}}$is introduced to represent correctly the concatenation of travel times. This requires to modify also the flow propagation model by explicitly taking into account the hyperarcs and their branches as in (6.25):

$$q_{idmg}(\theta) = d_{idmg}(\theta) + \sum_{a \in i^-} \begin{cases} \sum_{\check{a} \subseteq \left((a^-)^+ \cap A_m\right): \check{a} \in H} \frac{\frac{q_{a^- dmg}(\tau_a)}{\partial \theta_{a|\check{a}\ dmg}(\tau_a)}}{\frac{\partial}{\partial \tau}} \cdot p_{\check{a}dmg}(\tau_a) \cdot p_{a|\check{a}\ dmg}(\tau_a), & \text{if } a \in A^{div} \\ & \tau_{a|\check{a}} = \theta^{-1}_{a|\check{a}\ dmg}(\theta), \forall a \in i^- \\ \frac{q_{a^- dmg}(\tau_a)}{\frac{\partial \theta_a(\tau_a)}{\partial \tau}} \cdot p_{admg}(\tau_a), & \text{otherwise}, \quad \tau_a = \theta_a^{-1}(\theta), \forall a \in i^- \end{cases} \cdot \tag{6.85}$$

6.4.5 Representation of Service Frequency as a Continuous Vehicle Flow

Line frequencies are among the most important features of a transit network. In a within-day, dynamic context frequencies are not a constant input but rather the result of a propagation process, as shown in this section.

It is possible to represent frequencies as a flow of vehicles, for example, as an additional class of users who travel from the first stop of the line to the last one with disabled boarding and alighting arcs; the demand of this particular class is the frequency profile $f_\ell(\tau)$ from the first stop. By construction, this flow has no route choice and shall follow exactly the line route.

From a formulation point of view, nothing changes with respect to the fixed-point scheme of Fig. 6.15 as frequencies can then be treated as an additional flow variable. Frequencies are processed by the FPM based on the current travel times and in turn influence, jointly with passenger flows, the travel times of both passenger and line-vehicle flows through the NCM. Thus, they become part of the DNL problem, which includes the averaging process.

More specifically, this particular flow can be propagated according to Eq. (6.78) along the sequence of line stops. The arrival and departure frequency of the generic line $\ell \in L$ at each stop can then be obtained by applying recursively the following equation starting from the first stop:

$$\begin{aligned} f_{\ell s}^{dep}(\theta) &= \begin{cases} f_\ell(\theta), & \text{if } s = S_\ell^- \\ \dfrac{f_{\ell s}^{arr}(\tau)}{\dfrac{\partial t_{\ell s}^{dwell}(\tau)}{\partial \tau} + 1}, & \tau{:}t_{\ell s}^{dwell}(\theta) + \tau = \theta, \quad \text{otherwise} \end{cases} \\ f_{\ell s}^{arr}(\theta) &= \frac{f_{\ell s^{-\ell}}^{dep}(\tau)}{\dfrac{\partial t_{\ell s^{-\ell}}^{run}(\tau)}{\partial \tau} + 1}, \quad \tau{:}t_{\ell s^{-\ell}}^{run}(\tau) + \tau = \theta \end{aligned} \tag{6.86}$$

The result is a frequency which varies in time and is different from stop to stop depending on the travel time variation on running arcs and dwelling arcs. The latter is more relevant because it is an internal congestion and will be specifically addressed in Sect. 7.4.4.

6.4.6 Reference Notes and Concluding Remarks

As mentioned earlier, the development of macroscopic models for dynamic assignment has been casted mainly in the context of road networks; in particular, the proposed approach has been proposed in Bellei et al. (2005) and further developed in Gentile et al. (2005), Bellei et al. (2006), Gentile and Papola (2006) and Gentile (2015). Thus, the proposition of the framework presented in this section for transit networks, which introduces also sequential route choice through

hyperarcs to reproduce passenger strategic behaviour, is to be considered an original contribution of this book.

Over the past few decades, many models were developed to solve dynamic transit assignment problems. Sumi et al. (1990) proposed a stochastic approach to jointly model departure times and route choices of passengers on a mass transit system. Alfa and Chen (1995) developed a transit assignment model for forecasting the temporal demand distribution along a corridor under a random assumption of passenger boarding. But more recently, the introduction of schedule-based models diverted most of the attention from macroscopic models for dynamic transit assignment.

An exception to this trend is provided in Meschini et al. (2007), who proposed a macroscopic DTA model based on continuous temporal profiles with variable frequencies and passenger queues; this work will be further analysed in Sect. 7.3.4. In the future, the link transmission models, developed by Yperman (2007) and by Gentile (2010) for road networks, will be likely adapted to reproduce transit networks by introducing a suitable node model describing the stop, thus further pushing the proposing approach.

6.5 Simulation-Based Models for Transit Assignment

Oded Cats, Umberto Crisalli and Agostino Nuzzolo

Computer simulations have become a useful framework for numerical modelling and the analysis of complex systems in various domains. In particular, simulations provide a powerful and attractive tool for representing system dynamics. This is especially true for large-scale systems that involve several interrelated stochastic processes that could not be solved analytically.

Agent-based simulations, in particular, allow to model complex systems that involve numerous autonomous and responsive elements. The agent-based modelling approach is used in a wide range of disciplines where the system dynamics emerge from the execution of individual strategies and the interactions among agents, as well as between each agent and the environment.

How can the simulation approach be used in the context of transit assignment modelling? Can transit system performance be represented as an emerging rather than a derived process? This section addresses these questions, exploring the development of simulation-based models for transit assignment by focussing on their potential capabilities, rather than on the formulation detail.

6.5.1 The Simulation Approach and Its Advantages

The transit assignment problem is concerned with finding the passenger flows and the corresponding travel times on the public transport network for a given travel demand. This is typically solved by the iterative loading of an origin–destination matrix consistently with a route choice model and the subsequent update of network conditions that may be required due to congestion phenomena (see Chap. 7). Simulation models enable to mimic the development of a global spontaneous order from numerous inter-dependent local and/or individual decisions. This modelling approach implies that the emerging equilibrium conditions are the results of complex interactions among numerous agents and the transit network dynamics.

The simulation-based approach has intrinsic advantages in obtaining a realistic representation of transit dynamics and in supporting the development of models that are practical for large-scale networks. Moreover, simulation models natively incorporate multiple classes through the synthesis of individual agents that are extracted from any distribution of user attributes. A great flexibility is allowed in the representation of agent interactions on the transport network (e.g., queuing, mingling, discomfort, seating), as well as of information provision and the consequent decision processes. In particular, simulation models are intended to mimic the adaptive response of travellers to changing system conditions and the incorporation of en-trip information in rerouting choices, thus making them a proper tool to support real-time traffic forecast and fleet management.

The main drawbacks of simulation models are the inability to derive mathematical functions that describe the system properties and the intrinsic randomness of the results, which actually represent a possible outcome of the system rather than the expected value of the desired output.

The combination of event-based mesoscopic modelling where passengers are aggregated in flows or packets on the supply side (congestion), along with a disaggregate modelling of individual decision-makers on the demand side (behaviour), yields better conditions for analysing large-scale systems with respect to fully microscopic models. This is particularly true when considering applications to advanced traffic management systems and advanced traveller information systems, where algorithm performance is an issue.

The more mature developments in the field of (road) traffic assignment models point to the potential role that simulation models can play in the context of transit assignment models. Coupling dynamic network loading (for the supply) and multi-agent simulation (for the demand) has been identified as a promising approach for modelling transit systems along with performance uncertainties and adaptive user decisions. However, the evolution of transit simulation models into dynamic transit assignment tools is at its early stages.

The mesoscopic approach in transit assignment consists of a dynamic disaggregated representation of both demand and supply, while their interaction is achieved through more aggregated models (e.g., discomfort functions, instead of pedestrian microsimulation). Assignment results are obtained from the iterative

loading of individual passengers to individual-vehicles serving the runs of public transport. This stands in strike difference to both frequency-based assignment models on static network and schedule-based assignment models on diachronic graph which consider travellers in terms of aggregate flows or loads. In contrast, the simulation approach applied to transit assignment aims at reproducing traveller behaviour at a microscopic level where each autonomous unit is considered an agent. The progress of vehicles and passengers on the transit system yields the temporal and spatial distribution of demand over supply.

Each iteration of the simulation is usually regarded, not as a repetition of a same random outcome, but as the representation of a single day in the context of an evolutionary day-to-day process which may be continued until the system possibly reaches stable conditions (see Sect. 6.1.10). In this sense, the system performances at equilibrium (if any) emerge in a 'bottom-up' rather than a 'top-down' fashion.

The simulation-based approach is especially appropriate in cases where the design problem at hand is concerned with travel demand attributes and their distribution, such as the provision of information, possibly in real time and the consequent adaptive behaviour of passengers, where the resulting travel strategies differ considerably depending on heterogeneous preferences and socio-demographic characteristics.

Simulation is also useful when various travel decisions, such as trip departure time and mode choice, are jointly considered as part of the assignment model together with route choice.

The disaggregated representation of supply and demand dynamics (vehicles and passenger trajectories), but, on the other side, the aggregated representation of interaction between vehicles and passengers (e.g., dwell time, discomfort), facilitate the explicit modelling of service uncertainties. In particular, simulation is well suited for capturing the dynamic evolution of network performances under oversaturation conditions due to queuing and crowding on vehicles and at stops. Service reliability and passenger congestion are therefore endogenous variables which emerge from system dynamics and their impact on individual travellers can be explicitly modelled.

6.5.2 Agent-Based Models

Agent-based models have been first developed in the domains of computer science, artificial intelligence and cognitive science. The iterative process is therefore often presented as either a computational method for optimization problem or a learning algorithm. The former considers the assignment procedure in terms of an iterative convergent method that seeks to obtain travellers' flows under stable steady conditions, while the latter formulates the iterative loading in terms of a cognitive process. While these different perspectives may be associated with a different interpretation and even implementation of modelling components, the simulation framework can be ultimately summarized in the following agent-based assignment algorithm, which adopts a day-to-day evolutionary approach, rather than an equilibrium approach (see Sect. 6.1.10):

Step 0 *Initialization*: generate traveller population U; select $K_u \; \forall u \in U$; reset $\tilde{\mathbf{c}}_{KU}^1$; $n \leftarrow 0$

Step 1 *New day*: $n \leftarrow n + 1$

Step 2 *Network Loading*: perform within-day assignment $p_{ku}^n \leftarrow p_{ku}(\tilde{\mathbf{c}}_{hu}^n, \forall h \in K_u)$; obtain $\mathbf{c}_{KU}^n$ for $\mathbf{p}_{KU}^n$

Step 4 *Stop criteria*: check steady condition, e.g., $\left\| \tilde{\mathbf{c}}_{KU}^n - \mathbf{c}_{KU}^n \right\| < \varepsilon$

Step 5 *Update*: perform day-to-day learning $\tilde{c}_{ku}^{n+1} \leftarrow \alpha_u^{learn} \cdot c_{ku}^n + (1 - \alpha_u^{learn}) \cdot \tilde{c}_{ku}^n$

Step 6 *Return* to Step 1

where

- n is the generic day
- U is the set of travellers
- $u \in U$ is the generic traveller
- K_u is the set of paths considered by traveller $u \in U$
- $K \in K_u$ is the generic path considered by traveller $u \in U$
- $\tilde{\mathbf{c}}_{ku}^n$ is the forecasted cost of path $k \in K_u$ for traveller $u \in U$ on day n
- p_{ku}^n is the probability of path $k \in K_u$ for traveller $u \in U$ on day n (route choice)
- c_{ku}^n is the actual cost of path $k \in K_u$ for traveller $u \in U$ on day n (congestion)
- α_u^{learn} is the coefficient of the exponential learning filter for traveller $u \in U$
- ε is a small positive number

Note that in this scheme, the choice updating filter presented in Sect. 6.1.10 to reproduce the tendency of users to conform to habits is not explicitly introduced, but is indeed present is some of the implementations presented in the following.

Network loading (this is not the DNL of Sect. 6.4.1, but the NLM of Sect. 6.1.8) is hence performed in Step 2 iteratively at the individual level for the entire travellers' population at once, so that congestion phenomena can emerge and affect passenger advancement. If the assignment results have not reached stable steady conditions, then the experienced travel attributes are incorporated into traveller cost anticipations for the following day/iteration.

The steps of the above algorithm are discussed in the remainder of this section, starting with a description of how supply and demand are represented in the simulation-based assignment model of transit networks, followed by the presentation of within-day and day-to-day dynamics. The discussion will refer to the following models which share the overarching schematic algorithm described above but vary with respect to their development context and objectives:

- MATSim, where the transit assignment model is part of an activity-based model;
- MILATRAS, which is tool for long-term planning of the transit system;
- BusMezzo, which is a joint traffic and transit assignment model oriented to operations.

These differences are clearly reflected in how supply and demand are represented in each one of these models as well as their overall design.

6.5.2.1 Demand Representation

The conventional origin–destination matrix considers demand in terms of aggregate traveller flows. However, travellers vary in their travel knowledge and preferences. For example, some travellers are extremely familiar with service supply and network prevailing conditions, such as the timetable and typical travel times, while others may have a limited knowledge on potential connections. Individuals also vary with respect to their travel preferences concerning alternative modes of transport, departure time and the importance of various trip attributes. For example, some travellers are very reluctant to transfer between transit services due to the uncertainty as well as physical and mental effort it induces, while others may be willing to transfer whenever it results with time savings.

The simulation of individual travellers enables, not only the representation of various user groups or classes with heterogeneous characteristics and preferences, but also the modelling of utility perception, strategy and experience at the individual level.

The first initialization step of an agent-based model for transit assignment involves the generation of a synthetic population of travellers. Initial conditions matter in case the learning and evolution processes are of interest, rather than only the (possible) steady conditions obtained at the end of the iterative process. MILATRAS was developed with the latter approach as it considers agents to have perfect knowledge on network topology but lack prior knowledge on performance attributes.

The generation process converts the O–D matrix into a population of agents, based on conditional probability functions for various user attributes. In case of transit travellers, origins and destinations may correspond to any location in the study area that is within walking distance to a transit stop. In the context of transit assignment, system initial conditions correspond to the planned service and individuals' prior-perception of its performances. Agents could be distinguished with respect to attribute preferences, prior knowledge (e.g., commuter vs. occasional users), learning patterns (e.g., bounded rationality) or explorative versus habitual attitude. This determines the initially anticipated path attributes of each traveller. We can assume that all relevant attributes of paths are synthetized by a generalized cost. Cognitive science approach will imply that the initial conditions reflect agents' mental map which is then progressively articulated.

BusMezzo is a joint traffic and transit assignment model where public transport passengers and private cars are generated based on separate time-dependent origin–destination matrices. Each traveller is assigned with inherited attributes such as trip departure time, walking speed, access to personal mobile device (e.g., real-time updates on instantaneous journey time), travel preferences (e.g., disutility associated with in-vehicle time versus wait time, walking time and transferring) and decision protocols (e.g., non-compensatory filtering rules, level of adaptation). These inherited attributes are maintained throughout the day-to-day simulation.

MILATRAS transforms a given O–D demand matrix into random geographical origin and destination locations based on residential and employment proportions through a GIS platform. Similarly, the synthetic population generated by MATSIM

could also be obtained based on a probability function which reflects the distribution of various socio-demographic characteristics in the target population. For example, the population could be derived from the conditional probabilities that describe the relationships between fundamental travel decision determinants such as age, household composition and car availability. The activity-based simulation can hence link socio-demographic attributes to travel experience and guarantee the internal consistency of trip chains.

6.5.2.2 Supply Representation

Simulation-based models for transit assignment can vary with respect to the level of representation of the fundamental supply elements, namely stops and vehicles, as well as the movements associated with them. However, the disaggregate representation of travellers does not necessarily imply a microscopic simulation of transit vehicles and traffic dynamics.

Traffic simulation models are commonly classified based on their level of representation detail. Macroscopic models represent traffic as a continuous flow based on flow-density functions without the explicit modelling of lanes or vehicles. In contrast, microscopic models represent traffic at the most detailed level: individual-vehicle movements are represented and their driving behaviour depends on interactions with other vehicles, on the road geometry, on lane usage, etc. As a result of computational constraints, there is an inverse proportionality between the level of details and the possible size of networks under study. Mesoscopic models are an intermediate category, where individual vehicles are represented but detailed modelling of their second-by-second movement and interaction is avoided. Travel times on links are indeed determined by speed-density functions, while delays at intersections are calculated by using queue models.

Simulation-based model for transit assignment can conceptually use any of these levels of representation for passenger and carriers flow dynamics. A multi-agent transit assignment approach would typically imply the representation of individual transit vehicles and their movement would be governed by either microscopic or mesoscopic traffic flow principles. Occupancy on-board each vehicle can then be updated throughout the simulation and capacity constraints can be explicitly enforced. Moreover, passenger movements at stops and on-board can be represented in order to capture the impacts of crowding discomfort and queuing delays.

The available agent-based models for transit assignment differ considerably with respect to the level of integration they exercise with road traffic simulation models. MATSim is integrated into a larger transport model on the demand side, but transit supply is simulated separately from private traffic and a deterministic representation is adopted. MILATRAS is an advanced programming interface that allows the enhancement of PARAMICS, a mesoscopic traffic model. This implies certain restrictions on the extent to which transit dynamics could be explicitly modelled. BusMezzo is completely integrated into Mezzo, a traffic simulation model.

The progress of transit vehicles between one stop and the other is affected by the interaction with other vehicles. Even though MATSim includes a mesoscopic modelling of multimodal traffic flows, total traveller door-to-door journey time is deterministically assumed to be twice the corresponding free-flow times of cars. Transit travel times between stops are instead extracted in MILATRAS from link speeds that are modelled in a mesoscopic way by PARAMICS. Transit vehicles are thus propagated based on exogenous traffic conditions. BusMezzo also models traffic flows at a mesoscopic level where travel times of cars and transit vehicles depend on general (equivalent) traffic conditions. The explicit representation of background traffic allows capturing the impacts of congestion on transit operations. The level of interaction between transit vehicles and other vehicles depends on the right-of-way (e.g., buses running in mixed traffic, dedicated lanes, underground).

The explicit simulation of transit vehicles enables a rich and stochastic representation of public transport supply and its dynamics. Dwell times at stops are modelled in MILATRAS and BusMezzo as a function of passenger activity at stops. Trip dispatching times are modelled as a random variable in MILATRAS. Vehicle scheduling is modelled explicitly in BusMezzo which enables the propagation of delays through trip chaining. Different transit lines may be assigned to various vehicle types, running speeds and are operated with different control strategies. The explicit modelling of these processes and their inter-relations can facilitate a more realistic reproduction of the underlying sources of uncertainty and their joint impacts on reliability, compared with their generation based on independent statistical distributions. Specifically, emulating these dynamics allows modelling their impact on travellers' route choice decisions. Moreover, it allows mimicking the generation and provision of passenger information.

6.5.2.3 Within-Day Dynamics

Route attributes may include travel time components (walking, waiting, riding, etc.), travel costs, number of transfers, as well as quality of service measures such as punctuality, crowding level and probability of denied boarding. Anticipated route attributes evolve iteratively through the incorporation of realized travel attributes from assignment results.

The within-day activity corresponds to a dynamic network loading procedure with travellers' flow pattern determined by the decisions passengers make in reaction to transit conditions, such as vehicle arrivals at stops, experienced travel conditions and information provision. Throughout the day, travellers execute their trips and accumulate experience concerning various route attributes. For example, travellers gain experience on wait times for different lines, in-vehicle time between different locations using different routes or modes or even assess the reliability and crowding conditions on alternative services. This within-day learning allows travellers to exercise adaptive (or strategic) travel behaviour. At the same time, travellers' decisions affect transit performances through the effect of passenger loads and flows on crowding, dwell times at stops and their secondary implications on

service reliability. The dynamic interaction between supply and demand lies in the core of the within-day assignment.

The modelling of day-to-day dynamics allows both service users and service providers to adapt their strategy in order to optimize or improve their objectives. This occurs in the day-to-day update phase, where travellers integrate the experience of the previous days into their perception of the network cost pattern and choose the strategy that they will carry out in the following day. As travellers increase their experience with the transit system, their mental map extends and their expectations reflect more closely the actual performances. This day-to-day learning process can result in steady conditions that are equivalent under certain conditions to user equilibrium.

The interaction between network supply and passenger demand takes place in the within-day dynamic network loading. This is the core of the day-to-day iterative process, where system dynamics are simulated and transit performance is determined.

The other fundamental building block is the way in which individual agents decide how to travel towards their destination. This highlights an important difference from conventional models for transit assignment, as the choice probability is referred to individual decisions rather than to travellers flow. Here, choice probabilities determine therefore the likelihood that a certain travel alternative will be used by a single individual depending on his/her specific attributes, instead of the passenger flow share that is distributed over a certain path set. A choice-set generation model composes first the paths that will be further considered in the route choice phase based on a combination of various path search methods and heuristic filtering rules.

Various route choice models can be formulated and embed into simulation-based transit assignment, which may differ with respect to each of the above components. In particular, the linkage between these components could be based on alternative theoretical grounds, from rule-based computational processes to utility maximization econometric models. In general, we can apply Eq. (6.40) to the single traveller:

$$p_{ku}^{n} = p_{ku}\left(\tilde{c}_{hu}^{n}, \forall h \in K_u\right). \tag{6.87}$$

The within-day network loading in MATSim is the result of choices among alternative travel plans rather than paths per se. The utility function of transit alternatives is composed of static and deterministic door-to-door travel time. Hence, the utility value is uniform across the population. Probabilities are assigned to alternative travel plans based on a multinomial Logit. Paths are chosen at the O–D level with no within-day learning.

MILATRAS assigns at the origin a tentative travel plan to each passenger that will be followed unless the travel experience differs substantially from the expectations. Expected travel time components are based on previous experience (day-to-day learning) and real-time information provision; in case that no information is available, travellers' expectations depend solely on their experience. MILATRAS is a bounded rationality model with a deterministic utility function:

there is a rule that allows to decide between an exploitation option (a deterministic choice of the alternative with the maximum utility) and an exploration option (a random choice over the set of considered alternatives). This can potentially capture the process of habit formation and occasional deviations.

Route decisions in BusMezzo are based on agent's current expectations on future travel attributes. The anticipated travel attributes incorporate prior knowledge, previous experience (day-to-day learning) and real-time information provision. The model includes a phase of non-compensatory choice-set generation followed by a probabilistic path choice process. The progress of travellers in the transit network is considered as a sequential process of successive arc decisions; thus passengers do not choose at any point between door-to-door paths. The adaptive path choice model was developed within the framework of random utility.

More specifically, for each local choice, such as boarding versus waiting at the stop or alighting versus staying-on-board, the passenger evaluates alternative *actions* by assessing the joint (or expected) utility to reach the destination conditional on taking that arc using the logsum term for all the available paths, as follows:

$$w_{au}^{n} = Log\left(\sum_{k \in K_{ua}} Exp\left(-\tilde{c}_{ku}^{n}\right)\right), \tag{6.88}$$

where

- a is the travel action (e.g., 'walk to given stop', 'board a given line') associated with an arc
- K_{ua} is the subset of sub-paths of K_u from arc a to the destination of traveller $u \in U$
- w_{au}^{n} is the composite utility of the local action a for traveller $u \in U$ in day n.

Thus, the upper level of the choice model refers to travel actions rather than path, while the used path is merely an outcome of individual's successive decisions; again, passengers do not choose a path per se at any given point along their trip.

The generalized cost $\tilde{\mathbf{c}}_{ku}^{n}$ may synthesize in a linear form several attributes $c \in C$ of path k, whose expected values anticipated by the passenger for day n are denoted $\tilde{a}_{kcu}^{n}$; these are differently perceived by each user u who associates to them a specific weight β_{cu}, so that it is:

$$\tilde{c}_{ku}^{n} = \sum_{c \in C} \beta_{cu} \cdot \tilde{a}_{kcu}^{n}. \tag{6.89}$$

As an example, the set of attributes C may include the number of transfers, in-vehicle time, wait time and walking time. This path utility function can be extended by accounting for service reliability and on-board crowdedness. The attribute values are determined by the integration of various information sources.

The within-day dynamic network loading yields passenger flows on each arc for that day. In addition, the values of the attributes experienced by the passengers on the utilized paths are obtained and are used to update the anticipated values for next day.

6.5.2.4 Day-to-Day Dynamics

The day-to-day learning process updates system states between successive network loadings. This process continues as long as the stopping criteria are not satisfied. The stopping criteria typically refer to the marginal change in key assignment outputs. For example, the stopping criterion can be defined as the share of travellers that changed their path from the previous day. If passengers cannot improve their travel costs by choosing an alternative path, then equilibrium conditions have been obtained. A probabilistic path choice, such as in MATSim and BusMezzo, could also be conceived in terms of a strategy repeated game and the Nash equilibrium. The convergence of assignment results can also be defined as a stopping criterion by considering changes in obtained arc flows. From a behavioural perspective, this indicates that travellers' perceptions are consistent with their experience, so that they have not gained new information from their most recent trips. This can also be explicitly assessed by checking whether the vector of travel attributes (costs) experienced by each single passenger is significantly different from its estimation before travelling.

The day-to-day learning process updates traveller perception by integrating the experience c_{ku}^{n} obtained on the path that was followed on day n, denoted k, to the accumulated passenger memory $\tilde{c}_{ku}^{n}$. Then, we can apply Eq. (6.39) to the single traveller:

$$\tilde{c}_{ku}^{n+1} = \alpha_u^{learn} \cdot c_{ku}^{n} + \left(1 - \alpha_u^{learn}\right) \cdot \tilde{c}_{ku}^{n}, \tag{6.90}$$

where α_u^{learn} is the step size assigned to the most recent experience. Various learning functions can be specified for different segments of the traveller population to determine how the step size evolves over time. For example, in the method of successive averages, (MSA) the step size is a function of the number of days/iterations, but not of their corresponding performances, with the weight of new solutions gradually decreasing throughout the solution process. In contrast, a more behavioural approach may assign larger weights to latter experiences.

The updating process in MILATRAS is formulated as a Markov decision. Each possible departure time and path decision combination is expressed as a state-action pair, where passengers' current state contains sufficient information for determining the next state. The Markov process is a non-equilibrium framework; however, the decision process may fulfil the conditions for convergence to a unique and optimal solution in terms of passenger state-action choices.

The day-to-day learning in MILATRAS and BusMezzo includes also the update of real-time information credibility. The information provided is evaluated against the experienced performance and influences the weight given to real-time information in future decisions.

6.5.3 *Traveller Cognitive Process*

In simulation-based models for transit assignment, the route choice probabilities lie at the individual level, as shown in Eq. (6.87), and result in passenger flows only after aggregation.

The disaggregated agent-based representation of transit demand is well suited to represent a population of travellers that is not uniformly informed about transit supply. Moreover, the representation of traveller learning behaviour, with an emerging mental map representation, enables the incorporation of various information sources and their integration into the path choice that changes dynamically from day-to-day.

These notions are elaborated in the following while referring to their implementation in MILATRAS and BusMezzo. In contrast, MATSim performs a joint modal split and route choice assuming that all travellers have perfect knowledge and information of the transit system; this approach is inspired by co-evolutionary theories. Moreover, this model does not distinguish between anticipated and experienced travel attributes. MATSim is therefore currently not suitable for modelling information scenarios and rerouting.

The strategies of travellers in a simulation-based model for transit assignment could be determined as the outcome of three sources of information:

- prior knowledge—the static information for passengers has on the planned service (e.g., schedule, frequencies);
- past experience—the accumulated first-hand experience with service performance, and
- real-time information—with respect to the arrival times expected today, in case available.

The reminder of this section describes how each of the above information sources is modelled in the context of a simulation-based framework, as well as how these information sources are integrated in the within-day assignment model and evolve from day to day.

6.5.3.1 Prior Knowledge on the Transit Network

A synthetic population U of users is generated based on probability functions that reflect the distribution of travellers' characteristics in the population. Expectations

for day 1 on costs $\tilde{\mathbf{c}}_{ku}^{1}$ for each user $u \in U$ and each path $k \in K_u$ are static sources of information that travellers inherit upon initialization.

This prior knowledge can be limited to network topology, as in MILATRAS, without any information on travel attributes; e.g., a fully optimistic null cost $\tilde{c}_{ku}^{1} = 0 \; \forall$ $k \in K_u$ is assumed. This implies that the first simulation will result with randomly chosen paths. This is equivalent to starting an optimization process with a randomly sampled solution and then progressively improving it using an iterative update process. Note that this is nevertheless different from models that generate travellers that are 'tabula rasa' and let them explore the network by applying random walk methods.

Alternatively, travellers can be assumed to have certain expectations on alternative path costs based on their prior knowledge. This information can be derived from planned headways, travel distances, travel times between stops or even timetables, as in BusMezzo.

In any case, travellers have no prior knowledge concerning other path attributes, such as service reliability or crowding levels.

For example, the prior knowledge may imply that the anticipated wait time at a given stop is half of the expected headway of the considered line, while actual wait times are the outcomes of the dynamic progress of individual travellers and vehicles in the simulation; the anticipated in-vehicle time may reflect the schedule or the expected speeds of road links during the relevant time period, which may vary due to traffic conditions on different times of day.

6.5.3.2 Accumulated Experience

The travel experience of each passenger $u \in U$ is however accumulated at the level of each single attribute $c \in C$ of the path $k^n \in K_u$ used in each day n, whose value is denoted a_{kcu}^{n}. The way in which this continuously updated source of information is compressed into a single value $\tilde{a}_{kcu}^{n}$ anticipated by the passenger for day n is defined by the day-to-day learning function. This function can, for example, allocate a greater value to more recent experience or specify a limited memory horizon.

The experienced attributes are calculated based on simulation dynamics. In particular, the experienced wait times are calculated directly as the time difference between traveller arrival time at stop and the time at which the passenger boarded a vehicle. The latter is also used as the reference value for calculating the experienced in-vehicle time until the passenger alighted the vehicle. BusMezzo and MILATRAS also represent access, egress and transfer links and account for their experienced travel times in a similar fashion.

The experience in the same day made earlier during the trip can also influence in traveller path choice. If the perceived path attributes deviate substantially from those anticipated, then passengers may revise their choice. For example, if a passenger experiences a wait time that exceeds considerably from that anticipated, then the connection decision is reconsidered and the traveller may choose to walk to another nearby stop.

6.5.3.3 Real-Time Information Provision

The dissemination of real-time information (RTI) may influence travellers' attribute perception and ultimately passenger flows. The RTI that is available to a traveller when making a certain route decision is determined by the dissemination means and their locations, as well as by individual characteristics, such as the availability of a personal mobile device. The rapid increase in the penetration rate of smart phones may considerably change the dissemination pattern, since passengers are possibly provided with instantaneous access to RTI during their entire trip.

Agent-based models for transit assignment enable the generation of RTI for the single passenger, based on individual-vehicle progress and arrival prediction schemes, which are embedded into the simulation engine. The explicit modelling of real-time predictions and information generation as a function of dynamic supply conditions enables the analysis of alternative dissemination strategies. In particular, the impact of various information provision schemes on travellers' decisions and ultimately on travellers' flows can be assessed. Note that RTI is therefore not equivalent to modelling the impact of perfect information.

Information availability is uniform across the population in MILATRAS, which represents the impacts of RTI on vehicle arrival times, when available pre-trip or through public displays at stops or on-board. The RTI concerning wait times is calculated based on the average conditions during the previous 45 min.

The dissemination of passenger information simulated in BusMezzo is classified according to the following aspects:

- Type—wait times, in-vehicle travel times, crowding levels, service disruptions;
- trip stage—pre-trip, at stops, on-board;
- comprehensiveness—concerning the local stop, cluster of connected stops (i.e., transit hub), the entire system.

The share of individuals that have access to RTI by using a personal mobile device can be specified in the population generation phase. The combination of the above aspects determines the level of information that is available to a specific passenger at each trip stage regarding downstream travel conditions.

The approach adopted in the BusMezzo implementation is to generate RTI based on historical data as expressed in timetables and on real-time data as resulting from the vehicle propagation. For example, RTI concerning wait time is calculated based on the current schedule deviation of the next vehicle arriving vehicle and the remaining travel time to reach the stop based on historical average. This scheme is aimed to replicate the method that is commonly used by transit agencies for generating real-time information.

Given the above information sources, traveller decisions are modelled in the probabilistic framework of random utility choice models. The evaluation of local alternative actions, as in Eq. (6.88), depends on passenger preferences and

expectations with respect to forecasted travel attributes. The individual decision protocol specifies the forecasted attributes $\hat{a}^n_{kcu}$ as convex combination of the following information sources:

- $\hat{a}^n_{kcu}$, the prior knowledge,
- $\hat{a}^n_{kcu}$, the value anticipated by the passenger based on the accumulated experience,
- $\hat{a}^{RTI}_{kcu}$, the value resulting from real-time information.

We have then:

$$\hat{a}^n_{kcu} = \alpha^{PKn}_u \cdot \tilde{a}^1_{kcu} + \alpha^{TE\,n}_u \cdot \tilde{a}^n_{kcu} + \alpha^{RTI\,n}_u \cdot \tilde{a}^{RTI}_{kcu}, \tag{6.91}$$

where α^{PKn}_u, α^{TEn}_u and α^{RTIn}_u are the weights (that sum up to one) associated with prior knowledge (PK), travel experience (TE) and real-time information (RTI), respectively, in day n. These weights could be interpreted in terms of the credibility associated with each information source. Therefore, in presence of information Eq. (6.89) becomes:

$$\tilde{c}^n_{ku} = \sum_{c \in C} \beta_{cu} \cdot \hat{a}^n_{kcu}. \tag{6.92}$$

6.5.3.4 Day-to-Day Evolution of Information Credibility

The weights associated with the various information sources are determined through a day-to-day learning process and thus vary with day n. Hence, day-to-day dynamics influence not only the experienced travel attributes, but also the credibility assigned to various information sources. As the day-to-day assignment progresses, the weight given to prior knowledge is expected to decrease while the impact of experience increases. Moreover, the credibility associated with various information sources vary among travellers. However, the extent to which the memory of passenger u with respect to path $k^n \in K_u$ extends over time is determined endogenously as it depends on how relevant is the path that was followed in day n in that network loading iteration and cannot be defined a priori as a function of n.

Moreover, the weight given to *RTI* reflects its perceived credibility which is a function of the extent to which the information provided in advance accurately predicted the corresponding travel attributes actually experienced. For example, the day-to-day update function of the *RTI* weight can take the following form:

$$\alpha^{RTI\,n+1}_{u3} = \alpha^{cred}_u \cdot \frac{\left\| \tilde{a}^{RTI}_{kcu} - a^n_{kcu} \right\|}{a^n_{kcu}} + \left(1 - \alpha^{cred}_u\right) \cdot \alpha^{RTI\,n}_{u3}, \tag{6.93}$$

where α^{cred}_u is the step size assigned to the most recent experience and the attributes refer to k^n that is the path used in day n.

6.5.4 Mesoscopic Models for Schedule-Based Simulation

In Sect. 6.3, the schedule-based assignment is presented for uncongested networks with regular services (that perfectly adhere to the timetable) assuming that passengers make a fully preventive route choice. This approach is very limiting to model urban transit networks, especially when we need to take into account the effects of:

- vehicle capacity, with queue formation and fail-to-board events;
- service irregularity, with path attributes that change over time;
- passenger's en-route choices, due for example to the arrival of a run at a stop later than expected or the arrival of overcrowded vehicles.

In particular, the basic schedule-based models reported in Sect. 6.3 do not allow for the simulation of real-time conditions and short-term prediction.

Therefore, in the following another class of schedule-based models for transit assignment is reported. It uses a simulation approach, which allows to overcome the above-mentioned limits.

In particular, schedule-based assignment can be casted as an event-based simulation, in which events represent instants when passengers depart from origins or transit vehicles arrive and depart at stops.

Passengers depart from origins with a preventive path choice in mind. Once arrived at stops, the simulation of fail-to-board probabilities due to possible formation of queues induces rerouting choices, thus providing a better estimation of vehicle loads for each run. Note that rerouting (especially if queues are not a recurrent event) does not necessarily imply a strategic behaviour, where passenger would choose a hyperpath (and not a path) fully including expectations on the real-time events and their costs.

Moreover, the use of schedule-based simulation models allows to reduce the computational complexity of large networks. It can be defined in the context of a mesoscopic approach similar to that described in Sect. 6.5.2, which presents an aggregate representation of individual-vehicle performances, allowing to avoid the simulation of second-by-second vehicle movements and interactions. Specifically, while the mesoscopic models of Sect. 6.5.2 are characterized by a disaggregate representation of the demand at single passenger level, the simulation approach to the schedule-based assignment here presented considers also in aggregate way the demand as group of travellers with homogeneous features, called 'packets', i.e., passengers moving over the transit network and experiencing the same trip. Therefore, the demand–supply interaction of this class of models is based on a within-day dynamic network loading in which packets of passengers are propagated along the chosen transit routes.

6.5.4.1 Supply Variance

The network model is made of transit services represented by runs moving between stops with travel times that are external inputs. Therefore, the representation of transit services uses the run-based approach and the *diachronic graph* described in Sect. 6.3. In case of real-time simulation, each time a transit vehicle departs from a stop; the diachronic network is updated with the new forecasted travel times that are used for the next step of the simulation. Such travel times can be obtained through an Automated Vehicle Location (AVL) system in case of real-time and short-term modelling or they can be the result of realisations of multidimensional random variables with parameters and dispersion matrix obtained from experimental data (e.g., using Automated Vehicle Monitoring data). If experimental data are not available, departure times from the terminals can be assumed equal to the scheduled times and travel times (summing up, running and dwell times) as multivariate normal variables with the average equal to the scheduled times plus a given quantity and variance–covariance matrix $\mathbf{\Sigma}$ defined, considering correlation among running and dwell times of the same and different sections.

Starting from the vector of scheduled arrival/departure run-times at stops, we obtain the vector of scheduled run and dwell times $\boldsymbol{\theta}$, according to which vector $\boldsymbol{\theta}^n$ of actual run and dwell times in day n is generated by extracting values from multivariate normal random variable, $MVN(\boldsymbol{\theta}, \mathbf{\Sigma})$.

The obtained vector $\boldsymbol{\theta}^n$ must satisfy some feasibility rules including the congruence of generated times with the allowed speeds for transit vehicles and the congruence of possible bunching phenomena, for which runs passing on the same sections cannot pass one another, so the quickest vehicles must slow down and follow the slowest ones. The vector that satisfies the above feasibility criteria is used to update the diachronic graph for the next simulation step.

6.5.4.2 Hierarchic, Sequential and Adaptive Route Choice

In the framework of the simulation-based mesoscopic approach, instead of individual passengers, the model considers packets of passengers with homogeneous characteristics (at least, origin, destination and desired departure time from origin or desired arrival time at destination) moving on the transit network.

The segmentation over time of the demand can be represented by dynamic O–D trip matrices, from which packets of passengers can be generated for each minute of the simulation period.

The typical approach in modelling trip choices is based on random utility, where a set of alternatives is identified and each one is associated with a systematic utility plus a random residual with a given joint distribution (see Sect. 4.5). In this case, the path choice probabilities (6.87) apply to the packet of travellers, allowing thus to estimate the average number of passengers using a given run, and then their contribution to its on-board load.

As introduced in the previous section, the preventive joint (one shot) approach to modelling route choice, where the passenger decides the whole path from the origin to the destination before starting his/her journey, based on historic information obtained from previous trip experiences or supplied by a travel planning system, can be questioned from a behavioural point of view.

A different interpretation of user behaviour assumes that the actual route used by the passenger results from a hierarchic sequence of *stop choice* and *run choice*, until the destination is reached. Moreover, these choices can be based, not only on past experience, but also on the information regarding the current conditions of the transport system, possibly improved by real-time updates.

For the choice of the first boarding stop $s \in S$ where to access the next service starting from vertex $i \in B$ at instant $t \in T$, we can assume a *pre-trip choice behaviour*, based on the comparison of possible alternative considering expected characteristics, or attributes, which also include variables such as the inclusive utility. The choice set is defined by the stops that are reachable within a maximum walking time t_g^{wmax} on the pedestrian network, which differs for each user class $g \in G$.

The opposite of the systematic utility v_s^{idgt} of each stop $s \in S$ for passengers of class $g \in G$ directed towards destination $d \in D$ can be given as follows:

$$-v_s^{idgt} = \gamma_{sg}^{stop} \cdot t_g^{stop} + \gamma_g^{vot} \cdot \gamma_g^{walk} \cdot t_{is}^{walk} + w_{jdg}, \tag{6.94}$$

taking into account:

- the characteristics of the stop (e.g., ergonomy, presence of shops) which can be synthetized by the stop discomfort coefficient γ_{sg}^{stop} (introduced in Sect. 5.1.2) that multiplies in this case a reference stop time t_{sg}^{stop};
- the walking time t_{is}^{walk} on the shortest path from i to B_s^{stop} on the pedestrian network;
- the set of connection opportunities that can be found at stop s, synthetized by the inclusive utility (also called satisfaction) w_{jdg} of node $j = (s, e)$, which represents the attractiveness of the stop in terms of runs useful to reach the destination at the time when s is reached, $e = t^+(\tau_t + t_{is}^{walk})$.

The probability p_s^{idgt} of choosing stop s starting from vertex $i \in B$ at instant $t \in T$ is formally given by (Fig. 6.22):

$$p_s^{idgt} = p_s^{idgt}\left(v_{s'}^{idgt}, \forall s' : t_{is'}^{walk} \le t_g^{wmax}\right). \tag{6.95}$$

Once arrived at stop, transit vehicle boarding (e.g., run choice) is simulated through an *at-stop choice behaviour*, which describes how users respond to unknown or unpredictable events, such as the transit vehicle arrivals in a different sequence with respect to the expected one due to service irregularity.

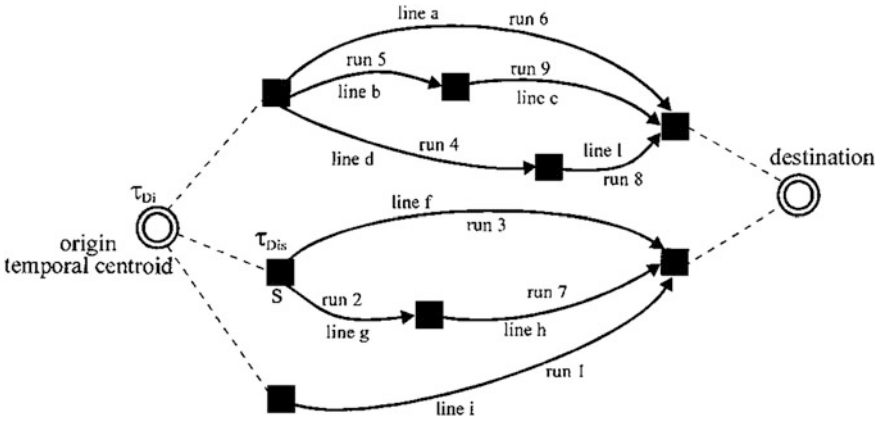

Fig. 6.22 Example of hierarchic approach to the subsequent choice of first stop and route

The overall choice set for a passenger arriving at stop s at instant t includes each run that directly or indirectly allows to reach to destination d and satisfies some predefined rules, such as the following:

- it is the first run of its line departing from the stop after the user arrival (this shall be removed in case of oversaturation queues when passengers may not be able to board the next-arriving run of each line);
- it is not dominated by another run leaving after arriving before with a lesser generalized cost;
- it implies less that a maximum number of transfers.

The choice of the run $r \in R$ to board at stop $s \in S$ for a user that reached it at instant $t \in T$ can be interpreted as the result of a sequential dynamic decision, where the passengers waiting at the stop probabilistically reject or accepts to board each arriving run $r \in R$ depending on their performance estimation of the run alternatives r' that are still available to reach the destination $d \in D$.

The estimation may take into account also the current conditions of the service (possibly provided by a real-time information system). Indeed, frequent users, who know from previous experience how the system operates, can react to en-route events (or to their information) to optimize their journey cost by adapting their route choice (the same result can be obtained through personal information provided by a real-time journey planner). However, this goes in the direction of a strategic behaviour which is treated in the next chapter (see Sect. 7.1); therefore, in the following, the role of information is simply included in the random errors associated with each run of the choice set.

The opposite of the systematic utility $v_{r|r'}^{sdgt}$ associated with each run r of the choice-set conditional upon arrival of run r' for passengers of class $g \in G$ directed towards destination $d \in D$ can be given as follows:

$$-v_{r|r'}^{sdgt} = \gamma_g^{vot} \cdot \gamma_g^{wait} \cdot \gamma_{sg}^{stop} \cdot \gamma_{sgt}^{crowd} \cdot \left(\theta_{rs} - \theta_{r's}\right) - \gamma_g^{vot} \cdot \gamma_g^{loss} \cdot \left(\theta_{r's} - \tau_t\right) + c_{L_r s}^{bfee} \cdot \gamma_g^{mfee} + w_{jdg}, \tag{6.96}$$

taking into account:

- the wait time for run r given by $\theta_{rs} - \theta_{r's}$;
- the time already waited $\theta_{r's} - \tau_t$, where the value of time γ_g^{vot} is further multiplied by a new discomfort coefficient γ_g^{loss} that weights the regret of the passenger on the past *lost opportunities* and the consequent 'loss of hope' in the future opportunities;
- the monetary cost $c_{\ell s}^{bfee}$ of boarding line $\ell = L_r$ at stop s;
- the expected cost to reach the destination once the passenger is on-board of the run r, which includes travel time, comfort and number of transfers, which is synthetized by the inclusive utility (also called satisfaction) w_{jdg} of node $j = N_{rs}^{dep}$.

The attributes composing Eq. (6.96), including those making up the inclusive disutility w_{rsdg}, can be differently estimated according to the information sources through Eq. (6.91).

It is worth noting that the choice set is modified over time during the wait because after each arrival, the corresponding run is eliminated (and possibly the next run of the line is added).

When a run r arrives at the stop a passenger may choose to board it if its perceived utility (given by the systematic utility plus a random residual) is greater than that of each other run $r' > r$ of the choice set that has not passed yet (with some abuse of notation). The resulting (conditional) probability of choosing to board the arriving run r' is denoted $p_{r'|r'}^{sdgt}$ and depends on the systematic utilities $v_{r''|r'}$ of each run $r'' \geq r'$. If the passenger does not choose run r, the choice is reconsidered when the next run r' arrives and so on.

Thus, a run r is boarded if it is chosen when it arrives at the stop while the runs $r' < r$ of each previous arrival where not chosen. If such events are assumed independent from each other, it is possible to evaluate the unconditional probability p_r^{sdgt} of boarding run r by a passenger of class g directed to destination d who arrives at stop s at instant t as follows:

$$p_r^{sdgt} = p_{r|r}^{sdgt}\left(v_{r''|r}^{sdgt}, \forall r'' \geq r'\right) \cdot \prod_{r' < r}\left(1 - p_{r'|r'}^{sdgt}\left(v_{r''|r'}^{sdgt}, \forall r'' \geq r'\right)\right). \tag{6.97}$$

The above model can be particularly difficult to solve. To reduce the computational effort of this approach, the following further assumptions can be made:

- only the choice of the first stop from the origin is considered as a separate hierarchic level, while the choice of intermediate stops is included in the run choice of a joint route to reach the destination;

- the choice-set restriction mechanism is not considered, while instead a new run of the line that just passed is added;
- the loss discomfort coefficient is assumed null.

In this case, the proposed sequential framework reduces to a stochastic arc-based model on the diachronic graph which can be solved easily through the equations and algorithms provided in the previous sections. Indeed, for each run departure from a stop, the diversion on the diachronic graph between boarding arc and (keep) waiting arc represents a binary probabilistic choice, in which all passengers have the same behaviour independently from their arrival time at the stop.

6.5.4.3 Dynamic Network Loading

The dynamic network loading allows to simulate the propagation of travellers on the diachronic graph and to obtain run on-board loads. It can be divided into several steps considering the simulation framework in which the previous choices of travellers moving on the network are updated at stops.

The first step consists of loading the pedestrian network from origins to access stops on the basis of traveller pre-trip choices consistently with Eq. (6.95). The second step allows defining the contribution to the on-board load of each run consistently with Eq. (6.97).

The loading process is carried out in discrete times, considering only the instants in which a run of transit services arrives at any of the stops and hence on-board loads could change. Passenger boarding a given run is obtained by summing up all contributions due to all paths of all O–D pairs.

If the loads of passengers willing to board a given run at a given stop exceeds the residual capacity, this implies a redistribution of this share to next-arriving runs and updating the path choice for the passengers who failed-to-board, using in a recursive way the loading process defined by Eqs. (6.96)–(6.97), and assuming a FIFO rule or a mingling rule at the stops (see Sect. 7.3).

6.5.5 Reference Notes and Concluding Remarks

The application of day-to-day dynamic assignment to transit networks for schedule-based models on diachronic graphs (see Sect. 40) is today further employed in most simulation-based approaches (e.g., Toledo et al. 2010).

Among simulation-based models for transit networks we can mention: MATSim (Balmer et al. 2008), where the transit assignment model is part of an activity-based model; MILATRAS (Wahba and Shalaby 2005), which is tool for long-term planning of the transit systems; BusMezzo (Cats 2013), which is a joint traffic and transit assignment model oriented to operations. In addition to the above models, an agent-based bus model was developed by Meignan et al. (2007). However, it is not

a simulation-based as passengers' decision is limited to choosing between the shortest path by alternative travel modes.

In any case, the reader should know that the development of simulation-based models for transit assignment is still in its early stage. Their development is inspired by a range of theoretical domains and their implementation is often part of a larger laboratory environment development.

The dynamic and disaggregate modelling of both transit supply and demand could potentially yield more realistic assignment results. The validation of simulation-based transit assignment model is a prerequisite for them to become more operational. The representation of traffic dynamics is already at a mature stage with sufficient validation studies. Transit vehicle trajectories and service variations were validated for BusMezzo (Cats et al. 2010, 2011). Moreover, MATSim traffic assignment and MILATRAS transit assignment were validated against standard assignment tools (Gao et al. 2010; Wang et al. 2010). These validation results provided positive indications. However, there is a need to further validate assignment results against actual time-dependent passenger flows at the individual-vehicle run level.

The performance of agent-based transit assignment model in terms of running times and convergence properties has not been carefully analysed yet. The availability of prior knowledge for example may provide a first feasible solution, which will improve the assignment solution process in terms of both quality and speed compared with starting with a random solution. The learning function parameters also presumably have important implications on the converging process. Further developments of dynamic path choice models and the underlying behavioural determinants will require a more extensive framework for representing memory construction as well as habit formation and risk assessment.

Until agent-based models will not reach acceptable calculation times, the use of schedule-based simulation models allows us to reduce the computational complexity, particularly in large network applications. The use of the schedule-based assignment approach can be very useful for real-time and short-term modelling, and today, it represents one of the frontiers of modelling and applications in this field, especially when effects of traveller information on short-term predictions about on-board loads should be deployed.

Simulation-based assignment models provide a natural common modelling platform for analysing complex urban transport dynamics. The co-evolutionary process which drives the assignment and the modular simulation environment could potentially accommodate additional travellers' adaptation strategies such as modal shift, trip departure time adjustments and even destination choice. Existing models already combine several decisions layers. This development is in line with the development of activity-based demand models and agent-based urban planning tools such as ILUTE (Salvini and Miller 2005) and PUMA (Ettema et al. 2007).

References

Alfa AS, Chen MY (1995) Temporal distribution of public transport demand during the peak period. Eur J Oper Res 83:137–153

Amin-Naseri MR, Baradaran V (2014) Accurate estimation of average waiting time in public transportation systems. Transp Sci 49:213–222

Andreasson I. (1976) A method for the analysis of transit networks. In: Roubens M (ed) Proceedings of the 2nd European congress on operations research, North Holland, Amsterdam

Balmer M, Rieser M, Meister K, Charypar D, Lefebvre N, Nagel K (2008) MATSim-T: architecture and simulation times. In: Bazzan ALC, Klügl F (ed) Multi-agent systems for traffic and transportation engineering. Information science reference, Hershey, pp 57–78

Bellei G, Gentile G, Papola N (2005) A within-day dynamic traffic assignment model for urban road networks. Transp Res B 39:1–29

Bellei G, Gentile G, Meschini L, Papola N (2006) A demand model with departure time choice for within-day dynamic traffic assignment. Eur J Oper Res 175:1557–1576

Bellman R (1958) On a routing problem. Q Appl Math 16:87–90

Bowman LA, Turnquist MA (1981) Service frequency, schedule reliability and passenger wait times at transit stops. Transportation Research A 15:465–471

Cantarella GE (1997) A general fixed-point approach to multimode multi-user equilibrium assignment with elastic demand. Transp Sci 31:107–128

Cantarella GE, Cascetta E (1995) Dynamic processes and equilibrium in transportation networks: towards a unifying theory. Transp Sci 29:305–329

Cats O (2013) Multi-agent transit operations and assignment model. Proc Comput Sci 19:809–814

Cats O, Burghout W, Toledo T, Kousopoulos HN (2010) Mesoscopic modeling of bus public transportation. Transp Res Rec 2188:9–18

Cats O, Kousopoulos HN, Burghout W, Toledo T (2011) Effect of real-time transit information on dynamic path choice of passengers. Transp Res Rec 2217:46–54

Chriqui C, Robillard P (1975) Common bus lines. Transp Sci 9:115–121

De Cea J, Fernandez JE (1989) Transit assignment to minimal routes: an efficient new algorithm. Traffic Eng Control 30:491–494

Dial RB (1967) Transit pathfinder algorithm. Highw Res Board 205:67–85

Dijkstra EW (1959) A note on two problems in connexion with graphs. Numer Math 1:269–271

Ettema D, Jong K, Timmermans H, Bakema A (2007) PUMA: multi-agent modelling of urban systems. In: Koomen E et al (eds) Modelling land-use change, pp 237–258

Fearnside K, Draper DP (1971) Public transport assignment—a new approach. Traffic Eng Control 13:298–299

Friedrich M, Hofsaess I, Wekeck S (2001) Timetable-based transit assignment using branch and bound techniques. Transp Res Rec 1752:100–107

Gallo G, Longo G, Nguyen S, Pallottino S (1993) Directed hypergraphs and applications. Discrete Appl Math 42:177–201

Gao W, Balmer M, Miller EJ (2010) Comparisons between MATSim and EMME/2 on the greater Toronto and Hamilton area network. Transp Res Rec 2197:118–128

Gentile G (2010) The general link transmission model for dynamic network loading and a comparison with the due algorithm. In: Immers LGH, Tampere CMJ, Viti F (eds) New developments in transport planning: advances in Dynamic Traffic Assignment (selected papers from the DTA 2008 conference, Leuven). Transport economics, management and policy series. Edward Elgar Publishing, MA, pp 153–178

Gentile G (2015) Using the general link transmission model in a dynamic traffic assignment to simulate congestion on urban networks. Transp Res Proc 5:66–81

Gentile G, Papola A (2006) An alternative approach to route choice simulation: the sequential models. In: Proceedings of the European transport conference, Strasbourg, France
Gentile G, Meschini L, Papola N (2005) Spillback congestion in dynamic traffic assignment: a macroscopic flow model with time-varying bottlenecks. Transp Res B 41:1114–1138
Hickman MD, Bernstein DH (1997) Transit service and path choice models in stochastic and time-dependent networks. Transp Sci 31:129–146
Jolliffe JK, Hutchinson TP (1975) A behavioral explanation of the association between bus and passenger arrivals at a bus stop. Transp Sci 9:248–282
Larson RC, Odoni AR (1981) Urban operations research. Prentice-Hall, Englewoods Cliffs
Last A, Leak SE (1976) Transept: a bus model. Traffic Eng Control 17:14–20
Le Clercq F (1972) A public transport assignment method. Traffic Eng Control 14:91–96
Meignan D, Simonin O, Koukam A (2007) Simulation and evaluation of urban bus-networks using a multiagent approach. Simul Model Pract Theory 15:659–671
Meschini L, Gentile G, Papola N (2007) A frequency based transit model for dynamic traffic assignment to multimodal networks. In: Allsop R, Bell MGH, Heydecker BG (eds) Proceedings of the 17th international symposium on transportation and traffic theory (ISTTT). Elsevier, London, pp 407–436
Moller-Pedersen J (1999) Assignment model of timetable based systems (TPSCHEDULE). In: Proceedings of 27th European transportation forum, seminar F, Cambridge, England, pp 159–168
Nguyen S, Pallottino S, Malucelli F (2001) A modeling framework for passenger assignment on a transport network with timetables. Transp Sci 35:238–249
Nielsen OA (2000) A stochastic transit assignment model considering differences in passengers utility functions. Transp Res B 34:377–402
Nielsen OA (2004) A large-scale stochastic multi-class schedule-based transit model with random coefficients. In: Wilson NHM, Nuzzolo A (eds) Schedule-based dynamic transit modeling: theory and applications. Kluwer Academic Publisher, Dordrecht, pp 53–78
Nielsen OA, Jovicic G (1999) A large-scale stochastic timetable-based transit assignment model for route and sub-mode choices. Transp Plann Methods 434:169–184
Nuzzolo A, Russo F (1998) A dynamic network loading model for transit services. In: Proceedings of TRAISTAN III, San Juan, Puerto Rico
Nuzzolo A, Russo F, Crisalli U (2001) A doubly dynamic schedule-based assignment model for transit networks. Transp Sci 35:268–285
Osuna E, Newell G (1972) Control strategies for an idealized public transportation system. Transp Sci 6:52–72
Pallottino S, Scutellà MG (1998) Shortest path algorithms in transportation models: classical and innovative aspects. In: Marcotte P, Nguyen S (eds) Equilibrium and advanced transportation modelling. Kluwer Academic Publishers, Dordrecht, pp 245–281
Salvini P, Miller EJ (2005) ILUTE: an operational prototype of a comprehensive microsimulation model of urban systems. Netw Spat Econ 5:217–234
Sheffi Y (1984) Urban transportation networks: equilibrium analysis with mathematical programming methods. Prentice-Hall, NJ
Sumi T, Matsumoto Y, Miyaki Y (1990) Departure time and route choice of commuters on mass transit systems. Transp Res B 24:247–262
Toledo T, Cats O, Burghout W, Koutsopoulos HN (2010) Mesoscopic simulation for transit operations. Transp Res C 18:896–908
Tong CO, Wong SC (1999) A stochastic transit assignment model using a dynamic schedule-based network. Transp Res B 33:107–121
TRB (2013) TCRP Report 165–Transit Capacity and Quality of Service Manual, 3rd Edition
Wahba M, Shalaby A (2005) Multiagent learning-based approach to transit assignment problem a prototype. Transp Res Rec 1926:96–105

Wang J, Wahba M, Miller EJ (2010) A comparison of an agent-based transit assignment procedure (MILATRAS) with conventional approaches city of Toronto transit network. In: Proceedings of the 89th transportation research board annual meeting, Washington DC

Watling D (1999) Stability of the stochastic equilibrium assignment problem: a dynamical systems approach. Transp Res B 33:281–312

Yperman I (2007) The link transmission model for dynamic network loading. PhD thesis, Katholieke Universiteit Leuven

Chapter 7
The Theory of Transit Assignment: Demand and Supply Phenomena

Guido Gentile, Klaus Noekel, Jan-Dirk Schmöcker, Valentina Trozzi and Ektoras Chandakas

This chapter addresses the modelling of various demand and supply phenomena emerging on public transport networks: passenger information, congestion at stops and on board, and 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 Chap. 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.

G. Gentile (✉)
DICEA—Dipartimento di Ingegneria Civile, Edile e Ambientale, Sapienza University of Rome, Via Eudossiana, 18, 00153 Rome, Italy
e-mail: guido.gentile@uniroma1.it

K. Noekel
PTV AG, Haid-und-Neu-Strasse 15, 76131 Karlsruhe, Germany
e-mail: klaus.noekel@ptvgroup.com

J.-D. Schmöcker
Department of Urban Management, Kyoto University, C1-2-436, Katsura Nishikyo-ku, Kyoto 615-8540, Japan
e-mail: schmoecker@trans.kuciv.kyoto-u.ac.jp

V. Trozzi
Strategy and Service Development, Transport for London, London SE1H 8NJ, UK
e-mail: valentinatrozzi@tfl.gov.uk

E. Chandakas
Transamo, Transdev Group, Paris, France
e-mail: ektoras.chandakas@transamo.com

G. Gentile and K. Noekel (eds.), *Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems*, Springer Tracts on Transportation and Traffic 10, DOI 10.1007/978-3-319-25082-3_7

7.1 Strategies and Information

Klaus Noekel, Guido Gentile and 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 and
- information provision to passengers.

In the basic framework for frequency-based assignment (Sect. 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 with respect 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 Sect. 6.2.2 for frequency-based models and the arc performances presented in Sect. 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 Sect. 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, and 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 Sect. 6.2.2 is depicted in Fig. 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* $H^{wait} = \{\breve{a} \subseteq s^+ : \forall s \in S\}$.

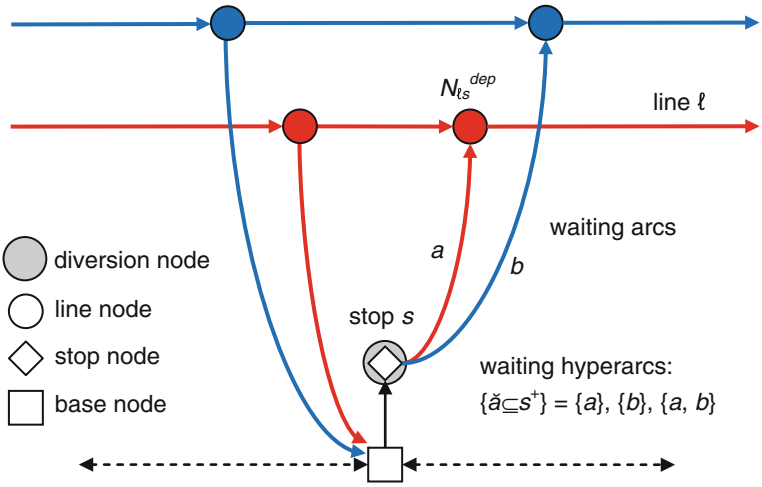

Fig. 7.1 Topology of the transit network with boarding hyperarcs exiting from the stop diversion node

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

The arc performances presented in Sect. 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 sub-graph ($N, \bar{A}_{dg} \in A$), 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 $i \in N$ defined 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 as follows:

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

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 Sect. 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 $t_a = 0$, 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 Sect. 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 subset is a combinatorial problem.

Rearranging the Eq. (6.30) based on Eq. (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 $i \in N^{div}$:

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

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

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 Eq. (6.67f)] and
- 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 Sect. 6.1.7).

Based on Eqs. (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, and
- 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 Eq. (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):

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

subject to:

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

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

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

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

where q_a is the flow on arc a, q_i is the total flow at node i, and d_i is the travel demand departing from node i, if any (these are flows of class g users directed towards destination d). At first sight, Eq. (7.9a) constitutes a mixed integer non-linear 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, to represent the total wait time at stop $i \in s$, as follows:

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

one obtains the equivalent problem:

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

subject to:

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

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

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

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

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

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

subject to:

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

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

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

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 as follows:

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

subject to:

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

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

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

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

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 as follows:

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

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 in Table 7.1.

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 variable f_i contains 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 $w_i = \infty$.

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 $c_a + w_{a+} \leq c_b + w_{b+}$ 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* $w_i \leftarrow \frac{w_i f_i + (c_a + w_j) f_a}{f_i + f_a}$; $f_i \leftarrow f_i + f_a$ 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)
	$q_j \leftarrow q_j + q_a$	

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 Eq. (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 Eq. (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 Sect. 5.1.3. The assignment graph is the same as in Sect. 6.2.5 and is also depicted in Fig. 7.2 along with the arc costs and the demand flows.

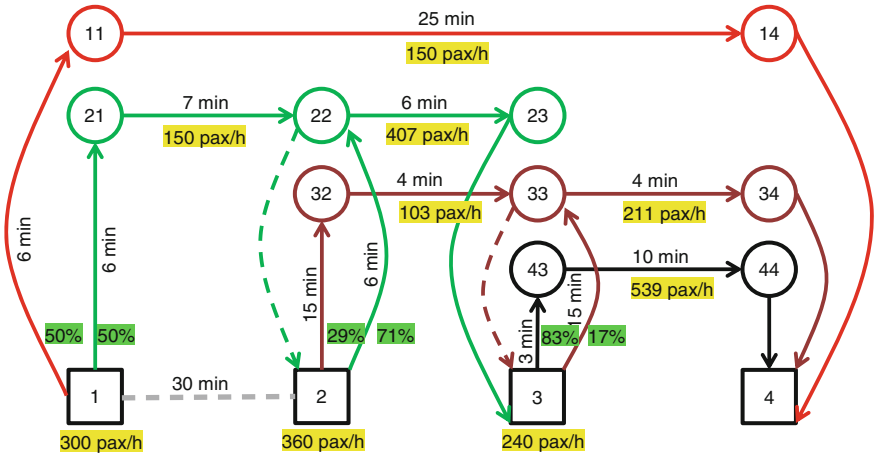

Fig. 7.2 Input data and results of AoN assignment to optimal strategies applied to the example network

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 numerical computation shown in Table 7.2 results from a slight modification of the first pass of the algorithm described in Sect. 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 Sect. 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.

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 that shown in Fig. 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 as shown in Fig. 6.8 for running arcs.

7.1.2 Regular Headways and Sequential Observation

As explained in Sect. 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 Eq. (6.66) spans from the exponential case $1/f$ (for $n = 0$) to the case of constant headways $0.5/f$ (for $n \rightarrow \infty$).

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 Eq. (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 Eqs (6.54) and (6.55); in the case of Erlang headways, these are given, respectively, by Eqs (6.52) and (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 $\breve{A}_i^+ = \breve{a} \subseteq i^+$ 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 as follows:

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

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 Eq. (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 Eq. (7.4).

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

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

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 $t_{a|\breve{a}}$ conditional on boarding line a is given by:

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

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:

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

By observing Eq. (6.65), which has general validity, some authors substitute the frequency of the line $f_{\ell s}$ with $2 \cdot f_{\ell s}/(1 - \sigma_{\ell s}^2)$ 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 Sect. 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 of 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 Eq. (7.4) to find the optimal solution:

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

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/he 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 Sect. 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 strategy 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, and
- a predefined cost attribute (e.g., the distance on the graph from the stop to the destination).

Finally, in Eq. (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 Sects. 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 min, the estimated wait time at the beginning of waiting is 10 min, but after t minutes of waiting, the remaining wait time drops to $(20 - t)/2$ min, until after at most 20 min the line must arrive with certainty.

Billi et al. (2003–2004) and PTV (2003) independently analysed 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 Eq. (7.4) for a constant attractive set $\breve{a}_{x(t)}$.

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

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

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 $w_i(t + \mathrm{d}t) \leq w_i(t)$ at any $t \geq 0$ for a small $\mathrm{d}t > 0$. 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 $\mathrm{d}t$ 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 $t + \mathrm{d}t$. 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 $\tau_a \geq 0$ 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}: $\tau_a \leq h_a^{max}$.

The dynamic attractive set can be then defined as follows:

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

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 $a1$, $a2$, …, an, with $n = |i+|$, be the lines in decreasing order of remaining cost, from the best to the worst; it is as follows: $w_{a1} \leq w_{a2} \leq \cdots \leq w_{an}$; $\tau_{a1} \geq \tau_{a2} \geq \cdots \geq \tau_{an}$. In the following, the index a is dropped for the sake of simplicity. The expected cost $w_i(t)$ can then be denoted as $w_i(t, \{\tau_1, \tau_2, \ldots, \tau_n\})$, thus showing explicitly its dependence on the dynamic set $\breve{a}_{x(\tau)}$.

The construction (Fig. 7.3) 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:

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

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:

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

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:

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

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

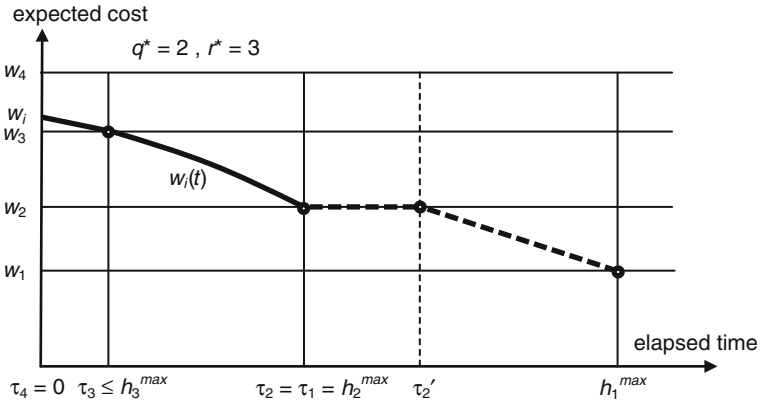

Fig. 7.3 Expected cost as a function of the elapsed time for a dynamic attractive set. Note that the instant obtained as intersection of $w_i(t, \{h_1^{max}, 0, \ldots, 0\})$ with w_2, denoted τ_2' is higher than h_2^{max}, and the procedure is then restarted from h_2^{max}

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^*.

7.1.3.2 Expected Cost and Line Shares

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

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

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

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

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 $h \le k$ is boarded at time τ is as follows:

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

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:

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

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 $\bar{A}_i^+ = \breve{a} = \{q^*, \ldots, r^*\}$, 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 $k + 1, \ldots, r^*$ 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 Eq. (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 $t_{a|\check{a}}$ conditional on boarding line a can be obtained like in (7.27), once $p_{a|\check{a}}(t)$ 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 $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_n$ 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 as described in Sect. 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 with respect to 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 proposed a label-setting algorithm for computing expected costs from each stop to a given destination and assigned 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 $\bar{A}_i^+ = \check{a}$ at stop i and then boards the first arriving vehicle of a line in $\check{a}$. 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 Sect. 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:

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

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:

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

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

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

As headways are constant, the general formula reduces to:

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

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

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

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 $t_{a|\breve{a}}$ 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 $\breve{a} = i^+$ and then find out which line have a positive diversion probability:

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

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 Sect. 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 $a \in \breve{a}$ 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 (Table 7.4) for the example network of Sect. 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 passengers board and alight the service at the same stop.

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

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	30 min 30 s	27 min 45 s	27 min 41 s	27 min 35 s

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

30 min 30 s to 27 min 45 s 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 is 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 as shown in Table 7.5.

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.

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

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	27 min 25 s	26 min 37 s

original example. The reason is that Line 5 is slow but frequent; this results in a gain of 10 s.

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 behaviour 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 as follows:

- 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 Sect. 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 countdown 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 Sects. 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 Sect. 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), which are all reachable lines, and the local alternatives physically connected with the node, which are all arcs of its forward star. In essence, the choice is modelled with respect to the first set of alternatives (the lines), and then, the results are aggregated with respect to the second set of alternatives (the arcs) to apply the sequential route choice paradigm of Sect. 6.1.5.

7.1.7 Optimal Strategies on Diachronic Graphs

Consider the space-time network introduced in Sect. 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 Sect. 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 Fig. 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 Eqs. (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 Eqs. (5.12) and (5.13) as in Sect. 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 Sect. 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 Eq. (6.15), but it is provided by the combined cost of the attractive set for the arc.

More specifically, with 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, then 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; and
- 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 rigour 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 Sect. 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.

7.2 Discomfort: Seating and Crowding

Jan-Dirk Schmöcker and 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; and
- seat capacity.

As explained in Sect. 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); and
- 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 Sect. 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 Sect. 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 Sects. 6.2.2 and 6.3.1. In Sect. 7.2.3, we describe the formulation of the equilibrium problem. Finally, in Sect. 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 Sect. 6.2, although the proposed formulation can be immediately extended to the case of schedule-based models on diachronic graphs presented in Sect. 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 Sect. 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 Eq. (6.67d). This can be done with a BPR-type function, similar to the cost functions considering the impact of road congestion to travel times:

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

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; and
- ${\alpha_g}^{crowd}$ 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 Sect. 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 Eq. (6.67f). 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 Sect. 6.2.3:

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

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; and
- 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:

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

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

7.2.1.1 Applications to the Example Network

The arc performance model of Eq. (7.50) to reproduce crowding congestion is here applied jointly to the classical route choice model of optimal strategies presented in Sect. 7.1.1. The resulting equilibrium problem has been solved for the example network of Sect. 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

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

Line	(pax/h)	Segment					
		1 → 2	2 → 3	3 → 4	1 → 4	2 → 1	Production (pax * km/h)
		3.5 km	3 km	3 km	10 km	3.5 km	
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

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

Line	(pax/h)	Segment					
		1 → 2	2 → 3	3 → 4	1 → 4	2 → 1	Production (pax * km/h)
		3.5 km	3 km	3 km	10 km	3.5 km	
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

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 shown in Table 7.7 are exactly the same of those resulting in the numerical example of Sect. 7.1.1.6.

Let us 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 with respect to 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 Eq. (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 Sect. 7.2.2 would instead ensure priority for passengers already on board).

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 penalizes 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 overcrowding.

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 Sect. 6.1.3) and to the introduction of 'fail-to-sit' probabilities, as described in the following.

A different specialization of line nodes (see Fig. 7.4) with respect to that proposed in Fig. 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

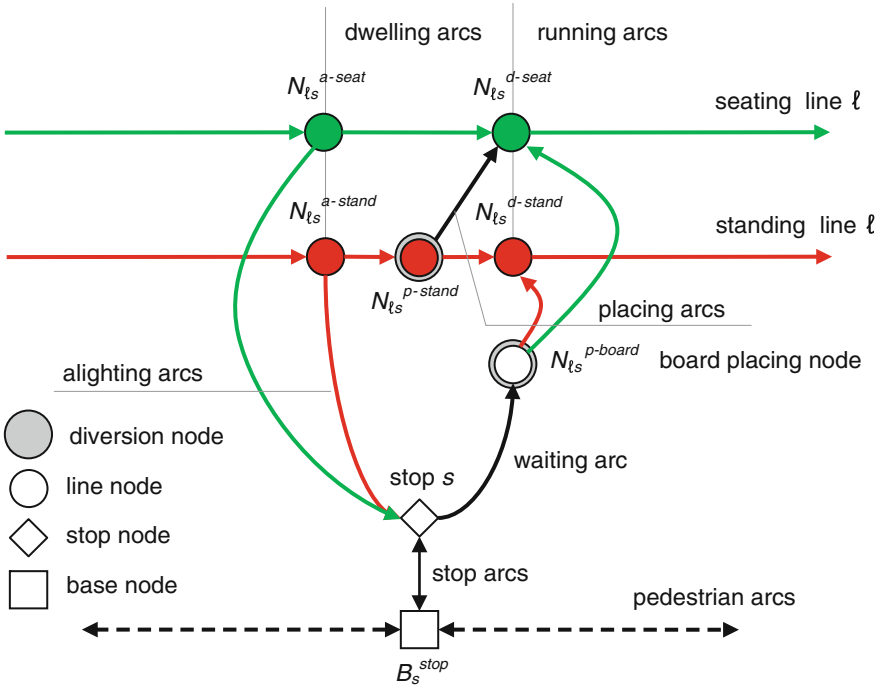

Fig. 7.4 Network topology to represent seat availability and priority

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

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* $A^{r\text{-}seat} = \{(N_{\ell s}^{d\text{-}seat}, N_{\ell s[+\ell]}^{a\text{-}seat}): \forall s \in S_\ell - S_\ell^+, \forall \ell \in L\}$;
- the *seat placing arcs* $A^{p\text{-}seat} = \{(N_{\ell s}^{p\text{-}board}, N_{\ell s}^{d\text{-}seat}): \forall s \in S_\ell - S_\ell^+, \forall \ell \in L\}$;
- the *seat dwelling arcs* $A^{d\text{-}seat} = \{(N_{\ell s}^{a\text{-}seat}, N_{\ell s}^{d\text{-}seat}): \forall s \in S_\ell - S_\ell^- - S_\ell^+, \forall \ell \in L\}$;
- the *seat alighting arcs* $A^{a\text{-}seat} = \{(N_{\ell s}^{a\text{-}seat}, s): \forall s \in S_\ell - S_\ell^-, \forall \ell \in L\}$;
- the *stand running arcs* $A^{r\text{-}stand} = \{(N_{\ell s}^{d\text{-}stand}, N_{\ell s[+\ell]}^{a\text{-}stand}): \forall s \in S_\ell - S_\ell^+, \forall \ell \in L\}$;

- the *stand dwelling arcs* $A^{d\text{-}stand} = \{(N_{\ell s}^{a\text{-}stand}, N_{\ell s}^{p\text{-}stand}): \forall s \in S_\ell - S_\ell^- - S_\ell^+, \forall \ell \in L\}$;
- the *stand placing arcs* $A^{p\text{-}stand} = \{(N_{\ell s}^{p\text{-}board}, N_{\ell s}^{d\text{-}stand}): \forall s \in S_\ell - S_\ell^+, \forall \ell \in L\}$;
- the *stand alighting arcs* $A^{a\text{-}stand} = \{(N_{\ell s}^{a\text{-}stand}, s): \forall s \in S_\ell - S_\ell^-, \forall \ell \in L\}$;
- the *waiting arcs* $A^{wait} = \{(s, N_{\ell s}^{p\text{-}board}): \forall s \in S_\ell - S_\ell^+, \forall \ell \in L\}$;
- the *switch seating arcs* $A^{p\text{-}switch} = \{(N_{\ell s}^{p\text{-}stand}, N_{\ell s}^{d\text{-}seat}): \forall s \in S_\ell - S_\ell^+, \forall \ell \in L\}$;
- the *keep standing arcs* $A^{p\text{-}keep} = \{(N_{\ell s}^{p\text{-}stand}, N_{\ell s}^{d\text{-}stand}): \forall s \in S_\ell - S_\ell^+, \forall \ell \in L\}$;
- the *boarding hyperarcs* $H^{board} = \{\{(N_{\ell s}^{p\text{-}board}, N_{\ell s}^{d\text{-}seat}), (N_{\ell s}^{p\text{-}board}, N_{\ell s}^{d\text{-}stand})\}: \forall s \in S_\ell - S_\ell^+, \forall \ell \in L\}$; and
- the *dwelling hyperarcs* $H^{dwell} = \{\{(N_{\ell s}^{p\text{-}stand}, N_{\ell s}^{d\text{-}seat}), (N_{\ell s}^{p\text{-}stand}, N_{\ell s}^{d\text{-}stand})\}: \forall s \in S_\ell - S_\ell^+, \forall \ell \in L\}$.

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 Sect. 6.2.3, the following changes:

- placing arcs ($A^{p\text{-}stand} \cup A^{p\text{-}seat} \cup A^{p\text{-}switch} \cup A^{p\text{-}keep}$) and hyperarcs ($H^{board} \cup H^{dwell}$) are dummy (null cost);
- Equation (6.67c) apply to all dwelling arcs and branches; and
- 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:

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

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:

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

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 Sect. 6.1.5.

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

For what concerns route choice, Eq. (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 Sect. 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 Fig. 6.3 can be applied considering the sequential model based on hyperarcs of Sect. 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 shown in Fig. 7.4 to each single run of the diachronic graph introduced in Sect. 6.3.1.

The example in Sect. 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 Eq. (7.50) in the expanded seat-availability network of Fig. 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 Eq. (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 Sect. 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 travelling on arc $a \in A$ to destination $d \in D$, whose objective function is the well-known sum of arc cost integrals (Beckmann 1956):

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

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

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

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

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

This leads to a convex optimization 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; and
4. check the distance to equilibrium (e.g., 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:

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

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 Sect. 7.3.

To address the combination of overcrowding congestion with the route choice model based on optimal strategies discussed in Sect. 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 travelling 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:

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

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 Eq. (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 nonlinear optimization program such as 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 Sect. 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 Sect. 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) described 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 are 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 Sect. 7.2.2 finds 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.

7.3 Passenger Queuing

Guido Gentile and 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 reboarding 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, such as 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 and
- mingling and fail-to-board probabilities versus 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:

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

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 onboard (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 (oversaturation 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 oversaturation queues, as the under-saturation queues have been analysed indirectly in Sect. 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; and
- 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:

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

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 Sect. 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 Eq. (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:

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

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 Sect. 7.1) and effective frequencies into a user equilibrium model (see Sect. 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 Sect. 7.2.1.1, by substituting the crowding congestion with the queueing congestion.

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

Line	(pax/h)	Segment					Production (pax * km/h)
		1 → 2	2 → 3	3 → 4	1 → 4	2 → 1	
		3.5 km	3 km	3 km	10 km	3.5 km	
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

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

Line	(pax/h)	Segment					Production (pax * km/h)
		1 → 2	2 → 3	3 → 4	1 → 4	2 → 1	
		3.5 km	3 km	3 km	10 km	3.5 km	
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

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

The arc performance model of Eq. (7.60) is here applied jointly to the classical route choice model of optimal strategies presented in Sect. 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 Line 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 has 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.

To overcome this drawback, we finally present the results of a similar equilibrium model with queue congestion where Eq. (7.60) is substituted with Eq. (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.

Yet, this is achieved through very high costs for boarding passengers, which may be unrealistic, and also requires many MSA iterations 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.

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 Eq. (7.62). When mingles queues occur at the stop, the probability to succeed in boarding the next approaching

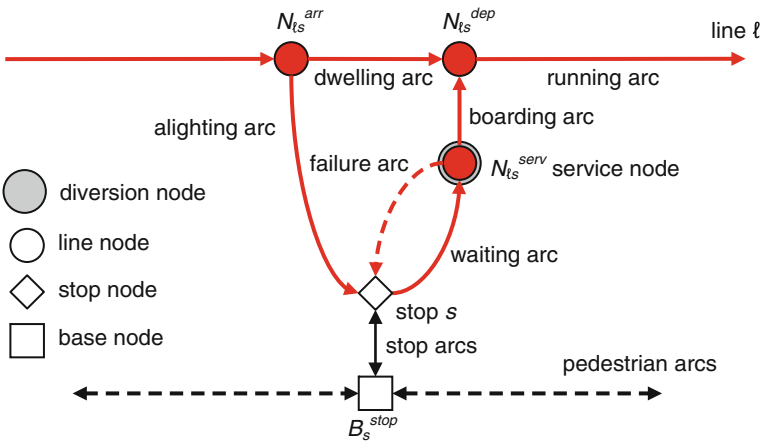

Fig. 7.5 The stop topology to reproduce the fail-to-board probability

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 Sect. 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 Fig. 6.6 are required in the stop topology to reproduce the fail-to-board probability (see Fig. 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* $A^{fail} = \{(N_{\ell s}^{serv}, s): \forall s \in S_\ell - S_\ell^+, \forall \ell \in L\}$; and
- the *service hyperarcs* $H^{serv} = \{\{(N_{\ell s}^{serv}, N_{\ell s}^{dep}), (N_{\ell s}^{serv}, s)\}: \forall s \in S_\ell - S_\ell^+, \forall \ell \in L\}$.

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 versus 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:

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

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 Sect. 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 Eq. (6.67f), where the non-temporal cost is given by the risk of failure:

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

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

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 $f_{\ell s} \cdot (1 - p_a^{fail})$ 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, Eq. (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 Eq. (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 Fig. 6.3 can be applied considering the sequential model based on hyperarcs of Sect. 6.1.5.

The example in Sect. 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 would 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 Sect. 6.3.1, the following arc and hyperarcs are introduced or modified:

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

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:

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

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 (oversaturation 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

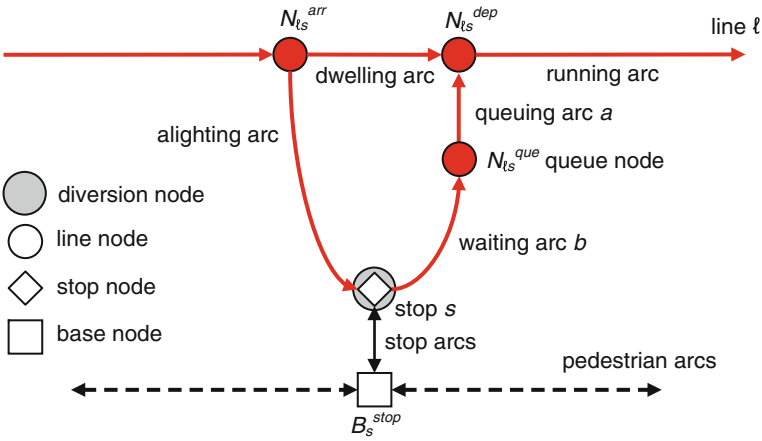

Fig. 7.6 The stop topology to reproduce FIFO queuing

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 Fig. 6.6 or Fig. 7.1 are then required in the stop topology to reproduce FIFO queuing (see Fig. 7.6). To separate the oversaturation 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* $A^{wait} = \{(s, N_{\ell s}^{que}): \forall s \in S_\ell - S_\ell^+, \forall \ell \in L\}$;
- the *queuing arcs* $A^{que} = \{(N_{\ell s}^{que}, N_{\ell s}^{dep}): \forall s \in S_\ell - S_\ell^+, \forall \ell \in L\}$.

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 Sect. 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 us 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 Sect. 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:

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

Let us 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 us define $\kappa_a^{cum}(\tau)$ as the cumulative remaining capacity:

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

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

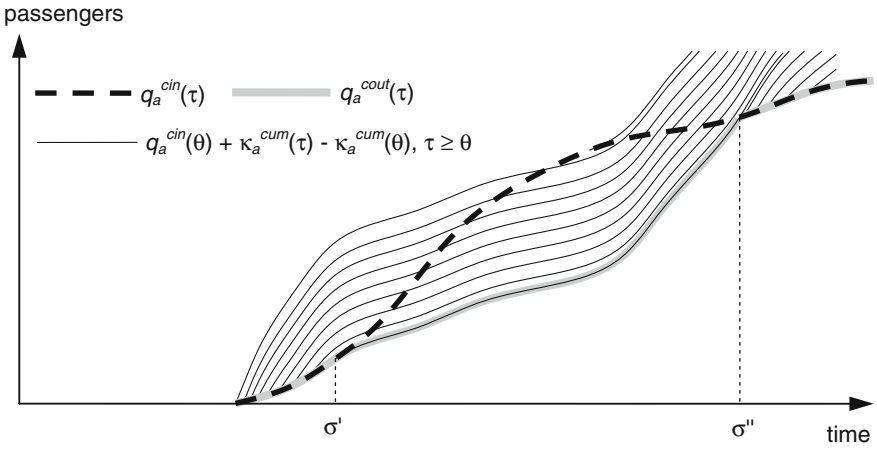

Fig. 7.7 Bottleneck with time-varying capacity. The cumulative outflow is the lower envelop of the profiles family for each θ, with τ ≥ θ, obtained from the vertical translation of the cumulative remaining capacity that goes through point (θ, $q_a^{cin}(\theta)$). No queue is present when $q_a^{cin}(\tau)$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

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 Fig. 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:

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

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 $\theta_a(\tau)$:

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

  If there is no oversaturated 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 us 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 the presence of oversaturation 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 Sect. 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 correspond to the best possible regularity:

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

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 σ_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 $\theta_a(\tau) - \tau$ 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 such as (7.72) and (7.73) in the equations of Sect. 7.1, the combined wait times $t_{b|b^{\check{}}}(\tau)$ conditional to take line $b \in b^{\check{}}$ and the line shares (diversion probabilities) $p_{b|b^{\check{}}}(\tau)$ can be obtained for each hyperarc $b^{\check{}} \in s^{+}$ of stop $s \in S$. Again, in the conditional exit times, the queuing times have to be deducted:

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

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 with respect to 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 Sect. 6.4.4. A possible approximation which may simplify the model is as follows: $t_{b|\check{b}}(\tau) = t_{\check{b}}(\tau)$.

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, such as the static transit network of Fig. 6.6, by adapting the macroscopic model for dynamic transit assignment of Sect. 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:

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

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 $\rho_{br} \leq \theta_{rs} - t_\ell^{board}$ 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:

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

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 us initialize $\rho_{b0} = -\infty$. The passengers willing to take line ℓ that arrive at the stop later than ρ_{br-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:

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

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 Sect. 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

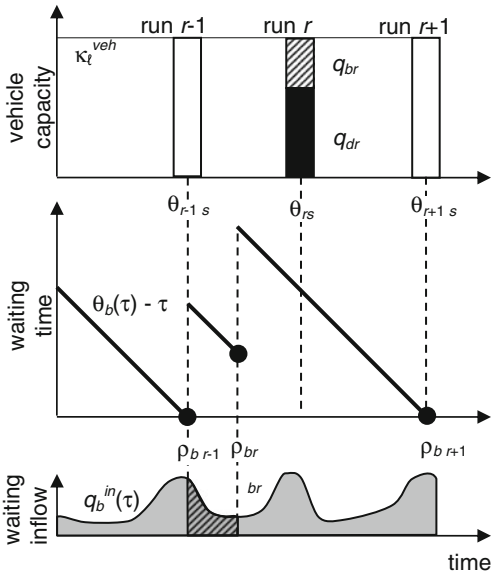

Fig. 7.8 Saw-toothed waiting time for given residual capacities and inflows; here $t_\ell^{board} = 0$

saw-tooth shape depicted in Fig. 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:

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

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 $q_{ar}/\kappa_\ell^{alight}$ 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:

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

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 Sect. 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 Sect. 6.4.5, with the possibility of representing also its modulation from stop to stop due dynamic phenomena, including dwelling congestion (see Sect. 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 instead the approach of fail-to-board probabilities to quasi-dynamic model; Schmoeker et al. (2008) extended 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 multi-commodity flow model with strict capacity constraints.

Tian et al. (2007) introduced in-vehicle congestion through a bulk-queue model and analysed 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 Sect. 7.3.3 finds its roots in this work.

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

7.3.6.2 FIFO Queuing

An early attempt to model FIFO queues was made by Bouzaiene-Ayari et al. (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 general 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 Sect. 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 Sect. 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, b) extended this approach to strategies and information.

7.4 Service Perturbations

Ektoras Chandakas, Moshen Babaei, Oded Cats, Pieter Vansteenwegen and 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; and
- reliability and robustness of transit networks also with respect to 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 disfunctionalities 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, and 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 the 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

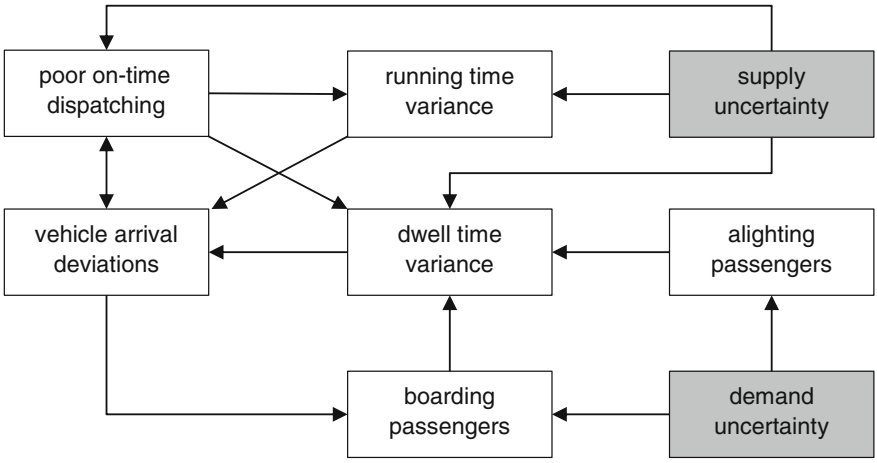

Fig. 7.9 The flow diagram of the interactions between service irregularity and the uncertainties on both demand and supply sides

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).

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 min 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) $\sigma^2 = \mathrm{Var}(h)/E(h)^2$, 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 timetable, 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 $s \in S_\ell - S_\ell{}^+$ 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 Sect. 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 (e.g., 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, as follows:

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

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

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 become 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:

$$E(n) = E_h(E(n|h)) = E(q \cdot h) = q \cdot E(h), \tag{7.87}$$

$$\begin{aligned} Var(n) &= E_h(Var(n|h)) + Var_h(E(n|h)) = E(q \cdot h) + Var(q \cdot h) \\ &= q \cdot E(h) + q^2 \cdot Var(h). \end{aligned} \tag{7.88}$$

Compared to the case of constant (non-random) headway of Eq. (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 Eqs. (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 Eqs. (7.85) and (7.86):

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

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

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 $E(t)$:

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

Compared to the case of constant (non-random) headway of Eq. (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 formalize 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., $h_\ell \leq 15$ min) 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 $t_{r\,i}^{run}$ and dwell times t_{ri}^{dwell} of previous stops $i \leq s$ as follows.

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

The (departure) headway $h_{rs} = \theta_{rs} - \theta_{r-1\ s}$ 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:

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

Let us 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 as follows: $E(n_{rs}) = q_s \cdot h_{rs}$;
- the preceding vehicle run followed the planned headway so that $\alpha_{r-1\ i} = 0, \forall i = j, \ldots, s$;
- running times between stops are assumed to be independent of headways and constant among runs; and
- 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 Eq. (7.93) to two consecutive stops i and $i + 1$ under the assumption of constant running time among runs, we get:

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

Dividing each side by the planned headway h_ℓ and using $t^{dwell}_{rs} = t_0 + q_s \cdot h_{rs}/\kappa$, 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 $h_{r-1\ i+1} = h_\ell$; 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 $i = j + 1, \ldots, s$, by induction, we get Eq. (7.95):

$$\begin{aligned}
\alpha_{rj+1} &= \frac{\alpha_{rj}}{\left(1 - \frac{q_{j+1}}{\kappa}\right)} \\
\alpha_{rj+2} &= \frac{\alpha_{rj+1}}{\left(1 - \frac{q_{j+2}}{\kappa}\right)} = \frac{\alpha_{rj}}{\left(1 - \frac{q_{j+1}}{\kappa}\right) \cdot \left(1 - \frac{q_{j+2}}{\kappa}\right)}. \\
\alpha_{rj+3} &= \frac{\alpha_{rj+2}}{\left(1 - \frac{q_{j+3}}{\kappa}\right)} = \frac{\alpha_{rj}}{\left(1 - \frac{q_{j+1}}{\kappa}\right) \cdot \left(1 - \frac{q_{j+2}}{\kappa}\right) \cdot \left(1 - \frac{q_{j+3}}{\kappa}\right)}.
\end{aligned} \tag{7.99}$$

Note that the product in Eq. (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 are 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 Sect. 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 timetable. 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; and
- 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 Fig. 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:

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

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 and TRB 2003) suggests a range of values for the parameters of Eqs. (7.100) and (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 s/pax for boarding and 2.1–3.3 s/pax for alighting; for the rail and metro, values in the range of 1.38–3.97 s/pass for boarding and 1.11–4.21 s/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:

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

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 Eq. (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 the interaction among individual vehicles and travellers are 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 Sect. 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 us 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 Eqs. (7.100) and (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:

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

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:

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

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 $1/h$):

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

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:

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

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 affect 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 realize 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, for example:

- the percentage of services arriving on-time;
- the percentage of services arriving more than 5 min late; and
- 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; and
- 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 (such as 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) 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 Sect. 7.4.3 is an original contribution of this book.

7.4.7.2 The Dwelling Process

Whereas Eq. (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.

First, 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 problems (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 and Iravani 2002). They are further influenced by physical factors, such as the vertical gap between the vehicle and the platform and the door width (Fernandez 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 Wirasinghe (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 data set 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 microsimulators 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 (1990) and Asakura (1996).

The distinction between punctuality and regularity and their determinants is discussed in Okrent (1974), Bowman and Turnquist (1981), Carey (1999), and 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. (2014).

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 and Kratz (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 and Tozzi 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) formulated 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 model 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) proposed a stochastic model for frequency-based assignment. In addition, Meschini et al. (2007) proposed 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 signalling 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).

7.5 Fares

David Watling, Guido Gentile, Klaus Noekel and 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, for example:

- 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, and
- smart cards with a maximum daily fare.

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 summarized 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 generalized 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 interrelated. 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 generalized cost. We assumed generalized 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 Sect. 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 generalized cost and will take effect in route choice.

Many fare schemes are not additive, however, including simple schemes such as 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 Sect. 6.3, this does not pose a problem, because the evaluation proceeds in three distinct steps:

1. search paths,

2. calculate generalized cost for each path, and
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 O–D 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 generalized costs per path, and we can apply a choice model in step 3.

Sections 6.2 and 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, and 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 Sect. 5.13 emended by the inclusion of Line 5—purple and Stop 5, as in Sect. 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, and 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).

Table 7.12 Line shares (%) with and without the effect of fares

Fares	Ignored	Exact solution	Exact solution
Origin of travellers	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

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 Sect. 7.1.5. The second column of Table 7.12 repeats the last column of Table 7.6, which ignores 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.

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. For 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

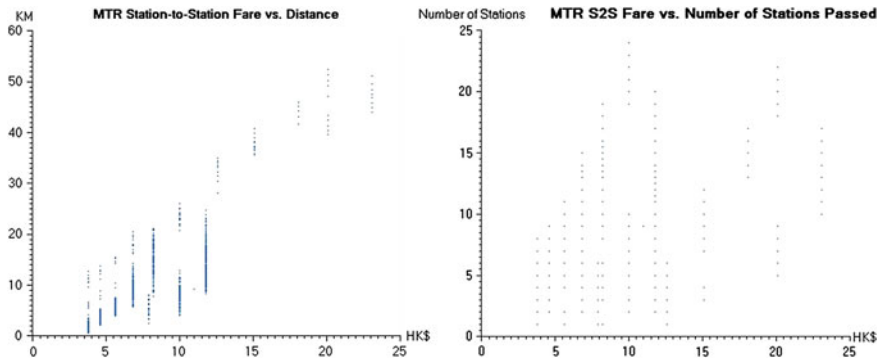

Fig. 7.10 From station to station matrix of Hong Kong MTR shows non-additive fares that are nonlinear with respect to distance nor with respect to the number of station passed (these plots were provided by Michael Florian, INRO)

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 non-trivial.

In particular, integrated fare schemes cannot be easily reproduced through the arc cost model presented in Sect. 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 Fig. 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.

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 (e.g., 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 depend 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 versus 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.

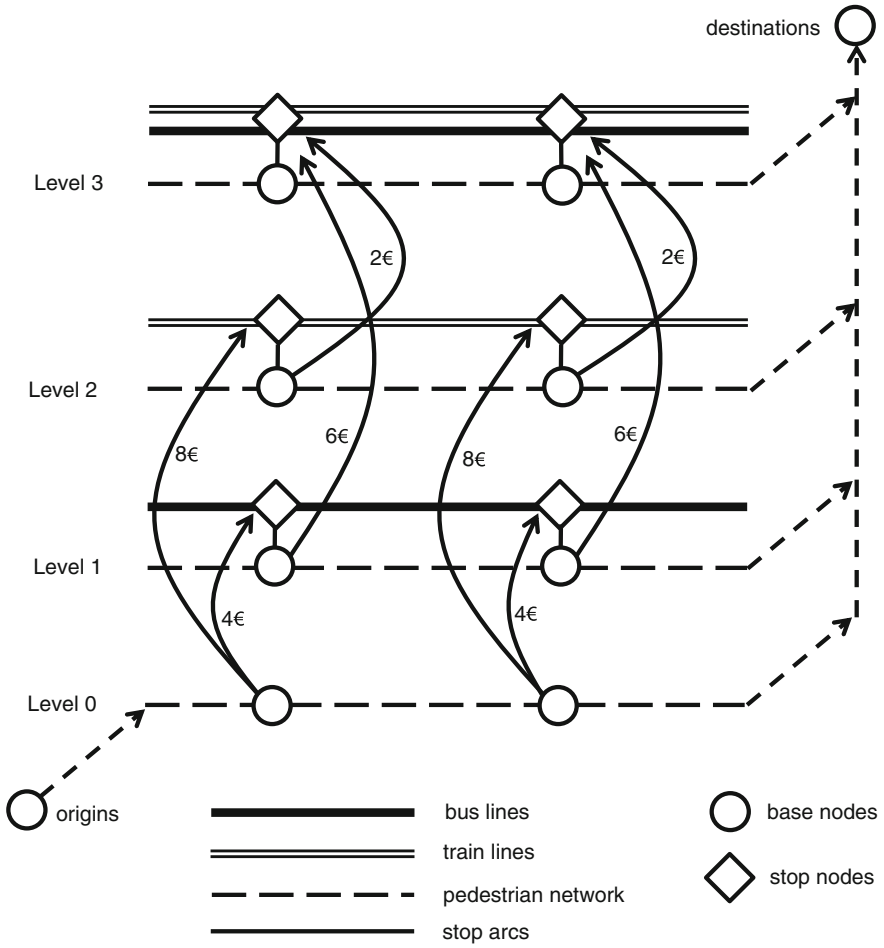

Fig. 7.11 Simulation of fare discount for getting both trains and busses through journey levels

- 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 Fig. 7.11).

Note that the journey-level approach can be seen as an extension of the multi-modal network approach presented in Sect. 5.1.1.2, based on the duplication of transport system sub-networks. 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 with respect to 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 and 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 (2015). The same approach has been applied to multi-modal journeys by Lo et al. (2003, 2004).

References

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

Abkowitz M, Tozzi J (1987) Research contributions to managing transit service reliability. J Adv Transp 21:47–65

Adebisi O (1986) A mathematical model for headway variance of fixed-route buses. Transp Res B 20:59–70

Arentze TA (2013) Adaptive, personalized travel information systems: A Bayesian method to learn users' personal preferences in multi-modal transport networks. IEEE Trans Intell Transp Syst 14:1957–1966

Asakura Y (1996) Reliability measures of an origin and destination pair in a deteriorated road network with variable flows. In: Bell MGH (ed) Transportation networks: recent methodological advances. Pergamon Press, Oxford, pp 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 ZH (2005) Algorithm for equilibrium transit assignment problem. Transp Res Rec 1923:227–235

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

Bell MGH (1995) Stochastic user equilibrium assignment in networks with queues. Transp Res B 29:125–137

Bell MGH, Schmocker J-D (2004) A solution to the congested transit assignment problem. In: Wilson NHM, Nuzzolo A (eds) Scheduled-based dynamic transit modeling: theory and applications. Springer, New York, pp 263–280

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

Bellei G, Gentile G, Papola N (2000) Transit assignment with variable frequencies and congestion effects. In: Proceedings of the 8th meeting of the EURO working group on transportation, Rome, Italy

Bertsekas DP (1999) Nonlinear programming, 2nd edn. Athena Scientific, Belmont, MA

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

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

Bouzaıene-Ayari B, Gendreau M, Nguyen S (2001) Modelling bus stops in transit networks: a survey and new formulations. Transp Sci 35:304–321

Bowman LA, Turnquist MA (1981) Service frequency, schedule reliability and passenger wait times at transit stops. Transp Res A 15:465–471

Carey M (1999) Ax ante heuristic measures of schedule reliability. Transp Res B 33:473–494

Carraresi P, Malucelli F, Pallottino S (1996) Regional mass transit assignment with resource constraints. Transp Res B 30:81–89

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

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. Transp Res B 40:437–459

Chen X, Yu L, Zhang Y, Guo J (2009) Analyzing urban bus reliability at the stop, route and network levels. Transp Res 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: Ahuja RK, Möhring RH, Zaroliagis CD (eds) Robust and online large-scale optimization: models and techniques for transportation systems, Lecture Notes in Computer Science, pp 28–60

Cominetti R, Correa J (2001) Common lines and passenger assignment in congested transit networks. Transp Sci 35:250–267

Constantin I, Florian D (2015) Integrated fare modelling with strategy-based transit assignment. In: Proceedings of CASPT15, Rotterdam

Cortés CE, Jara-Moroni P, Moreno E, Pineda C (2013) Stochastic transit equilibrium. Transp Res B 51:29–44

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

De Cea J, Fernandez JE (1993) Transit assignment for congested public transport systems: an equilibrium model. Transp Sci 27:133–147

Dewilde T, Sels P, Cattrysse D, Vansteenwegen P (2011) Defining robustness of a railway timetable. In: Proceedings of 4th international seminar on railway operations modelling and analysis, Rome, Italy, pp 1–20

Dewilde T, Sels P, Cattrysse D, Vansteenwegen P (2014) Improving the robustness in railway station areas. Eur J Oper Res 235:276–286

DfT (2007) Model structures and traveller responses for public transport schemes. Transport Analysis Guidance 3.11.1. UK Department for Transport, London, UK

Dial R (2006) A path-based user-equilibrium traffic assignment algorithm that obviates path storage and enumeration. Transp Res B 40:917–936

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

El-Geneidy AM, Horning J, Krizek KJ (2011) Analyzing transit service reliability using detailed data from automatic vehicular locator systems. J Adv Transp 45:66–79

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

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. In: Proceedings of TRB annual meeting, Washington, DC

Fritz M (1983) Effect of crowding on light rail passenger boarding times. Transp Res Rec 908:43–50

Garcia R, Marin A (2005) Network equilibrium with combined modes: models and solution algorithms. Transp Res B 39:223–254

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 (2014) Local user cost equilibrium: a bush-based algorithm for traffic assignment. Transp A 10:15–54

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

Goverde RMP (2005) Punctuality of railway operations and timetable stability analysis. Ph.D. thesis, Delft University of Technology, Delft

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

Hamdouch Y, Florian M, Hearn DW, Lawphongpanich S (2007) Congestion pricing for multi-modal transportation systems. Transp Res B 41:275–291

Harris NG (2005) Train boarding and alighting rates at high passenger loads. J Adv Transp 40:249–263

Harris NG, Anderson RJ (2007) An international comparison of urban rail boarding and alighting rates. In: Proceedings of the institution of mechanical engineers. Journal of Rail and Rapid Transit 221, 521–526

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

Hickman MD, Wilson NHM (1995) Passenger travel time and path choice implications of real-time transit information. Transp Res C 3:211–226

Holroyd EM, Scraggs DA (1966) Waiting times for buses in Central London. Traffic Eng Control 8:158–160

Horn MET (2003) An extended model and procedural framework for planning multi-modal passenger journeys. Transp Res B 37:641–660

Huang R, Peng ZR (2002) Schedule-based path-finding algorithms for transit trip-planning systems. Transp Res Rec 1783:142–148

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

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

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

Kroon L, Maroti G, Retel Helmrich M, Vromans MJCM, Dekker R (2008) Stochastic improvement of cyclic railway timetables. Transp Res B 42:553–570

Kurauchi F, Bell MGH, Schmöcker J-D (2003) Capacity constrained transit assignment with common lines. J Math Modell Algorithms 2–4:309–327

Lai YC, Wang SH, Jong JC (2011) Development of analytical capacity models for commuter rail operations with advanced signaling systems. In: Proceedings of the 90th annual meeting of transportation research board, Washington, DC

Lam WHK, Gao ZY, Chan KS, Yang N (1999) A stochastic user equilibrium assignment model for congested transit networks. Transp Res B 33:351–368

Lam WHK, Zhou J, Sheng Z-H (2002) A capacity restraint transit assignment with elastic line frequency. Transp Res B 36:919–938

Leurent F (2012) On seat capacity in traffic assignment to a transit network. J Adv Transp 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. In: Proceedings of the 88th annual transportation research board meeting, Washington, DC

Leurent F, Chandakas E, Poulhes A (2011) User and service equilibrium in a structural model of traffic assignment to a transit network. In: Procedia—social and behavioral sciences 20, Proceedings of EWGT2011, pp 495–505

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

Lin T-M, Wilson NHM (1992) Dwell time relationships for light rail systems. Transp Res Rec 1361:287–295

Lo HK, Chen A (2000) Traffic equilibrium problem with route-specific costs: formulation and algorithms. Transp Res B: Methodol 34:493–513

Lo HK, Yip CW, Wan KH (2003) Modeling transfer and non-linear fare structure in multi-modal network. Transp Res B 37:149–170

Lo HK, Yip CW, Wan QK (2004) Modeling competitive multi-modal transit services: a nested logit approach. Transp Res C 12:251–272

Marguier PHJ (1981) Optimal strategies in waiting for common bus lines. Master's thesis, Department of Civil Engineering, MIT, Cambridge

Marguier PHJ, Ceder A (1984) Passenger waiting strategies for overlapping bus route. Transp Sci 18:207–230

Meschini L, Gentile G, Papola N (2007) A frequency based transit model for dynamic traffic assignment to multimodal networks. In: Allsop R, Bell MGH, Heydecker BG (eds) Proceedings of the 17th international symposium on transportation and traffic theory (ISTTT). Elsevier, London, pp 407–436

Morichi S, Iwakura S, Morishige S, Itoh M, Hayasaki S (2001) Tokyo metropolitan rail network long-range plan for the 21st century. In: Proceedings of the 80th annual meeting of transportation research board, Washington, DC

Morosan CD, Florian M (2015) A network model for capped link-based tolls. EURO J Transp Logistics 4:223–236

Nguyen S, Pallottino S (1988) Equilibrium traffic assignment for large scale transit networks. Eur J Oper Res 37:176–186

Nguyen S, Pallottino S, Gendreau M (1998) Implicit enumeration of hyperpaths in a logit model for transit networks. Transp Sci 32:54–64

Nielsen OA (2000) A stochastic transit assignment model considering differences in passenger utility functions. Transp Res B 34:377–402

Nielsen OA (2004) A large-scale stochastic multi-class schedule-based transit model with random coefficients. In: Wilson NHM, Nuzzolo A (eds) Schedule-based dynamic transit modeling: theory and applications. Kluwer Academic Publisher, pp 53–78

Noekel K, Wekeck S (2009) Boarding and alighting in frequency-based transit assignment. Transp Res Rec 2111:60–67

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

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

Nuzzolo A, Crisalli U, Comi A, Rosati L (2013) An advanced pre-trip planner with personalized information on transit networks with ATIS. In: Proceedings of 16th international ieee conference on intelligent transportation systems, The Hague, pp 2146–2151

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

Osuna EE, Newell GF (1972) Control strategies for an idealized public transportation system. Transp Sci 6:52–72

Owen AD, Phillips GDA (1987) The characteristics of railway passenger demand: an econometric investigation. J Transp Econ Policy 21:231–253

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

Poon MH, Wong SC, Tong CO (2004) A dynamic schedule-based model for congested transit networks. Transp Res B 38:343–368

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

Rajbhandari R, Chien S, Daniel J (2003) Estimation of bus dwell time with automatic passenger counter information. Transp Res Record 1841, 120–127

Ren H, Gao Z, Lam WHK, Long J (2009) Assessing the benefits of integrated en-route transit information systems and time-varying transit pricing systems in a congested transit network. Transp Plann Technol 32:215–237

Schmöcker J-D, Fonzone A, Shimamoto H, Kurauchi F, Bell MGH (2011) Frequency-based transit assignment considering seat capacities. Transp Res B 45:392–408

Schmoeker J-D, Bell MGH, Kurauchi F (2008) A quasi-dynamic capacity constrained frequency-based transit assignment model. Transp Res B 42:925–945

Schobel A, Kratz A (2009) A bicriteria approach for robust timetabling. In: Ahuja RK, Möhring RH, Zaroliagis CD (eds) Robust and online large-scale optimization: models and techniques for transportation systems, Lecture Notes in Computer Science, pp 119–144

Shimamoto H, Kurauchi F, Iida Y (2005) Evaluation on effect of arrival time information provision using transit assignment model. Int J ITS Res 3:11–18

Shimamoto H, Kurauchi F, Schmöcker J-D (2010) Transit assignment model incorporating the bus bunching effect. In: Proceeding of 12th world congress on transport research, Lisbon, Portugal
Spiess H, Florian M (1989) Optimal strategies: a new assignment model for transit networks. Transp Res B 23:83–102
Sumalee A, Tan ZJ, Lam WHK (2009) Dynamic stochastic transit assignment with explicit seat allocation model. Transp Res B 43:895–912
Szeto WY, Solayappan M, Jiang Y (2011) Reliability-based transit assignment for congested stochastic transit networks. Comput-Aided Civil Infrastruct Eng 26:311–326
Szplett D, Wirasinghe SC (1984) An investigation of passenger interchange and train standing time at LRT Stations: alighting, boarding and platform distribution of passengers. J Adv Transp 18:1–12
Teklu F (2008) A stochastic process approach for frequency-based transit assignment with strict capacity constraints. Netw Spatial Econ 8:225–240
Tian Q, Huang H-J, Yang H (2007) Commuting equilibria on a mass transit system with capacity constraints. In: Allsop R, Bell MGH, Heydecker BG (eds) Proceedings of the 17th international symposium on transportation and traffic theory (ISTTT). Elsevier, London, pp 261–384
TRB (2003) Transit capacity and quality of service manual. On-line report prepared for the Transit Cooperative Research Program
Trozzi V, Gentile G, Bell MGH, Kaparias I (2013a) Dynamic User Equilibrium in public transport networks with passenger congestion and hyperpaths. Transp Res B 57:266–285
Trozzi V, Gentile G, Bell MGH, Kaparias I (2013b) Effects of countdown displays in public transport route choice under severe overcrowding. Networks and Spatial Economics (published on-line)
Van Oort N (2011) Service reliability and urban public transport design. Ph.D. thesis. Netherlands TRAIL Research School, Delft
Van Oort N, Van Nes R (2009) Regularity analysis for optimizing urban transit network design. Public Transport 1:155–168
Vuchic VR (2006) Urban transit: operations, planning and economics. Wiley, New York
Whelan G, Johnson D (2004) Modelling the impact of alternative fare structures on train overcrowding. Int J Transp Manage 2:51–58
Wu JH, Florian M, Marcotte P (1994) Transit equilibrium assignment: a model and solution algorithms. Transp Sci 28:193–203
Yang L, Lam WHK (2006) Probit-type reliability-based transit network assignment. Transp Res Rec 1977:154–163
Yin Y, Lam WHK, Miller MA (2004) A simulation-based reliability assessment approach for congested transit network. J Adv Transp 38:27–44
Zhang Q, Han B, Li D (2008) Modeling and simulation of passenger alighting and boarding movement in Beijing metro stations. Transp Res C 16:635–649

Part IV
Applications and Future Developments - Fabien Leurent

Chapter 8
Applications and Future Developments: Modelling the Diversity and Integration of Transit Modes

Ingmar Andreasson, Fabien Leurent, Francesco Corman and Luigi dell'Olio

Passenger transit modes typical of the urban setting, including bus, tram, metro, and train, have been described in Chap. 2, along with less conventional modes such as BRT and cable. Then, Chaps. 6 and 7 have provided network assignment models that address primarily the passenger side, dealing with route choice situations and behaviour, the individual exposure to traffic conditions, and the contribution of individual users to local flows. In these models, the transit mode is represented as a set of lines, each of which is abstracted into its topology (nodes and links) and some features of traffic operations: run time, dwell time, and some capacity parameters. In such an abstract setting, no distinction has been made between, for instance, bus and railway services, apart from their respective parameter values.

So our aim now is to focus on the side of transit supply and its diverse modes, in order to capture their respective operational features and also to address their joint modelling in an integrated system.

Indeed, transit services are delivered at best performance for both the users and the environment by high capacity modes with high quality of service. However, to bring high flows of passengers to high capacity services (guideway modes such as train, metro, and BRT) requires to feed them timely by less capacitated services,

F. Leurent (✉)
Laboratory on City, Mobility and Transportation, University Paris-East, Ecole des Ponts ParisTech, Paris, France
e-mail: fabien.leurent@enpc.fr

L. dell'Olio
University of Cantabria, Av. de los Castros s.n., 39005 Santander, Spain
e-mail: delloliol@unican.es

I. Andreasson
LogistikCentrum Göteborg AB, Osbergsgatan 4A, 426 77, V Frölunda, Sweden
e-mail: ingmar@logistikcentrum.se

F. Corman
Delft University of Technology, Mekelweg 2, 2628, CD Delft, The Netherlands
e-mail: f.corman@tudelft.nl

G. Gentile and K. Noekel (eds.), *Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems*, Springer Tracts on Transportation and Traffic 10, DOI 10.1007/978-3-319-25082-3_8

from bus to demand responsive in order to extend the range of walk access to guideway stations. Thus, large transit networks are made up of a variety of modes—often called transit submodes. The proper use of their hierarchy is key to deliver good performance throughout space and not only along major corridors. Intermodality is a key feature of multimodal integration, along with the coordination of service operations. In practice, it depends on specific arrangements—layout, service coordination—that must be planned and implemented to deliver good performance.

To meet its objectives, this chapter is organized in four sections. Section 8.1 addresses line-haul modes and emphasizes their operational features. The main conditions of operations are discussed, including the various interactions between passenger flows and service vehicles and also the operating mode (e.g., man-controlled versus automated), with their modelling implications. Complementarily, Sect. 8.2 deals with the coordination of line operations: their joint efficiency for trips that use a sequence of line legs involves efficient transfers, due to optimized layout as well as to timed transfers and coordinated timetables.

Then, Sect. 8.3 is devoted to the modelling of demand-responsive and paratransit services, including feeder and shuttle services, cable-propelled transport, bus-on-demand and special transportation services, taxi, and personal rapid transit (PRT). PRT is described in-depth as this transit mode combines characteristics of the car and the train; thus, it is well suited to act as a feeder/distributor to mass transit. The relevant strategies of operation are discussed; their application is based on the general dial-a-ride problem (DARP), which is presented along with its solution methods.

Finally, Sect. 8.4 addresses the integrated modelling of travel demand and transit operations. Of course, micro-simulation can be used to model the passenger and vehicle movements of all modes. However, less complex macroscopic models can be relevant to depict features of a massive kind such as those pertaining to flows and capacities. Integrated modelling is intended to take the best out of the two approaches, by suitable combination. The section presents first several bi-level models of line-haul transit modes, then some integrated models of multimodal systems—notably a hybrid transit system including line-haul and demand-responsive services—, and lastly some models that integrate transit assignment with the optimization of network design or control.

8.1 On Line-Haul Operations

Fabien Leurent

Line-haul along a predetermined route is the basic form of mass transit for passengers, be it on roadway or railway. A line-haul service (or mission) involves vehicle runs to serve a given route at given stops, with predetermined headways—at least at the line departure terminal. In bus rapid transit, as well as in railway transit,

a so-called line may involve several services with different routes and express services versus omnibus, or along different branches grafted in some common trunk. The sharing of operational facilities such as station platform or service way (e.g., railway track or dedicated lane on roadway) implies a high level of coordination between the services that make up the line.

This section addresses the modelling of a line in order to capture its traffic operations. It deals with a number of interactions between service vehicles and passenger flows, which have been identified in the previous chapter on the passenger side, so as to mark their incidences on the operator side depending on the transit mode.

8.1.1 Different Modes

In a classical mode of urban bus, most lines are operated with one service only and can be modelled as a couple of unidirectional services. Most stations' stops are relatively small, meaning that small numbers of passengers alight or board there so that the associated facility is often restricted to some road sign on the kerb side. The stopping there of a vehicle is made conditional on passenger request from on-board or on-street.

More important stops, including terminal stations of roadway lines as well as railway stations, are much more developed: a typical railway station includes passenger areas as line platforms for passenger alighting and waiting prior to boarding, as well as corridors for pedestrian access between platform(s) and the outside of the station.

In their present form, transit assignment models deal with each stop in a uniform way by associating it with a given dwell time which may be related to passenger flows.

Between two stops, buses run along roadway sections with mixed right-of-way (RoW) in the classical case, or segregated RoW in the case of bus rapid transit which is isolated from interaction with general roadway traffic. The section times are modelled as average quantities exogenous or endogenous in macroscopic assignment. Only micro-simulation can address the randomness in section running and the specific advantages of segregated RoW in this respect.

The same applies to bus priority at junctions, which reduces the randomness of junction crossing: this can be captured explicitly only by micro-simulation. Dynamic macroscopic assignment, dealing with time slices of 5–15 min, cannot detail the roadway performance of buses more than static macroscopic assignment —apart from the time variations of frequencies and passenger flows.

Table 8.1 Main characteristics of urban modes of passenger transit in greater Paris (2010)

Mode	Seats/veh	Nominal in-veh K	Streams per side	Peak frequency (veh/h)	Interstation spacing (km)	Mean speed (km/h)
Bus	30	100	3	20	0.3	16
Rapid bus (TZen)	45	150	6	20	0.5	18
Tramway	100	400	8–20	20–25	0.5	19
Metro	250	600	30–48	20–40	1	22
Regional express railway (RER)	500–800	1800–2800	30–64	20–30	3	25–30
Commuter rail (Transilien)	400–1600	800–2600	40–50	10	5	30–40

Guideway transit is more regular than roadway traffic and can involve much larger vehicle capacities. Every station along a service route will give rise to vehicles stopping there, even in the rare case of no changing passengers. The vehicle is a large one—say a train—due to its length and maybe also to double decks rather than single. Between stations, the RoW is segregated from any other flow in the case of a subway. However, many railway services use sections that are shared with other services, inside the same line or outside. In this respect, two lines that have distinct sets of services to users but share some track section should be modelled in some integrated way, since their joint operations influence their respective performance.

Table 8.1 indicates some characteristics typical of transit modes as of 2010 in greater Paris.

8.1.2 In-Vehicle Passenger Traffic

Recently, seat capacity has been recognized as an important feature of transit services in a number of assignment models (cf. previous chapter). Passengers are more comfortable at sitting than at standing. However, limited spacing between opposite rows of seats as well as seat crowding makes sitting less comfortable. The sensitivity of sitting discomfort to seat occupation is easy to model per unit of run time. So is that of standing discomfort to the number of standing passengers. Given a nominal capacity of 4 p/m^2, when the flow ratio compared to nominal capacity is varied from 0 to 1, then standing discomfort cost may change from, say, 1.2–2 times that of sitting discomfort cost.

The notion of standing capacity is questionable since in large cities of developing countries the western standard of 4 p/m^2 is often exceeded by far—up to 9 in

Indian cities. Even in developed countries, under severe conditions (high demand and/or disruption), values of 6 or 7 can be observed. So the modelling of total vehicle capacity on the basis of a stiff nominal capacity requires to exert caution in the analysis of results, especially so to compare the in-vehicle and platform conditions since every user will make their own trade-offs between in-vehicle discomfort and platform wait time.

8.1.3 Passengers on Platform

Let us assume that in a station, a given line has a given platform for passenger storage and waiting prior to boarding. This is typical of urban transit by its main modes, as opposed to large stations of interurban railways where train runs serving a reiterated daily schedule are assigned platform slots that may vary from day to day.

If the line comprises several services, then these share the same platforms: this is indeed a strong feature as it imposes a first in–first out discipline between the service runs that stop there. The headways between them are likely to be regularized by the operator: although this differs from the Markovian assumption which is classical in frequency-based assignment models, it will lead to essentially the same split of passenger flow between alternative services serving a given station for exit down the current one, i.e., proportionally to the relative frequencies in the absence of vehicle saturation.

On the passenger side, the physical wait time depends on the users' instant of arrival, on that of vehicle arrival and the availability of a place. In the uncongested case, waiting users may avail themselves of platform seats, thus reducing the discomfort of waiting. Dynamic traffic information (DTI) with accurate prediction of the time of vehicle arrival is useful to reduce the discomfort of waiting, too. If the headways are irregular, then the relative discomfort of wait time compared to in-vehicle time may amount to a factor 2 or 3: adequate DTI decreases it down to 1. In the congested case, however, users must compete to take their chance to get in the next incoming vehicle: so they have to stand up and maintain their relative rank. For a bus or cable line under congestion, there is a first in–first out discipline among the waiting users, i.e., they form a queue. For a railway line, as a train has a large number of doors, waiting passengers can be assumed to be mingled rather than queued—at least if the platform does not have automated doors. Automated doors will give rise to local queues under platform congestion, thus to FIFO in a decentralized fashion. Whatever the waiting discipline, under congestion the physical wait time is increased; furthermore, its relative discomfort is accrued since the users must stand up and keep aware of vehicle arrivals. The waiting conditions are all the more severe as platform space is more constrained, possibly leading to

platform crowding and passenger impedance both with other waiting passengers and with the flows of passengers at alighting (and also at boarding on vehicles that do not serve all of the exit stations required by the users).

All of these issues could be addressed in an assignment model that captures vehicle capacity through a wait discomfort function. Such a function should come from a field survey among concerned users. By default, the discomfort function of in-vehicle standing may be applied. The outcome is a penalty function by unit of time at waiting with respect to passenger density on platform. The main issue is to model the physical wait time correctly. The models that impose a total vehicle capacity (see previous chapter) are relevant for platforms without automated doors. Automated doors on platform play the same role in transit traffic as junction controls in roadway traffic, as they can be used to limit the dwell time (analogous to green time at a traffic light). In that case, platform congestion may stem from the limitation of the exchange capacity between vehicle and platform, rather than from total in-vehicle capacity (of which the relative nature deserves to be mentioned here again).

With automated doors and under fixed dwelling time, the exchange capacity amounts to the product of that time by the vehicle exchange capacity per time unit. The latter depends on the number and width of the vehicle doors: it takes about 80 cm to make one passenger stream through a vehicle door and about 1.5 s per passenger alighting or boarding to pass the door. Denoting by S the number of passenger streams on a vehicle side and by h the individual time of door passage, the exchange capacity per unit of dwell time amounts to S/h for that vehicle.

At a given instant, the stock of passengers waiting on platform is related to vehicle headways. Other things equal, a higher service frequency will reduce the size of the stock in an inversely proportional way.

8.1.4 On Dwell Time, Service Frequency, and Run Delay

At a station where a transit vehicle stops, the flows of alighting and boarding passengers require a dwell time that is proportional to them (if the operations are controlled by man) and inversely proportional to the exchange capacity. Automated operations may limit the dwell time strictly. The flow dependence of dwell time is a major phenomenon of transit traffic, which has been captured in assignment models long ago, notably so to make it contribute to the vehicle run time by leg (a pair of stations for passenger's entry and exit), hence to the user's physical run time along the leg.

Congestion may also affect the run time by section, especially so for a roadway service with shared RoW along its route.

Then, congested dwell times and run times can interact with frequency in two ways. First, as the service fleet is limited, an increase in-vehicle cycle time will entail an inversely proportional reduction in the frequency that is effectively delivered: this has been addressed by Bellei et al. (2000) and Lam et al. (2002) in

static models, then by Meschini et al. (2007) in a dynamic setting. Second, in guideway operations, the track element where vehicles stop for station dwelling is occupied also for vehicle clearance and safety margins between vehicles. If, over a given period, the service requirements of occupation time exceed the period duration, then the service frequency is decreased proportionally to the period duration divided by the required time (Leurent et al. 2011). This frequency reduction is propagated down the station for all of the services that pass there, and a related delay is induced for the vehicles upstream (Leurent et al. 2014).

8.1.5 Traffic Interactions Along a Transit Line

To sum up, it comes out that vehicle operations and passenger flows interact along a transit line in a number of respects. Vehicle operations involve vehicle features such as in-vehicle capacity and exchange capacity, as well as higher level set-ups such as the set of services that make up the line and their respective route and nominal frequency. To the user, quality of service involves the physical run time along the required leg—essentially a vehicle attribute—and the leg comfort which depends on vehicle attributes either predetermined (seat capacity, standing space) or endogenous (passenger load), as well as the wait time before boarding which depends on service frequency, platform crowding, and the residual on-board capacity.

It is important to model this variety of traffic phenomena in an integrated way, especially so in the design of a new transit line or of a major improvement to an existing line. The modelling may be restricted to the line (cf. the line model in Leurent et al. 2014). However, in a large network that serves many origin-destination flows and includes many route options for user path choice, it is relevant to model the transit network and its traffic assignment, so as to allow the traffic conditions of the targeted line (i.e., leg trip flows, leg generalized costs, and vehicle operations) to depend on the performance of not only the line but also the other services that can be used as substitutes or complements.

8.2 Service Coordination

Fabien Leurent, Luigi Dell'Olio, Maria Bordagaray and Ingmar Andreasson

A transit line that comprises several services requires coordination between them for the allocation of shared resources such as station platform and track elements. More generally, a transit network is made up of a number of lines which complement each other in order to render service to users in a large, bi-dimensional geographical space that extends far beyond the corridor along each line.

Transit operations can be optimized at several levels to develop the advantages of complementariness: by integrated fares (see Sect. 7.5 in previous chapter), by matched transfers and line coordination (see below), or by integrated management of assets (crew, vehicle fleet, depots).

This section addresses first transfer optimization taking a broad perspective, then matched transfers, and finally timetable coordination.

8.2.1 Transfer Optimization

For transit to be an effective travel mode between an origin (access node) and a destination (egress node), it requires that each line be efficient along its route(s) notably so by rapid runs (i.e., competitive crow fly speed from point to point) and sufficiently high frequency, and also that transfers between lines be efficient at the relevant places. This is achieved first by station design and the organization of pedestrian paths from platform to platform: such paths should demand little time and effort on the user, i.e., with short distance, unimpeded route, few changes of floor, and mechanical ways for that (TCQSM 2003).

The organization of efficient transfers between lines also involves the coordination of their respective runs in order to reduce the wait time of transferring users. If the frequency of each line is high (say above 10/h), then the coordination reduces to ensure the regularity of each line. Under lower frequencies, it is important to arrange for "simultaneous" arrivals and departures of transit units at the station of transfer. This is indeed a very demanding objective since each vehicle should dwell during a time that would enable users not only to alight and board but also to move from one platform to another and furthermore to wait for the other vehicle to arrive. As long dwell times would impair line efficiency, the quality objective requires to limit the transfer time to around 1 min—indeed an issue of station layout—and to enforce simultaneous arrivals as much as possible.

8.2.2 Matched Transfers

A matched transfers or timed transfer approach is a strategy to develop a transit network in which transit units arrive simultaneously at transfer stops to offer coordinated transfers in all directions (Systan Inc. 1983; Vuchic et al. 1983).

The idea underlying these models is the modification, control, and/or design of the intervals of the various transit vehicles for them to arrive at the same time to certain stops or terminals (Chung 2009) or within a given time frame (transfer slack time, TST), in such a way that "synchronization" between lines or even transit operators will be possible. Timed transfer-based systems are the most widely used systems in combined transport modes (for instance, rail + feeder services) where this has to be considered.

The first models arisen in the mid-1980 s focused on TST optimization under different objective functions. Thus, Hall (1985) determined the TST value that minimized the passengers' delays, obtaining the influence of the transit line interval on the TST value. This conclusion was confirmed by Lee and Schonfeld (1991) some years later, suggesting the conjoint TST-intervals optimization after analysing these times with three different interval distributions.

On this basis, Lee (1993) developed a combined model for optimizing the TST and the intervals, analysing several cases where he considered various ways for coordinating transit routes: no coordination, partial coordination, or total coordination. From this analysis, Lee concluded that a timed transfer system is advisable for services with high interval times and, as in the preceding studies, determined that the arrivals variability was determinant for the efficiency of this kind of systems. A similar approach was used by Ting (1997), with a cost function (user + operation) to be minimized, reaching to conclusions very similar to those of Lee.

Specific studies (Chien and Chowdhury 2002; Ngamchai and Lovell 2003) have addressed the problem by using different resolution methodologies, such as genetic algorithms, in which once more the same conclusions were obtained:

- The systems based on optimizing TST, or timed transfer systems, are advisable for services with high interval times.
- The less variability in the arrivals of the services, the more efficient its implementation.

8.2.3 Coordinated Timetables

The main objective of the coordination of the schedules of two different transport systems is the minimization of transfer time. In the case of just one transport system in a corridor, it is vital to avoid transfers and, if this is not possible, minimize the transfer time. In addition to ensuring acceptable transfer times, it is necessary to minimize the waiting time in the trip origins and the in-vehicle times.

These aspects take into account the user point of view, since they set up the total cost for travelling by public transport; however, the operator costs have to be also considered in the problem. The transport operator tries to minimize the total distance or time travelled by their vehicles while ensuring their highest possible occupancy.

The complexity of the problem is the conflict of interests between the users and the operator, who take part simultaneously in the problem to be solved: user costs decrease as the fleet and service frequency increase, while the operator tries to minimize both parameters. Many variables and constraints are involved in the scheduling design problem, which limit the available tools to address it. In the literature, we can find two prevailing techniques: computer simulation and optimization, the latter being the approach of most success in solving the scheduling

problem. In addition, it is a nonlinear programming problem where gradient-based methods can obtain a local optimum (not global) in the case of non-convex functions. On the contrary, genetic algorithms have showed to overcome these limitations and solve multiobjective problems such as the schedule coordination problem. Along with heuristic approaches (Shrivastava and Dhingra 2001), this is the most used technique (Shrivastava and Dhingra 2002; Shrivastava and O'Mahony 2006).

Genetic algorithms are based on random searches where decision variables are defined by sequences of binary elements which resemble mimic chromosomes and genes of genetic evolution in nature. Each sequence configuration, that is, the different variable combinations, is assessed in the objective function. In addition, three methods for generating new combinations are applied to each sequence: reproduction, crossover, and mutation. The first one gets the best sequence, the second one exchanges information between sequences, and the last one within a sequence. These phases make possible to look for a global optimum, not a local optimum.

The data required to address the scheduling problem are as follows:

- Vehicle routes and travel times between nodes or stations of the routes. The route design is also a problem in itself, and some studies can be found in the literature, which have addressed this problem conjointly with the scheduling problem.
- Demand, that is, the number of trips made between each pair of stations or stops. Since the demand is correlated with the supply, an iterative process is more realistic. Notwithstanding, in practice, the demand is usually set for different time ranges throughout the day and different types of day. In the case of scheduling coordination of two transport systems, the demand and usage patterns of both systems must be known through transfer, that is, one after the other. Another interesting approach is the design of scheduling coordination based on activities.

The constraints to ensure the logic conditions of the problem are as follows:

- Transfer possibilities are limited to travellers of lines sharing a same node or station.
- Each user can only transfer from one vehicle to another one at the same node or station.

Other factors that are usually considered as constraints of the optimization problem are as follows:

- Transfer time. The lower bound ensures that a minimum number of users have enough time to transfer in train stations where changing platform is required or, in the case of two transport systems, when a walking time is required to change from a service or station to another one. The upper bound proposes a time after which the transfer service would not be enough. It can also be defined a time for the vehicle to remain in the station.
- Vehicle occupancy. The maximum occupancy is an indicator of the limit level of service volume not to be surpassed. Another possibility is to consider the

maximum capacity of the vehicle, regardless of the derived level of service. In contrast, the minimum vehicle occupancy ensures the operator's benefit. This condition tries to avoid services not justified economically by the served demand.

- Satisfied demand. Preferably, the non-satisfied demand should be zero, though it could be upper limited.
- Road or arc capacity of the network upon which the transport supply works.
- Fleet limit.
- Imposition of certain service schedules previously in force.

Other possible constraints, which add realism but also complexity to the problem, are the impact of the level of service and of the quality on the demand, and the influence of overlapping lines.

In the light of the possibility of the absence of solution satisfying every constraint, penalties to the objective function are introduced, in terms of a constant factor or aspect function whose condition has been violated. It is necessary to adequately calibrate these factors through different tests in order to achieve a realistic solution in the line of the objectives of the scheduling coordination problem.

It is advisable to represent the solution in graphs. In particular in the case of trains operating in a given corridor, the space-time graphs make it possible to determine conflicts or incompatibilities in the service designed, which may inform about the necessity of adding new restrictions to the optimization problem. Other interesting representations are as follows: costs, satisfied demand, and transferring passengers with regard to the frequency or the different scheduling combinations.

The integrated modelling of scheduling and passenger traffic assignment is addressed in Sect. 8.4 hereafter.

8.3 Modelling Demand Responsive and Paratransit

Ingmar Andreasson and Selini Hadjimitriou

Section 2.6 describes various forms of demand-responsive transport, such as feeder and shuttle services, bus-on-demand, special transportation services, taxi, and PRT.

In this section, we will discuss modelling approaches for various forms of demand-responsive transport.

The underlying assumption in modelling passenger trips is that each passenger tries to minimize his or her "disutility" associated with the choice of modes and routes. Important components of disutility are travel time components (walk, wait, and ride) together with cost, discomfort, and penalties related to modes and transfers.

For scheduled transit services, it is possible to model average passenger flows in static models based on service frequencies or timetables together with driving-times. For demand-responsive services, modelling is often more complex.

We will discuss the modelling problems first and then in the last subsection describe solution methods for the general Dial A Ride Problem (DARP).

8.3.1 Feeder and Shuttle Services

In many occasions, there are clearly defined patterns of trips for which feeder and shuttle services would be the best option to deal with the demand requirements. These requirements sometimes come from the excessive number of automobiles in a train station or transfer point car park, or excessive car trips on the same route from an origin to a destination at specific times of a day. Shuttle buses commonly serve trips between a highly demanded origin and destination with no stops in-between them and are usually designed along a fast route. When the destination is a transfer or intermodal point, shuttle buses serve as feeders of other modes. The extreme shuttle service is designed with no stops other than the origin and destination. However, this restriction may be relaxed as much as it is required.

Fixed or flexible routes and schedules, together with direction (single- or bi-directional routes) and the possibility of short turns and cuts (direct shortest path to the destination or back to previous stops instead of an established number of stops and a fixed path) lead to strategies comprised of combinations of the mentioned characteristics. The strategies could successfully be implemented as long as the capacity restrictions are considered, and the potential demand has been assessed. All in all, the design of such services aims to achieve minimum waiting and travel times with the minimum number of vehicles. In the case of a shuttle service to a long-distance bus/train station, the on-time or established arrival should be assured.

Shuttles can also be cable-propelled transport (CPT) either suspended or on rails as described in Sect. 2.6. From a modelling perspective, they can be treated as fixed shuttle routes with schedule or on demand depending how they operate. Freedom from traffic congestion and high effective speed makes CPT attractive. Spectacular view may attract some, while some may be deterred by vertigo.

8.3.2 Bus-on-Demand and Special Transportation Services

Most on-demand bus services work by a schedule, and the “on-demand” aspect refers to the subset of stops to be visited in given sequence order and possible trip cancellations. In this case, waiting-times can be modelled as with ordinary scheduled services. Riding-times will vary depending on routing which in turn depends on who else ordered the same trip. For practical purposes, it may be sufficient to use an average travel time for each origin–destination relation. Then, conventional assignment models for scheduled transit can be used also for bus-on-demand.

If there is no natural order between stops, if more than one bus is available and at least some trip destinations are also known, then the choice of bus and the optimum route for each bus can be determined by solving the DARP (see Sect. 8.3.5 about DARP solution approaches).

Elderly or handicapped persons entitled to special transport are required to order known trips at least one day in advance. Combining these orders into vehicle routes is similar to planning freight pickups and deliveries with time window and capacity constraints.

During the day, passengers call for return trips where the time of the return trip is not known in advance. These demands are inserted into one of the preplanned tours (if capacity and time windows permit) or if necessary into a new tour. When some of the trips are not known in advance the problem is called dynamic or mixed DARP.

Disabled passengers cannot access most local transit modes, and then, their modelling can be separated from other transit flow modelling. Some trips may be fed to a train, and then, they need to be coordinated with regional or interregional train schedules. For the assignment on the train network, these trips may be considered as originating at a train station.

8.3.3 Taxi

Taxi services are by definition individual, on-demand and non-stop. The level of operational planning varies between competing owner and drivers without central dispatching in one extreme and in the other extreme sophisticated fleet management with order centre with AVL, central assignment of vehicles, digital communication, and empty vehicle repositioning based on predicted demand and positions of all vehicles.

Fleet management systems aim for short passenger waiting (closest vehicle) while at the same time maintaining some justice between drivers (longest free vehicle).

Modern taxi fleet management was introduced around 1980 when the market was monopolistic. Today, competition is the norm with smartphone apps for ordering from one of the operators. The customer may call several operators to get fast service. The service would have been more efficient if all vehicles worked under a common dispatching system.

Taxi fares are much more expensive than transit fares and also higher than the cost of driving a private car. Fares are also unpredictable since they are based on both distance and time, effects of congestion, and differing rates between operators.

Ride sharing on taxis is not very common, partly due to the fare structure making it difficult to share cost between passengers with different destinations.

For the purpose of passenger (not vehicle) flow modelling, it is anyway sufficient to model taxi trips with expected waiting-times (by time period and district) and driving-times (origin–destination trip time matrices) taken from road traffic models

or historic taxi driving-times (Bell and Wong 2005). Some taxi fleets can automatically collect drive-time data from AVL systems.

8.3.4 Personal Rapid Transit

PRT and group rapid transit (GRT), more generally referred to as automated transit networks (ATNs), is described in Sect. 2.6.2, offering non-stop transport on demand on a network separated from other traffic (Irving et al. 1978; Andreasson 1994, 2009). From a service perspective, PRT is like driverless taxis on dedicated rights-of-way. Being under central control, ATN operations can be efficiently managed.

The demand for transport is treated as being random and instantaneous. Fleet dimensioning and demand predictions are based on demand modelling or on historical demand (by 15-min period, weekday, season, etc.), corrected for weather and special events.

The objective is to minimize waiting-times (average and maximum) with a vehicle fleet dimensioned for a given level of service during peak demand, such as 1-min average waiting.

The level of service depends heavily on operations and control strategies. The most important component of service is waiting-times and their distribution. Waiting-times depend on the available vehicle fleet, empty vehicle management, and ride-sharing strategies, described in the following.

Static models cannot adequately represent queuing, congestion, capacity, dynamic routing, and empty vehicle management. ATN modelling relies on micro-simulation of stochastic passenger arrivals, stochastic boarding and alighting times, vehicle trajectories, and delays in the guideway network and in stations.

8.3.4.1 Stations

ATN stations are offline from the main guideway with three possible layouts (Fig. 8.1): linear, saw-tooth or parallel berths for loading/unloading. The choice of station layout has implications on performance and modelling.

Parallel stations have the highest throughput but take up the most space.

In saw-tooth stations (as in parallel), passenger boarding is independent and vehicles can be parked for charging without blocking others. Capacity is reduced from vehicles backing out.

Linear stations are the most space-efficient, but slow boarding passenger may delay others.

All station layouts have space for vehicles waiting to get to the platform. Simulation experiments have shown that there should be more waiting positions than platform berths in order for the platform berths to be efficiently utilized.

Loading and unloading usually take place at the same position although in large stations it may be advantageous to arrange unloading also at vehicle wait positions.

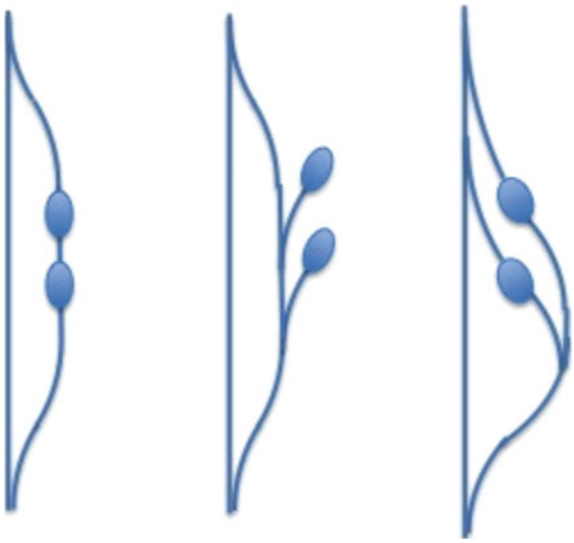

Fig. 8.1 PRT offline station layouts: linear, saw-tooth, and parallel

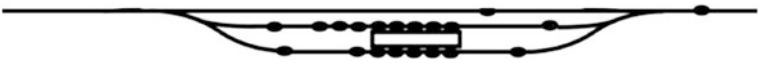

Fig. 8.2 PRT station with parallel station tracks

The capacity of a PRT station depends on the number of platform berths and the cycle time for door opening, unloading, loading, door closing, and advancing the next vehicle. This cycle time is typically around 30 s, giving a station capacity of up to 120 vehicles per platform berth and hour.

The throughput of long linear station declines beyond 7 or 8 platform berths due to variations in boarding times delaying following vehicles. Stations for higher capacity should be designed with parallel station tracks (Fig. 8.2).

8.3.4.2 Control Systems

Each ATN supplier (at the time of writing Morgantown, ULTra, 2getthere, Vectus, and ModuTram) has developed its own control system determining routing, merge priorities, and empty vehicle management. They have also proprietary simulation models adapted to their own system. Some generic, vendor-independent ATN simulation models are referenced at the end of this section.

ATN control systems are of two main types, with different implications (Anderson 1998). With *synchronous* or point-follower control vehicles follow virtual "slots" moving through the network with uniform headways between slots. Before a vehicle

can exit from a station to the main track, a free path (merge passage slots) is booked all the way to the destination. A central computer administers slot bookings and bookings of station space at the destination. Travelling is smooth without delays, but passengers may have to wait for departure, especially in large networks loaded to near capacity. Links cannot be used to their full capacity since then it would become impossible to book new trips spanning over several links. Scalability to large networks is limited, and a vehicle failing to maintain nominal speed may stop the whole system.

With *asynchronous* or vehicle-follower control, a vehicle can start as soon as the main track is free but may have to slow down or even queue before a merge, much like car traffic on roads. Congestion can be managed by dynamic routing and merge priorities. Conflict control and monitoring of safe headways are decentralised to local zone controllers, enabling unlimited scaling to large networks. Asynchronous systems allow changes of destination and routing so that they can adapt to disturbances.

Line capacity depends on the minimum safe headway between vehicles. With 3-s headway, the capacity is 1200 vehicles per hour including empty vehicles. This corresponds to the vehicle output capacity of a 10-berth station.

8.3.4.3 Dynamic Routing

Dynamic routing (in asynchronous systems) determines the quickest path for each trip considering the state of the network. Delays on each link are updated by every vehicle, and quickest paths are recalculated regularly, like every 5 min. Link delays can be amplified so that affected links are avoided before congestion gets worse.

Instantaneous reactions to disturbances are possible with so-called look-ahead over the nearest downstream links using the latest observed link travel time for each link. Without look-ahead, it would take until the next path tree calculation to react on link delays.

8.3.4.4 Empty Vehicle Redistribution

The empty vehicle redistribution (EVR) problem deals with the decision on which empty vehicle to move and where to move it. The optimal distribution of empty vehicles is a combination of dispatching orders that minimize passenger waiting and especially the longest wait. Running distances are less important since the cost of running is low.

The dynamic allocation of empty vehicles to stations can be formulated as an assignment problem. In the heuristic method called longest waiting passenger first (LWPF), the longest waiting passenger is assigned the nearest available vehicle whether it is running empty (regardless of destination), waiting at another station, or loaded on its way to the station in question. Remaining empty vehicles can be allocated based on minimum running distance and/or largest station deficit.

A three-step heuristic for empty vehicle management is described in Andreasson (2003). Station surplus/deficit is continuously updated as follows:

Station surplus =

- + vehicles in the station;
- + vehicles on way to that station;
- – passengers waiting at the station; and
- – expected new passengers during the expected call-time of empty vehicles.

Expected passengers are based on historical demand and planned special events. Call-times to each station are updated with sliding averages.

Each time a passenger arrives, an empty call is made if the surplus is below a desired level (e.g., 1 vehicle at each station). The call is made from a running empty vehicle or from a station with surplus.

Each time a vehicle enters a station, vehicles are sent off for any of the following reasons:

- Surplus over the desired level,
- Making room for the next loaded vehicle,
- Clearing the way to the passenger platform, and
- Clearing the station for continuing vehicles (with multiple destinations).

Vehicles are sent to stations with the largest deficit (smallest surplus).

A second strategy is applied at regular intervals (e.g., each minute) in that empty vehicles are reassigned based on longest waiting passenger in all stations. This passenger is allocated the nearest upstream empty vehicle (running or in a station). The next longest waiting gets his nearest and so on until all waiting passengers have been allocated vehicles.

When all waiting passengers have been allocated vehicles, the remaining empty vehicles are assigned to the largest deficits.

This heuristic has been proven for large systems. In a network with 850 vehicles, optimization with integer programming (Sjödin 2010) produced exactly the same solution.

Two other EVR heuristics are described in the references.

The fraction of empty vehicles depends on how balanced the demand is. Typically, 20–30 % of all moving vehicles are running empty.

8.3.4.5 Ride Sharing

Ride sharing in ATNs, as in any transportation system, serves to reduce the vehicle fleet needed to serve the demand (Andreasson 2005; Lees-Miller et al. 2009). For manually driven modes such as bus, LRT, and train, cost reduction drives towards larger vehicles and longer headways. Without drivers, there is less need for large vehicles. Many small vehicles operating at short headways can offer departure on demand and individual routing.

Fig. 8.3 Transfer terminal between train and PRT

Even without drivers, it is desirable to encourage ride sharing during times of peak demand. The motives now are to save on vehicle fleet and to increase link capacity by higher vehicle loads.

It is desirable to increase vehicle loads without compromising the attractive features of ATNs such as departure on demand and non-stop trips. During peak demand, it often happens that several passengers at the same platform are going to the same destination. The conditions for matching are especially good at points of transfer from mass transit modes (Fig. 8.3). This is also where ride sharing is most needed.

The control system identifies common destinations based on ticket orders or from history. Destination signs on or over each vehicle allow passengers to board into the same vehicle. If necessary, the departure may be held up to a maximum wait (1 min) so that more people can board.

To increase vehicle loads even further, one can allow an intermediate stop en route. This would be a compromise between non-stop trips versus economy and capacity.

Conditions for ride sharing are best during the morning peak when many passengers are distributed from mass transit. The typical afternoon pattern is more difficult to serve with many origins and few destinations. A strategy, which has been implemented in PRTsim, is to determine destination and path based on the first passenger (with possible ride matching). Then at one or more stations along the path, stop and pickup passengers with destination already being served by that vehicle.

Ride sharing during peak demand may double the average vehicle load (Andreasson 2005). An extreme case is at a mass transit transfer station where all vehicles are loaded and almost full. With 6-passenger vehicles and 3-s headway, the capacity on that link would be 7200 pass/h. Other links are not full of vehicles, some vehicles are empty, and the vehicle load is smaller.

Thanks to network effects, there are alternative paths in many relations so that the capacity between two stations can be larger than the link capacity (Fig. 8.4).

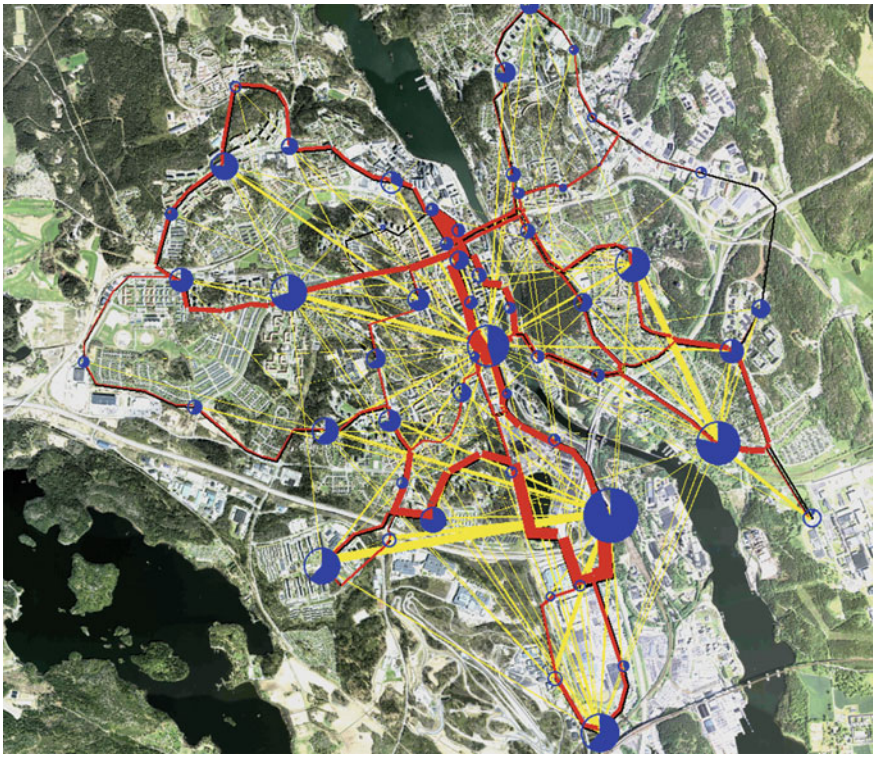

Fig. 8.4 Travel demand (*yellow*), trip ends (*blue*), and PRT passenger flows (*red*) in the city of Södertälje as modelled in PRTsim. The picture spans over 6.8 km east–west

8.3.4.6 ATNs as Feeder/Distributors

ATNs are well suited to act as feeder/distributors to mass transit, especially to enlarge the catchment areas around train stations. In order to attract more passengers to train, the first and last trip legs need to be fast and convenient. ATN vehicles can be collected at the station in time for train arrivals. Train passengers are distributed over a local network (with ride sharing in the peaks), and at the same time, passengers for the next train are collected to the station. On-demand and non-stop travelling can dramatically reduce access time to train stations, thereby increasing station accessibility (Fig. 8.5).

The improvement in accessibility is even larger when considering that waiting-times are valued higher than riding time.

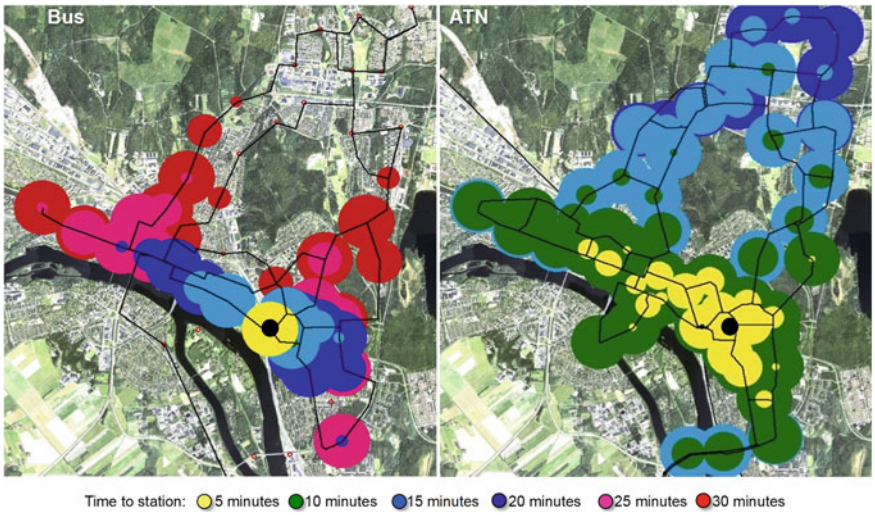

Fig. 8.5 Accessibility to a train station in Umeå with bus lines having 15–30 min headways (*left*) and the same with an ATN system. *Colours* indicate total travel time for walk, wait, and ride

8.3.4.7 Fares in ATN Systems

When the ATN system is a public investment, the fares will be harmonized with the local transit fare structure with free transfers to buses and trams. However, a private owner/operator may prefer to issue his own tickets at a different fare level. In any case, thanks to low cost of operation, the fare would be more akin to transit fares than to taxi fares.

It is possible to offer differing levels of service with different fares. At the normal rate, passengers would accept sharing rides involving some waiting or even an extra stop. A premium could be charged for individual transport or prebooked transport whereby an empty vehicle will be called in advance to be available at the departure station. If the passenger is late, he may be charged extra.

8.3.5 The Dial-a-Ride Problem (DARP)

The on-demand transport problem can be solved using linear programming (LP). LP is a particular class of problems in which a linear function is maximized or minimized subject to linear constraints.

The travelling salesman problem (TSP) is an integer linear programming (ILP) that is an LP with the additional constraint that some or all the variables have to be integer. The TSP is the problem of finding the minimum cost sequences of cities to visit such that each city is visited only once and the tour starts and ends at

the same site. The TSP is an NP-hard problem, i.e., the complexity increases very fast as the size of the problem grows.

The extension of the TSP to a fleet of vehicles is called the vehicle routing problem (VRP). The objective of the VRP is to design a set of routes, one for each vehicle, such that all customers are visited and the routes are of minimum cost. Each route must be generated and must terminate at a fixed depot; each vehicle can only serve one route; each customer can be visited only once and the capacity of each vehicle must not be exceeded. The VRP with capacity constraints on each vehicle is called capacitated VRP (CVRP).

If customers have to be visited during a specific time span, time window constraints are included in the model. Time windows can be hard or soft. Hard time windows must be adhered to. Soft time windows can be violated at a cost. If the service takes place outside the time window, a penalty is applied. The cost to be minimized is therefore a weighted sum of the total travel time over all routes plus the penalty applied in case of early- or lateness.

If a number of passengers need to be moved from a pickup location to a delivery location, the problem is called pickup and delivery problem (PDP). The PDP applied to public transport is called the DARP (Cordeau and Laporte 2007). The DARP is similar to the PDP except for more tight time windows and the presence of the maximum ride time constraints to ensure service quality. The problem consists in designing a route that accomplishes the greatest number of requested trips such that the total travelling distance is minimized. This problem has been formulated in a number of ways, usually depending on the underlying problem. The DARP is an adaptation of the three-index formulation for the VRP with time windows presented by Cordeau et al. (2002).

The formulation consists in an optimization problem, whose objective function minimizes the total travel costs subject to a set of constraints depending on the type of problem, ensuring consistency conditions in terms of flows and times, such as the following:

- Each node is served only once;
- For each vehicle, there is only one arc exiting the depot;
- Each vehicle returns only once to the depot;
- The number of vehicles exiting from a node is equal to the number of vehicles entering the same node;
- If an arc (i, j) is travelled, the time the service begins in j is greater than the time the service begins in i plus the service time in i plus the travel time from i to j;
- If arc (i, j) is travelled, the load of the vehicle in j is equal to the load of the vehicle in i plus the demand in j;
- The load of a vehicle is not negative;
- The load in i is at least equal to the demand in i;
- The load is not greater than the capacity;
- If a vehicle enters an origin node, the same vehicle will enter the corresponding destination node;

- The time of the service at the origin must be less than the time of beginning the service at destination
- The arrival time minus the departure time minus the time needed to serve each customer has to be less than a given maximum duration of the service;
- The service is performed between an earliest and a latest time of service;
- The departure time from the depot is later that the previous arrival time at the depot plus an operative margin; and
- The duration of the service is not greater than a maximum value.

The demand is negative in case of passengers alighting.

The DARP can be static or dynamic. If static, the demand for transport is known before the trip begins. If updated information or new requests for transport are received during the trip, the DARP is called dynamic. A mixed solution is also possible. In this case, part of the information is provided in advance—usually one day before, and additional requests are sent in real time.

The common objective of the static and dynamic approaches is the minimization of costs. Moreover, the dynamic formulation needs to take into account additional aspects. The most important one is the response time. That is to minimize the time between the service is wanted and the time the service takes place. There is in fact a trade-off between the need to provide good optimal results and the ability to provide the solution in a reasonable time.

Most transport demand models are deterministic, i.e., all information is available before the service takes place. However, many forms of uncertainty exist in reality. The most important ones are the dynamics of traffic conditions and the variability of demand. Other sources of uncertainty can be considered such as vehicle breakdowns, accidents, and weather conditions. Uncertainty is introduced into optimization models using random variables with probability distributions (Xiang et al. 2008).

In the static and stochastic DARP, all routes are designed before uncertain data are known. The uncertain information is represented by random variables, and their realization is revealed during the trip. Corrective actions are therefore taken at a second stage, as soon as the realization of the random variable is known.

The most used *exact method* to solve the static DARP is the branch-and-cut. This consists in solving branch-and-bound and using cutting planes to tighten the LP relaxation.

In case of a large number of requests, *heuristic* approaches are used to obtain good solutions reasonably fast (Jaw et al. 1986). The simplest heuristic for the static or the dynamic problem is called insertion heuristics. This method consists in including a new request in the "best" position of the current scheduling. As soon as a request arrives to the system, the least-cost feasible insertion is sought. If there is no feasible insertion and there are idle vehicles, a new route can be started. Many variants of the insertion heuristics for the DARP have been developed, often in combination with other methods.

Metaheuristic algorithms are usually used in combination with heuristic methods in order to improve the solution. Examples of Metaheuristic approaches are the local Search and the tabu Search (Attanasio et al. 2004). The local search algorithm starts

from an initial solution and moves, in each iteration, to a better solution. The problem of this algorithm lies in the possibility to incur in a local optimum. The current solution can be improved by listing forbidden solutions in a so-called tabu list.

In the dynamic and stochastic systems, uncertain data are represented by stochastic processes. For instance in case of demand modelled as a Poisson distribution, uncertain information is revealed during the execution of the service. Routes are built in real time, as soon as new requests are received. In addition, decisions must be taken concerning whether or not to include a new request in the current route.

The *stochastic* model is usually solved using a Markov decision process. However, ILP has been adapted in order to deal with stochastic information. For instance, in case of stochastic delays, the vehicle skips the absent customer and moves to the next customer. The approach consists in starting with a solution for the static problem. Successively, stochastic information is introduced into the model and fast insertion heuristics are used to continuously update the optimal solution.

8.4 Integrated Modelling of Travel Demand and Transit Operations

Ingmar Andreasson, Francesco Corman and Fabien Leurent

Real-world phenomena of travel demand, including mode and path choice on a network, and of traffic operations from local performance to network control and design, can be modelled at different levels. In the classical, four-step scheme of travel demand modelling, mode choice and network assignment are addressed separately, thus leading to separate assignment of car trips to the roadway network, on the one hand, and of passenger trips to the transit network, on the other hand. Of course, this is a rough approximation only, since in large cities the transportation network is multimodal and many users make intermodal trips combining a car leg and a transit path.

Complex networks can always be addressed by micro-simulation of all passenger- and vehicle movements of all modes—scheduled or on-demand. This comes out at the expense of much computation because there is much variability on both the demand side (trip generation, mode, and route choice behaviour under dynamic situations) and the supply side (the interactions of many vehicles at various places along the network and for each transit vehicle its interactions with passenger flows). The variability is addressed by specification of the underlying statistical distributions (indeed a bridge between disaggregate and aggregate treatment) and by the computation of a suitable number of replications, so as to establish average conditions. These are very relevant in the simulation of passenger travel since many travel decisions are based on a partial knowledge of dynamic conditions. When only aggregate effects are of interest and when they can be modelled with satisfactory realism for the modelling purpose, then a coarse

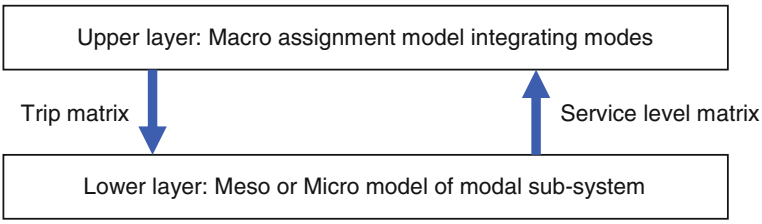

Fig. 8.6 Bilevel modelling of network route choice and traffic operations by mode

aggregate model is sufficient and efficient. Other times, the purpose is to study detailed behaviour or such behaviour is important for overall results.

Aside from full micro-simulation, structured, multilayered modelling can address a range of issues:

- In roadway assignment, signalized road intersections are often simulated while traffic flows of the larger network are modelled as static traffic flows and speeds. Micro-simulation of vehicle movements may be necessary to model interactions in a study area, while the surrounding network can be simulated at a more aggregate level with a simplified description of interactions.
- In the transit world, bi-level modelling can be used to study in detail a local complexity such as movements within a terminal or a waiting platform. Results are re-entered into the overall model as average delays or capacity constraints.
- Still in the transit world, bi-level modelling has been developed to deal with line-haul networks by integrating a higher level of passenger route choice and a lower level of traffic operations (Fig. 8.6): cf. first subsection below.
- The same kind of scheme has been applied to multimodal networks, combining car and line-haul transit or line-haul and demand-responsive services: cf. second subsection.
- Lastly, network assignment and network control or design have been integrated in joint models: these are addressed in the third subsection, which includes a special focus on railway modes.

8.4.1 Bi-Level Optimization of Line-Haul Transit Networks

The principle here is to integrate passenger travel and traffic operations in a modular, bi-level way:

- The upper level model simulates passenger route choice by origin–destination pair on the network and derives the local flow by network element, based on travel conditions taken from the lower level model in terms of section times either in-vehicle or pedestrian, discomfort notably due to crowding, wait time on station platform.

- The lower level model simulates the local traffic conditions on the basis of the local flows taken from the upper level model. It amounts to a complex form of a link travel time function with respect to link flow in roadway assignment.

This bi-level scheme has been applied to several kinds of line-haul networks.

In Teklu et al. (2007) and Teklu (2008), the upper layer deals with passenger route choice by probit-based stochastic assignment: the transit network is represented by route sections as in the leg-based model of De Cea and Fernandez (1993). However, local route choice is based not only on average run times but also on a variable wait time (with variance parameterized by the frequencies of the combined services) plus a Gaussian error component. Furthermore, a Markov process of learning is used to simulate day-to-day experience and the capitalization of average values for all kinds of times. On the lower level, micro-simulation is used to model individual passengers and individual vehicles in interaction along the transit line. This model pertains to bus lines as FIFO waiting is postulated.

An alternative blending of dynamic and static features is proposed by Schmöcker et al. (2008). The upper layer model for passenger route choice is frequency-based but in a quasi-dynamic way. The lower layer model of vehicle operations is partly schedule-based in order to deal with passenger storage on platforms under line saturation. As the waiting model involves mingled people, this model is applicable to railway lines (though with only one service per line direction).

The model by Sumalee et al. (2009) is macroscopic, dynamic, schedule based on the upper layer. It deals with vehicle operations in a mesoscopic way: by vehicle run, including dwelling operations of passenger alighting and boarding, and also by a seat allocation process.

The static, frequency-based model of Leurent et al. (2011, 2014) has a bi-level structure and enables a general line structure with one or several services, as can occur in railway transit. The upper layer addresses passenger route choice between hyperpaths on a network with pedestrian links and line sections (i.e., line legs constituted by the legs of the line service(s)) where the line frequency generalized run time and wait time by leg stem from the lower layer. Furthermore, route choice between alternative lines at a given node depends on their respective cost to destination and frequency, and on a specific correction for a saturated line where the passenger cannot be certain to board on the first incoming vehicle. The lower layer addresses traffic operations along each line: at the vehicle level, the flows alighting and boarding at every station along the service route are modelled, together with the on-board flows by exit station and comfort state (sitting/standing), thus yielding in-vehicle discomfort cost. At each entry station level, the stock of waiting passengers is modelled by station of exit: it depends on the leg flows and the on-board residual capacity by service (the product of service frequency and on-board residual capacity) on the basis of a bottleneck model by exit station, under the postulate of mingled waiting. Furthermore, in the static setting, the frequency is exogenous: on each track element, the requirement of occupation time by service vehicles (including dwell time, stopping manoeuvres and safety margins) is totalled over the

runs scheduled during the period under study. If the occupation requirement is larger than the period duration, the outgoing frequency is reduced and propagated downstream for all the concerned services. Lastly, frequency reduction induces a related delay for upstream vehicles, thus adding to physical times hence to passenger generalized cost.

8.4.2 Integrated Modelling of Multimodal Networks

Networks with several forms of transit and private transport can always be modelled by micro-simulation of passenger and vehicle movements of all modes—scheduled or on demand. Simulation requires all schedules to be specified, which may not be practical during a planning stage. In static models, it may be sufficient to specify service frequencies without schedules.

Passenger flows on scheduled transit modes can be calculated by macroscopic models for homogenous (supply and demand) time periods. Macroscopic assignment models are based on expected waiting and in-vehicle times for the (remaining) trip along each alternative path. Models include also walking distance, fare, comfort, and mode preferences, which apply equally to static or dynamic models.

The choice between modes is also affected by mode preferences as modelled by penalties or mode-specific constants. Most people with a choice prefer car over transit even if times and cost would be the same. Interviews indicate that some demand-responsive modes such as PRT are considered about halfway between car and bus in terms of mode preferences. If a trip involves more than one mode, the least attractive mode seems to determine the penalty for the whole trip.

Transfers between lines and/or between modes are associated with an extra penalty—larger for unreliable or inconvenient transfers, such as involving walking in unprotected areas. A cross-platform transfer without waiting would probably have a very low penalty.

When the level of service is improved, it is expected that trip-making will increase. If a demand-responsive mode is conceived as offering superior service compared to scheduled transit, then it will attract more passengers from private modes and may also induce increased trip-making. The impact of service level on mode choice and travel demand is discussed in Chap. 4.

Several multimodal models have been proposed which integrate line-haul services and private car. Wu and Lam (2003) proposed a network equilibrium model for the simultaneous prediction of mode choice and route choice in congested networks. The congestion effects of each mode and intermodal interactions were taken into account. An equivalent variational inequality problem was formulated to capture all the components of the proposed model, which allowed to achieve better travel times and network flow estimations.

Peric and Boilé (2006) presented a combined mode choice and assignment intermodal network equilibrium model, where combined automobile-to-bus and automobile-to-rail intermodal options were considered. Taking into account the

two-way interaction between automobile and bus transit (whose transit times depend on each other due to congestion) and its effects, the formulation of the network equilibrium considered the inverse demand functions of each mode in each level.

Li et al. (2007) modelled a multimodal network with constraints on both run segments and parking facilities under user equilibrium in a static setting.

Zhou et al. (2008) developed a dynamic micro-assignment and (meso) simulation system that incorporates individual trip-maker choices of travel mode, departure time, and route in multimodal urban transportation networks. The model maximized the stochastic utility considering multiple user decision criteria (such as travel time, travel cost, schedule delay, and travel time reliability). An efficient time-dependent least-cost path algorithm was embedded to generate an intermodal choice set that recognized time-dependent mode transfer costs and feasible mode transfer sequences. A two-stage estimation procedure is proposed, systematically utilizing historical demand information, time-dependent link counts, and empirically calibrated departure time choice models.

In urban transit networks where demand-responsive transit (DRT) modes can or could play a role, it is desirable to include DRT in a consolidated macroscopic model of all transit. Waiting and in-vehicle times are the factors needing special consideration for demand-responsive modes. For existing services, these times can be surveyed. Planned systems need to be simulated or approximated.

Where passengers have a choice between scheduled and demand-responsive modes, the performance of the demand part depends on the usage so that iterations may be necessary between network assignment and modelling of service levels. For well-dimensioned demand systems below capacity, the uncongested results may be a good-enough approximation. As a first approximation, in-vehicle times can be calculated without congestion and waiting-times can be set to a given average time, e.g., one minute at all stations for PRT systems. Based on static network assignment on combined transit services, new simulations of the DRT part provide updated waiting and in-vehicle times.

A bi-level model has been devised by Andreasson (2014) to deal with a large range of transit modes including line-haul and demand-responsive services:

- On the upper layer, a static flow model addresses scheduled services on fixed routes in a frequency-based assignment. Every DRT mode is included there as a generalized, matrix kind of line with origin-to-destination legs and their respective waiting and in-vehicle times. The upper layer assignment yields local flows on all links, particularly so on the DRT legs.
- On the lower layer, the DRT mode with origin–destination trip matrix taken from the upper layer is addressed by micro-simulation, yielding average waiting and in-vehicle times by leg. DRT modes such as taxi, dial-a-ride, and PRT need to be modelled dynamically in order to adequately represent effects of vehicle assignment, routing, empty vehicle management, and bottle-necks.

As for vehicle sharing systems (VSS) for one-way trips between stations, the model of Leurent (2013) has a three-layer structure as follows, from bottom up:

- The individual VSS station is modelled as a simple, Markovian system of double-ended capacitated queue—one side for the number of docks that are required by customers on exit and the other for the number of customers waiting for access (with a minus sign). The set of stations makes up the bottom layer.
- The network of VSS stations makes up a system of parking options for a customer willing to alight, with diversion from a saturated station to neighbouring alternatives if the expected wait time is too costly. This middle layer derives the station-to-station travel conditions of average wait time, average run time, and availability level from the station-to-station flows of users. It is a system of legs analogous to a transit line.
- By origin–destination pair and user class, path options are identified along the multimodal network and evaluated prior to the trip on the basis of their expected generalized cost, and every user is assigned to a hyperpath of minimum cost to him. A special kind of hyperpaths involving availability probabilities constitute the ex-ante travel options, from which stem the ex-post paths possibly with local adaptations due to cruising for parking.

8.4.3 Combination of Assignment with Control and Design

Micro-simulation or bi-level modelling of transit networks enables us to emphasize both route choice and service performance. In a similar spirit, some joint models of route choice and network control or design have been developed for line-haul transit.

Di Febbraro et al. (1996) proposed a simulation and control model of an intermodal urban transportation system, considering the transportation network as an oriented graph in which nodes represent single-mode stations or intermodal stations. The authors considered a discrete event model integrating different transportation services and proposed a special-purpose urban traffic simulator including two major modules: a Traffic Simulation Kernel (to consider the dynamic behaviour of the transportation system) and a Passenger Information Service (to provide the system users dynamic information about the different intermodal paths between any pair of origin/destination nodes). However, it did not fully integrate the public and private transport modes.

Abdelghany and Mahmassani (2001) presented a dynamic trip assignment simulation model for urban intermodal transportation networks, considering private cars, buses, metro/subways, and high occupancy vehicles (HOVs). The model considered the interaction between mode choice and traffic assignment and implemented a multiobjective assignment procedure in which travellers choose their modes and routes based on a range of evaluation criteria. This model may be used conjointly with real-time traffic management systems.

Uchida et al. (2007) considered travellers to choose their multimodal routes so as to minimize their perceived disutilities of travel following the probit stochastic user

equilibrium (SUE) conditions. Factors influencing the disutility of a multimodal route include actual travel times, discomfort on transit systems (to consider the in-vehicle congestion effect), expected waiting-times, fares, and constants specific to transport modes (the modes considered are car, bus, train, and walking). The authors use a sensitivity analysis to define linear approximation functions between Probit SUE link flows and the design parameters (used as constraints in the network design problem).

In railways, the (online) traffic control problem is to reduce delays (due to failure of components or vehicles, longer dwell time at stations) and, more importantly, their propagation along time and space due to the limited capacity of the railway network. Currently, traffic controllers base their decision on the measured delays, following prescribed rules or their experience. Several advanced ITS approaches that deal with passenger assignment or adjustment of schedule on which to perform a passenger assignment have been recently defined in the academic literature. In this case, control actions (change of times, orders, routes, services) need to be found quickly, within seconds or minutes. It is not common to have precise knowledge of the amount of passengers on the vehicles (automated passenger counters are very rare, and smart card data are not used in real time); thus, the expected passenger volumes considered have the same source and accuracy as those used for the offline horizon. Due to the importance of the schedule in operations, and the fact that the schedule is actually a variable to be determined, schedule-based assignment methods are commonly used. Approaches can be bi-level as in the previous section, or integrated.

The delay management problem focuses at determining which passenger connections should be kept in an updated schedule for delayed operations. Mathematical models assume known, non-elastic demand, a few hours of time horizon and tens of origin, destination, and connection stations. Mixed integer linear programs are mostly used that minimize travel time of passengers. Due to combinatorial complexity, capacity is considered in an approximated simplified manner at stations and along lines, basically considering only arrivals and departures at stations.

Simplified frequency-based assignment has been used by Ginkel and Schoebel (2007) based on the inherent periodicity of train services: basically passenger flows are assigned to a single line, and in case they miss a service, they have to wait for the whole period between two consecutive services of the line. Such frequency-based assignments are difficult to be generalized for larger networks, or heterogeneous services, as it is hard to define effective frequencies in such cases, and moreover, multiple routes might be available between an origin and a destination. More recent, schedule-based assignment models by Dollevoet et al. (2012) are able to deal with rerouting of passengers over large networks A mixed integer linear formulation taking into account rerouting of passengers aims at computing the flow of passengers following their path of minimum travel time. Simplified capacity constraints can be integrated to roughly approximate the capacity of a station area, or along railway lines. This results in a more computationally complex

problem of finding the underlying schedule over which passengers will be assigned. Heuristic procedures can then provide solutions in acceptable times.

The microscopic traffic control problem with passenger flows follows a complementary approach: instead of extending a simplified passenger model to include real-life constraints, a precise microscopic scheduling model (modelling explicitly the signalling system and operational restrictions) is extended to include passenger flows computed by a transit assignment. Also here, computing a schedule is the complexity core of the problem, and schedule-based assignment (based on the shortest path, or the least-cost path) is used. The interplay between assigning the passenger flows and generation of the schedule can arise at different levels. In Dollevoet et al. (2014), assignment is performed on a macroscopic schedule that is updated iteratively to take into account the precise microscopic movements of train operations. Sato et al. (2013) perform passenger assignment at a microscopic level, considering limited possibilities to update the microscopic schedule within a mixed ILP setting, and the possibility to change further the schedule by shifting train departures, if that decreases passenger travel time. The recent approach by Corman et al. (2013) integrates microscopic scheduling, delay management, and schedule-based passenger assignment in a single MILP model. A heuristic based on functional decomposition of the problem allows for solving real-life instances in practical times.

A reason for online control also includes major disruptions, i.e., blockades, trains, or facilities out of control. These situations require major changes in the train service provided, including short turning trains at intermediate points, cancelling planned transfers, or even cancelling full train services. Almodovar and García-Rodenas (2013) aim at modelling and controlling the predicted passenger flow during emergencies or exceptional events. Differently from all approaches here reported, they model passenger assignment to vehicles considering capacity, i.e., some passenger can be denied boarding on a service if the vehicle is full.

In traffic operations, traffic control must be performed on line: this imposes to simplify the assignment model that depicts the passenger-related issues. On the contrary, network design is performed off-line: this enables network managers to pay more consideration to the passenger side, both for timetable design and for rostering (explained in what follows).

The determination of a plan of operations (the timetable) in railway systems aims at distributing regular services over time (line planning, homogeneity); matching peak travel demand; securing headways between train services to make them resistant to delays; and synchronizing the arrival and departures for smooth interchange at major stations. Innovative objective functions for automated timetable computation can address passenger travel time. As the main computational complexity of the problem is the determination of the schedule of operations, a bi-level modelling is also used, i.e., first a reference schedule of operations is computed, and then, a schedule-based assignment model is solved, based on assignment to the shortest path for all origin–destination pairs.

From an academic point of view, Sels et al. (2013) define a mathematical optimization problem for timetabling, targeting reduced passenger travel time based

on estimated OD demand. The approach alternates a retiming phase-shifting tentative train paths in time to decrease passenger travel time, with a schedule-based assignment phase that computes the fastest travel time for each OD. The approach is iterated until convergence, leading to smaller travel times than a general timetable.

Rostering is the tactical process (horizon of one or few weeks) by which rolling stock and crew are assigned to trains. The assignment of train vehicles to services is based on the forecasted passenger flows and is influenced by vehicle availability. The former figure is mostly based on rules of thumb and time-series analysis, giving the average flow per link per day and hour of the day, rather than more sophisticated transit assignment models. Once the passenger estimate per service is given, the optimization in the rostering process can lead to substantial savings and performance improvements (see, e.g., Abbink et al. 2004).

Concerning bus services, control strategies can also be designed on the basis of a transit assignment model (Cats et al. 2010). High-frequency bus services are subject to service perturbations which if not mitigated may result in a vicious cycle deteriorating the level of service. In the case of high-frequency bus services, passengers arrive at stops without consulting the timetable and dwell times are flow-dependent. As a result, there is a positive feedback loop between the service headway (elapsed time between consecutive bus arrivals), the number of boarding passengers, and dwell times at stops and downstream headways. Hence, a relatively small perturbation from the planned service (e.g., delay at intersection, congestion, late dispatching, short dwell time) could propagate and lead to the well-known bus bunching problem.

Holding control strategies are among the most common measures to mitigate the deterioration of service reliability along the line by regulating bus departure times based on a certain service criterion. In a series of studies (Cats 2014; Cats et al. 2011, 2012, 2014a), BusMezzo, an agent-based public transport operations and assignment simulation model (see Chap. 6.5), was used to design a new control strategy for improving bus service regularity. BusMezzo is integrated into a mesoscopic traffic simulation model, Mezzo. Each vehicle, either personal car or public transport vehicle, is represented as an individual agent with its link running time determined by a speed-density function and a queuing time that is determined by stochastic queue servers. Dwell times at stops are determined by the number of boarding, alighting, and on-board passengers as well as bus stop and vehicle characteristics. Each bus is assigned to a sequence of runs. Run dispatching times depend on vehicle availability and recovery time constraints and hence enable capturing the potential propagation of delays from one run to the next.

Alternative holding strategies were formulated and implemented in BusMezzo, and their performance was evaluated by measuring their impact on service regularity, passenger travel times, operational efficiency, and vehicle scheduling. The impact of the following holding strategies was evaluated: (1) no control, (2) schedule-based control, (3) forward-headway control, and (4) even-headway control. The strategies were tested for a high-demand line in Tel Aviv metropolitan area, Israel, with two or three time point stops along each route direction. Passenger demand was modelled in this case study in terms of passenger arrival rates and alighting probabilities per

stop. The headway-based holding strategies outperformed the no control and schedule-based control scenarios in terms of passengers waiting time, albeit with an increase of in-vehicle time. A Pareto-front analysis indicated that applying the even-headway control with a maximum allowable holding time is especially efficient. Similar results were obtained for a case study on a trunk bus line in Stockholm inner city, where the share of bunched buses decreased and the regularity level of service improved dramatically. Moreover, the Stockholm case study demonstrated that the even-headway control also yields benefits in terms of operational costs due to more reliable vehicle trip times and more reliable arrival times at driver relief point, an important objective of crew management.

The design of a holding control strategy involves the determination of the number and locations of time point stops, locations where holding may take place, and the decision criterion for the holding time. BusMezzo was used to investigate what is the optimal number and locations of time point stops (Cats et al. 2014b). The combinatorial problem was solved by applying a greedy algorithm and a genetic metaheuristic. Each candidate solution was evaluated by multiple instances of BusMezzo where public transport performance is the outcome of numerous interactions between traffic dynamics, public transport operations, and control and passenger decisions. Simulation outputs were then used as part of an iterative optimization process with a multiple objective function considering total generalized passenger travel times as well as operational costs.

While the common practice is to hold buses at a predefined set of time point stops, continuously monitoring service regularity by either holding or speed adjustments could potentially help reduce and distribute the need to hold bus services. This hypothesis was tested in BusMezzo following empirical evidence that bus drivers adjust their speed in response to schedule deviations. Applying the even-headway control in a continuous cooperative manner outperformed its application in a limited set of time point stops. Furthermore, the simulation model was used to test the sensitivity of the results with respect to drivers' compliance rate and was found robust with respect to driver behaviour.

Irregular services do not only entail longer waiting-times and lower operational efficiency, but also yield uneven passenger loads. In high-demand services, this implies an inefficient capacity utilization and high levels of experienced crowding. The operations and assignment model in BusMezzo enable the assessment of the impact of the new control strategy on passenger on-board congestion. The dynamic assignment model in BusMezzo was used to calculate the on-board crowding which was used for determining the in-vehicle multiplier for each individual passenger. Furthermore, the enforcement of strict capacity constraints in the simulation model allows to identify the extent of denied boarding and the excess waiting time that is induced by failure to board.

The simulation analysis was followed by a series of field trials that led to a full-scale implementation of the continuous even-headway control. The design of this control strategy involves a paradigm shift towards regularity that requires the

consideration of a series of measures along the service chain including production planning, control centre, and incentive schemes. Since the summer of 2014, the entire trunk line network in Stockholm inner city operates with this control strategy.

References

Abbink EJW, Van den Berg BWV, Kroon LG, Salomon M (2004) Allocation of Railway Rolling Stock for Passenger Trains. Transp Sci 38:33–42

Abdelghany KF, Mahmassani HS (2001) Dynamic trip assignment-simulation model for intermodal transportation networks. Transp Res Board 1771:52–60

Almodovar M, García-Rodenas R (2013) On-line reschedule optimization for passenger railways in case of emergencies. Comput Oper Res 40:725–736

Anderson JE (1998) Control of personal rapid transit systems. J Adv Transp 32:57–74

Andreasson I (1994) Vehicle distribution in large personal rapid transit systems. Transp Res Board 1451:95–99

Andreasson I (2003) Reallocation of empty personal rapid transit vehicles en route. Transp Res Board 1838:36–41

Andreasson I (2005) Ride-sharing on PRT. Proceedings of the 10th APM conference in Orlando, American Society of Civil Engineers, USA

Andreasson I (2009) Extending PRT capabilities. In: Proceedings of the 12th APM conference in Atlanta, American Society of Civil Engineers, USA

Andreasson I (2014) Private communication on PRTsim model

Attanasio A, Cordeau JF, Ghiani G, Laporte G (2004) Parallel tabu search heuristics for the dynamic multi-vehicle dial-a-ride problem. Parallel Comput 30:377–387

Bell MGH, Wong KI (2005) A rolling horizon approach to the optimal dispatching of taxis. In: Mahmassani HS (ed) Transportation and traffic theory: flow, dynamics and human interaction. Elsevier, Oxford

Bellei G, Gentile G, Papola N (2000) Transit assignment with variable frequencies and congestion effects. In: Proceedings of the 8th Meeting of the EURO Working Group on Transportation—EWGT 2000, Roma, Italy, pp 525–532

Cats O (2014) Regularity-driven bus operations: principles, implementation and business models. Transp Policy 36:223–230

Cats O, Burghout W, Toledo T, Koutsopoulos HN (2010) Mesoscopic modeling of bus public transportation. Transp Res Rec 2188:9–18

Cats O, Larijani AN, Burghout W, Koutsopoulos HN (2011) Impacts of holding control strategies on transit performance: a bus simulation model analysis. Transp Res Rec 2216:51–58

Cats O, Larijani AN, Ólafsdóttir A, Burghout W, Andreasson I, Koutsopoulos HN (2012) Holding control strategies: a simulation-based evaluation and guidelines for implementation. Transp Res Rec 2274:100–108

Cats O, Fadaei Oshyani M, West J (2014a) Evaluation of RETT4 pilot study: empirical and simulation analysis of bus and passengers time savings. KTH, Sweden

Cats O, Mach Rufi F, Koutsopoulos HN (2014b) Optimizing the number and location of time point stops. Public Transport 6:215–235

Chien S, Chowdhury S(2002) Intermodal transit system coordination. J Transp Planning Technol 25(4):257–288

Chung E-H (2009) Transfer coordination model and real-time strategy for inter-modal transit services. Ph.D. Thesis. Department of Civil Engineering. University of Toronto, Canada

Cordeau J-F, Laporte G (2007) The dial-a-ride problem: models and algorithms. Ann Oper Res 153:29–46

Corman F, Sabene F, Pacciarelli D, Samà M, D'Ariano A (2013) Railway traffic control with minimization of passengers' discomfort. 3rd MTITS, Dresden, Germany

De Cea J, Fernandez JE (1993) Transit assignment for congestion public transport networks: an equilibrium model. Transp Sci 27:133–147

Di Febbraro A, Recagno V, Sacone S (1996) INTRANET: a new simulation tool for intermodal transport systems. Simul Pract Theory 4:47–64

Dollevoet T, Huisman D, Schmidt M, Schöbel A (2012) Delay management with rerouting of passengers. Transp Sci 46:74–89

Dollevoet T, Corman F, D'Ariano A, Huisman D (2014) A bi-level optimization framework for delay management and train scheduling. J Flex Serv Manuf (in press)

Fu L (2002) Scheduling dial-a-ride paratransit under time-varying, stochastic congestion. Transp Res B 36:485–506

Hall R (1985) Vehicle scheduling at a transportation terminal with random delay en route. Transp Sci 19:308–320

Irving JH, Bernstein H, Olson CL, Buyan J (1978) Fundamentals of personal rapid transit. Lexington Books, Washington DC

Jaw J, Odoni AR, Psaraftis HN, Wilson NHM (1986) A heuristic algorithm for the multi-vehicle advance request dial-a-ride problem with time windows. Transp Res B 20:243–257

Kroes E, Kouwenhoven M, Duchateau H, Debrincat L, Goldberg J (2007) Value of punctuality on suburban trains to and from Paris. Transp Res Rec 2006:67–75

Lam WHK, Zhou J, Sheng Z-H (2002) A capacity restraint transit assignment with elastic line frequency. Transp Res B 36:919–938

Lee KKT (1993) Optimization of timed transfers in transit terminals. PhD dissertation, University of Maryland, College Park, USA

Lee KKT, Schonfeld PM (1991) Optimal slack time for timed transfers at a transit terminal. J Adv Transp 25:281–308

Lees-Miller J, Hammersley J, Davenport N (2009) Ride sharing in personal rapid transit capacity planning. In: Automated people movers. American Society of Civil Engineers, USA

Leurent F (2011) Transport capacity constraints on the mass transit system: a systemic analysis. Eur Transp Res Rev 3:11–21

Leurent F (2013) Travel demand for a one-way vehicle sharing system: a model of traffic assignment to a multimodal network with supply-demand equilibrium. In: Albrecht T, Jaekel B, Lehnert M (eds) Proceedings of the 3rd international conference on models and technologies for intelligent transportation systems 2013. Verkehrstelematik, vol 3, TUD Press, Dresden, pp 523–534

Leurent F, Chandakas E, Poulhès A (2011) User and service equilibrium in a structural model of traffic assignment to a transit network. Elsevier Proc—Soc Behav Sci 20:495–505

Leurent F, Chandakas E, Poulhès A (2014) A traffic assignment model for passenger transit on a capacitated network: bi-layer framework, line sub-models and large-scale application. Trans Res Part C 47(1):3–27

Li ZC, Huang HJ, Lam WHK, Wong SC (2007) A model for evaluation of transport policies in multimodal networks with road and parking capacity constraints. J Math Model Algorithms 6:239–257

Liu Y, Bunker J, Ferreira L (2010) Transit users' route-choice modelling in transit assignment: a review. Transp Rev 30:753–769

Meschini L, Gentile G, Papola N (2007) A frequency based transit model for dynamic traffic assignment to multimodal networks. In: Allsop RE, Bell MGH, Heydecker BG (eds) Transportation and traffic theory 2007. Elsevier, London, pp 407–436

Ngamchai S, Lovell DJ (2003) Optimal time transfer in bus transit route network design using a genetic algorithm. J Transp Eng 129:510–512

Peric K, Boilé M (2006) Combined model for intermodal networks with variable transit frequencies. Transp Res Board 1964:136–145

Ropke S, Cordeau JF, Laporte G (2007) Models and branch-and-cut algorithms for pickup and delivery problems with time windows. Networks 49:258–272

Sato K, Tamura K, Tomii N (2013) A MIP-based timetable rescheduling formulation and algorithm minimizing further inconvenience to passengers. J Rail Transp Planning Manage 3:38–53

Schmöcker JD, Bell MGH, Kurauchi F (2008) A quasi-dynamic capacity constrained frequency-based transit assignment model. Transp Res B 42:925–945

Sels P, Dewilde T, Cattrysse D, Vansteenwegen P (2013) Expected passenger travel time for train schedule evaluation and optimization. In: Proceedings of the 5th international seminar on railway operations modelling and analysis, Copenhagen, Denmark

Shrivastava P, Dhingra SL (2001) Development of feeder routes for suburban railway stations using heuristic approach. J Transp Eng 127:334–341

Shrivastava P, Dhingra SL (2002) Development of coordinated schedules using genetic algorithms. J Transp Eng 128:89–96

Shrivastava P, O'Mahony M (2006) A model for development of optimized feeder routes and coordinated schedules—a genetic algorithms approach. Transp Policy 13:413–425

Sjödin P (2010) Allokering av tomvagnar I ett spårtaxinätverk via heltalsoptimering. MSc thesis at Department of Mathematics, KTH, Stockholm, Sweden

Sumalee A, Tan Z, Lam WHK (2009) Dynamic stochastic transit assignment with explicit seat allocation model. Transp Res B 43:895–912

Systan Inc. (1983) Timed transfer: an evaluation of its structure, performance and cost. Rep. No. UMTA-MA-06-0049083-6, Urban Mass Transportation Administration,Washington, D.C., USA

Szillat MT (2001) A low-level PRT microsimulation. PhD thesis, The University of Bristol, UK

TCQSM (2003) Transit capacity and quality of service manual. On-line report prepared for the Transit Cooperative Research Program of the TRB, USA

Teklu F (2008) A stochastic process approach for frequency-based transit assignment with strict capacity constraints. Netw Spat Econ 8:225–240

Teklu F, Watling D, Connors R (2007) A markov process model for capacity-constrained frequency-based transit assignment. In: Allsop RE, Bell MGH, Heydecker BG (eds) Transportation and traffic theory 2007. Papers Selected for Presentation at ISTTT17. Elsevier, Amsterdam, Netherlands

Ting CJ (1997) Transfer coordination in transit network. PhD dissertation, University of Maryland, College Park, USA

Uchida K, Sumalee A, Watling DP, Connors RD (2007) A study on network design problems for multi-modal networks by probit-based stochastic user equilibrium. Netw Spat Econ 7:213–240

Vuchic VR, Clarke R, Molinero A (1983) Timed transfer system planning, design and operation. Urban Mass Transportation Administration, DOT-I-83-28. Department of Transportation, Washington, D.C., USA

Wu ZX, Lam WH (2003) Combined modal split and stochastic assignment model for congested networks with motorized and non-motorized transport modes. Transp Res Board 1831:57–64

Xiang Z, Chu C, Chen H (2008) The study of a dynamic dial-a-ride problem under time dependent stochastic environments. Eur J Oper Res 185:534–551

Zhou X, Mahmassani HS, Zhang K (2008) Dynamic micro-assignment modeling approach for integrated multimodal urban corridor management. Transp Res C 16:167–186

Chapter 9
Applications and Future Developments: Modeling Software and Advanced Applications

Ektoras Chandakas, Fabien Leurent and Oded Cats

Situation. Part III and Chap. 8 are targeted to the theoretical modeling of transit systems. The challenge is to match the model's scope and contents to the real-world features that are relevant to the simulation purposes. As the theoretical development of models is oriented toward the development of scientific knowledge, it takes place mostly in the research arena—research teams, academic network, and scientific reviews.

Model application, however, takes place primarily in the consultancy arena, owing to consultant engineers that build applied assignment models out of a toolbox of modeling software that is available to them.

Objectives and contents. This chapter fulfills two objectives. First, we look at the main simulation software packages that are available commercially, in order to establish a state of the art of their functional capacity. Second, we demonstrate the added value of theoretical improvements and their usability to advanced planning issues on the basis of two instances.

Chapter organization. The body of the chapter is organized in two sections, each of which deals with a specific objective. The section on modeling software addresses in turn (i) an external overview, (ii) system representation, (iii) traffic simulation, and (iv) application frameworks.

E. Chandakas (✉)
Transamo, Transdev Group, Paris, France
e-mail: ektoras.chandakas@transamo.com

F. Leurent
Laboratory on City, Mobility and Transportation, Ecole des Ponts ParisTech, University Paris-East, Paris, France
e-mail: fabien.leurent@enpc.fr

O. Cats
Department of Transport and Planning, Delft University of Technology, P.O. Box 5048, 2600, GA Delft, The Netherlands
e-mail: o.cats@tudelft.nl

G. Gentile and K. Noekel (eds.), *Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems*, Springer Tracts on Transportation and Traffic 10, DOI 10.1007/978-3-319-25082-3_9

The other sections explore two advanced applications implemented in a research laboratory with proprietary software: a macroscopic and static simulation of the entire Greater Paris transit network with the CapTA model that deals with a range of capacity phenomena in a consistent way, and a mesoscopic and dynamic simulation of the Greater Stockholm bus network with the BusMezzo model, an agent-based model which is convenient to simulate the traffic dynamics of both vehicles and passengers and the disruptive effects of incidents.

9.1 Commercial Software as a Bridge Between Theory and Practice

Fabien Leurent and Ektoras Chandakas

In order to apply a traffic assignment model to a practical case, one needs modeling software, which is used to describe and simulate the case and to analyze and present the results.

The practical cases are traditionally network planning studies: The aim is to relate "transit supply," represented in terms of a plan of the lines and characteristics of service, to "travel demand," in other words an origin–destination (O–D) flow matrix combined with a model of choice behavior for each individual trip. The relation between supply and demand generates particular traffic conditions on the network, in particular travel times per section, together with price and quality of service conditions per trip (in terms of time spent and level of comfort).

The objective of this section is to describe the main modeling software commercially available on the international market for handling traffic assignment.

This software has been developed gradually, over the years, reflecting advances in theoretical knowledge and the progress of IT resources (in particular the shift to PCs around 1990), as well as the needs of the clients who commission studies—local authorities and network operators—and competitive pressures. All these factors have led to improvements in the range of commercially available assignment software. The products available today (2015) have become powerful and highly ergonomic toolboxes, capable not only of processing simulations but also handling input and output data, mapping networks, and zones in the areas studied, or calibrating parameters and adjusting input data using special utility functions.

In short, our aim above all is to show and measure the functional capacity of the main commercial software packages, i.e., to establish a current state of the-art.

Our presentation focuses on the major features of the software. These major features include essential characteristics that are common to the different packages. Wherever necessary, we will point out disparities that seem to us significant.

Our presentation is divided into four parts. We will begin with an overview of commercial software packages (Sect. 9.1.1). Then, we will discuss the representation of the system, distinguishing between the supply side and the demand side (Sect. 9.1.2). Then, we present the actual simulation functions (Sect. 9.1.3).

And finally, we describe the typical applications for which assignment software can be used (Sect. 9.1.4).

In order to get specific information on the available simulation functions, we conducted a survey with the major international software publishers: in alphabetical order, CUBE (Anglo-American publisher Citilabs), Emme (Canadian publisher INRO Consultants), OmniTRANS (Dutch publisher Omnitrans International), TransCad (American publisher Caliper), and Visum (German publisher PTV). The survey questionnaire is provided in the appendix.

9.1.1 An Overview of Commercial Software

We will begin by presenting the software from the outside, distinguishing between the user interface (Sect. 9.1.1.1) and the modeling functions (Sect. 9.1.1.2) before setting out the commercial conditions for acquisition and use (Sect. 9.1.1.3).

9.1.1.1 A Very Powerful and Highly Productive User Interface

All the big software packages available on the international market are presented as very powerful and highly ergonomic toolboxes for modeling a transport system and handling traffic system studies. The ergonomics are based on a very advanced user interface, giving users clear, intuitive, and powerful access to the data and modeling bricks.

The interface is integrated, with menus that provide access to a large number of functions, along with shortcuts to the main functions.

The interface allows data to be edited intuitively: through map type windows for editing a network or processing the study area in zonal divisions; or through spreadsheet type windows, for accessing or editing datasets. In some packages, the interface even makes it possible to compose a system of models graphically, with one node for every dataset or basic model (e.g., an assignment model on a modal network), and links between the data and the models (Fig. 9.1).

It is also possible to construct one's own model, one's own modeling bricks, by using the basic functions as primitives in a programming language, whether an "in-house" scripting language or an advanced programming language such as Python.

The ergonomics are also supported by several kinds of user help functions: extensive documentation, with a user manual, a tutorial, a help function within the application, and a link to the publisher's Web site which provides a FAQ and a helpline.

9.1.1.2 The Range of Functions in a Simulation Package

The user interface is the tip of the iceberg in a modeling application. Further down, a major software package offers four types of broad "modeling" functions:

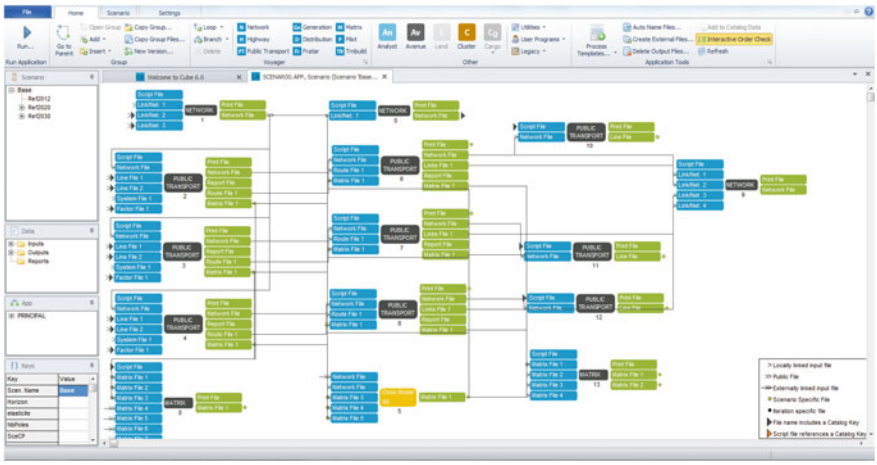

Fig. 9.1 An example of software interface for the composition of a system of models

- Data editing functions, for individual elements or batches. The user can even develop their own data types from preset types, by adding data fields, such as additional columns in the data row associated with an object in a table of objects of the same type. Each field can contain static or dynamic information (an elementary function), like in an Excel spreadsheet. The editing functions offer different ways of checking the validity of the data, in particular the connectivity and navigability of a network.
- Simulation functions: from elementary functions to composite ones for running complex processes. The minimum range of functions includes models for each of the four standard phases in the macroscopic modeling of travel demand: generation, distribution, mode choice, and assignment. We will take a closer look at the mass transit assignment category in the subsections which follow. Each big publisher now offers their own microscopic traffic simulator as a counterpart to their long-standing macroscopic package, with a common interface and special facilities to share data: Dynameq for Emme, Dynasim for Cube, TransModeler for TransCad, and Vissim for Visum. Furthermore, impact simulation functions are being added to the toolbox, in particular functions for assessing energy consumption or atmospheric or noise pollution.
- Functions for analyzing results: These are linked with the editing functions, but applied to specific data that can be organized in particular ways at the user's request. Some specific examples relating to the results of an assignment model: turning flows in a network node; flow trees, i.e., the trace of flows on the network that passes through a particular arc, so that their catchment basin can be mapped; or else highly sophisticated statistical summaries. These functions highlight the results of the model, greatly contributing to the analyst's work and interchange with the study's commissioning client.

- Inverse modeling functions, for exploiting data obtained by field observations or supplied by other models, for better simulation quality. In particular, counts of volumes passing through a point or through a couple of points, or observations of travel time between two points, can be compared with simulation results in order to assess the accuracy of those simulations. With inverse modeling, such observations can be exploited to revise the input data and parameters in the simulation model, using an ad hoc statistical method, for example, in order to estimate route choice parameters in a traveler's utility function, or to infer an O–D flow matrix that faithfully reproduces observed counts.

To sum up, the big commercial software packages offer a very rich range of functions. Simulation functions play an essential role, but they are supported, enhanced, and amplified by functions for data editing, results analysis, and inverse modeling.

9.1.1.3 The Market of Software Publishing Industry

The clients for applied studies are as follows: local authorities, planning agencies, and network operators; less frequently, user associations or environmental groups; and finally, research laboratories committed to develop knowledge (in particular for integrated transportation and environment simulations or student education).

The software customers come from these different bodies, or from specialist consultancies commissioned to conduct studies on behalf of a variety of clients.

The functional capacity and ergonomics of the commercial software are major contributors to the productivity of design engineers and their staff, fully justifying the cost of buying a package rather than developing bespoke in-house software.

Things are much the same in the academic community: Although a research laboratory may develop simulation software to try to improve the match between simulation and reality, it will generally also own a commercial package for editing data and running various additional processes.

The commercial terms for acquiring software are usually a combination of an initial payment to buy a license and an annual subscription, which provides updates and access to the helpline. The two prices can be adjusted, depending on the functions the client needs (in terms of network size and range of capabilities). The unit price can also diminish if a number of licenses are purchased. Additional services are proposed, for instance access to cloud computing and the hosting of models so as to ease the "user experience."

Market competition between software publishers has generated a common core of functions. This makes it easy to shift an application from one piece of software to another. The big consultancy firms with a large client portfolio generally have several software packages, so that they can adjust to each client and the software it uses in-house.

In addition, the publishers can cooperate industrially, by sharing software components, so that, for example, a given mass transit assignment model in TransCad comes from the Emme software.

At present, the international market seems stable: The Cube, Emme, TransCad, and Visum quartet dominate the global stage. There are "regional" packages such as Omnitrans in Europe, or Esraus and Tranus in Latin America, which have functions especially adapted to local needs, in particular for modeling certain intermediate forms of transit that are important in Latin America's big conurbations, or to link transportation and urban development (Tranus), which is important for network planning in a fast-growing conurbation.

9.1.2 System Representation

From here on, we will concentrate on traffic assignment on a mass transit network. In this subsection, we begin by presenting the general principles of computer representation (Sect. 9.1.2.1), then the specific representation of transit supply in the software (Sect. 9.1.2.2), and the representation of travel demand (Sect. 9.1.2.3). Finally, we explore the simulation functions, identifying typical results, i.e., data on the use of transit supply by transit demand (Sect. 9.1.2.4).

9.1.2.1 General Principles of Computer Representation

The simulation model deals with entities such as vehicles, passengers, and infrastructure elements. In the software, each type of entity is treated as an object in the IT sense of the term, with its data structure (a set of attributes each with its meaning and a format of variables) and specific processing procedures.

Objects belonging to a single type of entity are grouped into subsets, for example, the arcs of an infrastructure network for a given state of the system. Such a subset constitutes a database.

The different databases are linked together by a variety of relations, some of them standard in database management and others specific to simulation and to an assignment problem. In particular, the different databases relating to a single state of the transport system are handled together by a file manager.

Spatial characteristics—location, the spatial extension of the entities—are very significant in a simulation model. That is why the software has GIS (geographical information system) oriented functions to manage the spatial dimension:

- Mapping a set of geographical data: node set, arc set, arc and node network, zone set, etc.
- Carrying out spatial calculations on such entities, e.g., selecting elements on the basis of a distance condition in relation to a given point.

- Organizing the information in superimposed layers, from which the user can select a subset to be displayed on screen as a map that can be used in the analysis process.

The GIS orientation is achieved in Cube by special coupling with the ArcGIS software, whereas a specific GIS is embedded in TransCad.

9.1.2.2 Supply Data

Supply data are fundamentally organized into four levels: a lower level for infrastructure data, a second level for service data, a further level for the characteristics of the transit vehicles, and finally an upper level for sophisticated protocols.

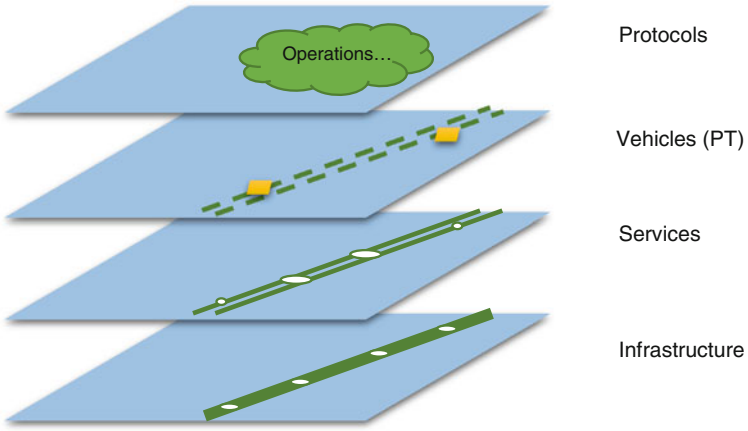

Fig. 9.2 Multilayer view of a public transport system

Infrastructure data are organized as a network, in the mathematical sense, in other words the pairing of a set of nodes with a set of arcs that refer to the nodes:

- A table of infrastructure nodes contains their identifiers and their respective characteristics, notably position coordinates, the location name, the mode(s) concerned, etc.
- A table of infrastructure arcs contains their identifier, their respective characteristics (links, nominal travel time, capacity parameters, etc.), notably with the modal type and, of course, the initial and final nodes.
- There is also a table of turns to describe the conditions of transfer between arcs that connect to a given node.

The software can map the network and in particular can show each arc by adjusting its width for a particular characteristic (e.g., capacity) and by linking its color to another characteristic (e.g., mode or travel speed).

The service data are constructed as a set of services: Each service is spatially situated in terms of the sections through which it passes (infrastructure arcs, with a reference) and the stations it serves (infrastructure nodes, which must be identified). For a static model, service frequency is also represented for each simulation period, along with the characteristics of the vehicles used (in particular, their interchange capacity at stops). For a dynamic model, in which vehicle runs are handled individually, the data entered for each run must include not only the vehicle type, but also its nominal timetable for arrival at (or departure from) each station. In a dynamic model that handles vehicle runs macroscopically, an intermediate representation is required: The time is divided into segments, and each segment is treated like a period in a static model (Fig. 9.2).

The modeling software provides special functions for editing services (and runs, if necessary). The basic representation is a trajectory shown as a straight line, to highlight the stations and sections. The dynamic model can be incorporated into a second dimension, which produces a space-time diagram for the trajectory of a run.

The "soft" components of the service are the most difficult to represent:

- The fare system, with categories of travelers defined by their payment package (single ticket or subscription), and specific spatial patterns.
- The presence of certain dynamic information at certain places.
- The relations between services that share infrastructure elements: Coordination within sections and its effects on the gaps between vehicles are represented in Visum, as well as coordination within a station between vehicles from several lines.
- The interactions with automobile traffic on the road network.

9.1.2.3 Demand Data

By restricting ourselves to traffic assignment, we reduce the representation of demand to a set of trips made by a set of individuals.

The set of individuals is represented as one or several categories of passenger: Each category is described in terms of physical behavior (nominal walking speed on flat terrain, which constitutes an aptitude parameter) and economic behavior, with a utility function formula containing parameters. The analyst specifies the utility function in the software and inputs the parameter values for a certain state of the system.

For each category of passenger, and each simulation period, trips are represented in macroscopic terms in space (distinguishing between the origin and destination zones, and identifying each OD pair) and also in time, with the intensity of trip flows for each OD and each period.

The spatial zoning and the OD matrix are fundamental features, which the software handles specifically.

Zones are characterized by their outlines in space. Zonal division can be mapped on the screen. Divisions can be made with several levels of precision. A centroid is attributed to each zone, if for no other reason than to put information on the map with a symbol whose shape (e.g., a disk with a surface area proportional to the number of trips generated) and color (e.g., based on the average distance of trips generated) are chosen by the analyst. The latter can even construct a composite symbol such as a pie chart or histogram in order to combine several pieces of information. The software can also connect the centroid to the physical elements of the infrastructure, by drawing "connecting" arcs.

Trips connect the set of origin zones with the set of destination zones. The OD flow matrix summarizes this interaction for each category and period. The matrix can be edited in the software, with matrix or table displays on the screen, and is also used to enter cell values (as in Excel). The software also offers specific representations. These include, from a selected emission or reception zone,

- a traffic "star" with a line to every zone in the network, with the thickness of the line adjusted to reflect the intensity of the OD flow;
- these kinds of "desire path" can be represented for all OD, applying a selection condition to eliminate small flows that might hide the signal.

Even in microscopic traffic assignment, macroscopic representation is important for specifying the spatial, temporal, and behavioral (passenger types) structures. The microscopic simulator then processes a set of elementary trips, each with its own specific origin and destination points and departure times. These elementary trips can be produced by random generator or imported from a demand model complementary to the assignment model (Fig. 9.3).

Fig. 9.3 The segmentation of a territory in zones (*in purple*) and the representation of the OD flows by desire lines (*in yellow*)

9.1.2.4 Data on the Use of Supply by Demand

The analyst has to specify certain conditions for passenger use of the transit supply: For each category of passenger, these conditions are the rights of access to particular services, the fare conditions and individual sensitivities to local travel conditions (discomfort function per time unit according to local physical state, sitting or standing, walking or still, with respect to passenger density), and also the conditions of route choice:

- Access to static information and dynamic information. The utopian assumption is usually that the user has total information on times and frequencies. In Visum, it is possible to adjust the level of information by a series of options reflecting specific physical conditions.
- The identification of decision points, i.e., places where passengers compare alternative options for reaching their destination. The software processes this choice at every through node on the infrastructure network—which is specially recoded for this purpose, though the coding is transparent for the analyst. With Visum, the bundle of routing options can be composed locally or for whole legs, which distinguishes the initial and final nodes of the legs from intermediate stations where no choice issues arise.

In addition to exogenous data for specifying the usage conditions, there are "endogenous" data, which result from simulation—hence they are produced synthetically. The results of an assignment model essentially relate to the use of supply by demand, at different levels that reproduce the layers of supply (see Sect. 9.1.2.1) or relate to demand. In fact, the assignment of a trip on the network is the basic element that characterizes the use of supply by demand. The basic result is a simple path between two nodes—respectively origin and destination—or a hyperpath or a multipath: whatever the case, a subnetwork in terms of infrastructure and missions (or runs in the case of a dynamic model) with a proportion of use on each arc.

The simulation also produces distances, physical times, and generalized costs between any node on the network and any destination zone. These results are important quality of service indicators. They can be collected for each passenger category and departure (or arrival) period, for all O–D relations, in a dedicated O–D matrix. From this, the software can be used to derive isochrone or accessibility maps.

Let us return to the constituent elements of supply. The stacking of elementary trips on the network produces local flow loads per location and per period, if necessary distinguishing between trip-maker categories. Typical results are as follows:

- A map of local flows on the framework of infrastructures or services; the thickness can be adjusted to reflect flow intensity, and the color can be set, for example, to indicate speed or the local ratio between volume and capacity.
- Locally, the representation of flows on the turning movements at a node.
- For a given section, a network map of its flow tree.

- A map of physical times for each network section, or of generalized costs.
- For each service (or run), the "load line graph" is a diagram of the sections through which the service passes, with the thickness of the line in each section being proportional to the volume borne at local level.
- For each service (or run), the physical time (or generalized cost) from the origin up to the destination can be illustrated at each point along the trajectory in a space-time or space-cost diagram. The representation by service is important to show dwelling times at the different stations as well as the times spent in the different sections.

By contrast, it is more difficult to map the times spent by passengers in stations, as well as the number of passengers waiting on the platform. Maps of the frequency of runs passing through stations are proxies, imperfect in several respects.

9.1.3 Traffic Simulation

Let us move on to the essence of modeling software: the capacity to simulate traffic, whether passenger traffic or vehicle traffic, and the interaction between them.

A model's expressive power consists in the range of phenomena that it can describe and the relations of cause and effect it can explain.

We will begin by looking at the features relating to passengers, i.e., demand, distinguishing between the individual or disaggregated features (Sect. 9.1.3.1) and the collective or aggregated features (Sect. 9.1.3.2). We will then discuss the features relating to vehicles, hence to supply (Sect. 9.1.3.3). Next, we will describe the treatment of variabilities and other special features (Sect. 9.1.3.4). Finally, we will discuss the scope and limitations of the main simulation software packages, as they stand in 2015 (Sect. 9.1.3.5).

9.1.3.1 Modeling the Individual Passenger

We have already seen that travel demand is modeled as a set of traveling passengers, for each OD relation and each period, and for each type of behavior (user category).

We have also mentioned the microscopic and macroscopic approach to modeling passengers. In fact, any assignment model handles passengers at two levels: the disaggregated level of the individual passenger, discussed in this paragraph, and the aggregated level of passenger flows, covered in the next paragraph.

At the disaggregated level, the model places an individual passenger in a situation of travel, identifying and characterizing the elementary traffic conditions in the places the passenger passes through in the course of the trip. The passenger's exposure to these local conditions is modeled firstly in terms of time spent, in relation to the passenger's individual characteristics (in particular walking speed),

and secondly in terms of the conditions of comfort, in particular with respect to local crowding density.

The degree of detail depends on the software used and the options selected by the analyst:

- Vehicle carrying capacity has to be stated in order to model crowding phenomena.
- Platform waiting time (WT) is better handled in a dynamic simulation than in a static simulation. The particular value for the passenger should depend on the statistical distribution of vehicle headways, but this influence is often modeled as a one size fits all value.
- As things stand in 2015, the different packages do not distinguish positions within a vehicle, or even the cars that form a train, or waiting positions along a platform.
- The distinction between sitting and standing positions in a vehicle is only partially incorporated, without reference to the order of precedence over passenger access to seats, or whether people remain seated until they leave the vehicle.

Passengers are also exposed to fare conditions: In this respect, the software packages offer a wide range of fare options, per arc, per line leg, per entry–exit on a network with integrated fare schemes, etc.

The passenger's exposure to local conditions produces an overview of the itinerary as a whole, encapsulated in a generalized cost. The software can provide the generalized cumulative cost along a path, weighting each period of time spent in a certain state of comfort by a unit discomfort cost that can be linked to the degree of crowding density (often by means of a script developed by the analyst). Certain packages, notably Cube and Visum, can process cumulative individual accounts, for example, the number of connections in the course of a total trip, or the distance traveled on foot, in order to eliminate route options that exceed a threshold value set by the analyst for the particular criterion.

Beyond representing the passenger's exposure and perception of disutility, the passenger's route choice is the essential purpose of the assignment model. The different packages offer a range of possibilities:

- Assignment to the shortest path, assessed as if for individual transportation with average WTs for access to a service: in particular the "Sketch Assignment" option in Visum.
- Assignment to the shortest hyperpath, especially for static assignment, and on the assumption that the choice is repeated at every node that is significant for the passenger (junction along a pedestrian route or transit station). There is no option for leg-based assignment. See the frequency-based models in the previous chapters.
- An alternative model to "standard" hyperpaths is based on the notion of a user preference set (UPS), i.e., an ordered list of paths from a standard node, pictured

by the user in an order that reflects his or her preferences, and whose availability depends on residual capacities and the numbers of potential local users. The UPS model is available in Cube and Emme.
- Assignment to a multipath: a bundle of elementary paths, with a local distribution model that depends on the respective costs (e.g., logit discrete choice model). This option is available in Cube, Omnitrans, and Visum. It is the standard option in Visum for schedule-based dynamic simulation.

In all these models, the passenger's choice is treated as a complex calculation that includes large quantities of information of different kinds. Modeling the amount of information that a passenger actually identifies and handles is an important goal. The existing options are still rudimentary:

- In Omnitrans, the scale factor in a logit model of local distribution can be adjusted from one node to another.
- In Visum, dynamic information can be represented with high precision, but this still assumes that the passenger's capacities for calculation are extensive.

So constructing the passenger's particular universe of choice is a real focus of complexity in the assignment model.

Two further aspects add further complexity: multimodality and departure time choice.

Multimodality is first modeled upstream of the transit assignment model, in a mode choice model. However, most of the software packages can model the choice of access mode from the point of origin to the transit network, between a pedestrian option and a park-and-ride option: by a specific model in Omnitrans and Visum, or by a script in Emme and in Cube. However, the other transit modes—individual or collective taxis, shared car, or motorcycle systems—are not taken into account, and passenger itineraries cannot combine mass and individual transit modes.

Departure time choice is also modeled upstream of the actual assignment model: The passenger's journey is assigned to a particular period. Visum's dynamic model characterizes the trip on the basis of the scheduled arrival times, in order to construct routing options that are compatible with that timetable and thereby to simulate some departure times that can be chosen by the passenger.

9.1.3.2 Modeling Passenger Flows

In an assignment model, passenger trips are aggregated in several circumstances:

- Exogenously, for each demand segment in terms of the O–D relation, the period and also the behavioral category.
- Endogenously, in terms of the places traveled through and the time segments, as in assignment on an individual transit network, and also in terms of the vehicle used at a given moment, which is specific to transit assignment.

With regard to exogenous aggregation by demand segment, the software can link the volume of demand to the generalized individual cost, in an elastic demand model.

Endogenous aggregations depend on how fine-grained simulation is as follows:

- A static simulation does not take account of the diversity of the runs used to complete a mission. It considers only an average load.
- Within a mass transit vehicle, the 2015 software can distinguish between sitting and standing spaces, but cannot accurately allocate them to passengers. The individual cars making up a train (or an articulated bus) are not distinguished.
- Passenger numbers on the platform and their spatial distribution are not handled with the same degree of attention as flows per section.
- Pedestrian elements, in particular along station corridors, are reduced to network links with associated average travel time, which does not reflect the passenger type nor the crowding density along the route.

This latter feature could be handled with specific scripts. However, the features above are inaccessible to the analyst with current software.

9.1.3.3 Vehicle Traffic and Supply Management

In a mass transit assignment model, the vehicle as a discrete entity is always modeled, at least notionally, even in a static model based on mission frequency. Each service is supplied in the form of a number of runs, and each passenger accesses one particular vehicle.

The simulation software can represent the types of rolling stock and therefore describe each vehicle by features such as the number of seats, the space for standing passengers, the number and width of the doors, and the relative height of the door threshold. Their capacity to contain passengers and exchange them within a station is fundamental characteristics of a mass transit vehicle. There are several processing levels:

- For each resource (interior space or passing point), the ratio V/C between a volume of passengers V and a capacity C is a congestion indicator that can be used as an explanatory variable in a passenger discomfort function.
- At greater depth, modeling seated places and their occupancy and assigning passengers to different states of comfort. In this respect, the existing modeling capacity in commercial simulation software has room for improvement.
- The total capacity of a vehicle can be modeled, with its effects on platform WT. The Cube software establishes a probability of boarding based on residual capacity and the number of passengers waiting to board. The Emme software can force the boarding flow to reflect constraints by increasing passenger station WT, by means of a fictional frequency associated with the service but specific to the boarding station.

- Modeling interchange capacities: Commercial software packages can link a vehicle's dwell time in a station with the flows of passengers leaving and boarding, on the basis of interchange capacity. At present, this phenomenon is represented as an aggregate, even in microsimulation models: This is because for a disaggregated representation, there would not only have to be distinctions between one passenger and another, but also between one door and another along the vehicle, as well as between the individual positions of passengers inside the vehicle and along the platform.

Obviously, dynamic models are more appropriate for modeling capacity phenomena.

So the interaction of the vehicle with each passenger can be modeled with a certain degree of validity. The model derives consequences for passengers in terms of WT and discomfort levels, and also for the vehicle in terms of crowding and station dwell time.

Travel time on a section can also be modeled endogenously,

- If the operator maintains regular gaps between successive runs on a given service: Omnitrans and Visum provide an ad hoc model.
- If the operator synchronizes the arrivals of several runs of different services in a single station, or spaces the passage of different service runs over time on a single section on the common trunk line: again available in Omnitrans and Visum.
- In addition to these interactions between vehicles which are specific to mass transit modes, there are interactions between vehicles on a road network. The software packages can incorporate travel times per arc into the mass transit assignment model, as well as the junction crossing times which result from traffic assignment on the road network.

In dynamic models, other interactions between vehicles and other traffic management measures by the operator can be modeled, provided that ad hoc scripts are developed.

For example, modeling the vehicle in the traffic assignment model tackles the interaction with passengers and movement on the infrastructure at the level of a single run. One additional capacity constraint could be handled with a specific script: the size of the vehicle fleet and its effect on service frequency, in relation to the cycle time, which includes the total station stopping time and the total section running time.

9.1.3.4 Special Features

Coordinated vehicle and line management require the operator to centralize information on the current situation of the resources concerned and reciprocally to send real-time instructions to those resources.

With regard to the informational interaction between users and the transit system, dynamic software can model the local supply of dynamic information, and from this, we can derive the consequences regarding path choice from the place in question. However, the model does not yet include customized route advice given to users by an information provider: This could be handled by a specific script.

The software can represent the diversity of demand segments and the diversity of transit locations, modes, and services. Over and above these major but obvious diversities, there remain variations and heterogeneities that the packages do not capture (at least not yet):

- The precise location of a passenger inside a vehicle or a particular functionalized space (station platform, station corridor, etc.) and therefore the spatial heterogeneity of passenger distribution within such a space.
- Demand fluctuations between days within a given type. Typically, microscopic traffic simulators use Poisson processes to generate passengers, which sharply restrict the variations between several repeated simulations. The statistical distribution of fluctuations over time needs to be modeled more closely for each segment, as well as the correlations between segments.
- The mechanisms whereby an individual passenger learns from the conditions experienced on previous days are not explicitly modeled.
- Disruptions that affect service operation can be simulated, starting from an initial event specified by the analyst. A statistical structure of disruptions could be simulated using an ad hoc script. However, the commercially available software packages do not provide an explanatory physical model for the causes of disruptions, although they can derive the consequences.
- The parameters for the regularity of a service can be set exogenously. On the passenger side, the reliability of a route can be given a value in the generalized cost function. However, the software cannot make regularity endogenous, nor the effects of regularity on passengers in terms of reliability.

9.1.3.5 Outreach and Limitations

The software packages described here all have extensive powers of description and explanation. They capture the travel infrastructures and the stations, the services and service runs, the passengers with their particularities, and their individual behaviors.

In other words, the software captures a significant and already very complex part of the reality of the system under consideration. Entity-based modeling achieves a high degree of consistency:

- The passenger is treated as an entity, in terms of the path followed. Integrity is maintained in a microscopic simulation better than in a macroscopic model, where the entity is distributed across a bundle of itineraries on the basis of local proportions.

- The vehicle is treated as an entity in microscopic simulation and at least as a notional entity in macroscopic models. Its individual trajectory is captured, as well as the continuity of passenger flow as people board and leave.
- The spatial elements—nodes or arcs—are also entities. Their explicit handling contributes to the realism of the model.

Nonetheless, the models remain simplifications of reality. We have identified various gaps or imperfections in the representation of physical phenomena and economic circumstances, for example, the lack of precision in the internal description of a vehicle, in the location of a passenger within a space, and therefore in the spatial and temporal relations between entities. These shortcomings are clearly potential avenues for future development of the models. One pragmatic solution is to combine the assignment model with external models: with mode choice or departure time choice models to simulate individual mobility behaviors better, or with pedestrian traffic models such as Legion, PedSim, Simwalk, and Viswalk for a more fine-grained simulation of passengers interacting with places and vehicles and of crowd dynamics.

Finally, the market software only allows "positive" modeling of the system, on the principle of user equilibrium with regard to the individual user. This positive modeling is incomplete, since it leaves out the economic behavior of the operators. There is also a lack of normative modeling, which would provide guidelines for operators and regulators in improving system performance: assignment based on the principle of system optimum is fairly well developed for road networks, but there is no equivalent for assignment to a mass transit system.

9.1.4 Application Frameworks

The functions included in travel modeling software are primarily intended for application studies. We will now describe the main types of application in terms of the objectives, the methodology, and the modeling options. We will tackle successively:

1. System simulation and diagnosis
2. Impact analysis and assessment
3. Cost–benefit analysis
4. System planning
5. Operational management
6. Commercial management.

9.1.4.1 System Simulation and Diagnosis

The aim is to simulate the operating state of a real system under exogenous conditions of supply and demand. The challenge is to reproduce the complexity of that system and to understand the respective influences of its components.

This process may be applied to a hypothetical situation for planning purposes, or to a real and observed situation, in order to calibrate the simulation model and/or enrich the observations with information generated by the model. In the latter case, the simulation contributes to system diagnosis.

In this kind of application framework, the analyst needs to reconcile the technical and financial resources available with the level of complexity of the application, which depends on the dimensions of the system (number of modes, number of lines, number of missions and stations, diversity of passengers). As things stand in 2015, static models broadly continue to be used for big conurbations with several hundreds of lines, to study ordinary operating conditions, particularly at peak hours, in order to set capacity levels. Dynamic models are used for smaller networks, with a few dozen lines, in order to analyze extraordinary conditions, interactions, or details which cannot be represented in static models.

Incident analysis adds a further dimension of complexity: Understanding all the effects requires a dynamic model, but the variety of possible incidents generally means that the model is restricted to one line or a handful of lines.

9.1.4.2 Impact Analysis and Assessment

Simulating a system involves cause-and-effect mechanisms and establishes a whole series of impacts. The direct impacts relate to the state of supply and to the state of demand, both of which are influenced by usage conditions: local vehicle load, infrastructure occupancy, quality of service, and price per trip.

From these direct impacts, a series of more or less direct consequences can be derived:

- The technical productivity of the operating resources, and their economic productivity, by relating production to operating costs.
- Surplus demand.
- Energy consumption, greenhouse gas emissions, and atmospheric and noise pollution.
- The effects of noise pollution on the local population.
- The global effects of energy consumption and greenhouse gas emissions.

The assessment of a variety of impacts can contribute to a multicriterion analysis of the system.

9.1.4.3 Cost–Benefit Analysis

For a mass transit system, a development operation can consist in the introduction of a new line, or the reconfiguration of existing lines, or an in-depth overhaul of service schedules, etc. This kind of operation not only has consequences for the operation of the system, but also entails specific investment costs to transform the system. These costs need to be assessed in a specific way, outside the system simulation.

Simulation of the system is important to establish the effects in different circumstances:

- During the construction phase;
- In ordinary operation, first on initial start-up, then as a series of subsequent milestones, in order to check for long-term returns on the investment.

This means that the simulation model needs to be applied several times, incorporating progressive changes in supply and demand, to be specified in incremental change scenarios.

This complexity over time is managed in the main commercial assignment software packages by special functions which specify scenarios and run a series of system simulations, as well as a multicriterion impact assessment for each state of the system.

The accumulation of impacts over time depends on the type of impact. For example, greenhouse gas emissions have an impact on the climate that depends on the date they occur, partly relating to the greenhouse effect at that time, but partly to the persistence of the chemical species in the atmosphere. Similarly, for noise pollution, the local population exposed may vary over time, as may social sensitivity to noise levels.

The usual economic method is to assign a monetary value to each impact and to measure the combined impact with a single aggregated indicator representing the socioeconomic and environmental impact of the investment.

This type of application demonstrates both the intense need for traffic simulation, the need to assess an array of impacts, and the benefit of application software that can integrate the different assessments to a maximum.

9.1.4.4 System Planning

Planning a mass transit system is much more complex than planning a road network, because action is possible at multiple levels: on the infrastructures and on the service lines, their particular mode and their combination of services, on the services in terms of trajectory and stops, and on the run rate in each service; on the stations, the conditions of outside access and the conditions of connections between lines; and finally, on prices, more an issue in mass transit networks than on the road network—at least in the urban setting.

In principle, a mass transit assignment model is well suited for specifying and simulating a transit scheme that involves all these components. However, real systems are highly complex, and each component of the scheme needs to be specified at its own spatial scale: In particular, the station layout and the specification of the connection conditions are a critical factor for a network model, which needs to be combined with a local model in order to simulate the features concerned with sufficient realism. This is also true for the transversal profile of the road axes that carry bus facilities, which need to be simulated with a microscopic road traffic model. Moreover, in order to design in a valid way the operation of a bus line, and even more of a railroad line, expertise in traffic engineering is needed; however, this kind of expertise is not included in the commercial network simulation software available in 2015.

9.1.4.5 Operational Management

Among system plans, the operating plans for a properly configured network constitute a fairly well determined subset. It entails designing the schedule of vehicle runs, the schedule of drivers, and the allocation of drivers to vehicle runs, or else, where applicable, adjusting the size of the vehicle fleet or the number of drivers.

In principle, a traffic assignment model can be used to test the effects of a run schedule on demand and usage. It does not deal with the question of drivers. The size of a vehicle fleet, typically for each line, could be handled by a specific script added to the modeling software.

An important area for future development is to involve assignment models into real-time traffic management on a line and on a network: We will return to this in subsequent sections.

9.1.4.6 Commercial Management

The commercial relationship between the mass transit network in general, run by one or more operators, and all its users considered as customers is a field that has so far remained little explored by simulation for an urban environment, by contrast with the interurban environment.

Traditional studies relate to setting fares for a new line or a general change in pricing for the entire network. Applied assignment models are not always very relevant in this respect, since fares represent only a minor part of the generalized cost to individual trip makers.

The distribution of fare revenues based on a subscription to multiple transit modes, for example, the Navigo card in Paris, which is divided between two railroad operators and several bus companies, is a subject better suited to assignment models, which can be used to estimate the contribution of each operator to the distances traveled by passengers under the relevant fare scheme.

An assignment model can also be used to estimate the impact on demand of providing dynamic information locally, by simulating the effects of a targeted system first on perceptions of the generalized cost and then on route choice.

Applications still need to be designed so that assignment models can contribute more to commercial management. One obvious area is traffic analysis in terms of the social class of users, because transit plays an important social role for the poorest households, for whom there may be a case for the provision of special fare arrangements.

Other avenues to explore concern yield management, in order to make urban mass transit supply more sensitive to needs on the demand side, to the diversity of those needs from one category of user to another, and to their variations over time.

9.2 Advanced Applications and Research Prototypes

Fabien Leurent, Ektoras Chandakas and Oded Cats

In this section, two research softwares are presented which include innovative features for the simulation of public transport networks and then allowed the development of relevant projects in important cities.

9.2.1 Simulation of Greater Paris Using the CapTA Model

Greater Paris has a dense public transport network that offers a wide range of road and rail options. With 8.5 million daily journeys (OMNIL 2012) made on public transport, the system is under constant pressure. Matching supply on the network to the demand for transport is therefore a crucial issue.

This study does not claim to offer a diagnosis of the Paris region's PT system. Although it identifies some of the sticking points on the Greater Paris network, its main aim is to illustrate the original capacities of the CapTA model and its behavior. We start by describing transport supply and demand (Sect. 9.2.1.1) and the model's variants and parameters (Sect. 9.2.1.2), which we used to carry out the simulation for the Ile-de-France region. We then examine the assignment of passenger flows on the network (Sect. 9.2.1.3) according to two alternative model specifications and the effect of capacity constraints on the lines' operation (Sect. 9.2.1.4). Lastly, we evaluate the impact of capacity constraints on passengers (Sect. 9.2.1.5), before concluding with a discussion of the salient features in this simulation (Sect. 9.2.1.6).

9.2.1.1 Demand and Supply of Public Transport in Greater Paris

A simulation involves assigning a demand for transport to a network according to the supply. We apply the simulation to the morning rush hour in 2008, as described by DRIEA (the state agency for regional and interborough department infrastructure and development). More precisely, Ile-de-France is split into 1305 Traffic Assignment Zones. The trip demand between the zones is described by a 1305 × 1305 matrix of OD flows and totals 1.15 million trips for 1 h. We want to highlight the model's behavior and capacities, and so we use the OD matrix homogeneously enlarged by 30 %.

The public transport network in Greater Paris is characterized by a wide range of modes of transport and types of service. Buses, trams, subways, regional express trains (RER), and railway lines run through the region offering different services in terms of frequency, commercial speed, and capacity. The transport supply we consider corresponds to the annual service in 2008, as described in DRIEA data, with several necessary modifications. Concerning vehicle characteristics, we defined 17 types of vehicle on guided transport systems and 7 types of bus and coach, which we then associated with services (Chandakas 2012).

Lastly, the service network is transformed into a calculation network. The process involves creating line legs that correspond to entry and exit station pairs along a line (see second subsection in second part). In total, the calculation network comprises 160,000 nodes and 307,700 arcs and service and line legs (of which 30,000 are line legs).

9.2.1.2 Specifying the Model's Sensitivity

The main aim of this study is to identify how capacity constraints affect the behavior of the CapTA model. This involves comparing two variants of the model, either with or without account of capacity constraints. Common to the model's alternatives are the parameters linked to the generalized time (GT). To calculate the GT of a trip, we multiply the physical time inherent to each component (according to a passenger's specific physical state) by a coefficient that indicates the discomfort. If we take 1 min of being seated in vehicle as a reference, the multiplicative coefficients for waiting on the platform, access, and transfer come to 2.

A set of parameters chosen for the local models and capacity constraints included in the CapTA model can be used to specify each model variant. The two variants used in this study are defined below:

- UC: This is the baseline alternative with no account of capacity constraints. Most concrete planning studies are still carried out in this way. In this case, passengers are concentrated on the most efficient routes, in terms of nominal performance (scheduled run time and frequency, no in-vehicle discomfort). We can expect the structural lines of the network to be heavily loaded—maybe beyond their nominal capacity.

- CVCW: This variant includes all capacity constraints. We distinguish rail transport modes from road transport modes. For the former, the constraints represented pertain to total vehicle capacity (platform and waiting model), seated capacity (in-vehicle comfort model), and line capacity (platform occupancy model). For the latter, the constraints concern the occupancy of seats and in-vehicle comfort. Concerning the in-vehicle comfort model for all modes, the discomfort coefficient varies depending on the density of standing passengers, ranging from 1.2 for the first standing passenger to 2 for 4 people/m^2, which corresponds to maximum density. Taking capacity constraints into account, passengers can choose between different lines, avoiding overcrowded lines by opting for longer routes.

On the UC variant, an equilibrium state is computed in one iteration only, whereas on the CVCW variant, the determination of an equilibrium involves a series of iterations. The simulation's convergence level is evaluated as the average gap in passenger flows on the arcs from one assignment to the next. An acceptable level of convergence is reached after 50 iterations, with an average gap that is reduced to 1 ‰ of the initial value.

9.2.1.3 Assignment Results: Passenger Flows on the Network

To system planners, the principal result of a traffic assignment model is the flow of passengers on the arcs of the network. The variant without capacity constraints (UC) corresponds to the choice of optimal hyperpaths without taking into account the effects of capacity constraints on passengers' route choice. The simulation results indicate that some structural lines are heavily loaded, notably so the RER A (east–west) and RER B (north–south) lines (the acronym RER stands for Regional Express Railways). More precisely, on RER A, westbound flows on the most loaded sections reach 100,000 passengers per hour, for a supplied capacity of 58,000 passengers: a ratio of 1.7 flow compared to capacity. This corresponds to an excess of 42,000 passengers who—taking into account capacity constraints—must choose an alternative route to get to their destination. In total, flow exceeds nominal capacity on 58 sections out of the 1750 sections of rail modes.

In the capacitated variant CWCV, passengers choose their route according to the generalized cost and local availability of lines, resulting from various local capacity constraints: the total capacity of the vehicle and the transit service, the occupancy of seats, and the vehicle capacity of the platform. Clearly, taking capacity constraints into account affects route choices. Passengers who face long WTs and low levels of comfort opt to transfer to alternative routes in order to reduce their perceived cost, as evaluated by the platform storage and waiting model and the seat capacity model.

9.2.1.4 Operating Services Under Capacity Constraints

The CapTA model (on the CWCV variant) is unusual in its sensitivity to how passenger flows affect the operation conditions of a service, followed by performance of the lines. The line model deals with this vehicle capacity constraint by reducing the service frequency. The simulation of the Greater Paris network leads to two remarks on this local model:

- We cannot faithfully represent this concrete phenomenon in an assignment without capacity constraints (UC variant). By including capacity constraints (as in CWCV), some sections are relieved and passenger flows are transferred to alternative lines, thus modifying the flow structure. Consequently, the sticking points in the respective results of the two variants (UC and CWCV) are not identical, and the redistribution effects influence the location of flow transfers and their size.
- Frequency adaptation is primarily triggered by an overflow of the service exchange capacity, since the dwelling times of vehicles depend on the product of boarding and alighting flows by elementary times that depend on the vehicle exchange capacity. This exchange capacity involves the number and width of the vehicle doors, which serve as passenger channels for boarding and alighting. On a given line, a frequency reduction may take place upstream from the most loaded sections and thus reduce the capacity available on sections where it is most strongly required.

The simulation of the Greater Paris network suggests that a dense network with massive flows of passengers undergoes a frequency reduction on some of the structuring lines. Figure 9.4 shows the drop in frequency simulated by the CWCV

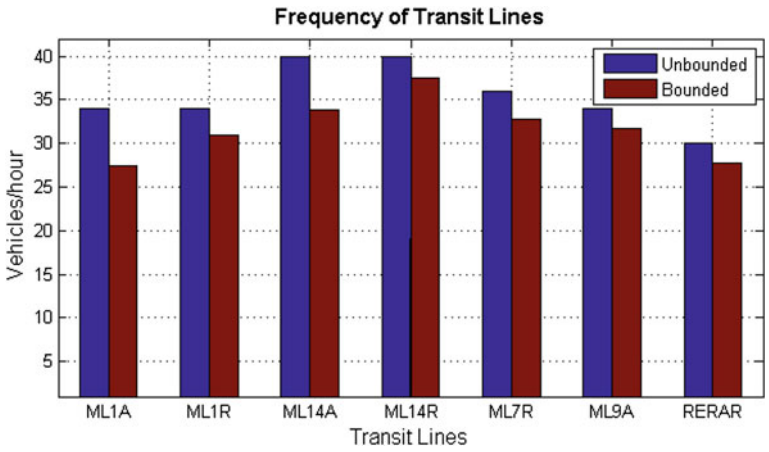

Fig. 9.4 Comparison between nominal frequency and adapted frequency at the terminus of several structuring lines on the Greater Paris network

variant of the CapTA model. We can see that the frequency of the M1 metro line (east–west) drops from 34 veh/h (nominal frequency) to 27.4 veh/h eastbound and 31 veh/h westbound. Similarly, line M14 (automatic) undergoes a reduction in nominal frequency (40 veh/h) of 6–15 % depending on the direction, while the frequency of the westward RER A decreases from 30 to 27.7 veh/h, or 7.7 %. This reduction in frequency has significant secondary impacts: 7.8 % less capacity on the line, or 4500 passengers and 1400 seats during the rush hour. The lost capacity is the equivalent of 1.7 double-decker trains.

9.2.1.5 Consequences for Users

The average generalized cost on the network includes WT and in-vehicle transport time (IVTT), as well as time spent on walking to transfer between two lines and times to access the network (at the origin or destination of a trip). Table 9.1 summarizes the average GT of variants of the CapTA model and details the components. Let us analyze the composition of the average GT of the CWCV variant. The WT corresponds to WT on the first line and WT included in transfers. The average perceived WT comes to 29.1 % of the total time. At 41.5 % of the total time, in-vehicle time constitutes the largest component of GT, whereas walking time during transfers comes to 5.6 %, and access time constitutes 23.8 % of GT.

It comes out that the GT of the CWCV alternative increases by 11.2 % compared to the UC alternative. Among the components of GT, in-vehicle time increases the most (23 %). In fact, ¾ of the increase in GT can be attributed to in-vehicle discomfort. On the other hand, the increase in WT is limited, which indicates that the lack in total service capacity applies to a subset of sticking points which is fairly limited on the network scale.

9.2.1.6 Discussion of Simulation Results at the Network Level

The Greater Paris public transport system, which includes 13 rail transport lines, 14 subway lines, 4 tram lines, and several hundred bus services with overlapping lines on the central sector, provides an ideal field to test an assignment model with capacity constraints. From the simulation results, it comes out that significant flows of passengers are concentrated on structuring north–south lines, such as RER B and

Table 9.1 Average generalized time (in minutes) on the Greater Paris network

Model variant	Optimal generalized time	Actual travel time	Perceived waiting time	Perceived in-vehicle time	Perceived transfer time	Perceived feeding time	No. of transfers per trip
UC	61.56	40.63	18.79	23.10	3.96	15.71	1.42
CWCV	68.45	41.70	19.90	28.40	3.88	16.27	1.35
%diff	11.2	2.63	5.9	23	−2.04	3.6	−5.13

metro line M13, and east–west, such as RER A and metro lines M1 and M14. By integrating capacity constraints, flows can be spread out to alternative routes in a more realistic way. Flows on the lines only occasionally exceed the nominal capacity.

The passenger flow on the network influences lines' performance and especially their service frequency. This frequency decreases by as much as 19 % for some lines and directions, with secondary effects caused by the drop in downstream capacity. For passengers, capacity constraints contribute to increasing the average WT by up to 14 and 22 % for metro lines M1 and M11, respectively. This moderate increase corresponds, however, to 230 additional hours of passenger waiting on line M1 during a single morning rush hour. In terms of quality of service in vehicles, the average standing time for a passenger depends on the structure of demand and the topology of the line. The longest times are mainly on long radial lines, such as metro line M8, with individual values of 5.5 and 7.5 min depending on the direction.

9.2.1.7 Focus on Line A of the RER (Regional Express Railway)

The CapTA model acts on two superimposed levels: At the upper layer of network route choice, the passenger flow is assigned to routes and lines depending on the passenger cost of these items; on the lower layer, the model evaluates the cost of entry–exit station pairs along the line according to the passenger flow by leg that are determined on the upper layer.

Let us from now on restrict our analysis of the simulation results to the RER A line, which is the busiest in the Greater Paris network, often accommodating over one million travelers per day. RER A comprises 46 stations spread over 5 branches and the central trunk. Two branches go eastward (Marne-la-Vallée (MLV) north–east and Boissy-St-Léger south–east) and three go westward (Cergy-le-Haut and Poissy north–west and St-Germain-en-Laye south–west), converging on the central section. As a result, eastern and western suburbs are connected to central Paris and the La Défense business area.

We shall describe more precisely the initial supply and demand data linked to RER A (in subsection 9.2.1.8). In subsection 9.2.1.9, we look at the results of the reduced frequency model, and in subsection 9.2.1.10, we present the behavior of the passenger stock on the line's central trunk, as it comes out from the local model of platform storage and waiting. We then highlight the consequences of integrating capacity constraints into the GT of journeys on the line (subsection 9.2.1.11) and end with a discussion of the salient points of the line model (subsection 9.2.1.12) (Fig. 9.5).

9.2.1.8 Supply and Demand on RER A

The subnetwork of the RER A line comprises 1200 service nodes and 182 initial arcs, transformed into 1702 line legs. Demand, resulting from the CWCV

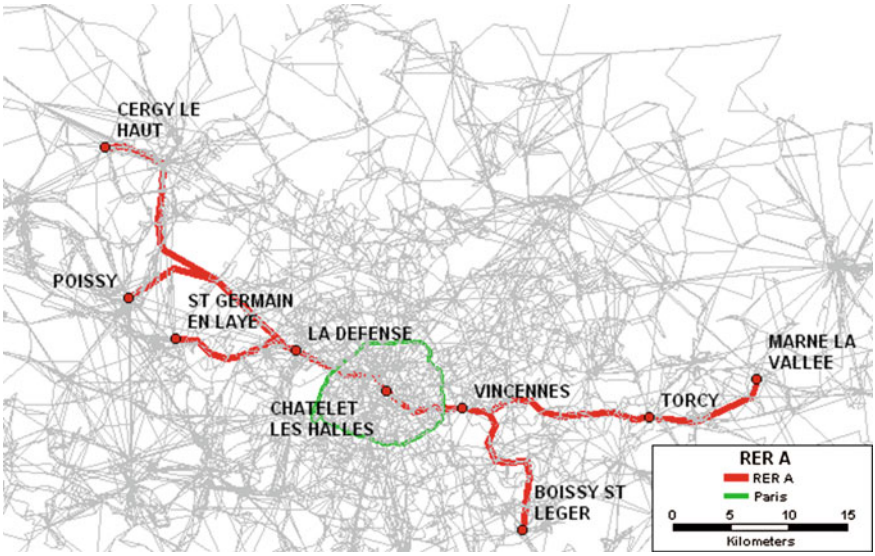

Fig. 9.5 RER A (*in red*) in relation to central Paris (*green ring*) and the La Défense business district

assignment (see Sect. 9.2.1.2), is made up of 107,000 eastbound passengers and 141,000 westbound passengers; the total is 248,000 passengers in both directions for 1 h in the morning rush.

We focus the analysis of the results in the direction of the morning rush, i.e., from eastern areas and central Paris toward La Défense. During the most intense hour of the morning rush, transport supply comprises 18 trains on the MLV branch and 12 trains from Boissy-St-Léger (BOI), which converge before the Vincennes station. Consequently, 30 trains/h are scheduled to travel on the central trunk to La Défense. Out of them, 4 trains/h terminate their service at La Défense, while the remaining 26 diverge on the three branches, including 5 toward Poissy (POI), 5 toward Cergy (CRG), and 16 toward St-Germain-en-Laye (STG). The westbound line comprises 15 different services, each of which is characterized by a service frequency, a subset of stations where it stops and a type of vehicle. The characteristics of the latter are given in Table 9.2.

9.2.1.9 Dwell Time and Service Frequency

A vehicle's dwelling time at the platform depends on the number of passengers at alighting and boarding, the type of vehicle, and the interface between the platform and the vehicle. In the CapTA model, the dwell time is calculated from the number of passengers boarding and alighting, according to the number of passage units (individual passenger channels through the door). The basic time for a passage is

Table 9.2 Types of vehicle on RER A and their characteristics

Type of vehicle	Seated places per train	Total train capacity	# doors per train side	Passenger channels per train side
MS 61	600	1888	36	72
MI 84	432	1760	32	64

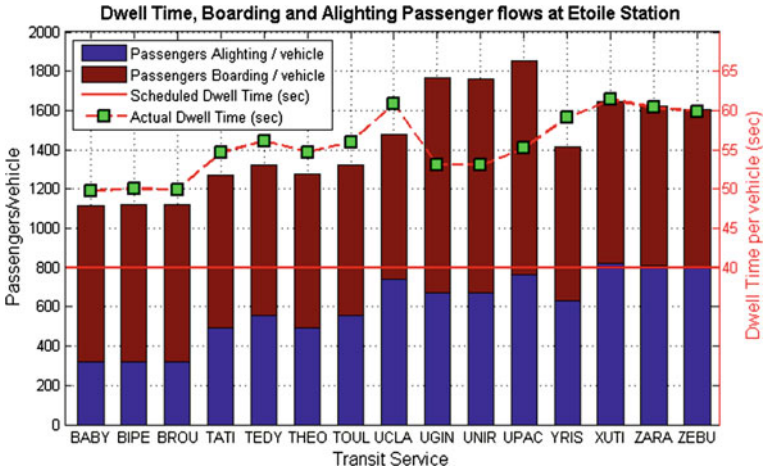

Fig. 9.6 Volume of boarding and alighting passengers and dwell time of westbound vehicles at Etoile station

1.55 s/passenger and is assumed independent to vehicle and platform congestion. In contrast, we simulate a suboptimal use of door channels (due to the use of folding seats and other operating features), which are reduced by one-third.

The line model provides results disaggregated by service. Figure 9.6 shows the dwelling time simulated versus planned at Etoile station (located before La Défense) for westbound vehicles. The volume of alighting passengers per vehicle is shown in blue, with the boarding passengers in brown. We can see that the scheduled dwell time of 40 s (continuous line) is not maintained and that vehicles dwell for between 50 and 61 s, depending on the exchange volume and the type of vehicle. The longest dwelling time (UCLA service) does not correspond to the services with the biggest exchange volume (i.e., UGIN, UNIR, and UPAC). This is because these three services use MI2 N vehicles with a larger exchange capacity than the UCLA service's M184. This example shows that the line operations need to be modeled at a sufficiently disaggregated level and that in practice, the assignment of vehicles to the line services should depend on the vehicle capacity to match the demanded flow.

9.2.1.10 Stock of Passengers

The platform storage and waiting model evaluate the impact of total capacity constraints on the average passenger WT and local choice of service (intra-line). This is done by explicitly setting out the stock of passengers on the platform. From a station of entry, the stock of passengers per exit station corresponds to the number of passengers who want to get into a vehicle on a service that directly stops at the exit station. The stock depends on the exogenous flow per exit station and the available capacity per serving vehicle. The stock of passengers who are candidates for a service is the sum of the stock of passengers for all of the exit stations stopped at by this service. Consequently, two services with identical downstream stopping policies will have the same stock of candidate passengers, whatever the supply in terms of available capacity.

Figure 9.7 shows the stock of candidate passengers for all of the services on the central trunk traveling toward the La Défense station. Two particular observations reveal the behavior of the platform model and its bottleneck submodel:

- On the profile of the stock of passengers per service: In the eastern part of the trunk, from Vincennes to Auber, the stock of passengers per service is similar, since passengers are traveling to stations in the central trunk (of whom a significant proportion alight at La Défense) and all services stop at these stations. However, at the Etoile and La Défense stations, the stock varies depending on the service because of their downstream stopping policies.
- On the evolution of stock along the line: When the total available capacity of the line is sufficient, but one or several services are saturated, the stock increases slightly compared to the flow/frequency ratio. This leads to an increase in the average WT (at nation, it is 3 min instead of the reference 2 min) since some of the passengers do not succeed in getting on to the first vehicle that arrives. However, when the total available capacity is insufficient, the stock accumulates

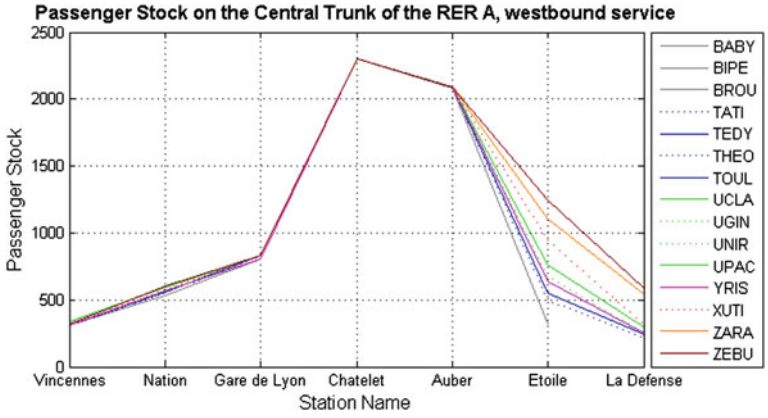

Fig. 9.7 Stock of passengers for services in the central trunk (westbound)

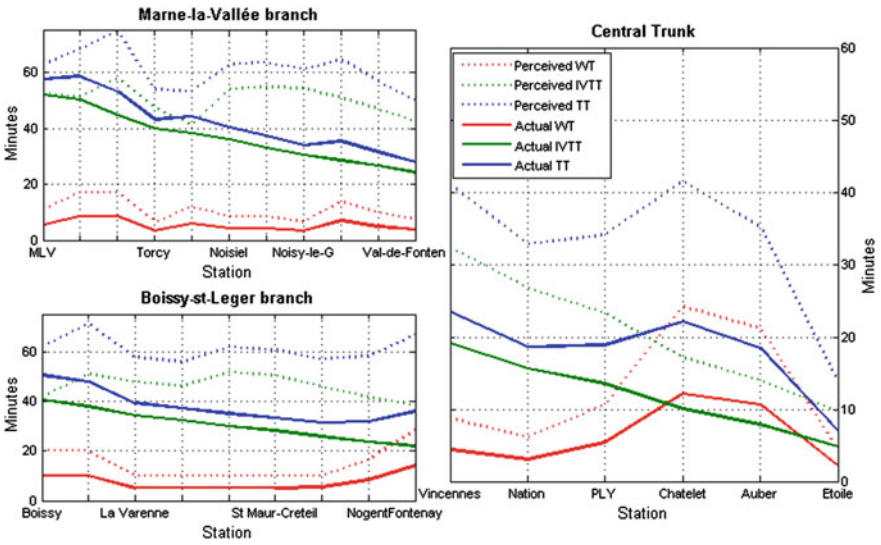

Fig. 9.8 Average actual and perceived time to reach La defense station from east

because the flow cannot be evacuated during the reference period. At Châtelet-Les-Halles, the stock reaches approximately 2300 passengers on platform and the average WT is 12 min, which is 6 times more than the reference situation.

9.2.1.11 Average Generalized Cost

The line model, included in the CapTA model, yields average passenger cost by trip leg along a line integrating capacity constraints throughout the leg. Figure 9.8 shows the actual time (continuous line) and GT (dotted line) of legs to La Défense from stations on the MLV and Boissy-St-Leger branches and stations on the central trunk. The time spent onboard a vehicle is show in green, the WT in red, and the total time in blue.

We observe that the time onboard a vehicle is reduced from upstream to downstream, in line with the distance between the boarding and alighting (La Défense) stations. In addition, the gap between the actual and perceived time spent onboard is related to the difficulty in finding a seated place and the density of standing passengers. Depending on the discomfort coefficients, this gap ranges from 40 to 70 %. A few remarks follow:

- Although leg distances differ significantly, we observe that the leg GT from a station on the branches is around 60 min. On legs beginning at the start of a

branch, access to a seat makes up for the low frequency and the long distance. Conversely, closer to Paris, legs are shorter, but passengers are subject to reduced comfort.
- If we compare a leg starting at Noisiel with another leg further up the line (Lognes or Torcy), the second leg emerges as more advantageous. This is due to comfort in the vehicle, because the likelihood of getting a seat at Torcy avoids having to stand for the rest of the leg, unlike at stations down the line.
- In addition, the actual leg time simulated from Châtelet-les-Halles (TT of 21 min) is longer than the time from the upstream stations Gare de Lyon and Nation (TT of 18 min). The main reason is the important stock of passengers changing at Châtelet-les-Halles.

9.2.1.12 Discussion

The simulation results pertaining to RER A, the busiest line on the Greater Paris network, illustrate the roles played on the field by the capacity phenomena that are captured in the CapTA model, and the need to treat them explicitly. We observe that:

- Taking dwell time and platform occupation constraints into account reduces the frequency of all services down the line. This drop occurs at Etoile station.
- The capacity constraint of vehicles determines the WT and stock of passengers. The value of wait time can shoot up when the candidate passenger flow is faced with insufficient residual capacity, like at Châtelet-les-Halles, where the total stock reaches 2300 passengers (the equivalent of a double-decker train).
- In-vehicle comfort plays an important role and has a significant impact on legs along the radial lines. On RER A, the perceived time of a leg to La Défense is around 60 min, whatever the starting point on a branch.

Lastly, the functional capacity of the CapTA model can be used to evaluate a project's socioeconomic impact, at least in terms of demand surplus. We simulated a capacity investment project for line A of the RER, replacing current vehicles with new, higher-capacity trains. Replacing single-decker trains (MS61 and M184) with double-decker trains (M109, similar to M12 N) adds 30–40 % of passenger capacity per train, as well as exchange capacity on the platform, thanks to the distribution of doors along the length of the train. This significantly improves the line's performance by limiting the decrease in service frequency and in-vehicle comfort. We simulated the investment project and compared the results of its assignment with those of the reference situation. The project would reduce the total GT by 24,150 h during one rush hour, i.e., almost 30 million hours per year. Giving a value to this user benefit of €10/h (conventional value for Greater Paris users), the benefit would be €300 M per year, which at an annual discount rate of 4 % would justify an investment of €6G by the community: Thus, the investment cost, at €2G, would be

repaid threefold, without counting the indirect benefits (i.e., production of value flows in the economic and social circuit, cf. chapter on territorial facilities).

9.2.2 *Agent-Based Simulation of the Stockholm Network Using BusMezzo*

The mass transit system in Stockholm County, Sweden, consists of commuter and regional train, metro, bus, tram, light rail, and local trains. The average number of transit trips per person per day in Stockholm County is 0.63 of which the lion share is performed by either metro (44 %) or bus (41 %). Moreover, 70 % of the inhabitants of the central part of Stockholm County use transit at least several times a week, and 54 % of all trips are carried out by transit, with this percentage increasing to 80 % for trips with a destination in the regional core. Stockholm's population increases by 1.65 % per year and reached 1.4 million inhabitants in the core urban area and 2.2 million inhabitants in Stockholm County in 2013.

The dynamics of the rapid transit system of Stockholm metropolitan area were analyzed in a series of studies with BusMezzo, an agent-based transit operations and assignment simulation model (see Chap. 6.5). The model has been used on a number of research projects as well as projects commissioned by Stockholm County transport administration, which is the regional public transport authority. BusMezzo represents individual vehicles and travelers and is therefore most adequate to applications where system dynamics, service uncertainty, and operations are of primary interest.

BusMezzo was used for a number of planning and operation purposes. Model applications included the analysis of service reliability, passenger congestion effects, the value of increased capacity, service robustness in case of disruptions, design of control strategies (see Sect. 8.4.4), and the impact of information. In the following, the public transport supply and demand representation is briefly explained along with the respective information for the case of Stockholm. Then, selected applications of the model are illustrated.

9.2.2.1 Transit Supply

The representation of transit supply consists of a physical layer, above which are a service layer and an operational layer, assigned to perform the planned services.

The *physical layer* of the transit network is represented by nodes and links that correspond to intersections/switches and road segments/tracks, respectively. The infrastructure layer is coded into BusMezzo by specifying the nodes and links and their attributes. Each node is governed by a queue server that determines the delay distribution and the throughout capacity for each turning movement from an incoming link to an outgoing link. Each link is characterized by its length and a

speed-density function. Each stop is identified by the link, and it is located on and its position on that link. It also contains information on its physical characteristics such as length, facility type, and distance from nearby stops.

The *service layer* is superimposed on the physical layer and consists of routes, lines, and trips. Each route is defined by the sequence of links it has to traverse, whereas lines are defined by the sequence of stops that they serve. In addition, each line is assigned to a control regime. The service is performed through individual vehicle runs. Each run is associated with a specific line and is assigned to a planned departure time for each stop along the respective line. In other words, the service layer contains information on the planned timetable.

The *operation layer* specifies how the planned service is realized. Each vehicle in the fleet belongs to a certain vehicle type. The vehicle type details the physical characteristics such as vehicle length, number of seats, standing capacity, and door configuration. In addition, each vehicle is assigned to a sequence of runs that it has to perform during the course of the simulation period. Finally, flow-dependent dwell time functions are specified for each line, and the parameters depend on the boarding regime (passenger flows and payment method), vehicle, and stop characteristics.

The Stockholm transit network represented in this study includes all lines with less than 15 min headway during the morning peak period (6:00–9:00). This results in a network of 70 lines consisting of all the metro lines, commuter trains, light rail trains, inner-city buses, and suburban trunk buses (Fig. 9.9, except for the northernmost trunk line and including the city center tram line). The metro and commuter trains are characterized by a radial structure and constitute the backbone of the network. Inner-city trunk lines provide high coverage in the inner city, while the suburban trunk lines and the light rail line function as orbital lines connecting major interchange stations strategically located along the southern and western edges of the inner city.

The multimodal network was coded in BusMezzo with detailed timetables, vehicle schedules, and walking distances between stops. In total, 1050 stops are served by more than 2400 runs. Each transit mode is simulated with distinguished vehicle types, vehicle capacities, operating speeds, traffic regimes (mixed-traffic, bus lane, segregated way), dwell time functions, and control strategies. These sets of operational attributes yield different levels of reliability and capacity depending on service design and right of way.

9.2.2.2 Transit Demand

Since BusMezzo is primarily concerned with high-frequency services, passengers are assumed to depart randomly from their origins without consulting timetables. BusMezzo enables several levels of demand representation to suit various application interests and data availability. Individual passengers are generated from a time-dependent OD matrix following a Poisson arrival process. When individual

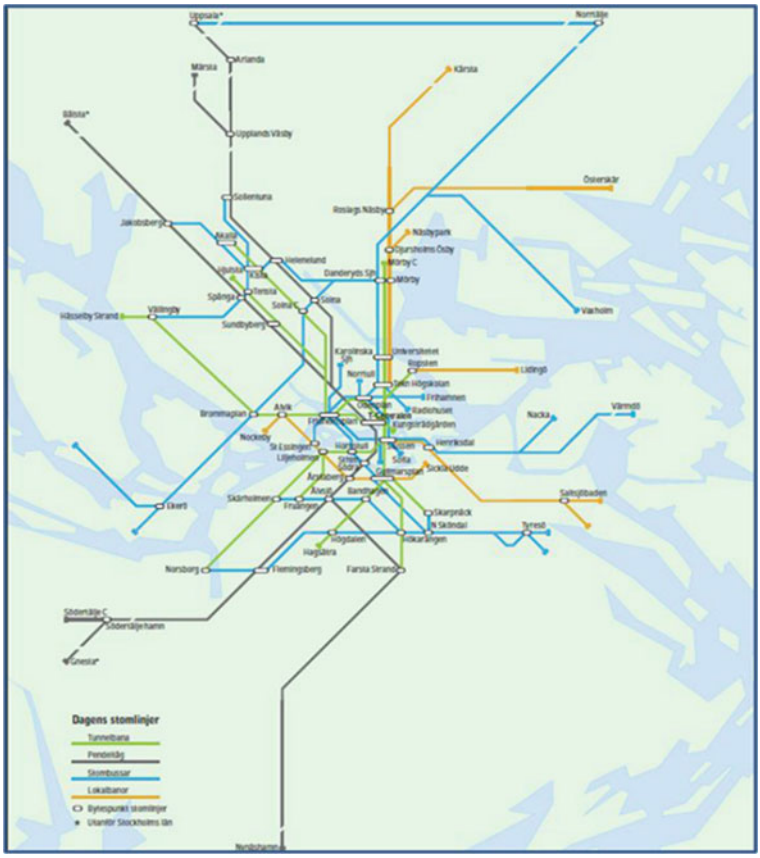

Fig. 9.9 Stockholm's rapid transit network, metro (*green*), commuter train (*gray*), trunk bus lines (*blue*), and local trains/light rail lines (*orange*); *Source* Stomnätsstrategi för Stockholms län, Trafikförvaltningen, 2013

passengers and their enroute decisions are considered, a transit assignment model is executed with two submodels: choice set generation and path choice process.

The *choice set generation* model is designed to reproduce the set of alternatives that are considered by passengers when traveling between a certain origin and a certain destination. Paths are defined by a sequence of transit stops connected by walking links and public transport lines, alternately. Alternatives with common transit stops and lines are merged to form hyperpaths. A static path search algorithm is performed upon initialization and results in a master choice set for each OD pair. The master set is obtained by performing a recursive backward search and applying a series of logical and dominancy rules. This choice set is then given as input to any network loading and is subject to dynamic filtering rules whenever retrieved.

The *path choice process* determines how passengers progress in the network. Passengers do not only select a path from the predefined choice set, but make a

sequence of path decisions as they progress through the network. All passenger decisions are based on random utility discrete choice models. Each decision is defined by the need to choose the next path element (stop, line, or walking connection) taking into account all the path alternatives associated with this element. The parameters of the choice set generation model and the dynamic path choice model were estimated based on a stated preferences survey on transit route choice decisions. The utility associated with a certain path is a function of the anticipated travel attributes (e.g., WT, in-vehicle time, crowding), depending on the information provision and passenger's experience.

Passenger demand for Stockholm transit network is generated in BusMezzo based on a demand matrix that was produced with SIMS, the strategic demand model that is used in Stockholm. SIMS produces travel demand for the whole region for 6:00–9:00 a.m. and for the analysis in BusMezzo, and a submatrix of 400 zones is used. Approximately 125,000 passenger trips are initiated during the peak morning hour traveling between 4576 nonzero OD pairs. The master choice set contains 615,111 hyperpaths for all O–D combinations in the network, approximately 4 hyperpath alternatives per OD pair. Trip fare is fixed in the Stockholm network for a given OD pair and therefore does not affect passenger path decisions.

The simulation generates a series of output files including the paths that were taken by each passenger and the corresponding travel time components. Passenger travel times are thus calculated based on the disaggregate demand representation and the time difference between simulation events. For example, passenger's WTs at stops are calculated based on the time difference between passenger's arrival time at a certain stop and the time when a passenger boarded a vehicle. Each scenario is analyzed based on the results of 10 simulation runs. This number of replications yielded a maximum allowable error of less than 1 % for the average passenger travel time in all cases considered. The execution time for a single run was less than 1 min on a standard PC.

9.2.2.3 Applications

Passengers' travel decisions depend on their expectations on downstream travel attributes. These expectations in turn depend on the information available to each passenger when taking a certain travel decision. BusMezzo enables to analyze the impact of real-time information (RTI) provision by specifying various levels of information coverage and comprehensiveness as well as the share of users who have access to this information. The RTI provision is not equivalent to perfect information but rather mimics commonly used prediction schemes for generating such information.

The impact of providing RTI, in the form of the expected remaining time until the next arrival of each relevant transit line, was evaluated for the case study network of Stockholm. In order to assess the potential benefits of information provision, it was assumed in this study that whenever passengers have access to information, this is used for updating their expectations, which are then explicitly

Table 9.3 Average passenger journey attributes (entire network/inner city) and the relative change compared with the base scenario (in percentages)

Scenario	Total travel time (s)	In-vehicle time (s)	Waiting time (s)	Walking time (s)	Number of boardings per trip	Standing time (s)
No-RTI (base)	2323/1399	1621/885	545/362	157/152	1.69/1.60	511
Stop-RTI	2317/1359 (−0.3/−2.9)	1619/876 (−0.1/−1.0)	540/333 (−0.9/−8.0)	158/149 (+0.6/−2.0)	1.61/1.44 (−4.7/−10.0)	529 (+3.5)
Cluster-RTI	2312/1354 (−0.5/−3.2)	1616/880 (−0.3/−0.6)	536/321 (−1.7/−11.3)	160/152 (+1.9/0.0)	1.60/1.42 (−5.3/−11.3)	544 (+6.5)
Network-RTI	2265/1334 (−2.5/−4.6)	1583/849 (−2.3/−4.1)	522/334 (−4.2/−7.7)	160/150 (+1.9/−1.3)	1.61/1.45 (−4.7/−9.4)	556 (+8.8)

incorporated into the utility function components when making en-route decisions. The case study considered 4 RTI levels—none (static information), local stop, cluster of stops, and entire network.

The additional information available in network-RTI yielded a considerable reduction in WTs and total travel times. Table 9.3 presents the average travel times and their components in the different scenarios. In the base (no-RTI) case, the average passenger journey time is almost 39 min for the entire network and 23.3 min for inner-city trips. The WTs, which include WTs at the origin stop and at transfer stops, account for 26 % of the total travel time in the inner city. Walking times refer to walking between stops either when making a transfer or in order to reach a nearby stop in the beginning or the end of a passenger trip. Walking times account for 11 % of the travel time in the inner city. In the other scenarios, the more informed passengers are, the more their journey times decrease. The travel time in the network-RTI case is 4.6 % lower than in the base case. This trend is expected as RTI enables passengers to make more informed decisions and improves the coordination of downstream transfers. The reduction is largest in WTs, which decreased by up to 7.7 %. In contrast, walking times did not decrease and even increased by up to 1.9 %. In-vehicle times decrease only in the case of network-RTI when passengers may choose to wait longer in order to take a path which is overall shorter. Interestingly, also the number of boardings per trip decreased from 1.60 to 1.45. The decrease of transfers is obtained already when RTI is provided at the stop level, suggesting that uninformed travelers make more unnecessary transfers instead of waiting longer for a direct service in this network. The topology of Stockholm's transit network plays an important role in limiting the opportunities to gain from more informed connection decisions. The network is designed so that there are only few cases where interchangeable rapid lines do not serve the same stops.

As passengers are more informed, passenger loads are subject to more fluctuations due to the adaptive passenger path choice process. A time-dependent analysis reveals that access to RTI affects passenger assignment results considerably. The fluctuations in passenger loads increase the probability of onboard and at-stops discomfort and capacity concerns. Moreover, it may negatively affect service

regularity as uneven passenger loads contribute to the bunching phenomenon. This trend is reflected in the increase in the average standing time per passenger shown in Table 9.3, which is a measure of crowding that is used as a proxy for the level of comfort. More uneven, loads on transit trips lead to an increase of 9 % in the average standing time per passenger.

Information provision can be particularly advantageous when things do not work as planned such as in the case of service disruptions. BusMezzo simulates the implications of disruptions on upstream queuing vehicles, stranded passengers, downstream passengers as well as spillover effects to other lines due to vehicle scheduling, capacity constraints, and rerouting decisions. The five most central segments were identified and the impacts of half hour disruptions on them induced a cost of between 1000 and 5000 additional passenger hours. Furthermore, service disruptions resulted in a larger share of the passengers experiencing long or very long travel times compared to the non-disrupted scenario. The impact on passenger welfare of moving from stop-RTI to network-RTI provision varies considerably for different disruption scenarios from a worsening of 2.5 % to an improvement of 4 %, although the general trend is that more comprehensive RTI provision results in higher passenger welfare.

The detailed dynamic representation of system supply and demand enables one to analyze the vulnerability of network components to disruptions. Each line in Fig. 9.10 corresponds to a rush hour vehicle run on bus trunk line 1 for the case of normal operations and a disruption on the southbound direction of the Blue metro line with RTI at the network level. It is evident that even for the case of normal operations (left), there are substantial temporal variations as a result of system dynamics and the interaction between inherent sources of supply and demand variability. This temporal variation is distorted significantly in the case of a disruption (right). Line 1 is an alternative for passengers traveling with the blue line toward the city center. Vehicle capacity is limiting for a couple of the runs that arrive at the affected station, Fridhemsplan, during the disruption (7:15–7:45) and the load is relieved at Hötorget which lies within the city center. As the system recovers from the disruption, passenger loads gradually resemble previous patterns

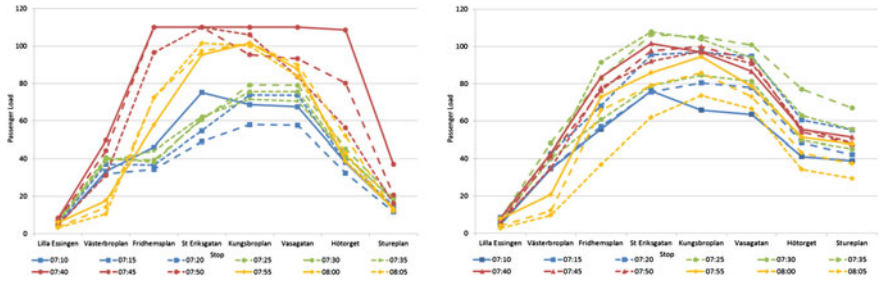

Fig. 9.10 Passenger loads at key stations along trunk line 1 eastbound during the rush hour—under normal operations (*left*) and a disruption on a partially alternative metro line. Each line corresponds to a single vehicle run

Fig. 9.11 Volume over capacity changes due to disruption scenarios—on the soutbound *Blue lines* 10 and 11 (*left*) and the southbound *Red lines* 13 and 14 (*right*). The disrupted sections are shown in *black*. Sections with more than 5 % relative increase in volume-over-capacity ratio are decipted in *red*; *yellow* corresponds to no significant change, and sections with a relative decrease of more than 5 % are shown in *green*. The link labels correspond to the line numbers, and the Lilheholmen transfer hub is marked with a *circle*

and levels. This stresses the importance of capturing how the system evolves over time when analyzing the impacts of disruptions on redundant capacity and mitigation strategies.

The cascading effects of disruption scenarios can be analyzed by considering the change in volume-over-capacity ratio across the network. Figure 9.11 presents the impact of two disruption scenarios on link saturation levels for the network core. In both cases, congestion increases significantly on a large number of links. Upstream links are negatively affected as well as links act as direct substitutes for passengers boarding at the upstream node of the disrupted segment. Note that these effects are not limited to links that are in geographical proximity to the failed segment as both upstream and downstream links of alternative lines are affected due to rerouting. Furthermore, network effects lead to a spillover to other lines that connect transfer hubs overtaking the failed segment. Interestingly, the saturation on some links is relieved following the closure of an important link, because the outgoing flow of the latter constitutes a significant share of the ingoing flow of these links. This is, for example, the case of the light rail line 22 in Fig. 9.11 (right) west of a major transfer hub (Liljeholmen) between the light rail line 22 and the disrupted red metro line (lines 13 and 14).

9.2.2.4 Assessment

One of the most common motivations for transit investments is increased capacity. The benefits from increased capacity are as follows: improved reliability, less congestion, and the capability to withstand a disruption. The analysis of disruption impacts was used for identifying potential services where allocating reserve

capacity could be an effective mitigation measure. The same evaluation method was also applied for assessing the robustness value of new lines. This was applied for the cross-radial light rail line 22 which demonstrated that the new line can relieve up to 73 % of the welfare reduction for certain disruption scenarios.

BusMezzo is an open access simulation model and is in early stages of applications to other transit systems in Sweden and elsewhere. Ongoing applications include the evaluation of a metro line extension and related onboard congestion effects, boarding regimes, and their impacts on service reliability, trunk and branch operations, and disruption management strategies.

9.2.2.5 Summary

Stockholm, the capital of Stockholm, is well known for its inseparable urban and transport planning, constituting a prime example of a radial transit system. Satellite towns were developed along the rapid public transport corridors and thereby creating a transit metropolitan area which promotes suburb to center commute (Cervero 1995; Börjesson et al. 2013). Statistics concerning current and projected population and travel patterns in Stockholm County are available in the annual statistical book of Stockholm County transport administration (SL 2013).

The conceptual framework of BusMezzo model is available in Cats (2013). The supply side representation including the representation of service uncertainties are provided in Toledo et al. (2010), which were then validated in Cats et al. (2010). Alternative control strategies were tested and evaluated with BusMezzo, supporting the design and full-scale implementation of a new strategy in Stockholm trunk bus system (Cats et al. 2011a, 2012a, 2014; Cats 2014a). The principles for modeling RTI are presented in Cats (2014b), and its impacts were evaluated for a case study in Stockholm (Cats et al. 2012b). The modeling framework was further extended to consider day-to-day learning of service and information reliability and was applied to Stockholm (Cats and Gkioulou 2015). The impact of service disruptions on system performance under various levels of information provision was studied in Cats et al. (2011a) and Cats and Jenelius (2014). BusMezzo was then used as a tool to evaluate the robustness value of changes to network design such as new lines in Jenelius and Cats (2014).

References

Börjesson M, Jonsson D, Lundberg M (2013) An ex-post CBA for the Stockholm Metro. CTS working paper 2013 34

Cats O (2013) Multi-agent transit operations and assignment model. Procedia computer science. In: The 2nd international workshop on agent-based mobility, traffic and transportation models, methodologies and applications (ABMTRANS), vol 19, Canada, pp 809–814

Cats O (2014a) An agent-based approach for modeling real-time travel information in transit systems. Procedia computer science. In: The 3rd international workshop on agent-based

mobility, traffic and transportation models, methodologies and applications (ABMTRANS), vol 32, pp 744–749 Belgium
Cats O (2014b) Regularity-driven bus operations: principles, implementation and business models. Transp Policy 36:223–230
Cats O, Gkioulou Z (2015) Modelling the impacts of public transport reliability and travel information on passengers' waiting time uncertainty. EURO J Transp Logistics
Cats O, Jenelius E (2014) Dynamic vulnerability analysis of public transport networks: mitigation effects of real-time information. Netw Spat Econ 14:435–463
Cats O, Burghout W, Toledo T, Koutsopoulos HN (2010) Mesoscopic modeling of bus public transportation. Transp Res Rec 2188:9–18
Cats O, Koutsopoulos HN, Burghout W, Toledo T (2011a) Effect of real-time transit information on dynamic passenger path choice. Transp Res Rec 2217:46–54
Cats O, Larijani AN, Ólafsdóttir A, Burghout W, Andreasson I, Koutsopoulos HN (2012a) Holding control strategies: a simulation-based evaluation and guidelines for implementation. Transp Res Rec 2274:100–108
Cats O, Burghout W, Toledo T, Koutsopoulos HN (2012b) Modeling real-time transit information and its impacts on travelers' decisions. In: Proceedings of the 91st transportation research board annual meeting, Washington DC
Cats O, Mach Rufi F, Koutsopoulos HN (2014) Optimizing the number and location of time point stops. Public Transp 6:215–235
Cervero R (1995) Sustainable new towns: Stockholm's rail-served satellites. Cities 12:41–51
Chandakas E (2012) Note sur la capacité du matériel roulant et son affectation sur le réseau Francilien. Working Document, Ecole Nationale des Ponts et Chaussées, Paris Est University, France
Jenelius E, Cats O (2014) The value of new cross-radial link for public transport network resilience. In: Second international conference on vulnerability and risk analysis and management, London, UK
Observatoire de la Mobilité en Ile-de-France (2012) Synthèse de principaux résultats de l'EGT 2010, Edition of the Observatoire de la mobilité en Ile-de-France. OMNIL, France
SL—AB Storstockholms Lokaltrafik (2013) Fakta om SL och länet 2012 (In Swedish—Facts on SL and the county 2012, Stockholm, Sweden
Toledo T, Cats O, Burghout W, Koutsopoulos HN (2010) Mesoscopic simulation for transit operations. Transp Res C 18:896–908
Software publishers contacted
Colby Brown and Michael Clarke (Citilabs)
Michael Florian (INRO Consultants)
Peter Kant (Omnitrans International)
Andres Rabinowicz (Caliper)
Klaus Noekel (PTV Group)

Chapter 10
Applications and Future Developments: Future Developments and Research Topics

Ingmar Andreasson, Fabien Leurent and Rosaldo Rossetti

Situation A mass transit system is complex, since it combines material resources (vehicles, infrastructures), operational personnel, operational processes, and action protocols. Information plays an essential role in coordinating resources (e.g., allocating routes to stations and drivers to vehicles), in the use of those resources by passengers (prior information on services, real-time information), or in the automation of payments.
By design, therefore, a mass transit system is an intelligent transportation system, founded on the use of a minimum baseline of information, developed during the design of the system.

Let us broaden our perspective by looking both at the day-to-day operation of the system—the usage phase—and at the design phase: Here, the use of a simulation model to plan the network contributes to the intelligence of the transportation system.

In the digital era, information and information processing are becoming increasingly important in the day-to-day operation of mass transit systems, even down to the nature of the services provided. This intelligence has successively penetrated into the control of vehicle traffic on the infrastructure, then into the interaction between operators and passengers (information, payment, reservation), next into the vehicles as subsystems (including into the engines themselves), and finally into the actions and practices of users, thanks to mobile terminals (smartphones), which allow direct and customized informational interaction. Within this

I. Andreasson (✉)
Logistik Centrum Göteborg AB, Osbergsgatan 4A, 426 77 V Frölunda, Sweden
e-mail: ingmar@logistikcentrum.se

F. Leurent
Laboratory on City, Mobility and Transportation, Ecole des Ponts ParisTech, University Paris-East, Paris, France
e-mail: fabien.leurent@enpc.fr

R. Rossetti
Faculdade de Engenharia, Universidade do Porto, Rua Dr Roberto Frias s/n, 4200-465 Porto, Portugal
e-mail: rossetti@fe.up.pt

G. Gentile and K. Noekel (eds.), *Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems*, Springer Tracts on Transportation and Traffic 10, DOI 10.1007/978-3-319-25082-3_10

broad technological—or rather, sociotechnical—process, simulation models are continually finding new applications: in real-time system management; in demand management, in particular directing demand through dynamic information; or for short-term forecasting, which is useful both to operators and users.

Objective This final chapter explores the potential of traffic assignment models for development, beyond traditional or advanced applications. The modeling of public transportation systems is a fertile terrain for research, especially as the digital era is seeing a proliferation of innovations: in the operation of existing systems and above all in the design of original, flexible, demand-responsive mobility services, which rely on different forms of resource pooling.

The principle of pooling, fundamental in mass transit for the sharing of infrastructures by vehicles and the sharing of vehicles by passengers, has now found a very wide range of applications, thanks to the presence of information everywhere in the mobility system, which includes the transportation system and its users.

Chapter content The body of the chapter is structured into three sections. First, we consider the new deal in public urban passenger transport that stems from the new order in the field of information: Ongoing or future innovations pertain to the management of line networks, to the provision of more flexible intermediate services, and to the sharing of vehicles, drivers, and parking spaces, together with the potential associated with autonomous (automated self-driving) vehicles.

Second, we identify a whole range of research topics on traffic assignment models and their inputs to their potential applications for system regulation, passing by (i) passenger behaviors and their statistical structures, (ii) the physics and control of traffic—both passengers and vehicles, (iii) the spatial features and their flow-oriented layout, and (iv) the organization and operations of specific travel modes.

Third and last, we open up a broad perspective onto the relation between mobility systems and simulation models: Models are becoming more and more modular, and they constitute a toolbox that is more and more powerful; a number of tools are implemented to bring augmented reality to the transit systems for all of its stakeholders (users, operators, regulators, general public); arguably, an Urban Mobility Living Lab should be an ideal framework to study system conditions, to design user-oriented innovations, and to test system's responses to them on the field.

10.1 A Forward Analysis of Public Transportation in the Information Era

Fabien Leurent and Ingmar Andreasson

The digital era is characterized by a new order in the field of information: Information has become much more abundant and diverse, very easy to transmit, receive, and use, and very widely available—in fact available virtually everywhere—at an extremely low price. So this is a revolution, in both functional and economic terms.

This revolution affects a wide variety of material products and, to an even greater degree, services: Existing services have been profoundly transformed, whereas innovative services are emerging, in some cases spreading massively.

The technological information ecosystem has progressed in giant steps: miniaturization of electronic components, information superhighways (optical fiber and ADSL), wireless networks (GSM, Wi-Fi), accurate and instantaneous geolocation (GPS), Internet with information, reservation, and purchase services, satellite imagery, and the smartphone as a portable terminal of astonishing functional richness and startling interactive capability. The different technological building blocks reinforce each other, thanks to the multiplicity of possible combinations, which enhance their reciprocal capacities.

This revolution in the technological information ecosystem affects all sectors of the economy, in particular the transportation sector. The notion of an "intelligent transportation system" broadly encompasses the penetration of transportation systems by the information revolution, a process that occurred in large waves, corresponding to the big steps in the development of information technologies.

In this section, we explore the consequences—whether existing or still potential—of the informational revolution for public urban passenger transport. The new technological order has brought technical and commercial renewal in the traditional collective modes of transportation by bus or rail, in the organization and day-to-day operation of the service, which is becoming more interactive, in the mobility practices of users, and also in the interaction between the operator and users (Sect. 10.1.1).

The new technological order is also leading to an industrial transformation in the supply of public transportation through a diversification of modes: The traditional trade-offs between vehicle size, service frequency, and interstation spacing (and therefore the terminal distance travelled to access the service) have been pushed aside and can be eradicated by the development of more flexible intermediate modes than line-haul services (Sect. 10.1.2).

The economic transformation is great enough to prompt transformations of an industrial nature: The roles of the traditional service actors are changing, new players have arrived to perform traditional or innovative functions, since combinations of production methods are facilitated by the availability of information. This industrial transformation is strongly marked by the pooling of resources in all the areas that constitute a public transportation service: vehicles, transportation and parking infrastructures, protocols for usage, information and pricing, with the development of reservation services (Sect. 10.1.3).

An additional major technological transformation is also in progress: automated self-driving vehicles, otherwise known as autonomous vehicles. In public urban passenger transport, this concept is not so new, since automatic subway systems have been developing since the 1990s. However, it is spreading to road vehicles, impelled by a profound economic transformation—economizing the driver's salary, which is the main cost in a taxi system—and increased technical potential (e.g., automatic repositioning or the end of driver rest requirements): We will set out the

technical and economic potential of autonomous vehicles for public urban passenger transportation (Sect. 10.1.4).

All these different changes have had the effect of reshuffling the cards between transportation modes. The two traditional pairings, on one side individual means combined with private use, and on the other side collective means combined with shared uses, are now experiencing the lure of different forms of exchange, with a range of hybrids reminiscent of gene recombinations. We explore their impact on transportation supply, on demand, and on the regulation of the mobility system (Sect. 10.1.5).

10.1.1 Line-Haul Public Transportation in the Information Era

Ongoing or upcoming changes relate to the different components of the system: vehicles (Sect. 10.1.1.1), stations (Sect. 10.1.1.2), infrastructures (Sect. 10.1.1.3), and especially operational coordination by centralized management (Sect. 10.1.1.4), as well as user practices (Sect. 10.1.1.5) and the interaction between the service and passengers (10.1.1.6).

10.1.1.1 Vehicle Changes

The new types of vehicles bring improvements of various kinds: Internal layout is optimized to contain more passengers in a given space, with more comfortable but fewer seats, thus leaving more space for standing passengers. Capacity optimization also affects exchange capacity, with more and wider doors, which open and close more quickly.

Driving is assisted by an onboard computer and various sensors, saving energy and giving a smoother ride.

Vehicles are fitted with various systems to provide passengers with information: audio and visual updates on journey progress and upcoming stops.

All these changes make travel more efficient, with better informed passengers.

10.1.1.2 Station Changes

Mass transit stations provide different levels of equipment, depending on their ranking in a hierarchy, from a simple bus stop through to an intermodal platform. The biggest stations are also the busiest; their layout is optimized across multiple levels, and internal passageways are shortened and efficiently signposted to reduce connection times.

For modes with the biggest capacity, BRT or rail, the installation of sliding platform doors improves passenger safety and the transfer of passengers between platform and vehicle, and controls the opening phases, such as green phases in traffic signals at a road intersection.

Whatever the hierarchical level, the generic functions of stations are gradually improving, through a combination of enhanced content and simpler ergonomics:

- Better identified waiting areas with greater protection from the weather, heat differences, and other disturbances;
- Automatic devices for validating tickets or for buying them if the station is big enough;
- Greater information content and increasingly dynamic information: time of arrival of the next vehicle for every service. Also under development is basic information so that passengers can transfer to other services in the event of major disruption.

So stations are better equipped and becoming smarter. The changes are making them more efficient and more comfortable to use, with better information and better oversight by the operator (and even control in the case of sliding platform doors).

10.1.1.3 Infrastructures: Open Sections and Junction Management

On way sections, movement is faster and more reliable in a dedicated corridor, with restricted or even controlled access. The presence of an additional lane, even on restricted sections, makes it possible to overtake and differentiate between express routes and standard bus services. Additional lanes are particularly useful at stations, to allow stationary vehicles to be overtaken. This arrangement makes the service more resilient.

Along the routes, junction management—intersections on roads and branches on railroads—is a major factor. Controlling junctions in order to give priority to mass transit vehicles and keep them moving is crucial to enhancing the efficiency of mass transit relative to individual transportation modes. It requires close and continuous coordination between vehicles and junction controllers by means of the information and line management system.

In the information era, integrated control of vehicle movement and junction traffic is becoming widespread. Real-time coordination of resources is a major technical challenge, a challenge that can be tackled with an efficient information and control system. Coordination remains dependent on spatial arrangements: Smart real-time operation combines with smart forward planning to allow the smooth and efficient movement of vehicles and ultimately of users.

10.1.1.4 Centralized Management to Coordinate Resources

So integrated line control requires the continuous exchange of accurate information on the location of vehicles and the condition of resources, whether controlled or simply monitored. Both the collection and transmission of information have been automated, which frees up operating staff and speeds up processing.

Whether on vehicles or in stations, or remotely by digital tracking, it is now possible to collect information in large quantities, in real time and at very low cost, in particular on the number of passengers on board and even their distribution inside the vehicle. These automatic readings give urban mass transit operators almost the same amount of information as operators of interurban mass transit systems, where seats are reserved in advance: The only differences in knowledge concern the seat the passenger is sitting in and the station where he gets out, and even these factors can be predicted statistically.

So close-grained, real-time information is now available to improve traffic management from a quality of service perspective—in particular by including the number of passengers in vehicles or stations into with management optimization criteria—to make the average service per passenger more efficient.

With an integrated real-time information system, operational management can be adjusted to circumstances and contingencies, making the service more efficient and reliable: It is the cornerstone of the resilience of the system.

In any management application, information obtained in real time is used for short-term forecasting: For example, vehicle trajectories are extrapolated in order to predict the time of arrival in stations or at controlled junctions. Real-time data on passenger numbers provide information on the needs to be met, both for onboard passengers and for those waiting in stations. The state variables of the transportation system are inherent to it: When monitored by the information system, these variables become observable indicators (in the sense of statistical physics), and when simulated in a traffic model, they can be anticipated. That is why the development of a powerful information system comes close to a simulation model: Ultimately, it is natural to include an assignment model in order to exploit real-time observations and thereby contributes to day-to-day operational management.

This convergence between information and management system and a simulation model is a major priority for operational improvement. Assignment models need to be integrated into day-to-day operations, as well as into off-line planning. A crucial step in this convergence is undoubtedly a line model that incorporates passenger numbers into online resource management. The next step is line coordination, which is more complex: This notably includes managing articulating nodes, which are junctions on the infrastructure and connecting stations for passenger flows.

The extrapolation from data collected in real time is useful for the operator, but also for users, whether they are present in the system or planning a trip and seeking remote information on current conditions.

10.1.1.5 The Passenger Side

Information system development benefits passengers, whether while traveling or when planning trips. Information has become dynamic and is also on the way to becoming predictive. The quality of prediction will depend on the capacity of the information system to incorporate both real-time operational adaptations and adaptations by users.

Major progress has already been made, thanks to the customization of information and its adaptation to the particular needs of the user in terms of times and places, and thanks to the selection functions provided by search engines. This informational upgrade costs the user nothing, apart from the cost of going online. In addition, with the smartphone, passengers have a portable terminal through which they can obtain the dynamic real-time information they need, wherever they are.

The effects of innovation are cumulative: The smartphone combines highly varied functions, a vast array of information, telecommunications, and games all in a format that is very easy to use on public transport. It replaces a newspaper and can be used even when standing. It enables people to use the time spent on the move, thereby reducing the inconvenience and masking the discomfort of travel. In other words, it transforms the quality of service, to the benefit of users. This can be expected to change uses and practices as well as, of course, provide feedback on time sensitivity and the trade-off between time spent and price. The possibility to use smartphones and tablets supports the mode shift from driving to transit.

10.1.1.6 Interaction Between Service and Passengers

In these conditions, the interaction between the user and the mass transit service is transformed: The physical presence of the user combines with a sort of mental presence, arising from real-time attention to traffic conditions and the continuous use of telecommunications. As a mobile terminal, the smartphone acts as a supplement to local information media, though without rendering them obsolete (provided that each medium is upgraded. In particular, the boards showing travel times need to become dynamic, shifting from paper to screen: This is a transition that has already happened in some advertising hoardings).

In other words, public transportation as a service is enriched by these informational changes, independently of the improvement in real-time management. In addition, there is a significant fund of synergy between these two respective changes in demand and supply: They make the service much more flexible in the sense that flexibility becomes the rule, rather than being confined to the handling of exceptional circumstances.

The stakes are notably to adapt the supply to the conditions of demand, for example the particular needs of certain days: professional or sporting or tourist events, or dynamic management of multimodal travel in the conurbation in response to a critical episode (bad weather, atmospheric pollution, etc.).

This system will become more flexible with vehicle automation, which relaxes the technical constraints associated with human drivers. It will require cooperation by users, i.e., compliance on their part: The better the ergonomics of travel—in particular with an "all-inclusive" smartphone service—the more spontaneous this cooperation will be.

The aim, therefore, is to change the routines of passengers, by providing an operating network that is sufficiently integrated in both informational and physical terms. These integrations need to go hand in hand with an integrated fare and sanction system.

In fact, the use of the smartphone should facilitate the introduction of positive sanctions: bonuses to stimulate individual cooperation for the smooth running of the system.

10.1.1.7 Summary

The development of traditional mass transit systems, whether bus or rail, has been marked by gradual improvements in their components, i.e., the vehicles, the stations, the infrastructures, the pricing and payment system, and, increasingly, the information system. The latter is becoming more and more powerful thanks to the neural net of local and remote sensors and actuators on the system: It constitutes a kind of nervous system for the line taken as an organized system and more broadly so for an entire transportation network.

This informational apparatus is amplified and boosted by the mobile terminals carried by users: The smartphone is the Swiss Army knife of mobility and also a digital companion that helps people to recycle the time spent in travel.

All that remains is to maximize the power of this nervous system, in order to develop a flexible, responsive, and anticipatory transportation network. The technical ingredients are in place, at least in planning terms, awaiting full deployment, which takes time and demands substantial financial investment. Flexibility will improve not only the system's responses to disruptions and congestion, but also its capacity to handle the diversity of demand situations. The integration of simulation models into the information and management system is necessary in order to make management more dynamic by improving its predictive ability.

User cooperation is equally necessary. This needs to be cultivated by offering services that are simple enough to continue bringing users the benefits of quality of service; by educating passengers; and by making them cooperators in public transportation with the same level of participation as in an individual transportation service.

10.1.2 The Diversification of Public Transportation Modes

Line-haul mass transit will retain certain rigidities, even if it becomes more efficient (Sect. 10.1.2.1). At the other end of the public transportation spectrum, the

individual taxi is adapting to the needs of users: This mode too should benefit from the new informational order (Sect. 10.1.2.2).

However, this order should above all stimulate the development of intermediate modes, collective taxis, or microbuses, which combine the flexibility of the taxi with the economies of scale of mass transit (Sect. 10.1.2.3).

10.1.2.1 Line-haul Transit Will Retain Certain Rigidities

We have mentioned the flexibility gains expected for line-haul mass transit, thanks to the increased capacity for response and anticipation afforded by a powerful information and control system.

However, line systems, with their routes and above all their necessary stops, remain restrictive for users in terms of spatial availability. In addition, service timetables impose time constraints that also limit the availability of such systems to users.

These constraints can be attenuated but not eradicated. On the spatial side, intermediate stops along a bus route can be skipped in the absence of stop requests from inside or outside the vehicle. The time constraint becomes softer with increases in frequency: For users, reduced and even headways are more important than strict adherence to timetables.

The traditional economics of mass transit should improve through increased flexibility of production. For users, some conditions will change little: The monetary cost will remain modest, provided that they use travel cards, terminal access will remain their responsibility, waiting time should diminish a little (outside saturation events), and individual reservations will remain unavailable. The use of time spent on board or in stations, by the simultaneous use of the smartphone, will change the conditions of modal choice, probably more than the conditions of route choice within the transit network.

10.1.2.2 The Individual Taxi Will also Benefit from the New Informational Order

The individual taxi, being available to all users, is also a transit mode, but at the other end of the spectrum of capacities compared with mass transit. The individual nature of the service eliminates the spatial rigidities of mass transit: The passenger decides the starting point and destination. On the other hand, the fare is markedly higher than that in mass transit.

With taxi firms, the taxi service is responsible for transportation, billing, and reservation. For the user, reservation entails a transaction cost, along with a waiting time, because the availability of vehicles is limited.

The new informational order offers the taxi mode significant advantages:

- At the level of the firm, centralized information and management system optimize efficiency in the allocation of vehicles to customers, which increases vehicle productivity and helps to reduce passenger waiting times.
- Upfront, integrated management makes it possible to plan the development of the service by adjusting the size of the vehicle fleet, subject to institutional constraints.

So progress in the information sphere provides benefits for the taxi on both the supply and demand sides. The commercial side of the service needs to be centralized in order to grasp this opportunity. This will combine with the advantages specific to a taxi firm, i.e., more integrated and productive management of vehicles and drivers, preferential conditions for the acquisition and maintenance of vehicles, and preferential access to credit.

10.1.2.3 The Consolidation of Intermediate Modes

With the individual taxi, occupancy can only be maximized in the case of small groups traveling together, which considerably restricts the possibility of economies of scale from greater vehicle capacity.

Collective taxi services, whether using vans or minibuses, represent an intermediate mode between the individual taxi and line-haul mass transit. It combines the ability to pick customers up and drop them off where they want, with maximized vehicle occupancy. Pooling the service between passengers reduces the unitary cost and hence the price per passenger, in the same way as for a group in an individual taxi, but with greater earning potential for the service provider.

The new informational order is a windfall for collective taxi modes:

- Firstly, as with an individual taxi, the centralized information management system matches reservation requests to available vehicles. Unlike individual taxi, however, requests that are close to each other in space and time can be grouped by the operator.
- Secondly, additional customers can be picked up in the course of the journey, provided that there is space in the vehicle. This increases the availability of the service, both in space and in time, and every point on a vehicle trajectory becomes an opportunity to recruit a customer.
- In turn, each customer benefits directly from increased availability and lower fares. In addition, if a potential customer receives continuous information on current opportunities, then her confidence in the service is increased: This allows more spontaneous, opportunistic, and unplanned uses, with reduced transaction costs.

Obviously, the need to pick up and drop off other passengers, and the associated route adjustments, reduces the speed of service for each user: The quality of service

is halfway between the individual taxi and the traditional bus, although with strong potential in terms of availability.

The enhancement brought about by the information system makes the mode more reliable and more appealing. Linked with the limited size of the vehicle, this can trigger the virtuous circle of the Mohring effect: More demand necessitates more vehicles, which increases the availability of the service (cf. the frequency of a line-haul service) and makes it more attractive, thereby further stimulating demand, etc.

The power of this positive feedback effect will depend on the area. In dense areas, it potentially makes the collective taxi a serious competitor to the bus.

In certain African and Latin American cities, the mass transit system consists entirely of collective taxis, and in other words, they have a bottom-up rather than a planned top-down system.

However, the sphere of relevance of the collective taxi is certainly not confined to dense areas. It is a very promising mode for transportation provision in low-density areas where the number of passengers for grouping is small, and there is therefore a significant need for intermediation to organize groups. Here, the opportunity is that the intermediation function can be automated by an information system and by centralized management, at negligible cost!

The infrastructural aspects of such a mode remain to be defined. In principle, it requires no access stations, which is an advantage. For access as for travel, the mode obviously uses the road network: But could it be allowed to use dedicated bus corridors (as is the case in some cities) and even granted priority at intersections?

10.1.3 Toward a Generalized Pooling of Transportation Means?

In the urban environment, mechanized transportation modes will retain their well-established physical composition: a vehicle, a driver, an infrastructure comprising stations and lanes, together with operational protocols: principles of use, highway code, service processes, operating processes, etc. Some of the different components are physical (hard) and others more organizational and informational (soft). The information revolution is expanding the role and functions of the soft components. In particular, it permits highly interactive protocols, offering the possibility of major recombinations in the way the physical components are used. Nonetheless, these recombinations need to maintain certain regularities; they take the form of particular mobility services, which bring considerable diversity to the transportation supply.

Innovative services are an addition to the traditional services; they still employ physical resources, but in a variety of forms of ownership and pooling between users.

All this takes place on the road network, which is the infrastructure for access, parking, and movement, since the general availability of the roads opens up the urban space to mobility needs and services. These therefore share the roads between them and with the pre-existing users.

We have already described the traditional forms of pooling: On the one hand, line-haul mass transit—which vertically integrates a vehicle, a driver, a service route and stations, and a management system—shares these resources between passengers in a single vehicle in real time; on the other hand, the taxi, which vertically integrates a vehicle, a driver, and a management system (in the case of a taxi firm), but where the sharing between passengers takes place by alternation, as does the occupancy of a public parking space.

Service innovations are reshuffling the modal components and offering new forms of organization, based on new ways of sharing physical resources, which of course still have to be amortized over a large number of uses. The scope, durability, and potential spread of an innovative service depend on its information management system. This system is at the heart of the new service, and it dominates the other components not functionally, but in the distribution of commercial power between the central system and the owners of the vehicles and the drivers.

This situation makes it possible for independent companies to form an organizational network, with a micro-franchise system for the basic suppliers of vehicles and driving, in the same way as a company with a network structure.

We will fill out this broad picture with a more detailed description of a series of innovative services—carpooling (Sect. 10.1.3.1), the multipurpose taxi (Sect. 10.1.3.2), car sharing (Sect. 10.1.3.3), shared vehicle systems (Sect. 10.1.3.4), car hire between individuals (Sect. 10.1.3.5), and shared parking (Sect. 10.1.3.6)—culminating in a brief synthesis (Sect. 10.1.3.7).

10.1.3.1 Carpooling: Emergence of the Intermediator

The ancestor of carpooling is hitchhiking, traditionally practiced by non-car owners, in particular young people, often with unpredictable success, long waits, detours, and multiple vehicle transfers. In order to increase the chances of success and reduce waiting time, hitchhikers are well advised to stand by the side of a large, busy road and carry a sign showing their destination in order to match up with a driver going the same way.

A more modern form emerged, which might be described as urban corridor carpooling: Stations are provided along certain big, carefully selected roads, to match up drivers and potential passengers. The advantage for society is to make better use of cars and promote passenger mobility. This kind of system makes particular sense in the case of high-occupancy toll (HOT) lanes reserved for collective vehicles and multioccupancy cars (another modern form: expressway carpool parking).

However, the new informational order offers a much more radical innovation: Over and above possible pickup and drop-off areas, they provide an intermediation

service between carpool providers and clients. Information and matching services have been developed by companies for their employees or by local authorities, but with limited success, because the client and provider base is small, the ergonomics is not simple enough (the smartphone had not yet appeared when most of these services were set up), and the service offered no rewards, except perhaps parking priority for the vehicle concerned.

The firm BlaBlaCar provides a similar service, but it is based on a principle of cost sharing and is quite remarkable in terms of ergonomics and commercial functions. Potential clients use a Web app on a computer or smartphone to enter their journey requirements, specifying their starting point and destination and preferred travel window. The application provides a series of appropriate options, each with a price, an indication of whether or not the journey will include toll roads, and feedback on the driver from people who have previously used the services.

On the driver's side, the interface is just as simple. In addition, the driver remains in control of planning the movement and organizing the journey, by accepting customers and negotiating the rendezvous.

Various tactics are possible, in particular reducing the price in order to attract more clients and maximize revenues.

The intermediation service brings together clients and providers. It provides reassurance about the provider, who is duly identified and assessed; for clients, who are personally entrusting themselves to a stranger in a dominant position, it acts as a trusted third party. In addition, the service receives the payment, takes its commission, and transfers the rest of the money to the provider, which facilitates transactions in the vehicle and also relieves the driver provider of the responsibility.

So the intermediator is not a simple intermediary: By centralizing requests, providing information, establishing trust, and taking payment, it creates value both for the client and for the provider. Carpooling thus becomes a three-tier operation, which has achieved considerable success. In the USA, services similar to BlaBlaCar are Lyft and UberPop.

10.1.3.2 The Micro-franchise Taxi

In the three-way relationship of carpooling with an intermediator, the providers are paid by the clients, but this is primarily a sharing of costs. UberPop has been the target of criticism in several countries, because the price exceeds the fuel costs, which brings it close to being a commercial activity.

Other services based on a powerful intermediation system are emerging. In particular, Uber matches clients to drivers for limousine services (Uber Black) and car-and-driver tourist services (UberX).

Compared with a taxi firm with a real-time management system, there is a major organizational difference: the commercial dissociation between intermediation and production. This makes for a particularly flexible combination: Each vehicle holder–driver is a self-employed person who joins a commercial network, but remains the holder of their physical means of production (another important

difference is the tax regime, which needs to be harmonized between products belonging to a single family of services).

This difference in distribution allows the intermediator to make profits with modest investment and the individual producer to conduct a paid business with fluid access to a wide client portfolio and to professional services, in particular for billing.

Ultimately, the users benefit from this new service and the efficiency of the intermediation.

10.1.3.3 Car Sharing

A car sharing service provides access to a car for individual use, by pooling the vehicle among different users over time (with alternating use). This service requires vehicles, parking spaces, a reservation and access system, and adherence to protocols of usage, ownership, and maintenance.

Originally, car sharing services grew out of community initiatives. The main organizational techniques developed gradually, and the system became professional, with the emergence of companies that provided protocols—in particular for intermediation—but also vehicles and parking spaces, often in concert with the regional authority in the service locality: cf. firms such as Caisse Commune in Switzerland, ZipCar in the USA, Communauto in Canada, etc.

Car sharing is a service that pools vehicles (in alternation) and parking spaces (permanently); users drive themselves. The pooling aspects relate car sharing to transit systems, while the self-drive aspect is closer to the private car mode.

The advantages to society are that users are less attached to cars and show a predisposition to multimodality—mostly practice alternative modes to the car—as well as more productive use of both parking spaces and vehicles. However, there is no sharing of seats in the car.

The car would seem to be the only vehicle versatile enough for sharing to work. Two-wheeled vehicles, which are much cheaper to buy and run, and much easier to park, tend to be chosen personally by their owners. As far as we know, privately owned bicycles have not been the target of significant community bicycle sharing initiatives.

10.1.3.4 Shared Vehicle Systems

There are now a variety of commercial services for the sharing of two- or four-wheeled vehicles for alternating use over time by different users: Halfway between car sharing and conventional short- to medium-term car rental, shared vehicle services offer very short-term hire within a circumscribed geographical area.

Some modes use stations for shared vehicles, where the vehicles have to be picked up and dropped off. The drop-off station may be different from the pickup

station in a one-way system such as Vélib or Autolib, but must be the same in a round-trip system such as Autobleue (in Nice).

Other services do not use stations, notably Car2Go (in Ulm, Austin…). Whatever the specific procedures, however, the system supplies the vehicles, insurance, maintenance and service, as well as the access protocol, perhaps with the possibility of reservation, and of course information system and a payment protocol. The service area needs to be large enough and dense enough for this form of sharing to be attractive and profitable. Cars need to be manually relocated to balance supply and demand.

The no-booking style of access to this mode of travel makes it comparable with transit, as does the sharing of the vehicles and parking. However, the use remains individual. Bicycle-based services (e.g., Vélib, Bixi) capitalize on the societal benefits of this mode—little or no environmental damage, and greater urban integration of both movement and parking. Certain automobile-based services use electric vehicles (e.g., Autolib) to adapt to the urban environment by avoiding local pollutant emissions and reducing noise pollution.

10.1.3.5 Car Hire Between Individuals

Car sharing and shared vehicle services provide a pool of dedicated vehicles, to which no user has privileged access except through the subscription arrangements. A different kind of service that has emerged is car hire between individuals (e.g., Drivy). The vehicle is made available for a period set by its owner, at a cost that undercuts any of the commercial competitors.

The intermediation system plays an essential role as a virtual sales counter between supply and demand, providing information, reservation, and payment.

Here again, pooling benefits both client and provider, as well as the intermediator.

The gains made by all three parties also benefit society, as does more intensive car use, which reduces the technical lifetime of the vehicle and brings forward its replacement by a more modern vehicle, likely to offer better environmental performance.

10.1.3.6 Park Sharing: Sharing of Parking Spaces

Car sharing and vehicle sharing services pool not only vehicles, but also parking between their users.

Shared parking is also developing independently, especially in dense urban areas where parking spaces are relatively scarce and therefore expensive. Intermediation services have developed in this field, targeting both individuals and companies, to encourage people to use each other's resources.

Partway between traditional shared parking systems, either roadside or in public car parks, and long-term renting of parking spaces between individuals, park

sharing works with privately held spaces rented for very short periods. In particular, some companies have large, underused car parks, which can be put to good use through services like this.

In France, the ZenPark and MobyPark firms offer to turn parking spaces into connected objects that are made available through a specific Web app: The car user can prospect some places and book one that is vacant. The ParkaDom service is targeted to individuals in order to ease hiring for any duration, from one hour to one month or more.

10.1.3.7 Provisional Synthesis

Powerful systems of intermediation between users and services have emerged in the sphere of urban passenger transport. Such systems centralize information and distribute it continuously to their users, according to the particular needs of the latter. They connect clients with providers and carry out all commercial functions.

The system's material medium is primarily IT, a combination of a central server relaying information by Internet to computers and especially smartphones, using a Web app specific to the service. For the user, the ergonomics are radically simplified: Setting up the service simply requires reading a page or two on a smartphone screen, clicking a few times—i.e., look and touch—according to a principle of immediate access. The transaction time is insignificant, so there is no disincentive to spontaneous use of the service, even if it proves unavailable.

The services provided by the intermediation systems broaden and strengthen the whole transit sphere. The most productive forms are those that optimize vehicle occupancy in real time: collective taxis—especially minibuses—and carpooling. Other forms rely on the alternating use of vehicles or of parking spaces: taxis, shared vehicles, car sharing, and park sharing.

In other words, pooling and sharing are on the move in urban transportation. It affects not just the uses, but also—symmetrically—the ownership of individual resources. Information sharing leads to the functional economy, to less individual attachment to mobility objects.

10.1.4 Autonomous Vehicles

Autonomous or self-driving cars have been demonstrated by several car manufacturers and IT companies. They can legally be used in several US states, the Netherlands, and the UK, albeit with a driver ready to take over, a systems engineer for remote assistance upon request and a bank guarantee. The car industry is investing heavily in autonomous cars, and it is an industry with powerful lobbies and often strong government support. It is expected that laws and regulations will be relaxed to allow driverless vehicles on dedicated lanes, on motorways, and eventually perhaps on all roads. Traffic safety is expected to improve, since over

90 % of traffic accidents are caused by driver error. More efficient car following through the use of "platooning" may improve roadway capacity.

This section discusses the implications of autonomous vehicles in transit, in taxi fleets, and on dedicated or shared guideways.

10.1.4.1 Driverless Transit

More than half of the operating cost for transit is driver wage. This fact has driven development toward ever larger vehicles and trains. In order to fill large vehicles, passengers need to be bunched in space and time by gathering them in stations and serving them at longer time intervals.

Driverless operation has already been introduced on dedicated rights-of-way such as subways, People Movers, and Pesonal Rapid Transit (PRT) or Automated Transit Networks (ATN). Technology developed by the car industry can be applied to trucks and buses as well. Some suppliers already offer conversion kits for standard vehicles.

The implications of driverless transit will be manifold:

- Without drivers, there is less motivation for large vehicles. More and smaller units running with higher service frequencies offer better service with shorter waiting times;
- Smaller buses are more suitable for adjustment to demand, e.g., by route diversions;
- Layover at terminals can be shorter without the need for driver breaks;
- There will be more freedom in scheduling when driver reliefs are not needed;
- Schedule/headway adherence will probably improve when bus runs are centrally monitored;
- Collision avoidance equipment will reduce accidents, repairs, and damage expenses for transit operators;
- Less damage should lead to lower insurance premiums or reduced costs of damage;
- Platoon driving can increase link capacity through bottlenecks such as the overcongested bus lanes in the Lincoln Tunnel between New Jersey and Manhattan New York.

10.1.4.2 Driverless Taxi "aTaxi"

The car industry is targeting the private car market by offering greater driver comfort. Commuters can work, communicate, relax, or be entertained during the ride in their autonomous private car. This in turn may generate further urban sprawl and more traffic.

However, a driverless car is also perfectly suited to taxi fleet applications:

- Taxi fares can be more competitive without the cost of drivers;
- Fares can be known in advance;

- Fleet management can be more efficient with the optimization of dispatching and empty movements, along with the reduction in parking times and parking spaces;
- Ride sharing in taxis can be planned with smartphone apps and central control.

aTaxi can become an alternative to transit at times of low demand and in suburban areas. It will be more attractive than transit by offering door-to-door service on demand, but will still be more expensive than transit because of a lower level of ride sharing.

It is expected that many households will give up their private cars or at least their second car when aTaxi services become available. This would free household capital for other uses. aTaxi would be a public system available to everyone, regardless of driving license, age, and sobriety.

The most visible effect of aTaxi in cities would be the release of parking spaces and car parks for other purposes. By contrast with the private car, which is parked 95 % of the time, an aTaxi simply goes on to pick up the next passenger. Parking spaces today take up more than 50 % of the land area in many cities.

Traffic will not necessarily be reduced with aTaxi replacing private cars. Empty vehicle movements need to be offset by higher degrees of ride sharing (Burghout et al. 2014).

Ride sharing together with empty vehicle movements is key to efficient aTaxi operation. We foresee customers ordering aTaxi with a smartphone app, specifying their destination and a time window for pickup. The pickup point can be given automatically by GPS or entered manually. The ride sharing possibilities will depend on the size of the time window and on the detours accepted to pick up or drop off copassengers. These factors could be set individually and determine the cost of the aTaxi fare.

In comparison with shared cars, aTaxi requires no driver's license and need not be picked up nor be parked.

Security in aTaxi is a factor that needs to be addressed. A taxi with driver is not always very secure, but unmanned taxis may still be a concern. We foresee pre-registration of users, pin codes to identify passengers, cameras in all cars, and emergency call buttons.

From a modeling perspective, aTaxi is equivalent to a conventional taxi except that empty vehicle movements and ride sharing can be optimized with one of the DARP algorithms.

For integrated transit modeling, the aTaxi part of the trip can be represented by average waiting time per zone and OD matrices of average ride time per time of day.

10.1.4.3 Autonomous Vehicles on Guideways and Roads

Autonomous vehicles are already operating on guideways in the form of People Movers, PRT, or Autonomous Transit Networks. Some of them run on rubber wheels and can be developed to leave the guideway under manual control.

The car industry plans to introduce self-driving cars on roads. Elevated guideways for cars may be attractive in areas with heavy congestion. Such guideways may be less costly than adding another paved lane with the dimensions and rigidity needed for heavy vehicles. In many places, there is no space for more lanes.

Elevated guideways can be self-financed by charging a toll. People are willing to pay wherever they can avoid congestion delays. Cars can more easily and safely be made to self-drive on guideways than on roads. Platooning would increase guideway capacity. The journey range of electric cars can be increased by charging systems (using sliding contacts or induction) on the guideway.

The same guideway can be used by guideway-bound transit and by dual-mode cars equipped with the proper communication and control. Transit stations can be elevated or at grade in conjunction with off- and on-ramps for cars.

Private dual-mode cars can be inserted into the transit system when not needed by the owner, instead of being parked, bringing substantial savings in terms of parking space and vehicle fleet sizes.

10.1.5 What Prospects for Urban Mobility and Multimodality?

10.1.5.1 Modes Are Diversifying…

Mobility for people in cities is the object of multiple service innovations. In the era of Web apps, transportation and parking services are being reinvented.

Powerful systems of intermediation provide services that match supply and demand and manage commercial functions. Logistical functions for making resources available, traditionally integrated into the commercial function in mass transit modes and also in the supply of private vehicles to individuals, can now be dissociated.

The result is a wide qualitative diversification in mobility services. Between the two traditional forms of line-haul transit systems and the private car, we are now seeing the insertion of services such as collective taxis, carpooling, micro-franchise taxis, shared vehicles, and park sharing.

Pooling for physical resources is on the agenda, stimulated by information sharing and easy interactions between potential users and providers of resources of all kinds.

10.1.5.2 But also Becoming More Homogeneous

Alongside the general trend toward modal diversification, two movements are underway that are simultaneously driving a degree of homogenization.

On the one hand, for a long time, the urban sphere was long a place of geographical singularity, marked by local customs, influenced by physical, climatic, historical, and cultural factors. True, the private car has spread worldwide, but with local adaptations in vehicles, infrastructures, and conditions of access and movement. Mass transit systems are equally marked by local contexts, in the importance of their role, in their technical and tariff conditions, and in their regulation. Individual or collective taxis show even more local differences. Now, the new mobility services are spreading internationally, some with significant local variants (e.g., shared vehicle systems), but others with adaptations limited to language (in the Web app interface) and the local currency units. If they continue to develop, they will generate a certain convergence between conurbations to a fairly standardized set of multimodal options. Business visitors or tourists will push for this uniformity, which will make it easy for them to transfer their day-to-day routines away from home.

In addition, all the modes and services are becoming smarter. Smart travel, thanks to dynamic information on the physical state of the system and assistance with the choice of route, mode or even schedule. Smart movement, with assisted driving and soon driverless cars. These intelligent functions are incorporated into the objects, which releases the user to carry out other activities at the same time.

As a result, individual transportation is coming to resemble mass transit, collective taxi and carpooling services are approaching the conditions of the private car, and the resemblance will further grow when vehicles become automated.

Alongside this kind of convergence between modes with regard to use, there will be a certain convergence in the technical principles of vehicle movement, for example, automatic travel on a linear infrastructure, or the platooning of driverless vehicles, to form a sort of road train.

10.1.5.3 Personalized Services, but Without People

Automated intermediation offers its users a very wide but at the same time customized range of information, with information that is targeted and dosed, since parsimony is also a factor of efficiency. This customization gives the service a sense of attentiveness.

In addition, intermediation can include community control of service providers (via the scores attributed by users and their feedback comments): This is an effective alternative to hierarchical control, since it stimulates providers to give good quality of service.

Vehicle automation will further promote the customization of mobility services, by bringing the vehicle to the client in the place and time of their choosing. The downside, however, is a profound dehumanization. Operational personnel will concentrate on maintenance and remote complaint handling.

10.1.5.4 Economics of the Services and Potential Demand

Innovative mobility services combine physical resources, which are in principle exploited more intensively than in private use, and a highly automated intermediation system. The unit cost of a service should be lower than traditional modes, in defined conditions of use: the private car for limited annual mileage (up to 10,000 or 15,000 km per year) or the traditional bus with very low average daily occupancy.

The transportation and parking infrastructure does not impose costs on these services: The roads already exist, and local authorities tend to favor parking for shared travel modes.

As a result, innovative services carry few fixed costs, which means that can be easily adapted to the specific local needs of each area. In short, this kind of service is flexible and resilient.

These in principle favorable economic factors will be further enhanced by the confidence of financial investors, who have already put huge capital sums into certain services, even early in their development ($41 billion at the end of 2014 for Uber, set up in 2009). This financial firepower enables them to invest for the long term and to introduce their intermediation service at a very moderate price, or even for nothing in the start-up phase. It seems out of proportion with the investment needs of an intermediation service and hence a striking contrast with line-haul mass transit systems, which traditionally find it hard to finance their large-scale investments.

One reason for this investor confidence in innovative services is the robustness of their business model: because of the specific value of intermediation, but also—more profoundly—because the mobility services provided will be priced in proportion to their use, much more directly than in traditional modes. There is a sort of techno-economic soundness in this, which indeed inspires confidence.

The difference in confidence between innovative services and traditional mass transit should ring bells with public transportation planners, in particular with regard to big projects for orbital lines around cities, like the Grand Paris Express: Would it not be more economical to invest in road corridors, possibly underground, and to encourage collective taxis?

10.1.5.5 Demand and Its Influences on the Multimodal System

Mobility demand is traditionally segmented according to the modes, which their users have become accustomed to and familiar with. Among habitual users of the private car, as among mass transit users, a significant proportion will not change their habits. A fraction will adopt an innovative mode as their preferred travel mode. Yet another fraction will be multimodal in their behavior: Every time they travel, they will first estimate the quality of service on the different modal options. The distribution of all users between the fractions will evolve with time, probably with a

gradual development of the multimodal type, stimulated by the availability of comodal trip planners with up-to-date information.

When users compare modes in order to choose one, this establishes conditions of potential substitution and therefore competition between them. By contrast, intermodality is a cooperative relationship between a sequence of modes on a single trip. The functional complementarities between individual modes for access to the mass transit system (walking, bicycle, car) and efficient collective modes for the main journey (railway or BRT) are already well established and widely practiced. One-way shared vehicle services seem to fit harmoniously into this cooperative system, whereas round-trip systems are better suited to the final city stage of interurban journeys. Similarly, carpooling and individual or collective taxis connect easily with efficient mass transit systems.

It is still uncertain, however, how intermodality will work between several innovative services: What can be the link between a carpool service and a collective taxi? Or between two collective taxis, each of which would prefer to remain inside its catchment area? Indeed, for a taxi mode, transfer between vehicles only seems natural to the passengers for access to a higher performance network.

Whatever happens here, demand will benefit from a diversified and globally better quality multimodal service, where prices will be held down by competition.

The advantages are obvious outside the usual mass transit sphere, i.e., outside operating hours or in the periphery of cities, where densities are low. Here, collective taxis and carpooling find a sphere of relevance that has already emerged, in a place where they were traditionally already present.

The diversification of services will be stimulated by the modal choices of clients; by choosing the mode most convenient for them, they will focus each service on its specific sphere of relevance, the expansion of which will depend on the specific location.

This stimulus will be all the more virtuous in that the charges of innovative services are directly proportional to actual use, so the price signal is right, and all the better for the fact that a shared service has a rental aspect, so the price will be geared to the full unit cost rather than a marginal cost.

In these circumstances, the pricing of innovative services should stimulate a tariff revision for mass transit modes, linking them more directly to actual use and harmonizing the unit price more fairly between regular users and occasional passengers.

The pooling of cars will raise questions about individual car ownership: What is the point of owning a car? In fact, a car that is used more intensely should cost less to use. So private car ownership will be justified when individual car use is very intense (annual travelled distance beyond 30,000 km), or in the case of customization that generates usage value: internal arrangements for personal needs (e.g., restricted mobility) or family needs (child seats), value of immediate access for private or professional purposes, and image factors (luxury and show-off).

10.1.5.6 What Future Structural Changes in Demand?

Innovative mobility services will be accessible at a moderate price and therefore to a large majority of the population. Modest but solvent households (e.g., individual returning to employment) should benefit particularly from car sharing, which will provide performances similar to those of a private car but at a much lower overall cost.

Another social advantage: Automated driving will make cars more accessible for people with disabilities and those without a driving license.

Expanding services, combined with falling prices, should stimulate overall demand, especially as payment will be easy and the time spent in travel will cost users less, as they will use it for other activities. However, the clearer pricing structure will make the cost of travel more obvious to users, who may therefore decide to manage their finances better by cutting their spending in transportation. Another factor of change will be the level of congestion on roads and on mass transit modes, which will itself depend on the development of vehicle sharing services, collective taxis, and carpools.

As regards urban forms: The expected benefits of the changes will be greater in the outskirts (unless there is a systemic effect of greater inner-city fluidity). They can be expected to stimulate urbanization and therefore urban sprawl. This is not harmful in itself: The important thing is that the price of transportation should reflect the costs for the user and the community.

10.1.5.7 Regulation: Issues and Strategies

The community regulates the transportation system by carrying out a set of functions:

- It plans the road system as an infrastructure to be shared between travel modes, and it manages traffic.
- It supervises transit, particularly mass transit but also taxis and parking. In cities, the community generally contributes heavily to the funding of mass transit.
- It regulates the multimodal transportation system, in other words the transportation service industry.
- It decides and implements mobility policies, in interaction with other sectorial policies, in order to meet social, environmental, and economic objectives.

Now, the new informational order is ushering in a new order in transit. For line-haul transit, a direct consequence is a—potentially profound—improvement in efficiency. In parallel, innovative services are emerging or spreading, primarily through spontaneous initiatives: They are competing with transit without seeking finance (at least for operations).

The public can gain from this on every front, through better use of mass transit, through the use of innovative services in their sphere of relevance, and even in the use of private cars as increased sharing leads to more fluid traffic conditions. In

addition, competition from innovative services is a great opportunity to revitalize line-haul mass transit systems. This competition will help to rationalize pricing, making it more clearly related to the service provided, at a level closer to the full cost, possibly reflecting the degree of pooling entailed (as in the Lyft service which proposes the sharing of costs between users)—all in a simplified form inspired by e-commerce (online payment) and with greater acceptability.

In addition, competition should stimulate improvements in the position of line-haul mass transit systems within their sphere of relevance, probably driven by a reconfiguration of bus lines for a better distribution of their in-route and feeder functions. The feeder function could in many cases be performed by collective taxis and other car sharing systems.

All these potentials constitute a real windfall for urban transit and therefore for the authorities responsible for managing the mobility system. It is up to them to grasp these opportunities boldly: by taking the right support measures and by steering changes in order to reap the benefits and reconcile the respective interests. In particular, the activities of the stakeholders need to be orchestrated so that they contribute to the harmonious development of the whole.

A key point concerns the market potential of innovative services and the refocusing of line-haul mass transit systems on their best sphere of relevance.

We will divide up the opportunities and the regulatory levers into their broad priority areas: first, environmental and urban quality priorities, then infrastructural priorities for traffic and parking, then public transportation service priorities, and finally economic and social priorities.

With regard to *the environment and urban quality*, an efficient multimodal transit system should reduce overall vehicle traffic and therefore reduce local traffic, energy consumption, and pollution emissions. An additional opportunity lies in the renewed appetite of investors for urban transit: This could facilitate the transition from the internal combustion engine to electric or hybrid modes, by mobilizing capital to finance the additional cost of acquiring these vehicles as compared with conventional vehicles. Electric vehicles should fit particularly well into automated driverless systems, since they can carry out their own recharging.

Less traffic, with a growing spread of autonomous and electric vehicles: This should lead to improvements in the urban environment (calmer driving, less noise pollution), health gains (reduce pollution emissions), and road safety gains (as for civil security, steps will need to be taken to prevent the risk of autonomous vehicles being hijacked by ill-intentioned people, by installing local security barriers around sensitive locations, and by fitting these vehicles with automated systems to ensure that these barriers are respected).

With regard to *infrastructure*, better vehicle sharing should reduce parking demand. At the same time, enough free parking spaces must be available for small transit vehicles to be able to pick up or drop off an individual client without disrupting local traffic. Focal points for passenger flows, such as railroad or bus stations, must also be designed to handle minibuses and other collective taxes, through specific forms of design that will reconcile vehicle size, passenger access, and specialization by destination area.

With regard to *traffic*, major arteries will need to be designed so as to avoid local disruptions from individual access needs. There will also have to be agreement about the granting of privileges for pooled vehicles in traffic management: priority at specific junctions, based on a certain vehicle occupancy level and/or deviations from timetables.

With regard to *transit services*, there is a need to think about recentering bus lines on the sphere of relevance specific to the size of the vehicles, both in terms of capacity and dimensions, and particularly by seeking a more efficient service for passengers: on more straight-line journeys, between more widely spaced stations and coordination with local junction control.

With regard to *economic and social priorities*, we would begin by pointing out that the benefits of calm, health, and security—in other words of public order—are key concerns for city dwellers, alongside the benefits to users. So local people, who are both residents and users, have double gain from the new transit order. In addition, on the supply side, recentering on the best spheres of relevance for greater efficiency should also lead to greater profitability.

The subsidiary question concerns the employment of operating personnel, particularly taxi drivers, who will be affected by new services and vehicle automation. Social support needs to be organized to facilitate the transition to related activities: For example, the driver of an individual taxi can become a collective taxi driver, or an urban delivery driver, since more and more people today order goods online for home delivery.

Finally, it is up to the community to orchestrate the overall process, by giving further impetus alongside the spontaneous initiatives of private actors and by setting the pace of change. To achieve this, they need to devise scenarios for gradual transformation, differentiated according to the final scale of transformation and the pace of implementation.

A simulation model will be valuable in seeking to anticipate the practical effects of each scenario, to compare the strategies for change, and to establish a preferential scenario which will support political decision-making.

10.2 Research Topics on Transit Modeling

Fabien Leurent and Ingmar Andreasson

An assignment model of traffic on a mass transit system reflects the encounter and interaction between a system of transportation that provides services and a population of passengers wishing to travel. The fundamental function of such a model is to represent the interaction between supply and demand, by capturing the essential features of that interaction: the flow of passengers at different places and times, the quality of service from point to point, and commercial revenues.

To do this, the scientific method is to *describe* each part, whether supply or demand, in its specific composition and operation (technical and commercial

operation for supply, economic and social behaviors for demand) and to *explain* their interaction as the combined set of their respective behaviors. This combined set of behaviors notably produces effects in the local formation of flows and in their influence on the quality of service and the economics of supply.

10.2.1 Background

In a half century of scientific research, assignment models have made significant cumulative advances. The first task was to understand the spatial structure of transit services (in the 1960s) while capturing their time-discrete nature and its impact on passenger route choice (in the 1970s and 1980s). Once these basics were established, it became possible (in the 1990s and 2000s) to model different macroscopic flow and capacity phenomena on the supply side and, on the demand side, to produce better models, more sensitive to the characteristics of passengers and to their particular travel conditions (quality of service, dynamic information). After this (since the 2000s), the focus of research shifted to the time dimension, the quest to capture the dynamics of the system's operation, both within a single day, notably by identifying traffic peaks, and between successive days, with an emphasis on learning phenomena for passengers and stabilization mechanisms for the state of the system. Please, note that our chronology is approximate. Each significant aspect received pioneering contributions, some of them well before the period mentioned, which reflects primarily the abundance of the work published on a given theme and the scientific community's grasp of the complexity inherent to that aspect.

The dynamic operation of the system remains an extremely lively field of study, with macroscopic, microscopic, or hybrid models (macroscopic for travelers, microscopic for vehicles) to capture normal traffic behavior, or disruptions and their consequences, and also the use of flow simulation to contribute to real-time service management (see Chap. 8).

10.2.1.1 Research Problems: Main Research Fields

Over time, therefore, the science of traffic in an urban mass transit system has reached a certain understanding of the spatial and temporal aspects in the operation of such a system. For the present (2015), this understanding remains partial, since the model's explanatory power is essentially focused on traffic flows.

Our understanding of the supply aspect requires further work: a better grasp of the particularities of each mode, from bus to train, not just in terms of vehicles but also infrastructure and the passenger interface; and also, making certain supply characteristics endogenous, in particular service frequency and dynamic traffic management.

On the demand side, the typology of passengers in terms of their needs and behaviors, and the associated statistical structure, is an important research area, as is discerning the route options and decision rules for a given class.

Another major research area is the empirical comparison between simulation results and observational data. Previous work on this subject has produced pioneering contributions: The challenge is to design statistical methods to assign confidence intervals to the different simulation results, as well as to the different observational data, and to make fine-grained estimates of the parameters and functional forms in the model.

In addition, this is a transitional period in the diversification and transformation of mobility systems: diversification in the technical and commercial forms of services, and transformation of uses and users. There is a need to model innovative forms of mobility system, to explore their market potential on a case-by-case basis in a diversity of territorial contexts.

Finally, a transformation is taking place in modeling itself. Computing power continues to grow. More and more observational data are available (big data). The state of the system can be described by variables that are simultaneously simulation results and real-time field "observables," which allow comparisons and reciprocal inputs between simulation and observation.

10.2.1.2 Objective: To Identify Research Paths

In this section, our objective is to identify research topics for the modeling of urban passenger mass transit systems. There are three types of challenges for such research, which cut across the big fields of investigation already mentioned:

- First, a theoretical challenge: to describe and explain generically the composition and operation of components, of subsystems: physical composition and technical operation of the service supply, composition of the demand, and socioeconomic principles of mobility behaviors.
- Second, an empirical challenge: the statistical or econometric capacity to reproduce and to reconstruct, to a given level of accuracy, the observable operation of a real system.
- Third, an operational challenge: to make the simulation model contribute to the real-world management of a system studied by means of the model. The model was originally applied to the planning of a network of lines, in a way that is primarily technical and secondarily economic. From now on, the model may encompass the management of traffic to optimize operational performance: Pioneering applications have already been implemented for road traffic management, but mass transit traffic management is a new frontier for research. Commercial management is another frontier, for the development of pooled mobility services, and a parallel economic rationalization of all transit modes, in particular mass transit.

10.2.1.3 A Systemic Approach

Modeling a sociotechnical system is essentially a form of operations research, in the original sense of the term; a system model enables the analyst to simulate the operation of the system and to adjust different action levers in order to optimize its performance. This is how a simulation model can contribute to the management of the system.

The different facets of modeling—physical and economical theorization, systemic structuring, mathematical formulation, algorithmic specification, econometric formulation, and statistical estimation—are all ingredients in the formation of an applied model. Rather than covering each facet successively, we give priority here to the system's structure, to its organization into two subsystems—supply and demand—and to their interactions.

10.2.1.4 Organization of the Section

The supply–demand structure of a transit system is the reason for the section's organization into seven parts.

Demand is covered in two parts: first, the aspects relating to individual passengers, in terms of situation, practices, behaviors, and decisions (para 1); second, the structural composition of demand into classes of travelers and also in terms of different territorial and temporal circumstances (para 2).

The next part concentrates on micro-local traffic: flow phenomena for travelers and vehicles, their effects on quality of service, and the impact of different management levers, including spatial layout (para 3).

Beyond the micro-local scale, the management of supply is covered in four parts. First, at the level of a transit line, simulating flows and costs, characterizing performances and optimizing management (para 4). Then at the level of a network of lines as a modal system (para 5).

Next, we look at the diversification of transport supply, through intermediate modes that pool vehicles between travelers in different technical and commercial forms (para 6). Finally, we tackle multimodal transit supply in a conurbation, in terms of functional and technical cooperation between modes, but also commercial competition and public regulation (para 7).

10.2.2 Individual Behavior, from Situations to Decisions Passing by Gestures

The individual traveler intervenes in the system and therefore in an assignment model, in several ways: first as a customer, a unit of service delivery; then as a user, exposed to local conditions of traffic and service quality; next as an autonomous

decision maker, a decisional unit in route choice and other smaller or larger decisions; and finally as a passenger, a unit within a flow.

Here, we consider three research topics, relating to:

1. The individual's situation within the system, as a user both exposed to ambient conditions and capable of agency.
2. The adaptive behavior of the passenger, who perceives the travel conditions and adjusts by making different kinds of choices.
3. Mobility practices and habit formation, through learning and its application.

10.2.2.1 The Individual User Within the System

As a user, the passenger may be located on a vehicle or in a boarding area or in yet another pedestrian element. He or she will be in a particular physical state, with varying degrees of comfort (notably either sitting or standing). Specific movements will instigate a shift from one physical state to another, either through very basic transitions (e.g., sitting down), or through transitions with greater impact on the journey, such as boarding or leaving a vehicle in a certain station.

In addition, although self-directed, a passenger performs no driving function and is therefore available for additional activities that depend on the particular conditions of the trip.

With a smartphone or tablet, even in congested mass transit conditions, passengers can occupy their minds with a variety of activities. High up on the list of these activities is verbal or written communication, provided that there is wireless access to a telecommunication network. This communication can be general in nature, or can relate to the travel situation in particular, allowing passengers to acquire real-time, customized information and also to produce information relating to their perception of the current situation.

In the latter case, the user becomes an information sensor: This function can be useful for the system, especially as the user is mobile, intelligent, and interactive, but only to a limited degree, since the user is autonomous and will not act as a roving informant who can be controlled remotely.

The research subjects on this topic relate to:

- The disaggregated observation of physical states and basic movements, accessory activities and instantaneous practices, and communication activities relating to the travel situation, throughout the trip and in relation to the traffic conditions experienced.
- The study of feelings and perceptions: sensations of comfort or discomfort? What is the user's particular sensitivity? How are local impressions consolidated or diluted throughout the trip? What conditions would the user prefer, and would he be willing to pay for them? Obviously, this leads to the utility function of a user for a route option.

The primary medium is obviously the user's smartphone, with its GPS functions and even, in certain cases, an internal gyroscope which can record basic movements, and above all with its interface and applications. With the right apps, users could provide information on their specific travel conditions, their basic movements, their perceptions, and their reasons for traveling.

Information could also be acquired about the user's average walking speed, use of staircases and escalators, etc.

10.2.2.2 Route Choice Behavior

Beyond local conditions and basic movements, a traveler makes a choice of itinerary. The core assumption in this respect in an assignment model is that the traveler has extensive and objective information in advance on services, their frequency and journey time, and their spatial configuration. Model variants may include some specified variability affecting certain descriptive attributes or dynamic information, particularly about the arrival time of vehicles in stations. In any case, it is assumed that the user identifies a range of options—a large number in the case of a long journey and highly interconnected network—and can assign a utility to each one, or even combine utilities for bundled options!

This postulate of the traveler as a hyper-informed hyper-planner constitutes a crucial, sensitive, and central point in an assignment model. In this respect, technological developments have brought significant advances:

- Above all, they have made the postulate more realistic, in the case of passengers equipped with a smartphone that provides customized and integrated information. The user's capacities are expanded through information and route advice services. In other words, with regard to information and route advice, the model was ahead of reality.
- In addition, more dynamic information is now available. In particular, information about passenger loads is becoming ever more accurate and can be passed on to users so that they can fine-tune their route choice or their waiting position on a train platform and therefore their position inside the train. The assignment model can contribute to this information. The use of information, the way it is accessed, and the rate of response need to be studied, analyzed in relation to certain factors, and integrated into the simulation model.
- Again for a traveler with a smartphone: Users of route advice services can specify preferences, e.g., the fastest journey, the cheapest, the fewest changes, exclusively by certain modes. The user's preference options correspond precisely to route choice behavior parameters in the assignment model, which is a good reason to investigate them, obviously in relation to the Web app offering the service. Similarly, surveys need to be done on the checking of options and the selection of one in particular, not only in terms of the result but also in terms of the method of use—at what points before or during the journey, etc.

- Again for a traveler with a smartphone: Individual itineraries can be recorded to detect repeat journeys and therefore to identify how common a situation is.

Investigating through a Web app with volunteer subjects seems much simpler, although less general, than other methods based on ticketing data or tracking the digital traces left by mobile phones.

Related research subjects concern the way users adapt to variations in travel conditions:

- Characterizing ordinary mobility conditions and detecting the intrinsic variability caused by users themselves in repeat journey patterns, for example, when weather conditions affect the pleasure of walking and therefore the choice of feeder station. Or transporting luggage, or group travel, in a certain proportion of repeat situations.
- Studying the user's perception of disruptions: What traffic situations do they consider normal, or abnormal but tolerable, or abnormal and intolerable, therefore requiring a change in travel plans?
- Studying how users manage disruption. If they have a smartphone and a dedicated mobile app, what information do they look for, how do they handle it, and what decisions do they take? People may assess route options differently in conditions of disruption than in normal travel conditions and give priority to optimizing their travel time in order to prevent delays in their schedules. It is important to know how people behave in situations of disruption in order to simulate them realistically (up to now, disruptions have been simulated with "ordinary" responses on the demand side) and to assess service delivery accurately with the right dynamic information.
- Studying to what extent informed users follow advice or recommendations from the operator, e.g., the recommendation to change route during a service interruption, either short and unexpected, or long, planned, and announced well in advance.

A closely related subject is payment protocols and practices, i.e., the selection and then purchase of a ticket, and its validation during travel. The disutility associated with this is known intuitively, but could be specified by detailed monitoring. In addition, more modern forms of payment—via smartphone, contactless validation—are economical in terms of the actions required and therefore represent a specific use value.

10.2.2.3 Mobility Practices and Habit Formation

Other research subjects concern the incorporation of mass transit travel into the wider framework of the activities and travel patterns involved in a round-trip, a daily activity schedule, or an intermodal travel chain. Users with smartphones could easily be surveyed on this subject, using the time they spend in mass transit systems.

The observation of intermodal practices, or multimodal practices entailing alternation between modes in a repeat travel situation, would be particularly useful in improving the modeling of individual uses. Users could employ a dedicated mobile application to describe their purpose of travel, any time constraints at destination and the impact of those constraints on their choice of departure time, the cost of any delay, etc.

The concepts of reiteration and standard situations lead to the issue of travel patterns. A regular user gradually becomes familiar with the potential and conditions of the transit system and tends to develop routines after a certain learning period. Rather than experimenting personally with route options, and choosing a satisfactory but not necessarily optimal solution, smartphone users have access to information that is in principle comprehensive and reliable, and to good advice (provided that the route search algorithm in the mobile app meets the same standards as the assignment models, which is worth checking).

It would also be good to investigate multimodal practices in relation to the possession of private mobility resources: car ownership, a subscription to a mass transit mode, etc.

Finally, the user's loyalty to his or her information service is also worth exploring and ultimately integrating into the system simulation.

10.2.3 Demand Patterns

Let us now move from the traveler as an individual, to the population of travelers in a mass transit system, in order to describe the structure of that population in relation to space, to time, and also to a typology of behaviors. Assignment modeling has to a large extent concentrated on representing the interaction between supply and demand in space and time: The composition of demand in terms of several types of behavior has been less explored, except in certain research to demonstrate that typological diversity can significantly influence assignment results.

So the modeling of demand patterns in terms of behavior is an important research area. We will begin by looking at behavioral diversity (Sect. 10.2.3.1), then spatial diversity and its link with behavioral diversity, notably in terms of specific generators (Sect. 10.2.3.2), next the time dimension (Sect. 10.2.3.3), and finally the generation of a population of travelers and journeys (Sect. 10.2.3.4). To finish, we will look at how to measure demand patterns (Sect. 10.2.3.5) and the elasticity of demand to traffic and price conditions (Sect. 10.2.3.6).

10.2.3.1 Disaggregating Demand in Terms of Behaviors

Since the 1970s, modal choice modeling has become very sensitive to the particular characteristics of travelers: In the utility function for a mode, the quality of service and price characteristics of a mode are compared with individual traveler

characteristics, such as age, gender, socioprofessional category, income, vehicle ownership, and purpose of travel. This disaggregated description of the passenger goes hand in hand with a fairly abstract description of a modal option, reduced to a few broad characteristics such as price, travel time, and number of changes.

Route choice modeling assigns much greater importance to space: Route options are embedded in space, on the network, between decision points, which gives them concreteness. This spatialization of services influences the representation of demand in terms of a primary dimension: the specification of trips with respect to the location of their end points and therefore their distribution in terms of origin–destination pairs, in an OD flow matrix.

In many assignment models, demand disaggregation is limited to this spatial representation, whereas behaviors are represented uniformly, with the same individual utility function for each traveler.

An important research topic for traffic assignment is to disaggregate demand in terms of behavior types, in order to obtain a multiclass model. The model's mathematical formulation is fairly easy to adapt. What is needed in research is a qualitative typology of travelers and therefore of individual journeys:

- In terms of the traveler's particular physical conditions: walking speed, reduced mobility due to disability or luggage, etc.
- In terms of the conditions of access to the transit mode: ticket type and possibility of accessing a particular subset of modes, walking distance at origin and destination, or private feeder mode, etc.
- In terms of access to information: ownership of a smartphone, familiarity with the route options.
- In terms of the route-finding process during the journey: pretrip route choice or dynamic choice at several successive decision points.
- Finally, in terms of economic preferences, trade-offs between quality of service and price factors.

Obviously, to combine all these analytical dimensions is complex and hence the need for research to design specifications, to tackle cases econometrically, and to identify the main descriptive axes within these different dimensions. The result will be an abstract specification (a segmentation of demand), which will need to be given concrete form in each particular application, depending on the territory concerned and the time frame simulated.

10.2.3.2 Spaces and Behaviors: Specific Generators

The structure of demand in terms of behavior types needs to be combined with spatial structure and specified in terms of the OD pair. A first approach is to specify an OD matrix for each behavior type. However, the interactions between spatiality and behavior can be more subtle. For example, some demand segments may be

more sensitive to certain quality of service or price characteristics on a short journey than on a long journey.

The specific traffic generators are another topic for research: At certain particular points in space, such as an interurban station or an airport, or a big facility like a stadium or university, and a hospital or shopping center, the purposes and profiles of travelers have marked specificities. Interurban access points or tourist hubs are characterized by a population of travelers who do not live in the conurbation, who have specific reasons for travel and relatively high incomes, who are fairly unfamiliar with the urban transit system and often travel with luggage or in groups. Traditionally, the modeling of specific generators takes into account the particular flows they elicit and adds them to the OD matrix of resident flows. Their specific behaviors need to be modeled, especially for big cities which attract a large number of external visitors.

10.2.3.3 Temporal Variations

Demand, available services, and conditions of use vary over time in interlinked ways. Within a working day, city traffic hits a sharp peak in the morning for mass transit systems, because of commuting and school travel, and a longer and less sharp peak in the evening. Mass transit operators raise their service levels in these periods, which increases capacity but also makes them more attractive compared with immediately pre-peak or post-peak periods, thereby limiting the latter's overflow role as alternatives to peak-period travel.

Each of the typical periods needs to be modeled with its particular conditions of supply and demand: Levels of use depend strongly on these, whether in local flows or in the quality of service of the options and therefore in route choices. Dynamic assignment models capture the dynamics of traffic in the course of a day: Here, we emphasize the need to include time variations in the OD matrix, not only in terms of flows, but also in terms of behavior types, linked with the considerations set out in the previous paragraph. In particular, the distribution of travel purposes varies considerably over the day and hence the distribution of traveler types and the behavior pattern segments.

At a larger timescale, the diversity of days also requires further modeling. Working days and weekends differ, and demand patterns are significantly different between vacation periods (in particular summer in the Northern Hemisphere) and non-vacation periods, tourist seasons, and tourist off-seasons. Supply patterns are often adjusted to some degree, but still in a fairly static way. In the era of information, adjustments should become more dynamic and more could be done on the supply side to anticipate demand patterns. These changes could be preceded by exploratory research into the diversity of demand patterns and their predictability, by supply modeling in order to optimize the system's utility as well as certain operational functions (vehicle maintenance, staff vacations), by the modeling of supply–demand interaction in a more dynamic and interactive way, and by the design of simple information services to facilitate that interaction.

10.2.3.4 Generation of a Population and Its Mobility

Up to now, we have emphasized the diversity of behavior types, especially at the trip level, by means of a macroscopic representation based on demand segments. In the last section, we wrote about the situation of a journey within a daily activity schedule, for an individual in a household taking part in different interactions.

These days, the modeling of individuals and their day-to-day activity schedules can now be addressed on the basis of agent-based models of mobility, a well-known example being MatSim. In this kind of framework, the individual is modeled as a particular agent, with beliefs (perceptions of the environment and subjectivity), desires, and intentions, who makes plans and chooses some to implement. This form of modeling allows us to refine the agent's characteristics, but also his or her interactions with external conditions, which can also be modeled through agents.

The multiagent paradigm offers great potential for the modeling of a mobility system. Its potential for vehicle traffic was discussed in Chap. 6 and demonstrated in Sect. 9.2.2. Its potential with respect to travelers remains to be explored. In this field of research, there are a number of important topics:

- The inclusion of a fine-grained simulation of the conditions of travel by the mass transit mode, resulting from an assignment model, with a fine-grained simulation of passengers and the feedback of quality of service on the organization of their activity and mobility schedule: choice of departure time, choice of mode, and choice of destination.
- In the organization of activity schedules, modeling the interactions between the traveler and the information services in terms of processes and effects. The information service, its route advice function, and its predictive capacity need to be modeled, along with the traveler's use of the service and the adjustments that may arise from it.
- The generation of a population of travelers with mobility needs that are sensitive to mass transit provision and its performance across the territory, from point to point and at different departure times. This kind of model would be situated upstream of a MatSim, running from mobility needs through to types of activity schedules.

10.2.3.5 On-Demand Measurement

The traditional instruments for measuring demand remain useful:

- Generalist socioeconomic surveys such as household travel surveys, to cover the population of residents in a conurbation and its daily mobility.
- Targeted social surveys on a mode of transport or population segment, often by Stated Preferences, increasingly via remote, automated interviews or Internet questionnaires have become standard.
- Counts carried out at certain points on the system.

Technological advances facilitate interview-based surveys, in particular with smartphone versions, as we have already mentioned. Advances have also improved counting instruments, in particular by video with automatic image recognition, to count travelers, measure crowding densities, etc.

Ticketing has become increasingly automated; every passenger interception produces more information, so that a single individual's passage through several successive points can be tracked. These close-grained results are instantaneously or almost instantaneously available. GPS tracking of an individual trajectory produces detailed information that can be exploited off-line or in real time, as Google and other companies do on road systems, using Web apps that capture and automatically transmit location data. Mobile phone operators can also provide counts, individual trajectories, and travel times, by means of GSM tracking.

There are multiple links here with assignment models:

- Synthesizing, inferring an origin–destination matrix from individual tracking field observations.
- Improving the OD matrix from local counts, through pairing with the assignment model which simulates the proportions of use on a segment in terms of the OD relationship.
- Similarly, estimating the parameters of route choice behavior from local observations.

Methodological research is needed to adapt these applications—which have so far been developed primarily for car travel—further to mass transit.

Applied research is needed for each modal network to exploit the data resources in specific ways.

Above all, modern information systems offer extensive time detail, which opens the way for real-time applications (see next paragraph) and also to the establishment of time-varying OD matrices, within a single day or between different types of day.

10.2.3.6 Demand Elasticity and Supply Management

Individual travelers react to the price or quality of service conditions of supply in their choice of route, departure time, mode, and even their choice of destination and journey generation. This is true both for ordinary situations, in relatively stable service conditions, and for conditions of disruption characterized by unplanned real-time variations.

If travelers are aggregated by macroscopic demand segment, a set of individual sensibilities can be summed up by a demand elasticity function which links the volume of the segment to exogenous factors (the supply conditions). Elasticities play an important role in economic models that are more aggregated than assignment models, intended to optimize a supply plan and price levels (ref. Wardmann). That is why determining elasticities from simulations is an issue of research, or exploratory application, in relation to assignment models.

A related issue is to analyze dynamic situations on the basis of a dynamic assignment model:

- In order to estimate the behavioral parameters of the assignment model, through the observation of real-time conditions, including passenger movements.
- In order to anticipate the short-term propagation of travelers on the network and to draw the consequences on the likely upcoming state of supply.
- In order to optimize supply management in real time, by reacting to disruptions.
- In order to draw up service plans that are resilient to disruptions.

Analyses of this kind are a kind of field experiment, although these experiments are passive and not constructed by the analysts who perform them.

Another direction for research is the construction of experiments, either field experiments in situ or simulations in the laboratory. In a conurbation, the establishment of an urban mobility Living Lab would provide a very effective framework, offering significant potential for changes in service management, modes of use, pricing, and commercial revenues (cf. Sect. 10.3.4).

10.2.4 Flow Physics and Traffic Management at the Very Local Scale

Passenger flows play an essential role in a mass transit system, since the essence of such a system is to concentrate travelers in vehicles in order to move them. Filling a vehicle with passenger up or close to its capacity is a key factor of viability in both economic and environmental terms.

The limit of this principle of maximization is the vehicle's capacity, which restricts the flow it can carry; in addition, the denser the flow, the less comfortable the passenger, which mitigates the attraction of the service.

A transit operator manages flows of travelers and of vehicles, which are themselves entities for traffic flows on roads or railroad tracks, with their own traffic phenomena and their own capacity constraints. A mass transit traffic engineer has to work with both types of traffic unit, travelers, and vehicles, in designing and operating a system: The Transit Capacity and Quality of Service Manual (TRB 2013) describes the physical and technical aspects in detail and provides design principles.

A traffic assignment model represents the system and therefore the two types of traffic unit, dealing with their interactions across an entire network. In reality, the interactions occur at several spatial scales:

- The network scale is wide and two-dimensional in space: It includes the local operation of all the elements of the network and the movement of traveler flows between the nodes, along a line and between lines.
- The scale of a route: A route extends in one spatial dimension for the trajectory of a traveler using one or more lines, or for the run of a vehicle serving one line.

- The local scale of a station, which is a subsystem with pedestrian access to and from the outside, pedestrian passageways and waiting areas, and platforms; or the scale of a section of infrastructure between two stations, which is also a subsystem, whether on a road or on a railroad track. A station and an interstation are elements in both a line and a network.
- The micro-local scale of a restricted passenger space, such as the inside of a vehicle, or a boarding area, or a waiting room, or a pedestrian element (corridor, staircase or escalator, etc.). Such spaces can be critical to the system's capacity: In particular, vehicle doors limit the flow of passengers moving between the vehicle and the station.

The research topics relating to traveler and vehicle traffic apply at these different scales. We mark them out in this section and the following two sections, from the most local to the most global. The present section deals with the micro-local scale. The research issues relate to the physical representation of traffic at this scale, in terms of flow, density, and physical time, but also user comfort, as well as the technical or economic management of the flow by different types of measure. We will begin by looking at the physics of traveler flows (Sect. 10.2.4.1) and the management of those flows (Sect. 10.2.4.2); we will then tackle the physics of vehicle flows at local scale and traffic management at the same scale (Sect. 10.2.4.3).

10.2.4.1 The Micro-local Physics of Passenger Flows

A flow of a given type is a quantity of mobile entities (traffic units) at rest or in motion in a given space. The direction and speed of motion are flow characteristics as important as the number of units.

Another important feature is the density of the flow relative to a certain space: its concentration in people per unit of area, or per linear unit along an axis of motion.

The flow dynamic relates to the movement of certain identified entities and also to changes in the traffic at a given point, in particular the variation in local concentration over time. In passenger traffic, and particularly in a mass transit station, the accessible spaces are limited and divided into functions: pedestrian passageway, waiting area, etc. Their physical layout determines their capacity and dynamic performance: What accesses are there between this space and the contiguous spaces, what thresholds or doors or security barriers are between two neighboring spaces, how wide are they, and what opening pattern is employed? Static performance (holding capacity) and dynamic performance (flow capacity) also depend on the composition of the pedestrian flow and its internal behaviors:

- Each pedestrian is a particular individual, varying in height, varying in bulk, and varying in speed depending on physical condition. The speed of which an individual is capable affects the people behind him and governs the movement of a columnar flow.

- Culturally and socially, individuals tolerate interpersonal proximity up to a certain threshold, which depends not only on individuals but also on their collective culture and the heterogeneity of the group. The threshold between homogeneous individuals is relatively lower.
- Each pedestrian has her own intentions and her own direction of motion. If the space is used for several directions of motion, the mobile entities may hinder each other and slow down. Even with a single direction, the available width forces the flow to move by order of precedence.

A model of traffic assignment on a network must represent the physical characteristics that influence the behavior of the traffic or the quality of service (time, comfort) for the passenger.

In this respect, research is needed on the following topics:

- For a passenger in a space, the sensitivity of the physical time at rest or in motion to the micro-local flow conditions: in other words a time flow function per pedestrian element. The element may be the door of a vehicle, a corridor, or a waiting area, etc. The flow is generally vectorial, with several directions of motion and different passenger categories. A microscopic simulation model can be used to establish a macroscopic function, through regression from a set of simulations for different factor values. It should be noted that, for a spatial element such as a boarding area, service frequency influences the rate of exchange between that spatial element and the vehicles and therefore the rate of change in the stock of passengers on the platform. This phenomenon needs to be integrated into the model, for investigation for each type of element and each configuration of vectorial flow.
- The macroscopic laws obtained for each spatial element, to link a physical time to vectorial flow, no doubt depend on the category of passenger exposed: there is a need here for the detection of statistical regularities and for research into a typology of passengers.
- Again for each spatial element, the greater the density of the flow, the more uncomfortable the time spent waiting. Discomfort laws need to be modeled that link the general cost per unit of physical time with the density of the flow and also no doubt with the movements that affect the flow. This will need to identify different passenger states in the spatial element (in particular sitting or standing places both in a vehicle and on the platform), categories of passengers exposed, and local cultural specificities for maximum density. Critical levels of density will also need to be explored, i.e., values beyond which certain movements within the flow become difficult, or even certain gestures, e.g., using a mobile phone, therefore affecting the possibility of accessing customized information.
- Pragmatically, a methodology of application to a transit network: What factors should be tackled as a priority, doubtless depending on the size of the flows concerned and the amplitude of the variations in time and comfort under the influence of the flow.

10.2.4.2 Local Passenger Flow Management

There are four kinds of mechanism for managing a passenger flow locally: spatial layout, use protocols, informational incentives, and pricing.

For a spatial element, layout sets the capacity dimensions, for both stock and flows. Within a space itself, the specific furnishings—notably including seats, benches, information panels, and automatic dispensers—influence the levels of comfort and capacity. Vertical and horizontal signage helps to direct and channel flows.

Use protocols determine certain operational processes, in particular, for access to a vehicle, the priority to alighting passengers onto boarding ones. Channeling flow direction by means of a dedicated lane is a combination of layout and protocol. Controlling vehicle doors in order to adjust opening time and limit dwell time is also a protocol, as is separating passengers between vehicles or between several cars on a single train, by pricing categories or gender.

Informational incentives can be used, to a certain extent, to manage the quantity and quality of a flow, depending on the compliance and specific interests of passengers. On a station platform, information on the waiting time before the next vehicle has the effect of increasing psychological comfort and therefore quality of service. Information on the arrival times and crowding levels of upcoming vehicles would make the option of waiting for a later, less crowded, vehicle more credible and attractive than taking the first, packed vehicle, and would help to manage flow quantities.

In rail transport, giving passengers information on the level of crowding in the different cars along the train would enable them to choose their waiting position on the platform and would doubtless have the effect of evening out the flow inside the train, improving the distribution of passenger movements between the doors and therefore reducing station waiting times.

Finally, pricing is a mechanism that is relatively little used. In air transport, business class passengers have access to dedicated areas in airports and special boarding conditions, as well as privileged conditions on board. In Tokyo's rail transportation, some "premium services" are available only to passengers who pay a relatively high price. At certain peak times, some stations on the London Underground are used only for connections and outside access to the station is prevented: Why not make people pay rather than excluding them?

All these practical forms of flow management have influences on how the system works and the quality of the service. They should therefore be integrated into the system simulation, using specific submodels that can be incorporated into an assignment model. Research is needed to model each practical arrangement, in terms of mechanism, factors, and effects, and to study the magnitude of certain effects at local level, as well as their impact on traffic at network scale. Here are a few examples:

- Using both sides of a train in stations, with one assigned to people leaving the train and the other to people entering.

- Fare discrimination in services.
- Evening out the flows of passengers waiting along a rail platform.

This latter option is particularly worth studying in relation to an assignment model, because a passenger's readiness to reposition herself undoubtedly depends on factors specific to her needs: not only comfort in the vehicle, but also position along the vehicle in relation to the destination station and also in relation to the corridor leading to the boarding platform. The structure of the flow on the platform in relation to the destination station is therefore likely to influence the level of compliance with these positioning recommendations. Walking from one end of a 200-m-long platform to the other takes two minutes for a person capable of walking at 6 km/h and four minutes for a slow walker who moves at 3 km/h: This is not insignificant in the economics of the journey. The passenger makes choices between detours on foot and onboard comfort; one information service about the journey may recommend certain platform positions, while another on local quality of service may suggest a different position. The information age offers a multiplicity of situations and decisions.

10.2.4.3 Vehicle Traffic at the Local Scale

The interactions between travelers in a pedestrian flow resemble those between vehicles in a highway flow. Vehicle traffic is also important in a transit system, with several fundamental interactions: with passengers, between vehicles on the same line, or with other vehicles in different directions. These interactions influence the journey time of the vehicles and therefore of the passengers they carry and indirectly the dwell time in the station. They need to be modeled to describe their effects and if possible to explain those effects by causal mechanisms and factors.

The research topics can be divided into four questions:

- The dwell time of a vehicle in a station. We have already pointed out that while multiple doors increase exchange capacity, they also make the process of passenger exit and entry more elaborate. Physical laws have been proposed to link the length of occupancy of a door with the number of passengers entering and exiting; the physical modeling needs to be extended, notably to take into account the layout of the vehicle, the layout of the platform, and the gap between the platform and the vehicle, and also to account for the distribution of passengers entering and exiting and of people remaining on board along the vehicle and those remaining on the platform. In addition, continuous trains without separation between cars allow passengers to move around inside the train, leading to a certain evening out of the internal load.
- At an infrastructure point, particularly the track on which a train stands in a station, the dwell time of the vehicle, the times for braking and reaccelerating on approach and departure, the safety margins, are successive uses that occupy the point and therefore consume its service capacity. The associated capacity constraint must be expressed in the assignment model. Its physical modeling

requires refinements and adaptations to local conditions. Reciprocally, among the different stages of track occupancy, the approach and departure maneuvers could be adapted to the state of the flows within the vehicle, with sharper braking and acceleration in absolute terms when all the passengers have a seat, which would speed up the interchange of vehicles.

- Headways and capacity. Between successive vehicles on a single route, headway (the time interval between vehicles) influences the number of passengers in the following vehicle and therefore its dwell time in the station and onboard comfort. This phenomenon needs to be investigated with a dynamic model. In addition, headway includes a safety margin between vehicles. On a road, it is easy to adapt the margin to the speed, reducing it at lower speeds. In rail transportation, this tactic is called "restricted speed": It is used to partially restore capacity after a disruption. It is worth modeling in order to assess its feasibility and effectiveness through simulation.
- Run time and delay. Models are needed for other local traffic phenomena relating to the interaction between a vehicle and exogenous factors. In particular, on an ordinary road lane, cars and other individual vehicles can hinder the movement of buses, not only when in motion but also when parked. To give an example, parking spaces situated just before a traffic light intersection can block it during a green phase: At a peak time when all the spaces are occupied, it is enough for a car in search of a parking space to stop behind a space that a car is vacating and in front of a bus; the successive maneuvers of the departing and arriving vehicle can occupy the entire green light phase. These events will further affect a following bus, especially one that stops to allow additional passengers to board… The resulting delays need to be quantified, at least the mean and if possible also the dispersion, to be integrated into the local run time of vehicles. A microscopic traffic model is appropriate for a fine-grained simulation of the phenomenon, in order to derive statistical summaries that can be used in a macroscopic assignment model. However, it would be better to incorporate fine-grained traffic simulation. In this vein, the application of an assignment model is a way of simulating the effects (benefits, costs) of local road layouts—location of the bus stop relative to the intersection and vocational parking spaces—relative to local traffic control.

10.2.5 Line Traffic, Management, and Economics

In a line-haul transit system, each line is a specific subsystem. This subsystem is a vertical composition consisting of a traffic and access infrastructure, a fleet of vehicles, and different technical and commercial protocols, in particular the service protocol, which includes vehicle movements.

Here, we consider an operating line consisting of one or more services in each direction of flow, each with its route made up of vehicle trajectories and stations

served. As a traffic system, a line with multiple services possesses horizontal in addition to vertical complexity. In particular, the topology of a rail transit line on a regional network can be complex, with one or more shared trunks and different branches.

In general, a large conurbation possesses both train lines on a regional network and simpler train lines, such as subway lines with a single service in each direction. The particular complexity of a line justifies a reinforced "infostructure," a centralized real-time management system. This dynamic line intelligence system constitutes an additional layer in the vertical composition of the system.

Traditionally, assignment models are used to simulate traffic on a new line on a network, or the effects of a substantial alteration to an existing line. They are particularly necessary with heavily cross-linked networks where passengers can put together more complex itineraries. However, up to now, there has been little modeling of the technical specificities in the operation of a mass transit line in the framework of traffic assignment. A historical reason for this is that in the early mass transit assignment models, all the lines were reduced to a graph of elements, so that the discrete nature of the vehicle runs is represented solely by the set of arcs of boarding from route decision points. Static or dynamic models that treat service frequency macroscopically do not really lend themselves to a fine-grained representation of traffic management. By contrast, run-based analysis makes it possible to model the operational process more closely.

The modeling of operational processes is an important research topic for traffic assignment models (Sect. 10.2.5.1). Reciprocally, the dynamic intelligence systems on existing lines would gain by incorporating a traffic assignment model, in order to relate traffic management to user needs (Sect. 10.2.5.2). Finally, the complexity of a mass transit system seems to have limited economic research, whereas there have been many contributions on marginal costs and pricing in road assignment models: Such research can be undertaken for mass transit on the basis of a single line, taking advantage of the relative simplicity of this subsystem (Sect. 10.2.5.3).

10.2.5.1 Line Traffic Performance and Its Factors

Let us concentrate here on line traffic in terms of vehicles and passengers. It is natural to model the topology of the line and its services by a particular network: Firstly, it is the topology of stations and routes that interest passengers, which justifies a specific graph model, i.e., let us say the passenger network; secondly, stations and routes are part of an infrastructure network, which also includes details of the sections used, the junctions crossed, whether the mode is road or rail (junction points): This justifies a specific graph model, called the vehicle network.

Each graph model can be used to model physical characteristics element by element: The most important are run times taken broadly, including station dwell times. These times depend on ambient traffic conditions and reciprocally determine them: indirectly via passenger route choices, therefore via passenger flows, and directly, because in each trip the downstream service depends on what happened

upstream, and because the different trips are managed by the operator in an integrated way.

This operational process needs to be modeled in the traffic assignment, distinguishing each traffic type and operating mode. Up to now, rail traffic has hardly been distinguished from road traffic in the assignment literature. However, it has specific characteristics (cf. TRB 2013; Hansen and Pächl 2014): The track is restricted to trains, whose planned speed can be revised downward or upward to offset upstream delays (e.g., during a stop at a station). This type of catching up is important for users: It can be applied to a significant degree if the stations are sufficiently far apart, i.e., much more for a regional train than for a tramline.

For a given traffic mode, there are several levels of centralized intelligence, as evidenced by fixed or mobile block section signaling systems in rail transport. Another factor is the degree of dynamic responsiveness, which depends on the communication system between the central post and the vehicles.

The different modes need to be modeled with an eye to their specificities. This will make the assignment models more realistic. It should be recalled that in a mega-city (e.g., greater Paris), a very busy train line can carry more traffic than the entire bus network in a mid-sized city (e.g., Besançon, France). Assignment models that are sensitive to operating modes could be used to assess the benefits of modernizing a line's operating system.

While certain operating modes can limit the influence of passenger flows on journey times, the concentration of passengers within the vehicle remains a factor of discomfort.

Centralized management systems are based on an information system and communication between the components of the service. A direct extension is to provide users with dynamic information. On many networks, this information is given for each line, following particular procedures: It therefore needs to be modeled at line scale, before shifting to network scale for higher-order functions (such as advice to shift to an alternative line).

10.2.5.2 User Needs and Traffic Management

The representation of traffic with its management, i.e., of the operational state of the system and quality of service factors, can be used to deduce the utility for each passenger, i.e., each passenger's level of satisfaction. The total utility for all passengers is an important criterion in assessing transportation projects and plans: This applies to the design of a line and could also be applied to real-time management of the traffic on the line, by replacing the current criteria, which are based on the status of service runs relative to the nominal schedule, neglecting the heterogeneity of loads between vehicles.

The use of an assignment model would also make it possible to add a predictive dimension to traffic management. This is because for each vehicle run, the model simulates passenger flows and their changes throughout the trajectory. At a standard point, the model "knows" at what station each passenger on the vehicle will get out.

The decision variables to optimize standard operation are notably:

- The interval between successive runs.
- The ordering between runs on different services which arrive at a confluence point and are in competition to access the shared section.
- Where applicable, for a given vehicle, a variation in the service scheme relative to the planned scheme.

Incorporating an assignment model into centralized traffic management on the line should therefore improve the service the line gives to its users. This is equally true for ordinary operating conditions and for disrupted services.

In order to simulate the system's response to a disruption, the adjustments made by all the actors concerned need to be taken into account, including users with their perception and their decision-making capacity (see Sect. 10.2.1) and of course the operator with its resource management systems.

Probabilistic incident modeling is a topic of research in this respect: Before simulating a particular incident, this involves describing the theoretical structure of the causes of disruption, using a combined physical and probabilistic model. Such a model will be sensitive to the state of wear of the system's components, because this determines their failure rate. It will make it possible to evaluate maintenance policies for the infrastructure and the vehicles, in particular preventative strategies to renew components before they fail.

10.2.5.3 Line Economics: Value Analysis and Yield Management

In transportation economics, there is a well-known effect called the Mohring effect, according to which, for a mass transit service, an increase in frequency improves quality of service (by reducing the wait before boarding) and therefore attracts more customers, which can justify a further increase in frequency and so on. This virtuous cycle is also productive in terms of commercial revenues.

In practice, however, urban mass transit systems are regulated by the authorities primarily in respect of the social utility of the services. Fares do not perfectly and completely reflect production costs, and only a fraction (often small) is covered by commercial revenues.

Improving the rate of cost covering, as well as the match between the service supply and use on the demand side, is an important economic priority for the mass transit system. Since the assignment model simulates the interaction between supply and demand, it could contribute further to economic analyses.

In this respect, it is natural to begin by treating a line as a subsystem, before subsequently tackling the greater complexity of the network. Here are a few research topics relating to this:

- Modeling production value: assigning a shadow price to each journey delivered depending on the route and the period, linked with the general costs for the passenger.

- Analyzing production costs: the cost of using the resources, of the associated wear and tear, and of consumption, particularly energy consumption. The evaluation needs to be done for each run and related to the number of passengers carried, in order to obtain unit costs.
- Looking for a level of frequency that will optimize production value net of costs, in other words the financial balance for the operator. Comparing this financial optimum with the socioeconomic optimum (i.e., the level of frequency that optimizes the service's socioeconomic balance). Analyzing the sensitivity of the results to the length of the service cycle and conjointly optimizing frequency and speed of travel (which influences cycle time but also energy consumption).
- The optimum level of service undoubtedly varies with demand conditions. This should be explored by simulation, with attention to the distribution of fixed costs between the different periods.
- Evaluating the marginal congestion cost of an individual journey: This can be calculated analytically if the model is macroscopic. Compare with the fare level, for different time periods.

These different questions have received much attention in individual transportation, but very little in mass transit, probably not only because of the complexity of the traffic model, but also because of the still somewhat uncompetitive nature of mass transit in many conurbations.

10.2.6 Line-Haul Network

Transit lines tend to become organized into networks in order to cover both dimensions of geographical space. Passengers expect these lines to be interoperable in terms of travel, information, and fares.

For the operator, interoperability applies at two levels:

- At the higher level, the interface with customers is the tip of the iceberg;
- At the lower level, the submerged part of the iceberg includes the functions of basic interconnection between lines, by means of specific technical components, stations for passenger traffic and infrastructure connections for vehicles, and traffic operations through cooperation between lines, here again distinguishing between the passenger perspective and the vehicle perspective.

As we have already covered the micro-local level (in Sect. 10.2.3) and lines as subsystems (in Sect. 10.2.4), we concentrate here on the aspects of a network that contribute specifically to cooperation between lines. We begin by tackling the specific technical components constituted by stations (Sect. 10.2.6.1) and infrastructure connections (Sect. 10.2.6.2), and then, we look at the operational processes that make the lines cooperate (Sect. 10.2.6.3). And then, we come to the informational aspects (Sect. 10.2.6.4) and economic aspects (Sect. 10.2.6.5). As before, our goals are to identify the importance of each item in a real system, to indicate the need for an assignment model to be sensitive to it, and to outline associated research topics.

10.2.6.1 Station Layout and Management

The purpose of a station is to enable passengers to access a service from outside or via a connection from another service, or to leave into the outside world. Here, we concentrate on its traffic functions, ignoring the accessory functions of obtaining information and buying a ticket and ticket inspection. Travelers go through different traffic or waiting areas, as described in the section on the micro-local scale: Each micro-local element imposes a capacity limit on the pedestrian flow. If this constraint has no slack left, it can be critical for the functioning of the station and the lines that serve it. This constitutes a first reason to model the internal layout of a station, i.e., the configuration of the micro-local areas. This includes the "hard" arrangement of spaces by necessary demarcations (partitions, barriers) or the "soft" arrangement by signage and direction systems.

The second reason for explicitly representing the layout of a station relates to passenger access and connection times. The following factors come into play:

- Travelers move around the station, each following a pedestrian route.
- The conditions of movement depend on the passenger flows, because of mutual hindrance, whether physical or visual (masking effect with respect to the information boards). Crowd dynamics are an active research topic, of particular interest for big stations, which concentrate very large flows in restricted spaces.
- Signage resources: The conditions of movement influence route choice and therefore the configuration of the station's internal flows.

All these characteristics are part of a pedestrian assignment model for the station, which we will call the station assignment model (SAM), a complement to the transit assignment model (TAM). The development of such SAMs constitutes an important research goal. Specific topics are as follows:

- Passenger movement in free flow, as affected by the hard and soft arrangement of the movement zones.
- The influence of the crowd on movement.
- On roads, adjacent stops between bus lines are used by passengers for connections. These should be modeled as an "unofficial" station, including the different possible routes between stops and the variability in the travel time between them (especially if this entails crossing streets).
- For any train station, if there are several access points to the platform, there will be a stochastic component in passenger itineraries, to be reflected in a probabilistic model.
- For train stations that combine urban and interurban lines, the allocation of platforms to interurban services can be variable, which means that the access point for the travelers concerned is stochastic.

A SAM is intended to be integrated into the TAM. The mutual relations between a SAM and a TAM are connected to the matrix structure of the origin–destination relations between the station access points, i.e., the platform accesses and the outside accesses:

- The TAM imposes on the SAM an OD matrix of flows between the points, for each demand segment, upon request for each destination on the network and with a travel cost at each access point from that point to the destination.
- The SAM provides the TAM with an OD matrix of the foot travel conditions, in terms of time taken (both mean and dispersion), possibly together with flow distributions from the entry points to the exit points, for each demand segment and destination.

The precise connection between a SAM and a TAM needs to be specified in each model integrated.

A related research topic concerns the actual location of route choice points in the station for the itinerary across the whole network. A traditional postulate is that several competing services are present in a single place, a decision point, where the traveler is considered to find out about the arrival of each vehicle. This postulate works for a boarding zone shared by several bus lines, but less so when the distances between the line stops are greater. For a rail mode, the postulate works for services that use the same platform, or that operate on either side of a shared platform.

The representation of the decision points is critically dependent on the fineness of the topological representation of the network: Modeling an intermodal station with a single node would imply that dynamic information and the possibility of instant access to each platform are available anywhere in the station. In practice, in a complex station, conjoint dynamic information on several services is only available to passengers at a few specific points, where the operator has installed information boards. The spread of smartphones and the development of dynamic information systems make information omnipresent for users, but they cannot move instantaneously to a given platform, and the travel advice service should estimate an access time that is comfortable for the user, allowing for his normal speed and the flow conditions.

Complementary research topics are as follows:

- The modeling of marginal congestion costs for journeys inside a station.
- The design of layout schemes, or information schemes, or even fare schemes, and the assessment of their impact on passenger traffic through simulation.

10.2.6.2 The (Infrastructural) Connections that Affect Vehicles

Chapter 8, Sect. 8.2, pointed out the challenges of synchronizing services that serve a single section, together with the technical operation principles and their modeling in the framework of traffic assignment. The associated techno-economic problem remains a research topic:

- The evaluation of spatial arrangement schemes and their impact on demand through simulation. Greater quality of service requires more optimal—and therefore undoubtedly more expensive—arrangements: What is the socio-economic cost–benefit outcome?

- Including the constraints of coordination between service runs among the operating constraints on each service in a line model.

Another form of connection between lines that concerns vehicles is another synchronization, but this time in order to avoid clashes and to organize the alternation of runs. The aim is to enable vehicles to cross paths safely. One solution is to work on a potential crossover site, by setting lines at different levels. However, the challenges go beyond crossovers alone. They include sharing the infrastructure by having several services traveling on a shared section, in order to make that section profitable. In this case, if the levels at the crossover point are different, access ramps need to be added and the access conditions organized: The "hard" technical solution generates a soft problem of traffic protocols. The "soft" solution for a crossover site is to coordinate the opposing traffic on a principle of alternation, which allows direct crossings and also turning movements.

For the rail mode, the crossover conditions need to be represented in the assignment model. The dynamic status of the junction points combines with block section signals on each section. Each junction point is a potentially critical resource, whose influence on each line must be modeled:

- Directly in a dynamic assignment model, in concert with the line model.
- In a static model, a hindrance function needs to be modeled for each flow current, in relation to all the other flows, similarly to the distribution of capacity between branches on a road intersection.

For road-based transit modes, crossover situations and the requirements of turning movement take place on the roadway, in interaction with other vehicle flows and pedestrian flows. The research topics concern:

- On each section, modeling the different "forms of coaxiality" between the different mobility modes—dedicated or ordinary lanes—and the consequences of the layout for the respective journey time of the modes and for the mutual hindrances between flow currents. Microscopic road traffic simulations are particularly suitable: They can be used to establish macroscopic laws for typical sections.
- For the intersections, modeling not only simple forms but also complex forms that connect several local subsystems—sequences of intersections as specific corridors. In a complex form of this kind, the local conditions are linked with particular correlations, which demands specific modeling, the vehicle analogue of a passenger station model.

10.2.6.3 Operating Protocols

We have just referred to the functional coordination between lines, in the organization of connections, and also local forms of technical coordination, through the sharing of a traffic infrastructure and coordination protocols.

Above and beyond local cooperation between lines, network operation is a way for lines to give each other mutual support in response to variations in the state of the system, following a rush of passengers (overflow situation) or a technical failure that affects a particular line, reduces its capacity, and causes transfers that generate rushes on alternative lines. The demand conditions on a line depend not only on the level of demand for the whole network, but also on the current situation on the other lines. Freeway operators now provide information for motorists a long way in advance of the decision points, directing them toward more reliable routes in order to avoid section closures or simply congestion delays in dense traffic.

This adaptive strategy needs to be transposed to line-haul transit and even enhanced in order to exploit the flexibility of such systems: Adjustments to service frequency allow rapid real-time capacity adaptations, whereas the only comparably strong mechanism available to highway operators is the dynamic allocation of road lanes to the directions of flow.

This establishes a few topics of research in traffic assignment:

- making the assignment sensitive to dynamic operating strategies;
- using simulation to explore dynamic operating strategies;
- integrating the assignment model into the centralized traffic management system at network scale, so that it can contribute fully to operational flexibility and responsiveness.

10.2.6.4 Network Information

By design, every mass transit network has static intelligence, for the composition and smooth operation of the service. Thanks to data sensors distributed around the system, to remote communication and centralized information processing, the operator possesses centralized dynamic information that it can use to manage its resources in real time, keep users informed, and encourage them to cooperate.

We have already mentioned the contribution of the centralized information and control system to the operation of a line and the advantage of representing its impact on an assignment model. In this kind of representation, the model must follow the specific dynamics of the chain of information and logistical operations, with their delays, inertia, and errors. In particular, an unexpected and sudden variation in instructions presents a human agent with the difficult task of transition between different operating conditions, requiring a change of routine and mental reprogramming. Such human factors are an integral part of the practical conditions of operation; they also affect users, each of whom is involved in the operation in an elementary capacity by their individual presence. In order to persuade them to cooperate effectively in sudden changes of operating mode, through appropriate adjustments in their route choice or departure times, the information must be presented to them simply and efficiently, as happens with dynamic freeway information signs. In addition, passengers can only be asked to make successive

adjustments if the overall sequential logic is sufficiently clear and does not cause excessive disruption to their own journey.

Since the assignment model specifies the service plan and intended routes of passengers, it is particularly well suited to act as a basis for the development of operational variants:

- Off-line, in the design of Traffic Management Schemes in response to particular circumstances (critical demand rush, major weather disruption, accident, etc.).
- In real time, through integration into an operational support system that monitors changes in the network and dynamically develops (or automatically triggers) very short-term action plans.

Such applications undoubtedly have great operational potential, but they are highly complex. Their development requires specific research targeting the following questions:

- Readjusting the simulation almost instantaneously from real-time data, so that the model comes as close as possible to reproducing the actual conditions of the system.
- Developing real-time operational variants that minimize user disruption.
- Formulating criteria for assessing user disruption: An overall indicator is easy to design, but it needs to be combined with indicators of the user's understanding and acceptance of the operational policy (compliance), or of fairness between users in terms of the respective efforts they are asked to make and their monetary contribution.
- Other criteria need to be developed to assess the stress imposed on operational resources, in order to avoid policies that focus solely on immediate effects but would subsequently weaken the system. Since this is a complex system with a very large human component, the operational variants must be robust.

10.2.6.5 Network Economics

In the previous subsections, we tackled several problems of an economic nature:

- The economics of demand, with regard to passenger preferences, their generalized costs, and fares.
- The economics of production at the scale of a line.
- The optimization criteria for traffic management, and the role of user information, employed as a management instrument.

These problems are all the more acute at the scale of a network, which is the primary scale of intervention for the authority that regulates the system and monitors the operators. The assignment model that specifies the operation of the system by incorporating the interests of users and the resources of operators is an excellent medium for informing regulatory decisions, as well as the management decisions specific to an operator, or planning decisions.

At network scale, the economics of production pertain to service lines, traffic infrastructure, and stations. Each station is a subsystem that requires resources and costs to operate, whereas usage produces traffic and therefore shadow revenue. In relation to this, here are economics research topics that can be based on an assignment model:

- The economic analysis of a station, in terms of the functions provided and the components of the supply (including the micro-local spaces) and in terms of their use on the demand side.
- The evaluation of the marginal congestion costs for the movement of pedestrians in the station.
- The allocation of (shadow) commercial revenues to the components of the travel process: runs on lines and movements through stations, taking into account their respective functions—access or transfer for a station, crossing of space for a line—but also access to a crossing line via a feeder line.
- The design of "decomposition and coordination" schemes for the techno-economic management of the parts of the system: line, station, and infrastructure.

With regard to the regulation of the system, an important principle is that the community should be involved through an organizing authority at conurbation scale, in order to tackle global issues and structural links. The relations with the operators are of two types: firstly awarding a service contract and secondly regulating its implementation. A familiarity with the supply–demand system based on an assignment model can be used for research on the following topics:

- Formulating criteria for the remuneration of operators that better represent the services delivered to users. The existing criteria are based on ticketing statistics and operational performance indicators relating to vehicle traffic. The challenge is to establish operational performance indicators that are sensitive to users in terms of their actual exposure to operational conditions.
- Determining the functional conditions (in terms of service quality provided to users) and the technical conditions of coordination schemes for a component-based operation of the system and identifying economies of scale and other externalities that justify unified management of a particular subset of components. This includes the distribution of the lines between one or more operators, but also vertical coordination between infrastructure management and service management and—why not?—the allocation of stations to specific operators, particularly in the case of big intermodal stations (e.g., Madrid). These broad industrial economic topics need to be revisited, with greater attention to the reality of demand as represented in an assignment model.
- Developing fare schemes that do more to follow broad User Pays principles and to cover production costs and Polluter Pays principles with regard to congestion costs.

With regard to fares, urban mass transit has long been limited by old methods of payment: station-based ticketing, with transaction times that are not insignificant relative to total journey time; moderate fare levels that did not cover the substantial costs of collection and payment. Automated ticket purchase and validation are major advances, as are commercial subscription packages, despite their often simplistic nature. Innovative information, payment, user identification, and customized relationship—for technologies—open the way to much more flexible fare schemes. An assignment model can be used to explore the potential demand and commercial revenues for a given fare scheme. Here are some avenues for exploratory research:

- Linking production costs to traffic disaggregated with respect to journeys and modeling an average journey cost for the operator.
- Modeling a marginal congestion cost per journey.
- Comparing the average production cost per journey, on the one hand, and the marginal congestion cost on the other and then the price level in terms of the commercial package with or without subscription, in order to describe the pricing system in terms of User Pays and Polluter Pays principles.
- Assessing the social equity of pricing schemes, between users with and without subscriptions, and also between social classes of users (in particular in terms of income and place of residence), by means of an assignment model that gives a fine-grained picture of demand segments and their respective consumption.
- Designing dynamic fare schemes, linked with average production costs and marginal congestion costs per period.
- Designing incentive fare schemes which reward efforts made by users—e.g., in following a recommended route or timetable—with credits on their next bill.

10.2.7 Pooled Transit Services (PTS)

Here, we define a PTS as a mechanized mode of travel in which vehicles are shared between passengers either simultaneously or successively. This definition covers several modal forms: individual taxi, collective taxi, carpooling, car sharing schemes, and shared vehicle services. The particular features of these options were described in Sect. 10.1.

Each modal form of PTS is a supply subsystem, whose existence and practical performance in an area depend fundamentally on the interaction with demand. The numbers and itineraries of the desired journeys, in other words the clustering of demand, relative to the number and productivity of vehicles on the supply side, determine the availability and speed of service. These performances in turn have a feedback effect on both demand and supply, in terms of levels of service and commercial conditions.

This interaction between supply and demand can be studied using a dedicated PTS assignment model. Here, we present the research requirements to model a PTS,

first in relation to demand (Sect. 10.2.7.1), then in relation to supply (Sect. 10.2.7.2), next its use in quantity and quality with the feedback effects on supply and demand (Sect. 10.2.7.3), and finally regulation (Sect. 10.2.7.4).

10.2.7.1 Representing Demand for a PTS

For each modal form of PTS, the demand model must represent (a) the quality of service characteristics, (b) a passenger's sensitivity to quality of service and price, (c) the nature of a travel option, (d) a service user's individual decision-making process, (e) the demand segments, and (f) for each demand segment, the OD flow matrix for every standard period.

Quality of service is described by the following characteristics:

- The length of time spent by the passenger on information and interaction before each journey, i.e., the time cost of the transaction;
- The conditions of access in time and space: the time required by the passenger to access the rendezvous point, vehicle waiting time, and total time to destination;
- The duration and comfort of the trip in the vehicle.

These different characteristics need to be specified for each modal form by a generic description, in other words by a descriptive theory.

An individual traveler's sensitivity to quality of service and price factors is modeled by means of a utility function, which combines those factors with the passenger's characteristics. Here again, a specific theory is required for each modal form of PTS.

The nature of a travel option needs to be modeled in terms of final access and the primary journey between the entry and exit points for the mode: What range of possibilities is open to the passenger for choosing the entry and exit point? Is the itinerary chosen by the traveler or determined by the operator?

The decision-making process depends on the nature of the options. The following features need to be modeled:

- The composition of a range of travel options;
- The user's knowledge of the potential options, notably in relation to the sales information system specific to the modal form;
- The choice of a particular option, depending on the dynamic status conditions of the service—e.g., the presentation of customized attractive options by an intermediator such as Blablacar or Uber.

The submodels so far mentioned together constitute the demand model for an individual customer. In addition, the segmentation of demand between different types of customer needs to be determined and an OD matrix to be modeled for each segment by period type (type of day, depending on the time of day, in particular distinguishing peak periods).

10.2.7.2 Representing Supply in a PTS

For each modal form of PTS, the supply model must represent (a) the infrastructure conditions, (b) the vehicle, (c) the usage protocols, and (d) the techno-economic model of the service.

The infrastructure conditions relate to (i) the road network available to the service, including any semi-dedicated lanes and (ii) the area served for entry and exit, whether or not it is focused around physical stations.

The service is based around a vehicle type: Its passenger capacity must be specified, along with the driving protocol, i.e., whether driven by the passenger or by the operator, manually or automatically.

The usage protocols relate to the physical use of the service, as well as the informational aspects: media and interfaces for giving information to potential users, the degree of customization, and ergonomics of those media, as well as:

- their impact on transaction time for a particular use;
- subscription packages in terms of price levels and service levels (e.g., separate premium packages).

The techno-economic supply model consists of the following components:

- The journey time function and the access time function for a boarding customer, in terms of the infrastructure conditions, the infostructure conditions, and the vehicle pooling mode;
- The transit operator's production function, whether integrated across a whole fleet for a company such as a taxi firm, or individually for a sole trader under a micro-franchise;
- The construction of the price in terms of demand level and production conditions;
- The intermediator's production function, in terms of the technical and commercial functions it performs between customers and providers: identification, information, reservation, payment, and price setting.

10.2.7.3 Modeling the Interaction Between Supply and Demand

The technical and economic conditions of the relationship between supply and demand have been anticipated in their respective models: Their characteristics are described by quantitative variables, the values of which are endogenous to the combined supply and demand model.

These are the actual conditions of the usage of supply by demand:

- Journey time, access time, and price: with respect to service type (premium or other), OD pair, and period;
- With their feedback on demand, which influences the volume of traffic;

- With their feedback on supply, which influences fare levels and service availability, in particular through the rate of use of each vehicle and the size of the fleet.

Here, the modeling requirement is to relate the demand model to the supply model: the connection between them and in particular:

- The total number of individual trips across all the demand segments;
- The total commercial revenues across all the service periods;
- The overall coordination of supply: depending on the type of company, either through integration or through concentration around an intermediator.

In addition to technical and economic modeling, mathematical, algorithmic, and IT processing needs to be developed in order to solve the model numerically.

10.2.7.4 Regulation

Mass transit modes are often subject to strong regulation: restricted access to the profession, vehicle quota systems, price controls, specific taxation, as well as the allocation of public spaces and access rights to certain dedicated transit routes.

In a particular conurbation, knowledge of the specific arrangements can facilitate the modeling of supply. However, for local authorities, the main priority is to maximize the population's well-being: An assignment model can and should help to integrate PTS into transit planning and to design regulatory arrangements not for restrictive purposes but for purposes of optimization and openness, taking advantage of the specific potentials of these services.

10.2.8 Multimodal Transit System

The different public transportation modes, whether mass transit or demand-responsive, need to cooperate for the benefit of travelers (see Sect. 10.1). That is why their integrated modeling is an important topic of research in traffic assignment (Sect. 10.2.8.1), together with the economic analysis and regulation of the system (Sect. 10.2.8.2), and the integrated design of such a system (Sect. 10.2.8.3).

10.2.8.1 Integrated Supply–Demand Modeling

Traffic assignment models on a mass transit network are necessarily multimodal, since they include walking and road or rail transit lines. The addition of demand-responsive modes involves a hard and a soft aspect, respectively, the physical conditions (hard) and the commercial conditions (soft).

With regard to the physical conditions, there are two research themes:

- Modeling intermodal transitions at a given point, in physical terms that include in particular the question of the availability of the second mode, and parking constraints for shared vehicle or car sharing services. Another factor to model and explore is the inconvenience of shifting between a demand-responsive mode and a mass transit mode, in terms of the traveler's preferences, as well as his or her propensity to follow a number of trip sequences in a certain subset of modes within a single urban trip.
- Modeling the local choice between a range of demand-responsive or mass transit modes. At what point(s) is the passenger in a position to make a decision, using what dynamic information, and what are the physical conditions for accessing each of the available modes?

The commercial conditions of the multimodal system encompass information, pricing and payment, and the associated subscription packages:

- Since multimodal systems are highly complex, static and dynamic information will be very useful in helping the passenger identify potentially convenient route options. So the information services available to passengers will need to be modeled, together with their particular conditions of use in a multimodal system with varying degrees of integration.
- To facilitate transitions between modes, the passenger needs not only local guidance, but also the minimum possible number of transactions to buy and validate a ticket. The more complex the network and the larger the number of modes that can be encompassed in passenger journeys, the more acute is this requirement. An integrated assignment model must explicitly include transactions and their effect on journey cost. In addition, the model should represent the fare conditions specific to each mode and the diversity of price levels between demand-responsive and mass modes: In principle, the more individualized the service, the higher the price. The price sensitivity of passengers, and their segmentation with respect to their respective sensitivity, is an important component in an integrated assignment model. However, this component has long been treated as a minor factor in line-haul transit assignment models.
- Subscription options for all or part of the available transit modes in a multimodal system are a related research topic. Subscription packages are more varied for shared vehicle modes than for mass transit, with a varying mix between a fixed component that provides access and is linked to the length of subscription and a variable component that is based on effective use. Combined subscriptions are therefore likely to be even more commercially optimized: their effects on demand and use need to be modeled. Furthermore, the ability to use a single physical medium (smart card, smartphone) for several modes is an important factor in integrating and reducing transactions, one that needs to be represented in a combined model.

10.2.8.2 Economic Analysis and Regulation

An integrated assignment model should represent the diversity of prices, price sensitivity on the demand side, and the effects of price on commercial revenues. It should also represent the economics of supply and the profitability of each mode or each operator, stating the costs of production and deducting them from commercial revenues in order to establish net profit.

Different impacts, and their economic valuation, also constitute modeling issues:

- The latent demand generated by multimodal supply: This needs to be broken down into segments and time frames;
- The specific contribution of each service to territorial access: For example, carpooling may have a significant effect on non-car owning individuals with poor mass transit provision and therefore contribute to social inclusion;
- The surface area occupied by the different modes in public space allocated to traffic and parking, notably on the streets: firstly, in physical terms of space occupied; then, interference between the modes and their flows, as well as with other traffic streams such as pedestrians, cars, and bicycles; and finally, the economic costs of this interference for those experiencing it;
- Environmental impacts: noise and atmospheric pollution, energy consumption, relative to vehicle routes, and the number of local people exposed to negative environmental impacts.

The evaluation of these different factors can be used to establish the overall value of the multimodal system for the community, in order to optimize the technical and commercial configuration and the operating modes for the community. Because of the different externalities, an ideal system optimum state necessarily differs from a user equilibrium, or rather an actor equilibrium when supply is endogenous.

The concept of a system optimum should play a fundamental role in regulation and planning.

An additional topic for economic research based on a multimodal assignment model is the distribution of added value (e.g., commercial revenues, social benefits minus costs) between the system's components:

- Between the different modes, particularly in terms of their functional combinations in journeys made by their joint users and further in terms of fare packages and their scope in space and time;
- Between mobility services, intermodal stations, and also the traffic infrastructure.

10.2.8.3 Multimodal Design

If sufficiently sensitive, assignment models can be used for exploratory applications intended for the design of transit systems that are efficient in all respects, in terms of demand, in terms of the environment, and in terms of the economic viability of the service.

Certain design topics relate to the technical aspects of supply (hard factors):

- Designing intermodal platforms that are optimized for demand, which minimize transfer and waiting times while offering associated services (information, payment, shops);
- Determining the specific added value of a certain level of development and facilities on a particular modal platform, for the overall operation of the services that use it;
- Drawing up "planning and coaxial operation schemes" for several transit modes along an urban artery: distribution of access rights to traffic and parking lanes and positioning of respective stopping points, all in relation to junctions and their specific distribution along the artery and also in relation to incoming and outgoing flows and to the requirements for connections between modes.

Other design topics relate to supply management, particularly on the commercial side (soft factors):

- Designing operating modes that are sensitive to the multimodal system's load conditions: for example, at peak times, directing demand-responsive modes in dense areas toward a premium service, but democratizing their use outside peak times and particularly at night, during periods when line-haul mass transit systems are closed;
- Dynamically allocating traffic lanes to different transportation modes;
- Designing multimodal Traffic Management Schemes, notably in order to compensate for the failure of a mass transit mode by an exceptional shift to demand-responsive modes: emergency use of carpooling, optimized use of self-service bicycle repositioning trucks, etc.;
- Ensuring that the assignment model contributes to multimodal information services;
- Designing combined fare packages for multimodal and intermodal uses.

These design topics are particularly innovative. A model could be used to explore their potential. It would also be desirable to test and optimize innovative approaches interactively with potential providers and users. An Urban Mobility Living Lab (Sect. 10.3.4) would be a particularly good framework for experiments of this kind.

10.3 System Simulation and Augmented Reality

Fabien Leurent and Rosaldo Rossetti

In this final section, we explore the relation between mobility systems and simulation models. In Chap. 8, as in Sect. 9.2, it was observed that assignment models are becoming more and more modular, with subsystems being identified and separately represented by a submodel.

In other words, modeling is becoming increasingly systemic, through assemblages of elementary models. However, elementary models themselves are continuously developing, forming an ever larger asset base as they accumulate. We will run through this asset base—a toolbox of models—and discuss the feasibility of a scale 1 assignment model (Sect. 10.3.1).

At present, models of this kind are still on the drawing board, although it is now conceivable that the location of all mobile entities could be determined on the move in real time. More concrete developments are already underway in two areas. Firstly, using IT techniques, elementary models are helping to bring augmented reality to the transit system for stakeholders in all categories: users, operators, regulators, and general public (Sect. 10.3.2). Secondly, the applications of assignment models are changing: Real-time applications are emerging, whereas new off-line applications are being developed (Sect. 10.3.3).

Finally, both model and application developments need to be related to the dynamic of transformation in the urban mobility system, as described in Sect. 10.1, and to the scientific dynamic described in Sect. 10.2. An urban mobility Living Lab should be an ideal framework for relating an actual system—operating in a territory and covered by a whole range of observations—to its operators and regulator, to industrial partners who design hard or soft innovations for such a system, and to academic partners who develop knowledge with potential operational outcomes, as well as users of the system (Sect. 10.3.4).

10.3.1 The Modeling Toolbox

An assignment model for traffic on a network reflects a transport system in a way that is simplified, but also broad, since it includes the main entities—vehicles, passengers, and infrastructure components—together with their interactions in traffic as a physical process. Yet there are now specific models for every type of entity, with potentially extensive detail that allows for very fine-grained simulation. We run briefly through these detailed models of entities (Sect. 10.3.1.1) before discussing the simulation of their interactions (Sect. 10.3.1.2). In addition, the diversification and spread of sensors now make it possible to observe entities and subsystems in detail: The collation of information relating to a single entity from different sensors also constitutes a computer model of that entity. There are natural convergences between a simulation-oriented model and an observation-based model. In addition, sensors and their observations can be simulated, and in other words, observations can be synthesized: The field of simulation and observation is wide open (Sect. 10.3.1.3).

In these conditions, simulation becomes an art of composition. A scale 1 empirical model of the transit system, giving real-time estimates of the status of disaggregated entities, now seems possible (Sect. 10.3.1.4).

10.3.1.1 A Review of Model Entities

A mass transit vehicle is a mechanical system that includes a cabin, a driving system, an engine, and wheels. It can be modeled with different levels of detail:

- In car production, each mechanical part is modeled as a subsystem with its own physical and technical behavior, by reference to its interactions with other parts in the vehicle's operation. In particular, engine operation is modeled in detail, in order to optimize energy efficiency;
- It is also possible to model the driving seat inside the vehicle, to simulate the position and working conditions of the driver, who interacts with traffic and controls the movement of the vehicle and interaction with passengers;
- There are simple mechanical models for the vehicle's motion dynamics and its energy balance: The vehicle is described as a material body with certain dimensions;
- And there are geometrical models for cabin layout and passenger comfort.

Each of these models is a considerable expansion on the basic representation of the vehicle included in an assignment model. A much more detailed simulation could be substituted, although it would be costly in IT resources.

The same is true for the modeling of a passenger as an entity. What exists now are as follows:

- Fine-grained biomechanical models, used to study road traffic accidents;
- Dynamic models of the passenger as a material body in a spatial configuration, in interaction with the limits of that space, street furniture, and other travelers;
- Psychological models of mental load, elementary physical actions, and the cognitive processes that determine them;
- Agent models to represent the beliefs, desires, and decisions of the traveler in motion.

Agent models are beginning to spread to assignment models, giving a more accurate representation of the travel situation in terms of all the needs and mobility conditions of the passenger as an individual.

Numerous models have also been developed for infrastructure "elements":

- Wear and tear models for each material component with respect to utilization processes, as input into asset management and predictive maintenance;
- Mechanical performance models for each track section, for vehicle movement;
- Still for each track section, technical traffic process models for controlling flows;
- For each element of pedestrian movement, geometrical layout models to simulate the conditions of use and the dynamic interactions between travelers.

The latter two model types tackle factors and processes that are important for traffic assignment. They can be used jointly with an assignment model, in order to specify the operation of the system for certain elements that seem sensitive—especially if the simulated flows coming out from the assignment model reach or surpass macroscopic physical capacities.

10.3.1.2 The Modeling of Situations, Behaviors, and Interactions

In Sect. 10.2.1, we took a detailed snapshot of the passenger situation and the passenger's behaviors in terms of elementary acts or decisions, in interaction with traffic conditions (local or remote) with regard to available information and individual perception.

The associated models capture physical and informational aspects of the situation. The psychological and behavioral aspects are much less deterministic and vary much more between individuals or even, for a single individual, between repeated situations. So, for example, the modeling of a passenger's secondary activities in a vehicle is necessarily stochastic.

Similarly, in Sect. 8.3, we wrote about the stochastic aspects of the passenger and vehicle movements. Computers can now tackle very detailed simulations of very specific and highly disaggregated interactions.

As far as possible, the structural relations between the elements need to be captured, in particular the topology of the infrastructure elements, of the vehicle runs and of passenger journeys. We also need to capture the two-way physical conditions between infrastructure and mobile entities, between vehicles and travelers, as well as the informational conditions.

Stochastic factors limit the ability of a particular simulation to mirror the operation of the physical system: There is no direct correspondence, and we can only interpret the specific enactment relative to a set of simulations that sample the stochastic process through which the system was modeled.

10.3.1.3 Collecting and Synthesizing Information

More and more sensors are now incorporated into the elements of a transit system: on the infrastructure, in the vehicles, and on travelers themselves through the smartphone. These sensors gather information that can be collected and centralized. Information about a particular entity can be used to backtrack its activity, and in fact, for an entity that is being observed by a variety of sensors, the collection of this information constitutes an empirical computer model of that entity.

A lot of information is produced in this way—we call it Big Data—especially as information generates information:

- Travelers react to a situation they have learned about, by themselves producing information for the operator or for other travelers with whom they interconnect in different ways (social networks);
- The messages sent by a telecommunication network can be analyzed by dedicated software, which will in turn generate further information.

The information can be processed in a multiplicity of different ways, potentially resulting in an infinite proliferation of information that reflects reality to varying

degrees. In addition to information that draws on real observations, there is synthetic (in the sense of virtual) information.

In particular, the results produced by a model constitute synthetic information. A traffic assignment model generates such synthetic information in massive quantities. It is also possible to model sensors, to simulate situations that the sensor would perceive, and to synthesize information comparable to what a real sensor would capture.

10.3.1.4 Virtual Reorganization of the Transit System

The quantity of information and the power of the resources for generating and processing it widen the possibilities for simulation. For an urban mass transit network, it is technically possible to have a network of sensors on the infrastructure, a set of sensors in the vehicles, and, conceivably, a sensor for each traveler. The total information on the real-time status of each entity, in particular the location of mobile entities, would constitute an empirical computer model of the system, with the same level of detail as an assignment micro-simulation of vehicles and travelers. An instantaneous disaggregated status model of this kind would, in itself, only have very limited predictive capacity, achieved by extrapolating the motion of each mobile entity on the basis of its previous positions.

The transit service schedule is a much more predictive model: The two models can be combined and joined with an assignment model that will include predictions of passenger movements and adjust the prediction of vehicle movements and infrastructure-related operations.

By adding other sensors, in particular for vehicle operation, or equivalent models to simulate vehicle energetics, the system could be represented in depth. As regards passengers, an empirical model of all the entities would ideally also represent their respective situations: In practice, however, not all real-time positions will be captured, and the state of the system in this respect can only be assessed by simulation, except in specially prepared and instrumented experiments.

10.3.2 Augmenting Reality

In the era of information, therefore, progress in observation and simulation has made it possible to substantially increase our knowledge of a real transit system. This enhanced knowledge can be of practical benefit to the different stakeholders in the system: (1) users, (2) operators, (3) regulators, (4) analysts, and (5) the general public. This is about augmenting reality, increasing its informational component and converting its physical component.

10.3.2.1 Augmented Reality for Users

For users, the information, pricing, and payment functions are simplified physically, through automation, while their informational aspect is amplified. Dynamic information, route advice, the delivery of different alerts from the operator, but also from other providers and users, augment the reality of the service delivered. Reciprocally, users themselves become information providers, even disseminating messages themselves.

In addition, physical reality is augmented by the emergence of new mobility services (see Sect. 10.1).

Service augmentation is further facilitated by the fact that the physical time spent in a vehicle or on the platform is available for informational interactions, in a way that can be compared with the availability of certain station areas for use as advertising media.

10.3.2.2 Augmented Reality for Operators

Different actors want to be involved in the design, organization, and shaping of the augmented reality offered to users: not only operators of new services, but also operators of traditional services seeking to encourage users to cooperate in the smooth operation of the system, or else information service providers (e.g., the Moovit service). All this is because the time available to the user is also a window of commercial interaction for the use of augmented services.

For a transit operator, the increased real-time complexity of the system is primarily an opportunity to improve operational monitoring and responsiveness. It also helps to strengthen the capacity for anticipation at several timescales, from real-time through to long-term planning, passing by very short-term scheduling adjustments and also by the medium-term planning of adaptations to demand levels.

As a result, the managerial capacity of operators should be substantially increased. This impacts on the management of production resources, from the automation of elementary functions (local vehicle management, traffic actuators on the infrastructure) through to the optimization of high-level functions for line and network management. It also impacts on commercial management, both for the elementary functions that benefit from automation (payment, pricing information, customized advice) and for high-level yield management functions: This offers a whole area of progress for urban passenger transit, with the potential for the development of specific functions and professional expertise.

Overall, the development of technical functions and commercial functions brings augmented operational reality by reinforcing the role of soft inputs. Hard inputs also develop in support, through the installation of sensors, transmitters, and telecommunication networks in vehicles and on infrastructure. These new commercial—or simply informational—services have the effect of augmenting the overall reality of operations.

The proliferation of information in this wider transit system, in particular between users through social media, constitutes for the operator a big opportunity with regard to traffic management but also a challenge, in that external recommendations may replace its own advice and prove counterproductive for the performance of the system. To pre-empt this, the operator needs to take on a further role: community management, obviously with an emphasis on incentives rather than coercion.

10.3.2.3 Augmented Reality for Regulators

Augmented realities on, respectively, the demand and supply side also apply to transit system regulators:

- More functions provided: mobility services and information services;
- Traditional operators and also providers of innovative services;
- More information available for knowledge and therefore for regulation of the system.

All this increases the scope, the needs, and also the resources of regulation:

- There needs to be informational interoperability between the different operators' information systems (this problem also arises for a network operator between the separate information systems on its lines);
- Cooperation needs to be maintained between operators in the dynamic management of services and flows, not only through mutual exchange of information, but also in the organization of the respective modal infrastructures, in particular for the introduction of new physical mobility services;
- Information needs to be regulated: its availability to users not only in terms of physical infrastructure and software capacity, but also in terms of authorized providers, as well as exchanges between users;
- Regulation can now be based on a more fine-grained description of system performance.

On the latter point, it is now technically possible to track the detailed operation of the system in real time. This should encourage the establishment of a Territorial Traffic Information and Management Center, as exists for road traffic, which channels information messages to the general public through traditional media (radio, television, Web services).

10.3.2.4 Augmented Reality for Analysts

The augmentation of reality through the increase in soft inputs to the physical system is particularly helpful to the development of knowledge, in action research and other studies, both through the emergence of new research topics and through the enabling of investigation methods. The research topics were set out in detail in

Sect. 10.2. Here, we will concentrate on the facilitation and reinforcement of investigation methods, which contribute to augmented reality for analysts, helping them to develop their understanding of demand and supply.

On the demand side, interaction with users is made easier through hardware (smartphones) and software (Web apps, automatic information processing). This transforms possibilities for surveys based on directive or more open questionnaires, while reducing survey costs and enlarging samples. Time spent in vehicles or on platforms offers a good opportunity for interaction. The information ecosystem also lends itself well to off-line surveys, owing to the availability of data and media for interactions, and also techniques for preparing, conducting, and processing interviews. This process can go as far as immersing volunteer users in virtual reality environments.

On the supply side, for the design of systems—in particular services and operating modes—the development of models and information systems facilitates and reinforces simulation-based experiments. The applications reported in Sect. 9.2 relate to this. Increased computing power also makes it possible to add optimization functions, in order to improve real-time operation or off-line design. Immersion in a virtual reality environment would also enable analysts to identify different potential real-world improvements in many areas—such as the design of services and spaces, vehicles, and infrastructures—aided by an understanding of the system in its real complexity.

10.3.2.5 Augmented Reality for the General Public

Mass transit services are available to everyone living in a territory. In general, their users constitute a significant proportion of the general public. As for mass transit information services, they are addressed both to users and to potential customers.

With information and simulation resources, it is now possible for people to prerun a journey in virtual reality, which would offer the general public the experience of augmented reality travel. Route advice services already offer a significant degree of customization. Nonetheless, the outline of an itinerary on a map, with details of timetables, travel time, and fares, still remains a fairly abstract description.

For motorists, some information services provide local 2D and even 3D views to illustrate the travel conditions encountered on a route: Remote access offers an experience of augmented reality for any individual. Equivalent services will undoubtedly emerge for mass transit, in order to reproduce the sensory experience of travelers at successive stages in a journey, in a vehicle, on the platform, and in the station. In particular, it will be important to convey the ambient traffic conditions.

10.3.3 Toward What Typical Applications for Assignment Models?

Assignment models have long been used for planning transit systems. In this role, they help to augment reality for users, operators, and regulators alike.

In the course of the sections in this chapter, we have pointed out the trends in the direction of systems and avenues of research for the improvement or application of assignment models. We will now recapitulate the types of application that appear both useful to stakeholders in the system and technically feasible. Before discussing the applications for long-term planning (Sect. 10.3.3.3), we will cover the real-time applications for traffic management and user information (Sect. 10.3.3.1) and then the medium-term applications for making supply and demand more flexible and reciprocally better matched (Sect. 10.3.3.2).

10.3.3.1 Toward Real-Time Applications

Real-time traffic management on a transit system can employ a centralized information platform. Centralized traffic management has a long history on interurban master networks, or for train management on an interurban rail network. Today, in 2015, however, such systems are still to be applied to urban mass transit passenger networks. It would seem just as technically feasible and equally socially useful in a big city as on an interurban system, since the traffic scales in terms of passenger kilometers per day or passenger hours per day are comparable.

At the end of Sect. 9.5.1, we highlighted the idea of using a scale 1 assignment model to make a real-time estimate of the state of the system, broken down into individual mobile entities and technical infrastructure elements. Establishing such a model through the pooling of different sources of observed or synthetic information is a first step.

The second step is to give this model predictive capacity by incorporating scheduled operations on the supply side and journey progress on the demand side. In this way, demand forecasts can be incorporated into the real-time adaptation of supply, alongside forecasts of the system's state in the dynamic traffic information communicated to users.

The third step will be to integrate the dynamic reactions of operators and users consistently into ongoing and anticipated situations. This is where the real scientific challenge lies, because there will be mutual cross-referencing between model and system, and in other words, the model will be self-referential. In particular, it will need to include the role of dynamic information vectors, i.e., the flow of information within the real system via the operators and other information services.

10.3.3.2 Toward Medium-Term Applications

While transaction costs (information, pricing, payment) remained high in urban mass transit systems, the service schedules were set statically, with few regime differences: peak and off-peak on working days, working days or weekends, and seasonal periods. With the general reduction in transaction costs, much more dynamic service schedule programs can be envisaged, as is already done by interurban transit operators, whether in rail or in air transportation.

Although the traffic constraints remain high—in particular peak hours on working days are determined by commuting trips which are very rigid—there are two obvious fields of application for yield management in urban mass transit:

- Journeys taken by passengers without subscriptions. It would be good to introduce a targeted customer management system for them, with a specific type of subscription and incentives for flexibility, in particular price adjustments for certain timetables and types of day, like travel cards for the elderly in interurban rail transportation. This customer segment should become the foundation of a new kind of commercial management in urban mass transit. The techno-economic principles of the commercial conditions can be established via an assignment model that will show the direct and indirect costs and benefits for the system.
- More generally, customized pricing has become as technically possible as customized information. Of course, people will need to be convinced and users will have to accept the changes.

An assignment model could be used not only to establish pricing principles, but also to inform negotiations between the regulator and representatives of the population and users, who could work together to explore future pricing and service scenarios.

Apart from (though not unconnected with) yield management, the medium-term applications include the preparation of Traffic Management Schemes in order to anticipate critical conditions, for example, a massive influx of additional passengers or a serious service breakdown. By analogy with interurban road systems, the priority would be to draw up such schemes for each major artery or, even better, each major corridor around a large artery, by incorporating alternative routes. Every large capacity line should have a Traffic Management Scheme, based on the use of an assignment model for simulation purposes, which anticipates passenger shifts within the network.

10.3.3.3 Long-Term Planning of the Transit System

The planning of a transit system is fundamental, whether in the development phase or at the more stable stage, during which both renovation and maintenance in good operational condition need to be managed. Assignment models are traditionally used as tools to guide planning decisions. This role will undoubtedly develop as a

result of two major factors: first the transformation of the system in the information era (see Sect. 10.1) and secondly scientific development of models, in both theoretical and algorithmic terms.

The applications to planning are likely to develop in four major directions: (1) refining and enhancing representation, (2) extending the scope of planning, (3) reinforcing techno-economic and commercial optimization, and (4) interacting with stakeholders and supporting consultative forms of planning.

(1) Refining and enhancing representation:

- Differentiating periods and capturing intra- then interday dynamics;
- On the demand side, differentiating between types of user on the basis of physical travel situations and economic trade-offs, in particular by describing the actual uses, the benefits obtained, and the costs experienced by each category of user;
- On the supply side, endogenizing the techno-economic management of services;
- Modeling a diversity of information services, as well as the receptiveness and sensitivity of travelers to information in its different forms and according to their usage practices.

(2) Extending the scope of planning to reflect the expansion of the urban domain for transit systems:

- Extending the assignment model to multimodality and taking account of the different service modes and user navigation on a multimodal network;
- Capturing the organization of individual mobility over a single day and also over the week, or even more;
- Incorporating a range of pricing scales, as well as the corresponding price sensitivities of users;
- Through the model, capturing the diffuse nature of services based on the pooling of small vehicles, analyzing their sphere of relevance, and synthesizing their overall contribution to mobility.

(3) Greater techno-economic and commercial optimization: Once techno-economic behavior on the supply side has become endogenous to the assignment model, the model can be reinforced by optimization functions that will automate (at least partially) the quest for a more efficient service schedule, for pricing conditions that generate higher revenues and greater fairness between the different categories of user, etc.;

(4) Interacting with stakeholders and supporting consultative forms of planning:

- Sharing and publishing the theoretical foundations and other scientific hypotheses implicated in the establishment of an applied model;
- Sharing the results of different scenarios, including details of local traffic conditions;

- Making the model's computational capacity available to volunteer or selected groups, whether to construct and process a scenario, or simply to customize the results of a certain scenario;
- Above all, developing planning scenarios in concert with the stakeholders concerned, with the oversight of regulators, involving the operators, user representatives, taxpayer representatives, and environmental groups.

10.3.4 Toward Urban Mobility Living Labs?

Assignment models should therefore find increasingly varied applications for the stakeholders in the transit system, whether users, operators, or regulators. Each category of stakeholder has a specific role to play in the operation of the system, in interaction with the others through relations of cooperation, complementarity, and, indeed competition, in providing services or infrastructure. Beyond the interplay of particular interests, the common objective is that the system should supply users with a good quality service combined with acceptable financial, social, and environmental performance.

All the different stakeholders can benefit from an overall understanding of the real system and its development prospects, in order to target their own activities and gradually adjust their own strategy. It is a good thing to share technical and economic understanding based on scientific theories. It is also a good thing to work together to anticipate the future and share potential scenarios for change: This is a fundamental priority of regional transportation master plans.

An additional form of cooperation between stakeholders with different interests-recently emerged (originally in the 1990s, rising to a few dozen and then hundreds of implementations between 2000 and 2015): the Living Lab (LL), a form of cooperation that stresses knowledge and innovation:

- Cooperation between complementary actors, working together in a collective project that generates a strong group dynamic;
- Around a real-world site where each carries out their respective activities;
- This site is handled as a study field where the system is observed and its operation theorized and also as a test site for innovations. The combination of observation and experiment provides a way of observing the effects of each experiment.

In this final subsection, our goal is to present the principles and potential benefits of an urban mobility Living Lab, and also to situate the role and contribution of an assignment model in the activities of such a laboratory.

After setting out the general principles of a LL (Sect. 10.3.4.1), we explore its advantages for each stakeholder category in the particular case of urban mobility: for users (Sect. 10.3.4.2), regulators (Sect. 10.3.4.3), operators (Sect. 10.3.4.4), and

designers and innovators (Sect. 10.3.4.5) and for researchers (Sect. 10.3.4.6). We shall finish by recapitulating the findings (Sect. 10.3.4.7).

10.3.4.1 Key Living Lab Principles and Methodology

According to Umvelt (2014), a LL is a framework for the design, development, and in situ experimentation of innovative services or products in which different stakeholders (public bodies, private partners including industrial companies and operators as well as start-ups, research bodies, and in particular users) cooperate closely in an open innovation process.

So an "Urban Mobility LL" would be an integrative platform based around an experimental site used for the creation and development of multimodal and mobility-oriented services.

The term first appeared in the 1990s in the work of Professor William J. Mitchell, from the MIT Media Lab and School of Architecture, in reference to the development of innovations through user-centered research methods applied in real-world environments. However, the concept received further impetus from the European Union, which employed the term to promote and finance a new kind of innovation. The developing power of ICT and the opportunities provided by such technologies suggested the possibility of user-centered development. The foundation of the ENoLL (European Network of Living Lab, "international federation of benchmarked Living Labs in Europe and worldwide") in CoreLabs (2007) marked the expansion of a movement which has since grown steadily. As of 2010, there were 370 certified Living Labs, in addition to structures that are Living Labs according to the definition, without being aware of or presenting themselves as such (Picard et al. 2011).

The ENoLL specified LL methodologies on the basis of five characteristic principles, namely (i) spontaneity and value creation, (ii) user empowerment, (iii) realism, (iv) openness, and (v) continuity.

Spontaneity and value creation: For new products and services to succeed, they have to offer not only more and objectively better functions but also subjective functions that inspire use, fulfill personal desires, and fit in with and contribute to the fulfillment of social and societal needs. For innovations to satisfy such clusters of needs and desires, it is not enough to explore and address the initial needs expressed by users. They must also have the capacity to detect, aggregate, and analyze spontaneous reactions and ideas over time, throughout the full life cycle of a product or service.

User empowerment: The role of users (primarily end users of the product, but also people who indirectly benefit) is fundamental: LL methodologies are user-centric, by contrast with traditional R&D approaches that are techno-centric.

Moreover, users are invited to co-create the product or service, with the result that innovations are user-driven. Product design in a LL focuses on the understanding of emerging customer needs and aims to use this knowledge to develop products that yield additional value for customers.

Realism In order to generate valid results, the Living Lab has to provide a "natural environment" where users and stakeholders behave realistically. In this way, innovative products or services can be tested in real-world conditions. This principle differentiates the way Living Labs work from other types of open co-design environments.

Openness The innovation processes must be as open as possible. This is essential to ensure that multiple perspectives are included and to attract sufficient input to achieve rapid progress. Openness enhances the appeal of Living Labs. It is a principle that also includes concepts such as open data or open innovation. Open innovation can be defined as "the use of purposive inflows and outflows of knowledge to accelerate internal innovation, and to expand the markets for external use of innovation, respectively" (Van de Vrande et al. 2009). Open innovation in a Living Lab methodology is about facilitating the flow of knowledge and accepting ideas from multiple stakeholders or citizens. It also raises the question of how the Living Lab deals with the profusion of ideas, how relevant knowledge is selected, etc.

Continuity Since cross-category collaboration is based on trust, which takes time to develop, a Living Lab has to be a structure that remains stable over a certain period of time. The continuity principle enables stakeholders to develop collaboration and to plan and run experiments in a durable environment.

Along with the five key methodological principles that characterize LLs, a LL is a framework that comprises five main components: users, partners, application environment, technology and infrastructure, and organization and methods (cf. Bergvall-Kareborn et al. 2009):

- The *application environment* is the context in which users interact with products and services. It includes the real-world conditions as well as the particular conditions of use, the product's user interface, and the usage protocol for the service.
- *Technology and infrastructure* refers to the LL's stock of sensors, telecommunication facilities, and software for analysis or design, or for interaction with users.
- *Organization and methods*. A LL is a platform where different partners join forces to increase the efficacy of their innovation process. Different organizational and management methods can be developed in order to achieve the objective of global collaboration. The LL's structure should also have the capacity to evolve in response to the changing needs and desires of both partners and users. In fact, a LL needs to have a governance system that focuses on the production process inside the structure and that also handles economic issues or other support functions, such as the communication and exploitation of results.

Ideally, there should be close partnership between private companies, the public sector, industries, research centers, associations, citizens, etc. The goal is that the

different stakeholders should share their skills and take maximum advantage of their mutual capacities to drive innovation. The stakeholders may have different interests in the Living Lab (Umvelt 2014), which are set out in Table 10.1 (Fig. 10.1).

Table 10.1 Benefits of a Living Lab

Stakeholder	Benefits of a Living Lab
Private sector: industries, start-ups, SMEs, etc.	• Being able to test new products or services directly on users and to obtain feedback based on real use; • Reducing costs and enhancing development processes; • Sharing knowledge with other partners and feeding their experience into the innovation process: Better knowledge of existing practices stimulates creativity and inventiveness
Public sector: public authorities and public services	• Providing the best living conditions for the population; • Keeping abreast of innovations; • Improving the dialogue between public and private, and public and citizens; • Allowing companies to test innovations in their territory to foster its attractiveness
Research centers, R&D clusters, and universities	• Observation and analysis of phenomena; • Real-world experimentation; • Developing new educational tools
Users and citizens	• Expressing needs and desires; • Contributing to a process of innovation in which their interests are central

Fig. 10.1 Key components of a Living Lab (from Ståhlbröst 2008)

10.3.4.2 Importance of Users

In a LL, attention is focused on the users of products and services, with the aim of satisfying needs and desires. Over time, this attention goes into the design of products, while in real time, the focus is on observing uses. In addition, users are invited to express their impressions and suggestions, which feeds back into the design process over time and then into the real-time interaction with the product/service.

The user-centered approach is entirely suited to urban mobility, which is a basic individual need in which each individual has his or her own particular goals. Usage entails physical presence, which is relatively easy to observe, while allowing time for interaction on the informational side. Moreover, smartphones are a powerful and efficient medium for interaction with the designers of products/services, as well as fulfilling the functions of information, payment, etc.

The physical description of uses can be effected by different sensors, traffic measurement systems, cameras, ticketing apparatus, digital tracking of mobile phones, and GPS tracking for smartphones. The smartphone can also feed data into informational activities.

The choice of the experiment site is very important: The geographical position determines the user population, and the site should be big enough to observe large parts of every trip that takes place in the site. Individual user sensors are a way to extend the spatial range of observation and experiment. Another important thing is to specify a set of individual and mass transit modes, with potential for multimodality and intermodality.

As regards the products/services to be designed and tested, let us summarize the ideas highlighted in the previous sections:

- innovative mobility services, collective taxis, etc.;
- innovative management methods: flexible options, dynamic pricing, customized pricing based on use, etc.;
- journey facilitation: physical facilitation of movement through the layout of station areas, traffic management, travel assistance through information and signage, simplified information and payment transactions, etc.;
- Not forgetting mobility practices, repeat journeys by individuals, travel choices, usage routines, and messages about travel that users exchange with others through various channels (including social media).

The introduction of a traffic assignment model, if possible at the heart of a multimodal mobility model, constitutes a major advantage for an urban mobility LL:

- Describing a transit system is a way to manage its technical complexity (construction of a service from infrastructures, vehicles, and commercial protocols) as well as demand patterns;

- The model enables one to design a service on first principles, to test it virtually, and to pre-test its commercial targets and conditions of use. It also enables to analyze the value of the service by comparison with alternative services;
- Through experimentation, a model offers a way to qualify uses retrospectively, to describe them relative to all the different journeys and the user population, and therefore to derive a typical customer profile.

Reciprocally, a LL system provides a whole ensemble of information that can be fed into the model, in order to improve its empirical accuracy and to develop theories (see para 6).

10.3.4.3 Public Stakeholders

Mobility has significant social implications, as a service for the population, with an important contribution to quality of life, and also as a physical phenomenon that presents an accident risk as well as health and environmental impacts. That is why the experimental aspect, as much as the territorialized aspect of a LL, is relevant to local authorities, who are responsible for quality of life, safety, health, and public security, and also for regulating the technical and economic performances of the transit system.

In principle, the mobility products and services designed within the framework of a LL should make a positive contribution to the public interest as fostered by public authorities, by improving the service delivered to users, reducing environmental damage, creating value, and subsequently generating economic activity and jobs.

On this principle, the public authorities should encourage an urban mobility LL, support, and even contribute to it:

- By facilitating the formation of partnerships between stakeholders, by supporting its establishment and operation, and by assisting with administrative formalities;
- By encouraging design and experiment initiatives from private actors;
- By giving innovation projects the benefit of their knowledge of local conditions and users, and also of the legal conditions applicable to the service scheme;
- By promoting cooperation or at least harmonious co-existence with existing services and their operators;
- By making technical and financial resources available: contributions “in kind” regarding the use of public spaces (access, parking and travel conditions for vehicles, meeting points for travelers, installation of specific street furniture, signage), or else the use of private telecommunication networks, and also the dissemination of information to local populations and sponsorship of initiatives.

In addition to these support functions, local authorities should apply and manage a multicriterion assessment system for the proposed services, in line with their oversight prerogatives (in addition, the observational system implemented on the

experimental site should also be used to assess the general state of the system on a period-by-period basis in order to monitor trends):

- impacts on traffic behavior: contribution to efficacy, reliability, and the reduction of incidents and other disruptions;
- effects on the quality of life of users and local people;
- environmental balance;
- financial balance and socioeconomic balance.

This oversight role would benefit greatly from a multimodal assignment model and environmental impact models, so that simulation could be used to evaluate effects not covered by the LL's observation system and also to carry out forward assessments of all impacts on the service and on its broader environment.

Local authorities could also use the LL to promote innovations in pricing, taxation, and subsidies in mobility services, in order to improve the economics of public transportation, while maintaining social equity.

For local government, and also at a larger scale, a LL is another way to channel innovation: rather than being a passive observer to the anarchic emergence of disparate innovations, encouraging their testing in optimum conditions both for themselves and for the community. This latter reason should be enough to recruit the local authorities who together hold the different sectorial responsibilities relating to the LL.

10.3.4.4 Transit Operators

Transportation in a territory is a well-established economic activity, carried out by operators who are linked to local authorities either by direct affiliation or through an operating license. The operators are fully engaged in the technical aspects of transportation:

- Highway operators, for infrastructure and traffic management;
- Mass transit operators manage their services and the resources they put into them.

Operators are involved in an urban mobility LL in several respects:

- For the establishment of in situ observation methods, on highways, on railroads, or in transit vehicles;
- By every innovation in mobility products or services, because of the technical and social interactions: technical interactions in the operation of the service, its use of a mobility infrastructure but also an access infrastructure (intermodality), its coordination with existing services, its influences on passenger practices in terms of modal combinations or shifts, or indeed in changes of practice, etc.;
- And of course, they can themselves be the drivers and beneficiaries of innovations, in their operational resources (including infrastructure and vehicle

resources, and operational processes) and in their interactions with users, in particular with regard to information, sales, or pricing.

Some may be tempted to see innovation as a threat, in its ability to bring change and therefore to undermine the status quo. In a globalized world, local opposition would be counterproductive: At worst, a stance of alert neutrality is advisable and at best, a position of welcoming cooperation, in order to anticipate the potential of innovation as quickly as possible and consider how to adapt to it: in other words, to co-construct one's environment.

The integration of a multimodal assignment model would seem crucial in this respect, in order to accurately interpret the scope of a local experiment, by relating it to the structure of the client base and also to the composition and operation of the multimodal service.

10.3.4.5 Industrial Firms, Designers, and Innovators

In a Living Lab, the private partners may be industrial companies (manufacturers of products with a life cycle much shorter than the longevity of a system), SMEs with a particular product line, and start-ups built around an innovation.

Centering on users, and therefore on customers, as well as on value creation, is entirely consonant with the aims of a private company. Regardless of the parent body, however, be it private partner, public partner, or research center, our focus here is the designers of innovative products or services.

For them, a LL is a framework for cooperation both in design and in experiment. With regard to design, the participation of users, operators, and a public regulatory body ensures a comparison of viewpoints and perceptions, and strategic positions and respective interests, within a constructive dynamic that helps to bring concepts to fruition more quickly. With multiple partners involved, design becomes co-design.

The positive involvement of partners who in practice represent complementary functions relating to the service being designed is a major advantage for the designer.

With regard to experimentation, user participation is crucial to the development of the concept: validating functions, adapting the ergonomics to patterns of use, through the observation of usage patterns, impressions, and habit formation, and specifying commercial and pricing structures.

Embeddedness in a given territory makes it possible to test not only the concept, but also its spread through the population, customer volumes, and therefore its market share and sales revenues.

The site's specialization generates economies of scale in the design and implementation of experiments; the observational resources in place facilitate the collection of information on uses and the conditions of use.

Overall, a LL framework shapes innovation through three specific processes: first, an analysis of value, by the designer from a user-centered perspective; then exposure to the partners; and finally field testing.

With a multimodal assignment model, a further process is possible: virtualization of the service and testing through simulation, which allows a large number of variants to be explored at relatively modest cost.

The availability of "immersive" virtualization facilities would obviously increase simulation capacity, both in the preliminary design phase and in the preparation of field experiments.

10.3.4.6 Researchers

The involvement of researchers in the LL brings scientific capacity and knowledge focus to the partnership, along with certain neutrality toward the economic interests at stake. In terms of content, scientific capacity includes expert knowledge in the domain concerned, including theories, stylized facts, and state-of-the-art international knowledge, combined with scientific curiosity (challenging) and investigative capacity. In terms of form, research methods prioritize objectivity, logic and rigor, critical doubt, and relativization.

This standpoint and approach are valuable to both design and experiment. Since theories and methodologies are organized into disciplinary segments, an urban mobility LL would benefit from the involvement of an interdisciplinary research pool:

Human and social sciences: for human factors regarding users or operational staff: psychology of perception, of attention, of mental load and emotion; sociology of needs, motivations and ways of life; sociology of organizations; microeconomics and econometrics of behavior and decision-making…

In engineering sciences: physical theory of flows, control techniques, mathematical and algorithmic modeling, computer science and computer engineering, statistical theory, massive data processing, etc.

Planning and development is an intermediate discipline, focusing on spatial organization and functions, and on the arrangement of spaces for the presence and movement of individuals.

Interdisciplinary cooperation fosters a holistic understanding of the system concerned, since mobility is a sociotechnical phenomenon. The advantage of a systemic understanding is obvious in the design of an innovative product or service:

Anthropologists and sociologists can identify the targets of a product/service in advance and its capacity to fulfill needs, its qualities (in particular congruencies with other services), and also its weaknesses;

Engineering and urban planning specialists can discern the technical and spatial characteristics of a product, and in the case of a service, its respective functional, applicative, and technical architectures, as well as its modes of functioning and interactions with the transit infrastructures and services already in place in the territory.

In other words, an interdisciplinary research center provides knowledge that is essential to innovation in the system concerned. Innovation also requires a desire, an aspiration, that some researchers may individually possess, so that they join the ranks of designers-innovators. Another favorable condition is cooperation between research and education: The student body may include learners with a vocation for innovation and entrepreneurship.

Since experimentation is a fundamental component of knowledge development, researchers are also competent and well equipped to contribute to experiments:

- For the design of an experimental protocol, its embeddedness in the territory and in the sphere of information: what information to collect, from what sources and under what practical conditions, while ensuring that observation does not disrupt the performance of the system;
- For the analysis of results: eliminating noisy observations, exploiting the remaining observations using statistical methods, and identifying essential features, factors of success or failure, qualification in terms of particular conditions.

Reciprocally, a LL offers researchers an extremely stimulating opportunity: as a big localized structure for observation and experiment, with both scientific and service components. For urban mobility, this is a truly unique opportunity, since concentration on a site makes it possible to grasp the relations between mobility and its urban context.

The observation of mobility as a phenomenon is an opportunity to identify regularities and stylized facts, which lays the groundwork for the development of theories.

Intensive observation also lends itself to the development of methods of analysis. One priority application is the calibration of simulation models—particularly economic and technical submodels—within a traffic assignment model.

The innovations to be designed may relate to the end users of mobility, which is of interest to the human sciences, or the intermediate stages, i.e., technical subsystems, which is of interest to engineering sciences. Management sciences have an interest in every stage.

Interdisciplinary cooperation leads to cross-fertilization, the effects of which develop over time to the benefit of all parties.

Finally, the cooperation between partners is enriching: The in-depth knowledge of the terrain and users held by local authorities, the specialist expertise of operators, and the creative spirit of designers stimulate the emergence of topics for research projects.

10.3.4.7 Recapitulation

To sum up, we have set out the general principles of a LL: fruitful cooperation between partners with complementary skills and roles, around a location site chosen as a place of observation and experimentation, for the analysis and design of innovative products or services.

We have concentrated on the case of urban mobility, identifying the specific interests of each category of stakeholder: The framework of a LL offers opportunities for each category, and reciprocally, the involvement of those stakeholders is a key factor for the success of the LL.

Mobility as a sociotechnical phenomenon is a valid focus for an urban transit LL, rather than a single particular mode of transportation. This is because, with regard to their mobility needs, end users see travel options as a multimodal package. In addition, transit modes run in technical interaction, and the more massive the respective exchanges and flows, the stronger the connection.

Reciprocally, the territorial embeddedness of a LL is particularly fruitful in relation to urban mobility, allowing it to be understood in context.

We have also pinpointed the benefits of partnership: complementarities and synergies between the respective competences of local authorities, including transit regulation, road and mobility service operators, designers-innovators, and researchers.

For the public authorities and private operational partners, a LL is useful at any time, if for nothing else than as a snapshot of the current status of the system that they contribute to manage. The same is true for researchers, for purposes of empirical observation and theorization.

Innovation and its dynamics constitute further motives, both because existing systems can undoubtedly be improved and because the digital era is rich in potential for change. Changes will include the renewal of existing technical forms, the emergence of new technical forms, also the renewal of the social forms of mobility, and the composition of new logistical lifestyle patterns (e.g., online shopping and delivery, instead of buying and carrying).

Finally, we highlighted the potential of a multimodal assignment model and more broadly of an arsenal of simulation models:

- To provide qualitative and quantitative knowledge about the operational status of the system in question, assessing it, and identifying the sequence of cause and effect, and structure, composition, and system effects;
- To conduct virtual experiments on products/services in terms of potential customers and consumption, and impact on the overall operation of the system. And therefore, to contribute to the forward development of the product/service;
- To complete the experiment in situ, deduce its impact on the rest of the system, and contribute to its retrospective assessment against a set of criteria;
- Reciprocally, the availability of an instrumented site is an ideal opportunity to improve the empirical validity of models. A comparison of real observations and simulation results is one obvious application. Going further, a LL is a way to observe reality in individual components or subsystems, i.e., at a more fine-grained level.

References

Bergvall-Kareborn B, Stahlbrost A (2009) Living lab: an open and citizencentric approach for innovation. Int J Innov Regional Dev 1(4): 356-370

Burghout W, Rigole PJ, Andreasson I (2014) Impacts of shared autonomous taxis in a metropolitan area. In: Proceedings of the transportation research board annual meeting 2015

CoreLabs (2007) Living labs roadmap 2007–2010: recommendations on networked systems for open userdriven research, development and innovation, in Open Document, Luleå University of Technology, Centrum for Distance Spanning Technology, Luleå

Hansen IA, Pachl J (eds) (2014) Railway Timetabling & Operations. Analysis - Modelling - Optimisation - Simulation - Performance Evaluation, 2nd edn. Eurailpress, 332 p

Picard R, Poilpot L (2011) Pertinence et valeur du concept de «Laboratoire vivant» (living lab) ensanté et autonomie. French Ministry of Economy, Finances and Industry

Ståhlbröst A (2008) Forming future IT: the living lab way of user involvement

TRB (2013) Transit capacity and quality of service manual, 3rd edn. Transportation Research Board, Washington DC

Umvelt (2014) Qu'est ce qu'un living lab? http://www.umvelt.com

Van de Vrande V, DeJong JP, Vanhaverbeke W, DeRochemont M (2009) Open innovation in SMEs: trends, motives and management challenges. Technovation 29(6):423–437