

PUBLIC TRANSIT PLANNING AND OPERATION

THEORY, MODELING AND PRACTICE



AVISHAI CEDER



Public Transit Planning and Operation

This page intentionally left blank

Public Transit Planning and Operation

Theory, modelling and practice

Avishai (Avi) Ceder

Civil and Environmental Faculty, Transportation Research Institute,
Technion-Israel Institute of Technology, Haifa, Israel

Cartoons: Avishai (Avi) Ceder



AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Butterworth-Heinemann is an imprint of Elsevier



Butterworth-Heinemann is an imprint of Elsevier
Linacre House, Jordan Hill, Oxford OX2 8DP
30 Corporate Drive, Suite 400, Burlington, MA 01803

First edition 2007

Copyright © 2007, Elsevier Ltd. All rights reserved

The right of Avishai (Avi) Ceder to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; email: permissions@elsevier.com. Alternatively you can submit your request online by visiting the Elsevier web site at <http://elsevier.com/locate/permissions>, and selecting *Obtaining permission to use Elsevier material*

Notice

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein.

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

For information on all Butterworth-Heinemann publications
visit our web site at <http://books.elsevier.com>

ISBN: 978-0-7506-6166-9

Typeset in 10.5/12 pts Times New Roman by Charon Tec Ltd (A Macmillan Company), Chennai, India
www.charontec.com

Printed and bound in the UK

07 08 09 10 10 9 8 7 6 5 4 3 2 1

Working together to grow
libraries in developing countries

www.elsevier.com | www.bookaid.org | www.sabre.org

ELSEVIER

BOOK AID
International

Sabre Foundation

Contents

Preface	xiii
Chapter 1 Introduction to Transit Service Planning	1
1.1 Motivation	2
1.2 The operational planning decomposition process	4
1.3 Service and evaluation standards and their derivatives	9
1.4 Viability perspectives	13
1.5 Outline of other chapters	16
References	18
Chapter 2 Data Requirements and Collection	21
2.1 Introduction	23
2.2 Data-collection techniques	24
2.3 Data requirements	27
2.4 Basic statistical tools	30
2.5 Literature review and further reading	37
References	41
Chapter 3 Frequency and Headway Determination	49
3.1 Introduction	51
3.2 Max load (point check) methods	52
3.3 Load profile (ride check) methods	56
3.4 Criterion for selecting point check or ride check	60
3.5 Conclusion (two examples)	64
3.6 Literature review and further reading	73
Exercises	79
References	80
Chapter 4 Timetable Development	81
4.1 Introduction	83
4.2 Objectives, optional timetables and comparison measures	84
4.3 Even headways with smooth transitions	90
4.4 Headways with even average loads	94
4.5 Automation, test runs and conclusion	98
4.6 Literature review and further reading	113

Exercises	115
References	116
Chapter 5 Advanced Timetables I: Maximum Passenger Load	119
5.1 Introduction	121
5.2 Even max load on individual vehicles	121
5.3 Optimization, operations research and complexity	127
5.4 Minimum passenger-crowding timetables for a fixed vehicle fleet	130
Exercise	137
References	137
Chapter 6 Advanced Timetables II: Maximum Synchronization	139
6.1 Introduction	141
6.2 Formulating an OR model for synchronization	142
6.3 The Synchro-1 Procedure	145
6.4 The Synchro-2 Procedure	146
6.5 Examples	148
6.6 Literature review and further reading	154
Exercises	159
References	161
Chapter 7 Vehicle Scheduling I: Fixed Schedules	163
7.1 Introduction	165
7.2 Fleet size required for a single route	168
7.3 Example of an exact solution for multi-route vehicle scheduling	170
7.4 Max-flow technique for fixed vehicle scheduling	172
7.5 Deficit-function model with deadheading trip insertion	176
7.6 Depot-constrained vehicle scheduling	188
7.7 Literature review and further reading	193
Exercises	196
References	198
Appendix 7.A: The maximum-flow (max-flow) problem	200
Chapter 8 Vehicle Scheduling II: Variable Schedules	207
8.1 Introduction	209
8.2 Fleet-size lower bound for fixed schedules	210
8.3 Variable trip-departure times	214
8.4 Fleet-size lower bound for variable schedules	220
8.5 Fleet-reduction procedures	223
8.6 Experiences with bus schedules	229
8.7 Examination and consideration of even-load timetables	233
Exercises	237
References	240
Appendix 8.A: Deficit-function software	241

Chapter 9 Vehicle-type and Size Considerations in Vehicle Scheduling	249
9.1 Introduction	251
9.2 Optimization framework	251
9.3 Procedure for vehicle scheduling by vehicle type	253
9.4 Examples	257
9.5 Vehicle-size determination	266
9.6 Optimal transit-vehicle size: literature review	268
Exercises	276
References	278
 Chapter 10 Crew Scheduling	 279
10.1 Introduction	281
10.2 Vehicle-chain construction using a minimum crew-cost approach	282
10.3 Mathematical solutions	290
10.4 A case study: NJ commuter rail	294
10.5 Crew rostering	300
10.6 Literature review and further reading	305
Exercises	309
References	312
Appendix 10.A: The Shortest-path problem	315
 Chapter 11 Passenger Demand	 319
11.1 Introduction	321
11.2 Transit demand, its factors and elasticity	322
11.3 Example of a demand forecasting method and process	330
11.4 Multinomial logit (MNL) model	336
11.5 Literature review and further reading (O-D estimation)	338
Exercises	341
References	341
 Chapter 12 Route Choice and Assignment	 343
12.1 Introduction	345
12.2 Route choice using waiting-time strategy	345
12.3 Proportion of passengers boarding each route	349
12.4 Proportions derived for regular vehicle arrivals	352
12.5 Passenger assignment based on route choice	354
12.6 Literature review and further reading	358
Exercise	362
References	363
 Chapter 13 Service Design and Connectivity	 365
13.1 Introduction	367
13.2 Service-design elements	368
13.3 Scheduling-based solution for operational parking conflicts	374
13.4 Optimum stop location: theoretical approach	381

13.5	Connectivity measures and analysis	388
13.6	Literature review and further reading	399
	Exercises	400
	References	404
Chapter 14	Network (Routes) Design	407
14.1	Introduction	409
14.2	Objective functions	410
14.3	Methodology and example	419
14.4	Construction of a complete set of routes	428
14.5	Multi-objective technique	438
14.6	Literature review and further reading	445
	Exercises	449
	References	452
Chapter 15	Designing Short-turn Trips	455
15.1	Introduction	457
15.2	Methodology	458
15.3	Candidate points and example	458
15.4	Excluding departure times	462
15.5	Maximum extensions of short-turn trips	467
15.6	Literature review and further reading	476
	Exercise	478
	References	480
Chapter 16	Smart Shuttle and Feeder Service	481
16.1	Introduction	483
16.2	Minimum fleet size required for a circular (shuttle) route	485
16.3	Routing strategies	487
16.4	Simulation	490
16.5	Case study	494
16.6	Customer survey	498
16.7	Optimal routing design: base network	503
16.8	Optimal routing design: algorithm	507
16.9	Implementation stages	509
16.10	Literature review and further reading	513
	Exercise	516
	References	519
Chapter 17	Service Reliability and Control	521
17.1	Introduction	523
17.2	Measures of reliability and sources of unreliable service	525
17.3	Modelling of reliability variables	530
17.4	Passenger waiting time at a stop	537
17.5	Advanced reliability-based data and control	542
17.6	Techniques to resolve reliability problems	546

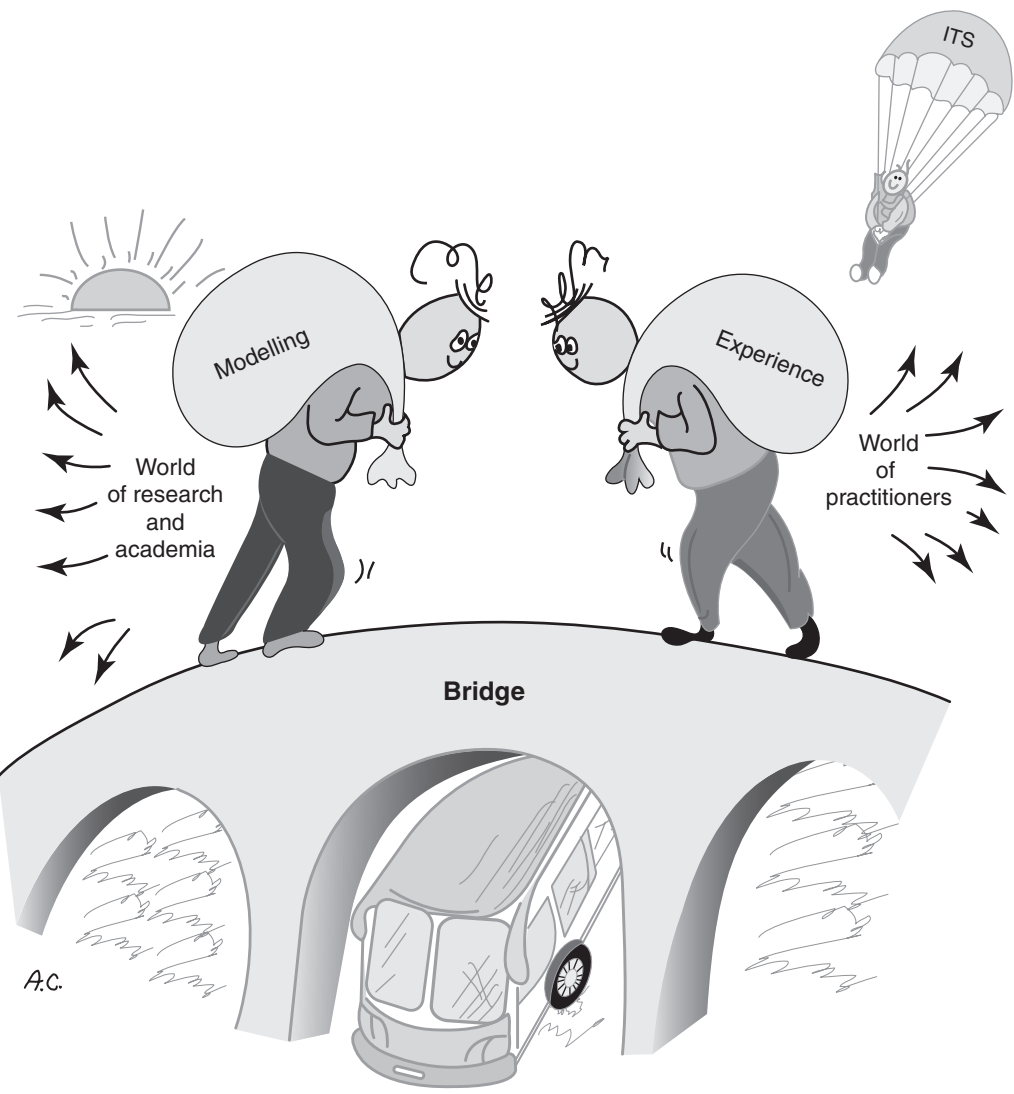
17.7	Literature review and further reading	555
	Exercises	563
	References	566
Chapter 18	Future Developments in Transit Operations	569
18.1	Introduction	571
18.2	Multi-Agent Transit System (MATS)	575
18.3	Vehicle encounters on road segments	578
18.4	Developments in transit automation	584
18.5	Literature review and further reading	586
18.6	Concluding remark	591
	References	591
	Answers to Exercises	595
	Author Index	607
	Subject Index	615

This page intentionally left blank

To my late Dad, Samuel (who worked for a large bus agency as a driver and treasurer for over 30 years), to my Mom, Anna, with wishes for many good years, and to my triumvirate, Roy, Ohad and Dror

This page intentionally left blank

Preface



Preface

Chapter outline

Personal motivation
Purpose and intended audience
Organization
Website and remarks
Acknowledgments

Personal motivation

The following story may serve to help understand the stimulus behind this book: A ship is sailing through a stormy ocean, and a little girl, who happens to be the captain's daughter, is playing on the deck when all of a sudden a large wave carries her overboard into the sea. The captain, who sees this from his post, immediately orders his sailors to jump into the ocean and save the girl, but no one dares for fear of risking his life. In desperation, the captain turns to the passengers and asks them for help while promising that whoever saves his daughter will get anything he or she wants as a reward. Again, no one reacts. But then suddenly a man with a long beard, who had been standing by the railing, lurches overboard into the sea. The sailors throw him a life preserver, and fortunately he manages to lift the little girl safely back on deck and into the arms of her father. The captain then hugs the man, who is thoroughly drenched, and says he will give him anything he wants, just name it. The hero's response: "I don't want anything. I just want to know who pushed me. . . ."

What pushed me, actually started some time between 1967 and 1971, when I was a bus driver for EGGED (the National Bus Company of Israel), whose 4000 buses make it one of the largest bus companies in the world. Before gaining a bus driver's licence, I had a theory about the way one should drive a bus; now I have a bus driver's licence – and no theory. The second motivation for writing this book came from my consulting work at EGGED from 1975–1985, when I was exposed daily to planning and operational problems in the public transit field. The third motivation came in 1981. I was at Massachusetts Institute of Technology (MIT) in Boston where, together with Professor Nigel Wilson, I was to give a new summer session course on Transit Operations Planning. This course became an annual offering until 2003.

From all these foregoing activities, I internalized the following realization: experience is what you get when you are expecting something else. My teaching of transit operations planning has taken place at universities in Adelaide, Boston, California, Hong Kong, Israel, Melbourne, Rome, and Sydney. Indeed, it has been my more than 30 years of teaching, research, and hands-on experience that has pushed me to write this book and to construct it in such a way that it will help not only teachers, researchers, and students in the area, but also practitioners in the field.

Purpose and intended audience

This book uses the concise term 'transit' to refer to public transit or public transportation or public transport; all three expressions are commonly used.

A major goal of this book is to establish a bridge between the world of practitioners and the world of research and academia for the purpose of narrowing the gap between these two worlds. I hope that such a bridge will also lead to opportunities for collaboration and interaction in order to improve public transit service and, no less important, its image. Henry Ford once said: “Failure is only the opportunity to start all over again more intelligently.” With this in mind, the book intends to introduce a few new ways of thinking about: (a) already implemented and investigated transit themes, while combining retro-perspective thoughts and cumulative experience; and (b) new concepts and ideas.

One of the main features of the book is its stand-alone (self-contained) capability, obviating the need to look back and forth at other publications for comprehending the text. At the same time, every chapter contains a literature review and a further reading list. Practitioners may have some difficulty in comprehending the sections with mathematical notation, but hopefully they can grasp the substance of the material and its practical implications. Researchers and academic professionals may find some of the sections unnecessarily detailed, but they should be aware that the book is also aimed at practitioners and undergraduate students, thus requiring more explanation. In summary, this work follows the notion that: (1) it is better to ask twice than to lose your way once; and (2) clarity is no less important than certainty.

Organization

Each chapter opens with a section containing information and remarks for practitioners, called ‘Practitioner’s Corner’. In fact, one can never tell which way the train went by looking at the track: for a practical decision, one needs more information. The purpose of these Corners is to impart guidance about sections of the chapter that are appropriate and sections that are perhaps too difficult for practitioners, thus allowing the less academically inclined to flow with the book and capture its substance.

The organization of the book is described in Chapter 1. Generally speaking, five groups of themes are addressed:

1. Overview of transit planning and data collection needs (Chapters 1 and 2).
2. Design and optimization of transit timetables, and of vehicle and crew scheduling (Chapters 3–10).
3. Passenger demand and assignment analysis (Chapters 11 and 12).
4. Transit service, network and route design, encompassing scheduling elements (Chapters 13–16).
5. Transit reliability and future operations planning developments (Chapters 17 and 18).

All the quantitative chapters offer exercises for practising the methods covered; of the book’s 18 chapters, only Chapters 1, 2 and 18 are without exercises. The answers to these exercises appear at the end of the book.

The literature review of papers relevant to the topic(s) covered in a chapter appears as the last numbered section of each chapter, except for Chapters 1 and 8 (the review for which actually precedes it, in Chapter 7). The reason for this order – instead of the traditional pattern of

starting a scientific article with a literature review, is to focus on each chapter's essence from the beginning, and only at the end to give the reader who may wish to broaden his or her knowledge of the particular topic, a kind of annotated reference list and an extended bibliography.

Website and remarks

The success of a professional book can be evaluated by the extent to which it succeeds in introducing new and improved ideas and methods. It is not only a matter of learning the book's content; it has to do, as well, with how much the volume can inspire the reader's imagination to think further. This concept has served as the guideline for the development of this book.

Indeed, the process of writing this book motivated the formulation of an interactive-software program, which may be found at this website: www.altdoit.com. This site (instead of the publisher attaching a CD) provides a highly informative graphical technique with which it is simple to interact. The user can interject practical suggestions, whose effects on the vehicle's schedule are immediately described. This useful tool, which relates more specifically to Chapters 7–9 and Chapters 12–15, will also assist the reader in solving some of the exercises and practical problems outlined in those chapters.

Finally, when lecturing this transit course, I tended to use humour at times because I believe in the insight captured by the English playwright George Bernard Shaw, who once said: "When a thing is funny, search it carefully for a hidden truth." More than once I have been asked to employ some of this humour (including the cartoons that I have also drawn) if I ever wrote a book. I have done this to some extent, especially in the Practitioner's Corners.

Acknowledgements

Many people have contributed to this book through their constructive feedback and encouragement. My views and understanding of the importance of public transit planning, service, and operations greatly benefited from my discussions with Professor Nigel Wilson of the MIT, with whom I annually shared the teaching of a summer course on the subject at MIT for 22 years.

I would like to acknowledge and thank Professor Hai Yang of the Hong Kong University of Science and Technology for his course material, including exercises on demand modelling in public transit (from which some of the exercises in Chapter 11 were drawn); Professors Yoram Shiftan and Shlomo Bekhor of the Technion-Israel Institute of Technology for their comments on demand modelling and transit assignment; Dr Yechezkel (Hezi) Israeli, who was my PhD student, for his contribution and remarks on transit-route modelling; the majority of Chapter 14 and part of Chapter 12 are based on his dissertation; and Moshe Flam for his contribution to future transit developments in Chapter 18. My appreciation to Yaron Hollander, a doctoral student at the University of Leeds, who contributed to the literature review of this book; and to my PhD student, Yuval Hadas of the Technion, whose thesis supported part of the last chapter of the book. Many thanks are also due to my Master's degree students, Shirin Azzam, Gali Israel and Shai Jerby, for their useful work and comments on the subjects of Chapters 9, 13 and 16,

respectively; and Yana Shnirman, an undergraduate student, for her practical comments on Chapter 10. Let me express my gratitude to Asher Goldstein, who provided me with editorial support. In this stream of acknowledgements, it will be only fair to thank the inventor of the treadmill, which has helped me stay in shape throughout the writing of this book.

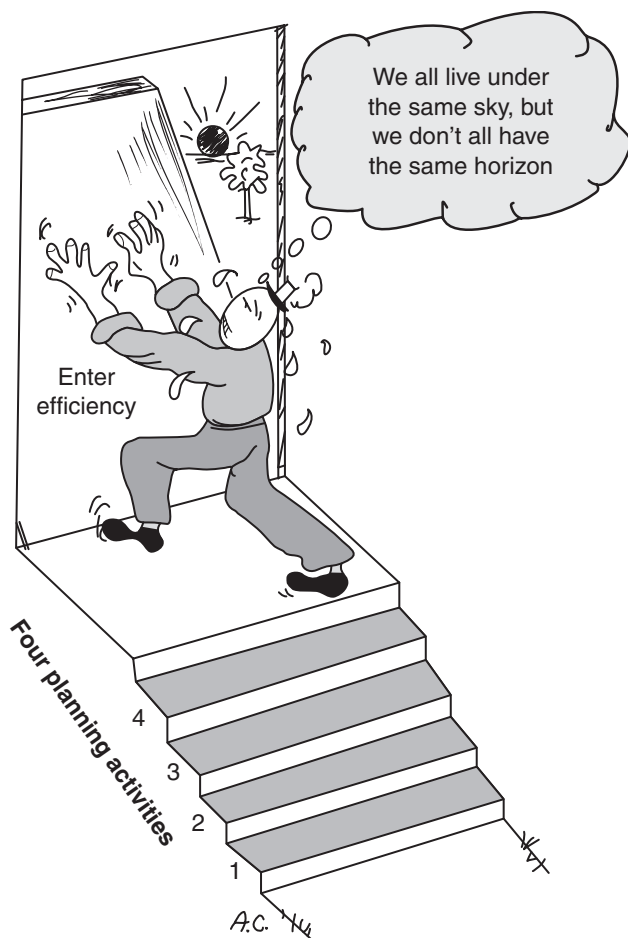
Lastly, I offer my heartfelt thanks to my wife, Patricia (Shira) Tolentino Ceder, for her great encouragement, love and understanding. But also, and not least, for making dedicated use of her talent as an architect in doing the artwork for the figures in this work. Finally, a bouquet of affections goes to my three sons and daughter-in-law, Roy and Roni, Ohad, and Dror, as well as to my Mom, Anna, and my brothers, Tuli and Hagai and their families, all of whom helped me get through the difficult periods when writing this book.

I retain, of course, sole responsibility for any errors. I would be very pleased to gain feedback.

Avishai (Avi) Ceder
Haifa, Israel

This page intentionally left blank

1 Introduction to Transit Service Planning



Chapter 1 Introduction to Transit Service Planning

Chapter outline

- 1.1 Motivation
 - 1.2 The operational planning decomposition process
 - 1.3 Service and evaluation standards and their derivatives
 - 1.4 Viability perspectives
 - 1.5 Outline of other chapters
- References
-

Practitioner's Corner

As stated in the Preface, each chapter opens with information and notes to guide practitioners. This introductory chapter will show that a successful transit service is not permanent and that failure in a new service need not be fatal. It contains neither mathematical formulations nor a complicated literature review.

The motivation behind the writing of this book is to bridge the world of the researcher and the practitioner. This chapter first presents a real-life description of the planning process for transit operations. It then provides an overview and a critique of currently used standards and guidelines for transit service, pointing out those standards that warrant research and those for which administrative decisions are sufficient. The main components affecting the viability of a transit service from both the passenger and agency perspective are described, as is the link between some problematic elements and the possible solutions advanced in the book.

The chapter ends with an outline of the other chapters and the links between each chapter and the main core of the transit-operation planning process. This book attempts to show how to deal with, and to solve in a practical manner, current transit-service dilemmas and confusions follows Pablo Picasso, who said: "I do not seek, I find".

1.1 Motivation

A basic premise of this volume is that using transit service should be like eating potato chips – once you start, it's hard to stop. The fundamental question, though, is how do we design and approach such a service while knowing, because of the competition offered by cars, that an adequate transit design must provide superlative service in order to be good enough. This book will demonstrate methods and procedures not only for improving transit design through the provision of an attractive, viable service but also for reducing its cost and increasing efficiency from the transit-agency perspective.

One of the main motivations for writing this book is the persistent use by transit agencies (especially in the US) of the *Bus Scheduling Manual* of August 1947, reprinted in 1982 (Rainville, 1947). The manual opens as follows: “This volume is viewed today [i.e. 1982] as a classic reference on manual scheduling practices. It represents a comprehensive effort made over 30 years ago to pull together the state of the practice in this field, and many of those practices still remain valid today”.

The manual’s cover and an inside picture of a device to measure distances appear in Figure 1.1. Those transit agencies that may still be using this manual might be asked a question: is the treatment of passenger loads and running-time data, frequency determination, and vehicle and crew scheduling really ‘classic’? Or is it obsolete? The conclusion implied by the latter served as a trigger to investigate why the treatment of these elements in the manual is not classic and to submit operations planning options that exploit existing, computerized computation power and advanced modelling. On the positive side, it may be said, as it was in the 1947 manual, that some basic concepts behind the planning process for transit operations do indeed remain the same.

In 1998, the US Transit Cooperative Research Program (TCRP) sponsored *Report 30* (Pine *et al.*, 1998) on transit scheduling with the idea presumably of advancing knowledge in the field for training purposes. However, *Report 30* neither contains nor refers to any research on transit scheduling (despite the many studies that existed); furthermore, it does not mention a single reference. The report actually provides very basic training material for new schedulers; in concept, it is almost a replica of the 1947 manual, while emphasizing that scheduling is a craft.

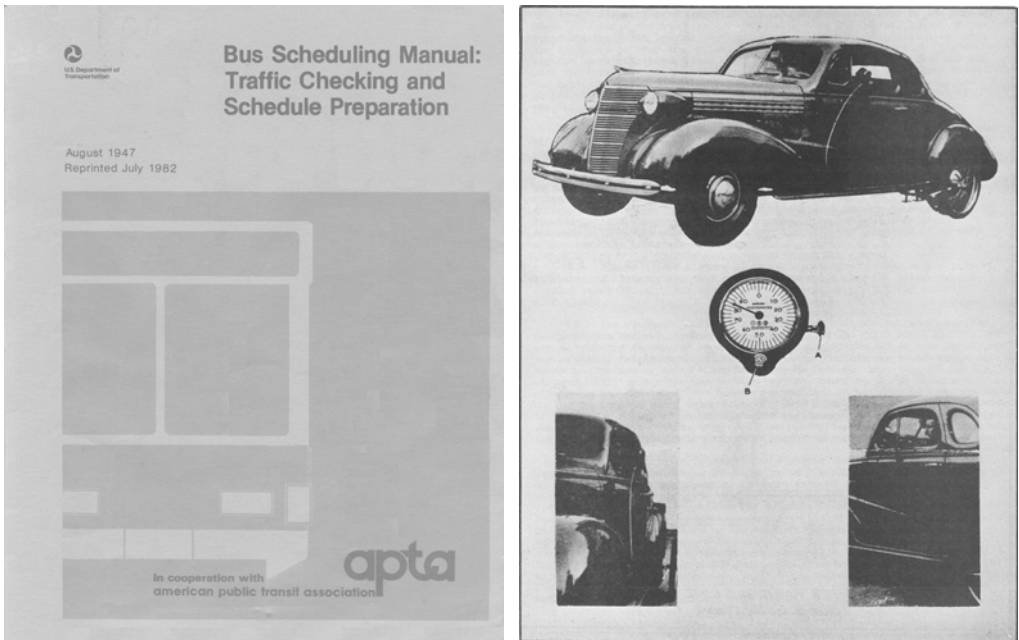


Figure 1.1 The cover and a page from the well-known, 1947 *Bus Scheduling Manual*, wrongly viewed as a ‘classic’ reference

The following story may sum up the motivation for writing this book: two marketing people from a European shoe factory are sent to investigate the potential sale of shoes in Africa. A few days later, faxes from each of them arrive at the manager's office. The first fax reads: "No chance, everyone is barefoot". The second fax urges: "Lots of chances, everyone is barefoot". This book will adopt the "lots of chances" view.

Another stimulus for this book is a long list of standards and guidelines (see Section 1.3) for transit-service planning, operations and control that determine the quality of service seen in practice. In the event that current transit-service quality has reached a satisfactory level, then these standards and guidelines are apparently effective. Nonetheless many aspects of transit service can stand improvement, necessitating changes or replacement of some of these standards and guidelines. Like those in the 1947 manual. We can imagine that the majority of the standards and guidelines to be covered in the present volume are intuitive-based or administrative-based instead of being research-based; hence, an opportunity is presented to develop new research tools and methods for improving some of these standards and guidelines (as discussed in Section 1.3).

This book uses the term **transit** as shorthand for public transit or public transportation or public transport, all terms in common use. This book describes transit-operations planning problems, a summary of different approaches that have been proposed for their solution, and a description of new methods and approaches incorporating some of the positive aspects of prior work. In essence, the new methods and approaches are all applicable to the four major modes of transit operation: airline, rail, bus and passenger ferry. Although the operations-planning process for bus, rail and passenger ferry is similar, that for airline has several special features, but these are not covered in this book. Moreover, rail and passenger ferry can be perceived, to some extent, as special cases of bus service because of the variety of activities and problems offered by the latter. Finally, the methods and approaches proposed are intended to be easier to implement and more sensitive to the risks of making changes than are those currently in use by practitioners and researchers.

1.2 The operational planning decomposition process

The transit-operation planning process commonly includes four basic activities, usually performed in sequence: (1) network route design, (2) timetable development, (3) vehicle scheduling, and (4) crew scheduling (Ceder and Wilson, 1986; Ceder, 2001, 2002). Figure 1.2 shows the systematic decision sequence of these four planning activities. The output of each activity positioned higher in the sequence becomes an important input for lower-level decisions. Clearly the independence and orderliness of the separate activities exist only in the diagram; i.e. decisions made further down the sequence will have some effect on higher-level decisions. It is desirable, therefore, that all four activities be planned simultaneously in order to exploit the system's capability to the greatest extent and maximize the system's productivity and efficiency. Occasionally the sequence in Figure 1.2 is repeated; the required feedback is incorporated over time. However, since this planning process, especially for medium to large fleet sizes, is extremely cumbersome and complex, it requires separate treatment for each activity, with the outcome of one fed as an input to the next.

The quantitative treatment of the transit planning process is reflected in the welter of professional papers on these topics and in the development of numerous computer programs

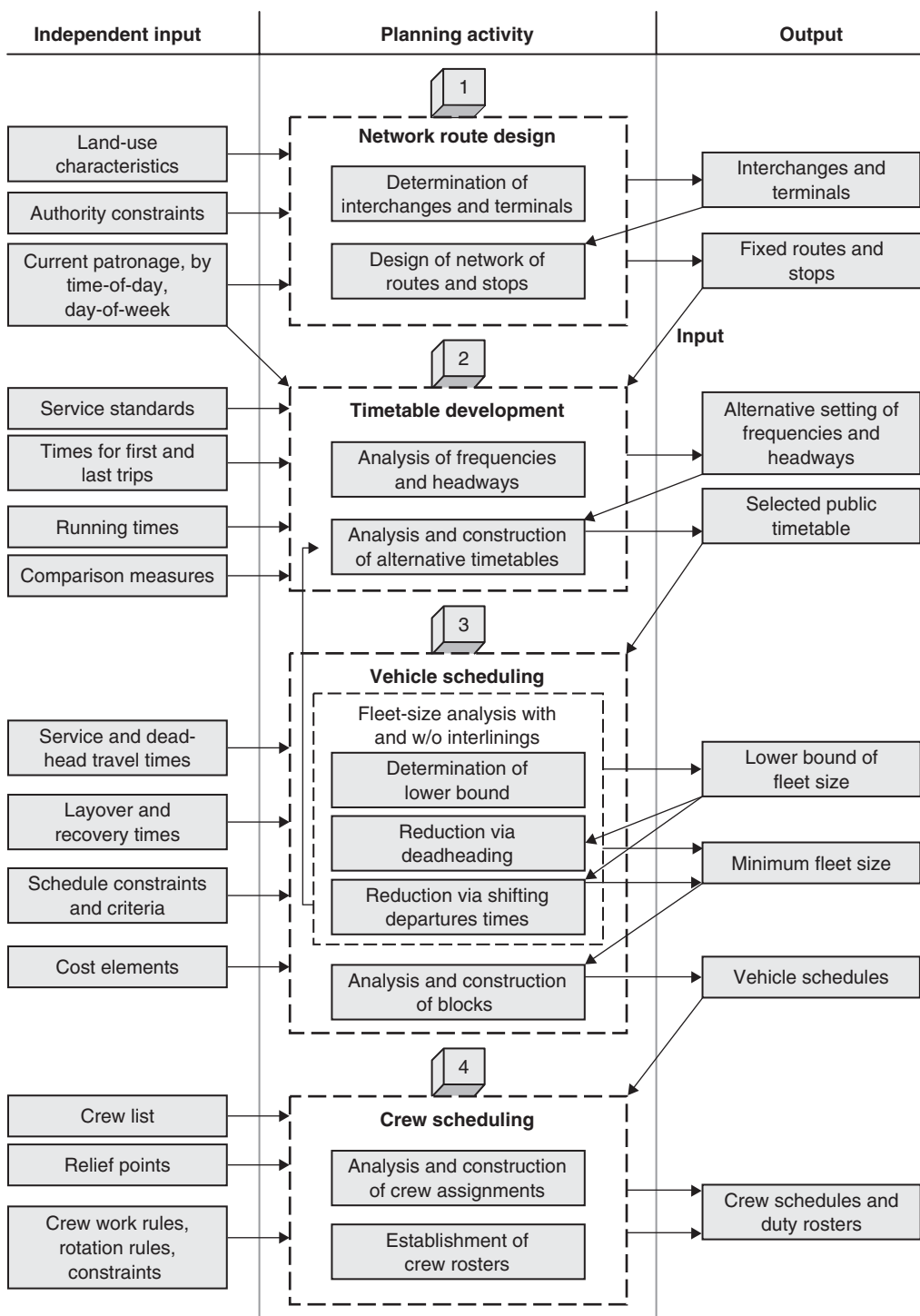


Figure 1.2 Functional diagram (System Architecture) of a common transit-operation planning process

to automate (at least partially) these activities. In the last 25 years, a considerable amount of effort has been invested in the computerization of the four planning activities outlined in Figure 1.2 in order to provide more efficient, controllable and responsive schedules. The best summary of this effort, as well as of the knowledge accumulated, was presented at the second through ninth international workshops on Computer-Aided Scheduling of Public Transport (CASPT), which in 2006 changed its name to Conference of Advanced Systems of Public Transport and in books by Wren (1981), Rousseau (1985), Daduna and Wren (1988), Desrochers and Rousseau (1992), Daduna *et al.* (1995), Wilson (1999), Voss and Daduna (2001) and Hickman *et al.* (2007). There is also some commercially available software in the area of transit scheduling, such as (in alphabetical order): AUSTRICS (www.austrics.com.au), HASTUS (www.giro.ca), ILOG (www.ilog.co.uk), MERAKAS Ltd (www2.omnitel.net/merakas), PTV (www.ptv.de), ROUTEMATCH (www.routematch.com), ROUTEMATE (www.nemsys.it), ROUTELOGIC (www.routelogic.com), SYSTRA (www.systra.com) and TRAPEZE (www.trapezesoftware.com). These software packages concentrate primarily on the activities of vehicle and crew scheduling (activities 3 and 4 in Figure 1.2) because, from the agencies' perspective, the largest single cost of providing service is generated by drivers' wages and fringe benefits. Focusing on activities 3 and 4 would seem to be the best way to reduce this cost. However, because some of the scheduling problems in these software packages are over-simplified and decomposed into sub-problems, a completely satisfactory or optimal solution is not assured, thus making room for decisions by experienced schedulers. After all, experience is what we gain when expecting something else; said another way, the exam is given first and the lesson after.

An argument in favour of automating activities 3 and 4 is that this scheduling process is extremely cumbersome and time consuming to undertake manually. In addition to the potential for more efficient schedules, the automated process enables services to be more controllable and responsive. The cost and complexity of manual scheduling have served to discourage adjusting activities 1 and 2. Only with automated scheduling methods, which are becoming more widely accepted, is it feasible to focus on higher levels in the planning process. Nonetheless, a case can be made that these higher levels have received short shift by both researchers and practitioners.

The network route-design activity in Figure 1.2 is described in Chapters 13 and 14. Planning practice in terms of transit-route design focuses almost entirely on individual routes that, for one reason or another, have been identified as candidates for change. Occasionally sets of interacting (e.g. overlapping or connecting) routes are subject to redesign, usually after a series of incremental changes to individual routes has resulted in a confusing, inefficient local system. Although it is difficult to predict the benefits that will result from redesigning any transit network without conducting a detailed assessment, it is reasonable to believe that they will be large compared with the benefits of additional efforts aimed just at problematic scheduling activities (2, 3 and 4 in Figure 1.2). The approach described in Chapters 13 and 14 generates all feasible routes and transfers connecting each place (node) in the network to all others. From this vast pool of possible routes and transfers, smaller subsets are generated that maintain network connectivity. For each subset thus generated, transportation demand is met by calculating the appropriate frequency for each route. Next, pre-specified optimization parameters are calculated for each subset. Based on the specific optimization parameter desired by the user, it is then possible to select the most suitable subset. This method has been designed as a tool for the planning of future transit networks, as

well as for the maintenance of existing networks. The method presented ensures flexibility by allowing the user either to input their own data or to run the analysis automatically.

The timetable-development activity in Figure 1.2 is described in Chapters 3, 4, 5 and 6 for establishing alternative frequencies and timetables. The aim of public timetables is to meet general public transportation demand. This demand varies during the hours of the day, the days of the week, from one season to another, and even from one year to another. It reflects the business, industrial, cultural, educational, social and recreational transportation needs of the community. The purpose of this activity, then, is to set alternative timetables for each transit route in order to meet variations in public demand. Alternative timetables are determined on the basis of passenger counts, and they must comply with service-frequency constraints. In Chapter 4, alternative timetables are constructed with either even headways, but not necessarily even loads on board individual vehicles at the peak-load section, or even average passenger loads on board individual vehicles, but not even headways. Average even loads on individual vehicles can be approached by relaxing the evenly spaced headways pattern (through a rearrangement of departure times). This dynamic behaviour can be detected through passenger-load counts and information provided by road supervisors. The key word in the even-load cases is the ability to control the loading instead of being repeatedly exposed to an unreliable service resulting from an imbalance in loading situations.

The vehicle-scheduling activity in Figure 1.2 is described in Chapters 7 and 8, and is aimed at creating chains of trips; each is referred to as a vehicle schedule according to given timetables. This chaining process is often called 'vehicle blocking' (a block is a sequence of revenue and non-revenue activities for an individual vehicle). A transit trip can be planned either to transport passengers along its route or to make a deadheading trip in order to connect two service trips efficiently. The scheduler's task is to list all daily chains of trips (some deadheading) for each vehicle so as to ensure the fulfilment of both timetable and operator requirements (refueling, maintenance, etc.). The major objective of this activity is to minimize the number of vehicles required. Chapters 7 and 8 describe a highly informative graphical technique for the problem of finding the least number of vehicles. The technique used is a step function, which was introduced as far back as 20 years ago as an optimization tool for minimizing the number of vehicles in a fixed-trip schedule. The step function is termed deficit function, as it represents the deficit number of vehicles required at a particular terminal in a multi-terminal transit system. In Chapter 7, the fixed-schedule case is extended to include variable trip schedules, in which given shifting tolerances allow for possible shifts in departure times. This opens up an opportunity to reduce fleet size further. The deficit function, because of its graphical characteristics, has been programmed and is available at www.altdoit.com. In this book, the deficit function is applied and linked to the following activities: vehicle scheduling with different vehicle types (Chapter 9), the design of operational transit parking spaces (Chapter 13), network route design (Chapters 13 and 14), and short-turn design of individual and groups of routes (Chapter 15). The value of embarking on such a technique is to achieve the greatest saving in number of vehicles while complying with passenger demand. This saving is attained through a procedure incorporating a man/computer interface allowing the inclusion of practical considerations that experienced transit schedulers may wish to introduce into the schedule.

The crew scheduling activity in Figure 1.2 is described in Chapter 10. Its goal is to assign drivers according to the outcome of vehicle scheduling. This activity is often called driver-run cutting (splitting and recombining vehicle blocks into legal driver shifts or runs). This crew-assignment process must comply with some constraints, which are usually dependent on a

labour contract. A brief summary is given of the conceptual analytical tools used in the modelling and software of this complex, combinatorial problem. The crew-rostering component of this activity normally refers to priority and rotation rules, rest periods and drivers’ preferences. Any transit agency wishing to utilize its resources more efficiently has to deal with problems encountered by the presence of various pay scales (regular, overtime, weekends, etc.) and with human-oriented dissatisfaction. The crew-scheduling activity is very sensitive to both internal and external factors, a feature that could easily lead to an inefficient solution.

The essential inputs for the transit operations planning process illustrated in Figure 1.2 are listed in Figure 1.3. These mainly independent elements are arranged by activity number. It should be emphasized that their values differ by time-of-day and day-of-week.

Planning activity	Input element	
Timetable development	(1)	Route (line) number
	(2)	Nodes (stops and timepoints on a route)
	(3)	Pattern (sequence of nodes on a route)
	(4)	Average passenger loads between adjacent stops on a route
	(5)	Load factor (desired number of passengers on board the transit vehicle)
	(6)	Policy headway (the inverse of the minimum frequency standard)
	(7)	Vehicle type
	(8)	Vehicle capacity
	(9)	Average running time (travel time between stops/timepoints on a route)
Vehicle scheduling	(1)	Trip recovery-time tolerances (maximum and minimum time to be prepared for next trip)
	(2)	Trip departure-time tolerances (maximum departure delay and maximum advance departure)
	(3)	List of garages (names and locations)
	(4)	List of trip start and end locations
	(5)	Average deadhead times from garage locations to each trip start location (pull-outs)
	(6)	Average deadhead times from trip end locations to garage locations (pull-ins)
	(7)	Average deadhead time matrix between all trip end and start locations (by time-of-day)
Crew scheduling	(1)	Relief-point location (stops, trip start and end points, garages)
	(2)	Average travel times between relief points
	(3)	Trip-layover time (minimum and maximum rest times between two adjacent trips)
	(4)	Type of duty (early, late, split, full, tripper, etc.)
	(5)	Duty length (maximum spread time)
	(6)	Number of vehicle changes on duty
	(7)	Meal breaks
	(8)	Duty composition
	(9)	Other work rules
	(10)	List of drivers by name and type (e.g. part-time, full-time, seniority)
	(11)	Driver priority and equality rules
	(12)	One-day-on, one-day-off work pattern

Figure 1.3 Typical input elements for the three schedule-planning activities

1.3 Service and evaluation standards and their derivatives

Tremendous and monotonous in transit service can be perceived as synonymous. It is best to have a monotonous, automated transit system that will always be there for the passenger; e.g. a modularized Personal Rapid Transit (PRT) system conducted like a horizontal elevator. This observation serves as an introduction to the need for transit service standards and guidelines. On the one hand, standards have to do with maintaining and improving existing service levels; on the other hand, they are often a source of fiscal pressure on transit agencies. Service standards are also linked to any evaluation effort aimed at improving the efficiency, effectiveness and productivity of a transit service. The greater these measures, the higher the level of service that can be offered.

1.3.1 Service standards

The need for dynamically updated standards in the transit industry is described in an article by Mora and Chericoff (2005). This need deserves attention, especially since the rapid introduction of advanced technologies in bus and rail transit and services. While standards, regulations and best practices are justified for supporting safety and security applications, in particular, they are also crucial for creating satisfactory transit service.

More than 50 per cent of the transit agencies in the United States (TCRP, 1995; METRO, 1984) and in Europe (QUATTRO, 1998) employ formal standards in service planning. The three reports discussing standards and guidelines in the US bus industry (TCRP, 1995; METRO, 1984) and in transit service in Europe (QUATTRO, 1998) are based solely on surveys conducted among transit agencies. For example, 109 of 345 agencies responded in the 1984 report, and 11 of 297 agencies in the 1995 report. The objectives of each survey were predominately based on the 'What' type of questions, and only a few on the 'How' type of question. (Examples: What service standards are currently used? What data are collected and used to evaluate the various performance criteria utilized? How is it collected?) The surveys related only to the effectiveness of transit-service evaluation efforts. They were conducted on the basis of how the agency perceived effectiveness (10 and 5 gradation scores in 1984 and 1995, respectively) in improving: (i) service delivery, (ii) equity, (iii) ridership, and (iv) operating costs. The highest ratings (i.e. standards perceived to be more effective) were given to (i) and (iv), while (iii) received the lowest score in 1984 and (ii) the lowest in 1995.

The main standards currently utilized can be partitioned into two categories: (i) route design and (ii) service design. These can be divided further into route level and network level standards, and into planning level and monitoring level standards. The resulting 20 standards appear in Figure 1.4 by category, group, number, standard name, typical criteria range and remarks. This figure, as elsewhere in the book, uses abbreviations for compactness of presentation: passengers (pass), population (pop), minimum (min), maximum (max), meter (m), origin-destination (O-D) and central business district (CBD). Figure 1.4 shows typical bounds for each standard, which can be used informally as a guideline. In the Remarks section, we provide extra information for each standard that will enhance its implementation characteristics.

Although the standards in Figure 1.4 are survey-based, using informative questions, we think that two, more elaborated and important questions, are missing: (1) What are the specific

Category	Group	No.	Standard item	Typical criterion range*	Remarks
Route design	Route level	1	Route length	Max 40–100 minutes one-way	Longer length for larger agencies
		2	Stop spacing	120–400 m (urban area)	Depends on pop density, land use
		3	Route directness**	Upper limit on deviation from car's shortest path, of 20–50%	Seeking higher productivity on deviated segments
		4	Short-turn**	At peak period only	Aim at reducing operations cost
	Network level	5	Route coverage	Min 800–1000 m route spacing (urban area); Max 400–800 m walk to stop	Min 50–95% of residents to have below max walk distance to stop
		6	Route overlapping	Overlapping allowed only on approaches to CBD	Avoiding confusion and balancing route dispersion
		7	Route structure**	Max of 2–3 branches per route/loops around terminals	Reducing confusion by different route number
		8	Route connectivity	Min of 1–3 routes that intersect a given route (transfer points)	Especially for a new route in an existing network
Service design	Planning level	9	Span of service	Min operating hours by day-of-week, between 5–6 a.m. and 10 p.m.–2 a.m.	Later ends for larger agencies
		10	Load (crowding) level	Max load factor 125–150% of seat capacity at peak hours/segments; 100% at off-peak	Higher load factors (150–175%) for short-hand service (shuttle, feeder)
		11	Standeers**	Max 50% of number of seats	Depends on the interior vehicles configuration
		12	Headway, upper limit**	Max (policy) headway, 15–30 minutes at peak period; otherwise, 20–60 minutes	Varies by type of service and day-of-week
		13	Headway, lower limit	Min headway, 2–3 minutes	More use in smaller agencies
		14	Transfers	Max of 1–3 transfers for any O-D	Larger agencies permit more transfers
		15	Passenger/shelters**	Min of 65–100 daily boarding pass	Attention to locations of elderly, and centres, such as hospital
	Monitoring level	16	Schedule adherence**	Min 80% 'on time' (0–5 minutes behind schedule) at peak period, 90% otherwise	This guideline is usually relaxed for short headway
		17	Timed transfer	Max of 3–8 minutes vehicle wait at transfer point	More use in smaller agencies
		18	Missed trips**	Min 90–95% of scheduled trips are OK	Missed trips can also not comply with trip reliability criterion
		19	Passenger safety**	Max 6–10 pass. accidents per 10 ⁶ pass; Max 4–8 accidents per 1.6 × 10 ⁵ vehicle-km	Depends on updated safety data
		20	Public complaints	Limits on number of complaints per driver/pass/time period	Public comments and complaints always received

* Reflects data mainly from US.

** Standards commonly in use.

Figure 1.4 Available service standards and their typical criteria ranges

purposes of a given standard? (2) Are the purposes in (1) fully and optimally attained? In answering these two questions, the transit agency may find out that the purposes are not clear enough to warrant a standard. Furthermore, even for well-defined purposes, the transit agency may realize that answering questions requires some search/research. The latter case indeed deserves well-founded research to justify the use of standards in terms of their range. Figure 1.5 outlines, for each standard, whether it is necessary to undergo research or whether it is sufficient to determine its criteria administratively. Six standards necessitate research, three standards could be administrative-based, and research findings can be input for administrative decisions for the remaining 11 standards.

The research-based column in Figure 1.5, lists the recommended objective for 17 of the 20 standards. The administrative-based column shows the recommended criteria for 15 of the 20 standards, and the remarks column introduces some helpful notes for each analysis recommended.

Finally, other chapters illustrate some of the required research findings and methodologies for certain important service standards. Chapter 3 demonstrates standards 10 (load-crowding level), 11 (standees), and 12 (headway, upper limit). Chapter 6 shows standard 17 (timed transfer). Chapters 13 and 14 show standards 1 (route length), 3 (route directness), and 15 (transfers). Chapter 15 illustrates standard 4 (short-turn), and Chapter 17 illustrates standard 16 (schedule adherence).

1.3.2 Evaluation standards

An assessment of ridership productivity and financial performance of any transit agency largely relies on five variables, determined on a route basis: (i) vehicle-hours, (ii) vehicle-km, (iii) passenger measures, (iv) revenue, and (v) operating cost. These five variables form the base for seven economic and productivity standards that are in use in the US and Europe (TCRP, 1995; METRO, 1984; QUATTRO, 1998): (i) passengers per vehicle-hour, (ii) passengers per vehicle-kilometer, (iii) passengers per trip, (iv) cost per passenger, (v) cost-recovery ratio, (vi) subsidy per passenger, and (vii) relative performance.

The main evaluation standards currently utilized can be divided into two categories: (i) passenger-based and (ii) cost-based; the first relates to ridership productivity criteria, and the second to financial criteria. These seven standards are given in Figure 1.6 by category, number, standard name, typical criteria range and remarks. Two more abbreviations, for vehicles (veh) and for kilometres (km) have been added to Figure 1.6. Whereas the service standards are self-explanatory, the evaluation standards need interpretation. Standard 1 in Figure 1.6, pass/veh-hour, is the most widely used productivity criterion, mainly because of the fact that the operating budget is paid out on an hourly basis. It is based, to the greatest extent possible, on unlinked passenger trips (i.e. each boarding adds one to the amount of passengers) and service (revenue) hours. Standard 2, pass/veh-km, reflects the number of riders boarding a vehicle along a unit of distance rather than a unit of time; it, too, is based on unlinked passenger trips and on service kilometers. Standard 3, pass/trip, is the number of boarding passengers per single (one-way) trip; its advantage lies in its simplicity. Standard 4, cost/pass, is a financial criterion attempting to ascertain the productivity of a route. Standard 5, cost-recovery ratio, is the ratio between direct operating costs (wages, benefits and maintenance costs) and the share recovered by the fares paid by the route's riders. Standard 6, subsidy/pass, is usually the difference between cost/pass and revenue/pass. The

No.	Standard item	Research-based suitability	Administrative-based suitability	Remarks
1	Route length	Max utilization	_____	User and operator variables
2	Stop spacing	Max utilization	_____	User and operator variables
3	Route direction*	Demand sensitivity to route deviation	Max % deviation from car's shortest path	Demand contains potential users
4	Short-turn*	Max vehicle saving with min short-turns	_____	Operator and user variables
5	Route coverage	_____	Max walking distance to stop for majority of residents	Adequate route-spacing will result
6	Route overlapping	Max utilization	_____	To comply with O-D demand
7	Route structure*	Max utilization	_____	To comply with O-D demand
8	Route connectivity*	Max utilization	_____	To allow for all O-D movements
9	Span of service	_____	Min operation hours, by day-of-week	Can be extended through lowering the frequency
10	Load (crowding) level	Determine acceptable criteria	Max load at the max load-point, by time-of-day	Flexible range for certain loading figures
11	Standeers*	Determine acceptable criteria	Max standees as % of seats for each vehicle type	Flexible range for certain loading figures
12	Headway upper limit*	Determine criteria	Max (policy) headway, by time-of-day	Depends on previous and next headways
13	Headway lower limit	Determine criteria	Min headway, by time-of-day	Depends on size of vehicles
14	Transfers	Determine acceptable criteria	Max number of transfers for a given O-D	Depends on transfer smoothness and waiting time
15	Passenger shelters*	Determine criteria	Min daily boarding to warrant shelter	Site-specific (% of elderly, handicapped, etc.)
16	Schedule adherence	Determine criteria through modelling	Min 'on-time' performance	Depends on real-time information provided
17	Time transfer	Determine acceptable criteria	Max waiting time at transfer points	Depends on real-time information provided
18	Missed trips*	Determine criteria	Max allowed % of missed trips	Vehicle breakdown separate from reliability problems
19	Passenger safety*	Determine acceptable criteria	Max pass-accidents per pass and km driven	Depends on safety data
20	Public complaints	_____	Max number of complaints per driver/pass/time period	Unimportant separate from serious complaints

* Standards commonly in use.

Figure 1.5 Suitability of standards to be determined by research results, administrative decision, or both

revenue/pass criterion reflects different fares and is used as a measure for one of the comparisons of routes. The last standard in Figure 1.6 is the relative performance of a route compared to other routes usually having the same characteristics, such as its percentile rank in either an overall ranking of system routes or in a group of routes associated with the same

Category	No.	Standard item	Typical criteria ranges*	Remarks
Passenger-based	1	Passengers per** veh-hour (PVH)	Min of 8–40 PVH; Min 50–100% systemwide PVH average	Depends on type of service, time-of-day and day-of-week
	2	Passengers per** veh-km (PVK)	Min of 0.6–1.5 PVK; Min 60–80% systemwide PVK average	Depends on type of service, time-of-day and day-of-week
	3	Passengers per trip	Min 5–15 riders per trip; Min 15 pass average load for all routes	Min average load on express trip, where this standard is most useful, is 20–30 passengers
Cost-based	4	Cost per passenger	Max 1.4 of system average	This standard is often included in a composite score with other criteria
	5	Cost-recovery ratio**	Min 0.15–0.30 ratio; Min 1.0 ratio for express-type services	Larger agencies differentiate between types of services and use a composite score
	6	Subsidy per passenger**	Based on revenue per pass. with Min 25–33% of system average	Used as a difference between cost and revenue, it is simpler than cost to explain to the public
	7	Relative performance	Min 10–20% across all routes of a composite productivity score	Route performance is measured against that of other routes

*Reflects data mainly from the US

**Standards commonly in use

Figure 1.6 List of evaluation standards and their typical criteria range

type of service. The exact measure of this ranking varies across transit systems. Figure 1.6 shows typical criteria ranges for each evaluation standard that can be used informally or in a more formal manner. The Remarks column in Figure 1.6 provides extra information on each standard that will enhance implementation simplicity.

1.4 Viability perspectives

Following the discussion of the framework and derivatives of service and evaluation standards that focus on tools to set adequate and improved transit service, a more general view of how the basic transit goals should be approached is presented. Current practice shows that transit agencies are experiencing increasing fiscal pressures caused by a decline in transit patronage, increased operational costs, and decreased government support. In response to this trend, transit agencies have more often been re-examining the manner in which their limited resources are allocated. This concentration on savings should not overlook some amenities (Ceder, 2002) that are concerned with the viability of service.

The decline in transit patronage is the result of two main factors: (a) poor level-of-service and (b) better competitors. New roads, bridges and tunnels serve the automobile, and to some extent the railroad, whereas the investment in transit enhancements has been at a relatively much lower level. On one hand, there is no need to promote transit service in a free market environment; on the other hand, transit has the best land-transportation safety record, and it can relieve some traffic congestion, as well as help preserve the environment.

Transit viability can be looked at through the perspectives of the passenger and the agency. Orderliness is the key to success. Figure 1.7 shows three cases when passengers will

use transit service: (i) when there is no alternative, (ii) when the service offers more comfort than does the automobile, and (iii) when the service offered is acceptable. Acceptable services, as shown in Figure 1.7, rely basically on reasonable and acceptable travel time, walking distance to and from the transit stop, waiting time, tariff, transfers, and timetable adherence, as well as seat availability and reliability and availability of information. Overall walking distance in the O-D or door-to-door chain and comfortable transfers are central points in understanding passenger travel behaviour, which has to do with the integration of transit service end points and the passenger’s origin and destination points. Comfortable transfers (vehicle to vehicle, using either the same transit mode or different modes) ought to be smooth (escalator, same platform, etc.) and synchronized (a match between arrival

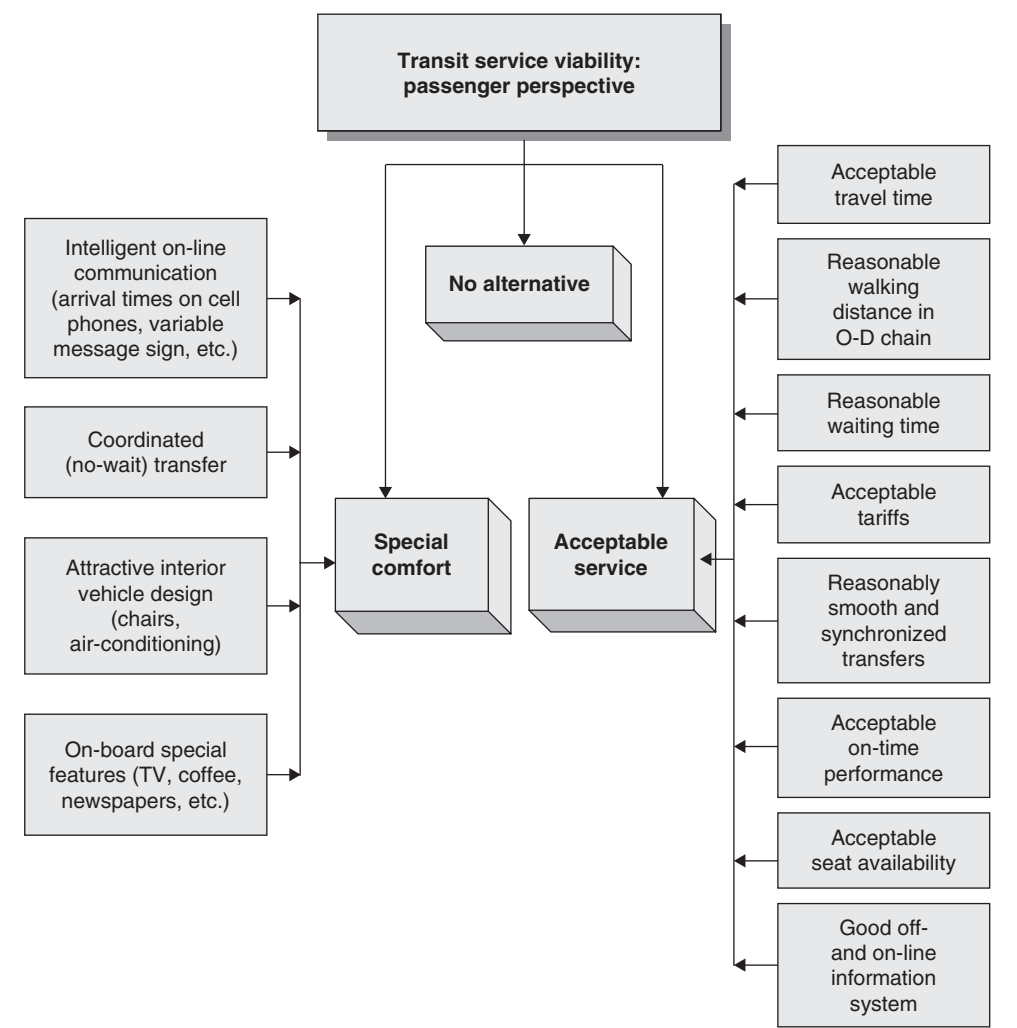


Figure 1.7 Viability of transit service from passenger perspectives

and departure times as stated by on-line information). Case (iii) in Figure 1.7 refers to passengers who regularly use their automobile, but would switch to transit if given special comfort features. Such features can include intelligent on-line communication, coordinated transfers without waiting, attractive transit vehicle interior design (à la advanced aircraft), and on-board services (again à la aircraft), none of which can be obtained in an ordinary automobile.

Figure 1.8 presents the agency perspective for viability. The provision of subsidy and commercial concessions is self-explanatory; however, the approaches to attain increased patronage and revenue and a reduction in operating costs deserve some elaboration. Increased patronage and revenue can be realized through remedying the current transit service illness; that is, obtaining a better match between service and demand, reliable service, coordinated network (à la network of routes in a good metro system), and improved comfort and information. Chapter 4 presents some remedies for a better match between service and demand; for solving reliability problems see Chapter 14; and for coordination among transit routes see Chapters 5 and 12. Decreased operating costs can be attained by reducing fleet size (less capital cost), total drivers' wages, and the average travel time in the transit network system without changing the service (timetables) offered. Techniques and methods to reduce fleet size are introduced in Chapters 6, 7, 8 and 13; to reduce total drivers' wages in Chapter 9, and to reduce travel time through a better match between service and demand in Chapter 4.

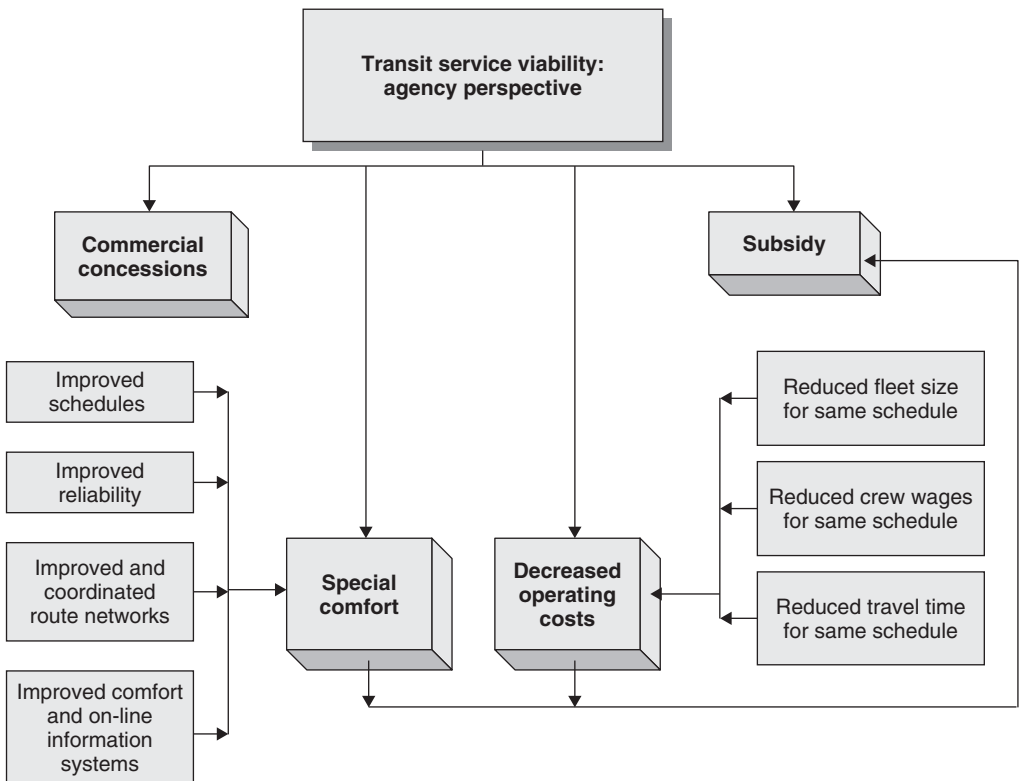


Figure 1.8 Viability of transit service from agency perspectives

1.5 Outline of other chapters

This introductory chapter has emphasized four major activities, as seen schematically in Figure 1.2, as the core of transit planning. Figure 1.9 describes the remaining themes of the book around these four activities. Each block in Figure 1.9 lists the chapter(s) dealing with the content specified.

Chapter 2 covers transit data-collection systems including manual and automated techniques, automatic vehicle monitoring sampling considerations, and some notes on passenger surveys. Chapter 3, which is part of the second activity in Figure 1.9, analyses passenger-load and running-time data, four methods of frequency and headway determination, and cost-effectiveness criteria for data gathering. Chapter 4 (part of the second activity in Figure 1.9) introduces current practices in timetable construction and optional timetables. The optional-computerized timetables are characterized by evenly spaced headways or by average even loads, with unevenly spaced headways, at the hourly maximum load point.

Chapters 5 and 6 present advanced methods of constructing timetables. One of these (Chapter 5) creates timetables with even passenger loads at different maximum load points for individual vehicles as opposed to a single maximum load point (across all vehicles in one hour). Chapters 5 and 6 tie in with the second planning activity in Figure 1.9. A second

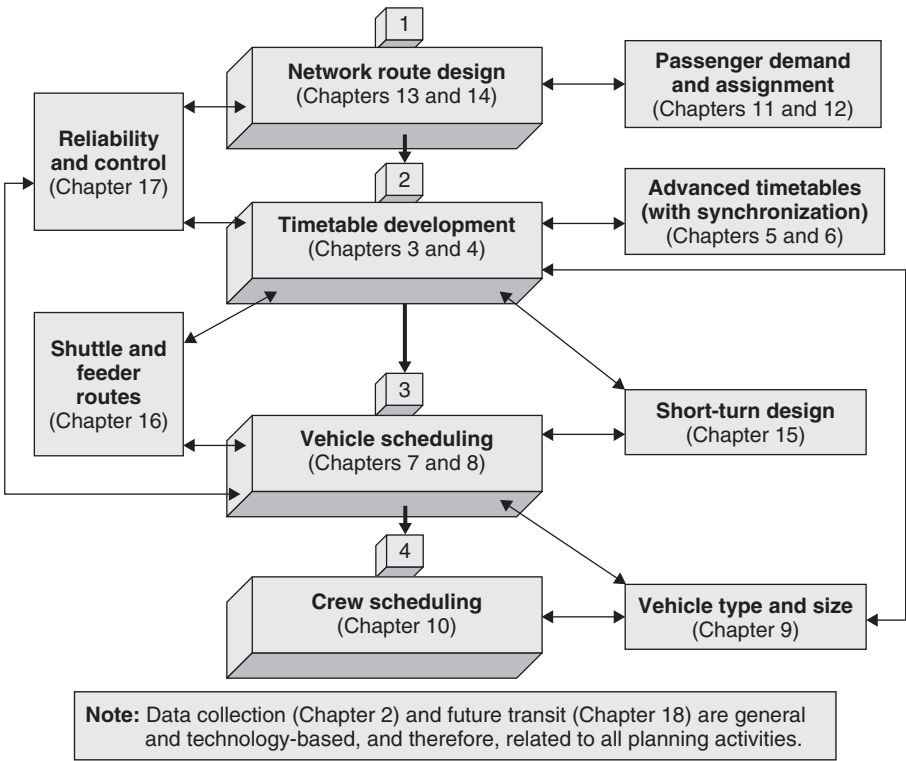


Figure 1.9 Relationship of the main four planning activities to the remaining book themes

method, in Chapter 5, allows an examination of optimal timetables for a reduced fleet for a network of transit routes. Such a problem can arise in a practical context when the fleet size is reduced because of vehicle-age attrition and budget-policy decisions. A third method, in Chapter 6, describes optimal procedures for designing timetables with maximum simultaneous arrivals at given stops. This synchronization solution is based on a given range of the headways determined.

Chapter 9 introduces different vehicle types with scheduling requirements, examines various scheduling scenarios, and presents a new method to create the best schedules with vehicle-type constraints. In addition, Chapter 9 describes different models to determine the required size of a vehicle. This chapter ties in with both the second and third activity in Figure 1.9. Chapter 10, which is the fourth activity in Figure 1.9, demonstrates some of the optimization concepts behind the problem of assigning crew (mainly drivers) to vehicle schedules with the minimum cost involved. Following a review of some existing models, we present a method to create a good framework for optimal duties. This chapter interacts to some extent with procedures described in the previous chapter for different vehicle types.

Chapters 11 and 12 provide an overview on passenger demand, route choice and assignment. In Chapter 11, we give an overview of the factors affecting passenger demand and their sensitivity (elasticity) and methods designed to predict passenger demand; we also show estimation methods for predicting O-D matrices and how best to forecast ridership. Chapter 12 presents passenger dilemma in route choice among alternative routes; the route choice is then incorporated into transit-assignment modelling at the network level. The modelling in Chapters 11 and 12 interacts with the first activity in Figure 1.9.

Chapters 12, 13 and 14 constitute the first planning activity outlined in Figure 1.9. Chapter 12 establishes objective functions for designing a network of transit routes while complying with passenger, agency, and government perspectives. This chapter also presents basic service-design elements for different purposes and introduces a new method for dealing optimally with operational parking limitations. This on-street bound on maximum parking spaces calls for solutions to avoid traffic congestion at some route end points. Chapters 13 and 14 delivers a complete solution for the creation of transit routes while taking into account the activities of vehicle scheduling and timetabling. These chapters construct routes and transfers, assigning O-D demand and initial frequencies, and provide optimal criteria and best route-network solutions for decision-makers.

Chapter 15 defines a cost-effectiveness approach for operating single routes with short-turns (not all trips start and end at the end points of the route). Employing some of the developments resulting from the second and third activities in Figure 1.9, this chapter provides an optimal method to maintain the required level of service while minimizing the fleet size through creating timetables with minimum short-turn trips and minimum required vehicles. Chapter 16 presents the need for and the concept behind an integrated transit system, with an emphasis on short-hand transit service (shuttle and feeder routes). This integrated service is presented as an attractive, reliable, rapid, smooth and synchronized system. Ten different routing strategies are developed for bus shuttle and feeder routes while using a simulation tool for analysis along with real-time communication possibilities. Chapter 16 interacts with the second and third activities in Figure 1.9.

Chapters 17 and 18, especially the latter, focus on new technology principles to assist in improving transit-operations planning. Chapter 17 discusses the essential part of transit reliability and its impact on operations planning. It covers the subjects of variability of concern

among passengers and in the agency, the bus-bunching phenomenon, methods to improve reliability, the calculating of passenger waiting time and vehicle running time, and features and benefits of AVL (automatic vehicle location) systems. Chapter 17 interacts with elements of the first three activities in Figure 1.9. Finally, in the last and generalized section, Chapter 18, we present a bridge between new technologies and transit operations, the concept of flexible routing based on distributed computing and electronic user-operator communication. The book ends with an outline of some future research needs.

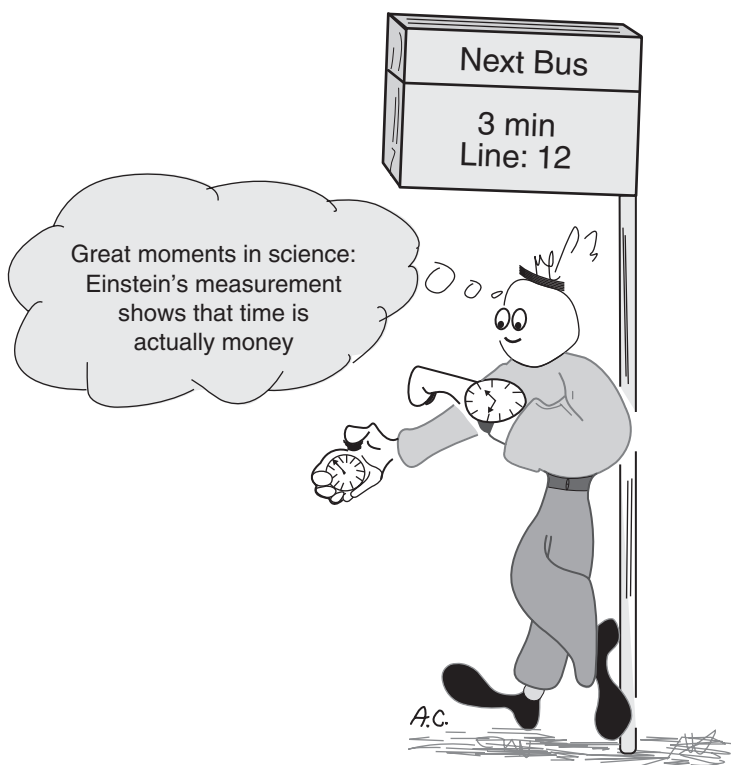
References

- Ceder, A. (2001). Public transport scheduling. In *Handbooks in Transport – Handbook 3: Transport Systems and Traffic Control*. (D. Hensher and K. Button, eds), pp. 539–558. Elsevier Ltd.
- Ceder, A. (2002). Urban transit scheduling: framework, review, and examples. *ASCE Journal of Urban Planning and Development*, **128** (4), 225–244.
- Ceder, A. and Wilson, N. H. M. (1986). Bus network design. *Transportation Research*, **20B** (4), 331–344.
- Daduna, J. R. and Wren, A. (eds) (1988). *Computer-Aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **308**, Springer-Verlag.
- Daduna, J. R., Branco I. and Paixao, J. M. P. (eds) (1995). *Computer-Aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **430**, Springer-Verlag.
- Daduna, J. R. and Paixao, J. M. P. (1995). Vehicle scheduling for public mass transit – an overview. In *Computer-Aided Transit Scheduling*, (J. R. Daduna, I. Branco and J. M. P. Paixao, eds), pp. 76–90, Springer-Verlag.
- Desrochers, M. and Rousseau, J. M. (eds) (1992). *Computer-Aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **386**, Springer-Verlag.
- Hickman, M., Voss, S. and Mirchandani, P. (eds) (2007). *Computer-Aided Scheduling of Public Transport*. Lecture Notes in Economics and Mathematical Systems, Springer-Verlag (forthcoming).
- METRO (1984). Metropolitan Transit Authority of Harris County. *Bus Service Evaluation Methods: A Review*. Urban Mass Transportation Administration, Washington, DC, DOT-1-84-49.
- Mora, J. G. and Chernicoff, W. P. (2005). Transit's complex route to improved standards and codes. *TR News*, Transportation Research Board, **236**, 3–7.
- Pine, R., Niemeyer, J. and Chisholm, R. (1998). Transit scheduling: Basic and advanced manuals. *TCRP Report 30*, Transportation Research Board, Washington, DC.
- QUATTRO (1998). Quality approach in tendering urban public transport operation in Europe. *Transport Research Fourth Framework Programme*, Urban Transport, VII-51, Office for Official Publications of the European Communities.
- Rainville, W. S. (1947). *Bus Scheduling Manual: Traffic Checking and Schedule Preparation*. Reprinted 1982, American Public Transit Association, US Department of Transportation, DOT-1-82-23.

- Rousseau, J. M. (ed.) (1985). *Computer Scheduling of Public Transport 2*. North-Holland Publishing Co.
- TCRP (1995). *Bus Route Evaluation Standards* (Report), Synthesis of Transit Practice 10, Transit Cooperative Research Program, Transportation Research Board, Washington, DC.
- Voss, S. and Daduna, R. (eds) (2001). *Computer-Aided Scheduling of Public Transport*. Lecture Notes in Economics and Mathematical Systems, **505**, Springer-Verlag.
- Wilson, N. H. M. (ed.) (1999). *Computer-Aided Scheduling of Public Transport*. Lecture Notes in Economics and Mathematical Systems, **471**, Springer-Verlag.
- Wren, A. (ed.) (1981). *Computer Scheduling of Public Transport: Urban Passenger Vehicle and Crew Scheduling*. North-Holland Publishing Co.

This page intentionally left blank

2 Data Requirements and Collection



Chapter 2 Data Requirements and Collection

Chapter outline

- 2.1 Introduction
 - 2.2 Data-collection techniques
 - 2.3 Data requirements
 - 2.4 Basic statistical tools
 - 2.5 Literature review and further reading
 - References
-

Practitioner's Corner

This chapter offers a number of potential benefits to transit agencies in the areas of data acquisition, data merits and data utilization. The major objective is to help develop a comprehensive, statistically based data-collection plan that will enable transit agencies to collect proper data and the right amount in a cost-effective manner.

The chapter starts with the three core keys representing the objectives of this book: (1) to know, (2) to plan and decide wisely, and (3) to operate in a smart manner. Achieving the first key, through the gathering of comprehensive data, is the objective of this chapter. The first section describes the various data-collection techniques and methods, including definitions and interpretations. The second section continues with the linkage of the data-collection method, the results expected from the data analysis, and enhanced service elements gained from the results. These two sections of the chapter are most suitable for practitioners, who usually understand the difficulties and conflicts involved in real-life data collection, as shown in the following account.

While trying to maximize their comfort, drivers are more sophisticated than one can imagine. The following account concerns gaining extra planned travel time by arriving early and having extended layover time. The checkers in a bus agency who boarded the vehicle used to announce to the driver that they were measuring average travel time with a rate factor. If the bus is too slow (i.e. encountering more overtaking vehicles than it overtook), the rate is below ten, otherwise more than ten. One driver found out when this check was going to take place and asked members of his family to drive their cars at a slow pace just in front of his bus in order for him to earn a rate of below ten and . . . extended planned travel time.

The third and longest section of this chapter introduces basic statistical tools that are employed in almost every phase of the data-collection undertaking. There is no prerequisite for understanding the statistical tools, although practitioners may want to skip the mathematical formulations and concentrate on the examples.

The chapter ends with a literature review and recommendations for further reading. Overall, the data-collection effort should be handled professionally, with a precise understanding of the purpose of each data element. Finally, since we want high-quality data, but not in excessive amounts, it is good to recall what Galileo said: "Science proceeds more by what it has learned to ignore than what it takes into account".

2.1 Introduction

Current practice in transit agencies shows that sufficient data seldom exist for service-operations planning. Manual data-collection efforts are costly and, consequently, must be used sparingly. Automated data-collection systems, although growing rapidly, are not yet perfectly linked to the requirements of planning data. Extraction of data from an automated system, which is ostensibly a simple task, may turn out to be rather complex. Having too much data is often as bad as having too little. The only concept we can trust is that in any collection of data, the figure most obviously correct beyond all need for checking is . . . the mistake. At the same time, the data are essential for responding to basic passenger needs; namely, the route: Where is the closest stop? What time should I be at the stop? The data are certainly crucial, too, for responding to the operations-planning needs of each transit agency: How can the network of routes, stops and terminals be improved? How can each route be improved? What is the best timetable to deliver? How can fleet size be minimized while maintaining the same level-of-service? How can crew cost be minimized without service changes? No doubt, a common element for all transit agencies is their pursuit of data to aid in improving the efficiency, productivity and effectiveness of their systems.

This chapter demonstrates data requirements and applications for enhancing transit-service productivity and efficiency. Fundamentally, there are three keys for achieving a successful transit service: (i) gathering and comprehending adequate data, (ii) using the data collected for intelligent planning and decisions, and (iii) employing the plans and decisions for astutely conducted operations and control. These three keys appear schematically in Figure 2.1. This chapter, which introduces the framework for attaining the first key, consists of four parts: first, different data-collection techniques; second, the data requirements, with attention paid to their coordination; third, basic tools in statistics needed to comprehend the mechanism of data analysis; fourth, a literature review and list for further reading.

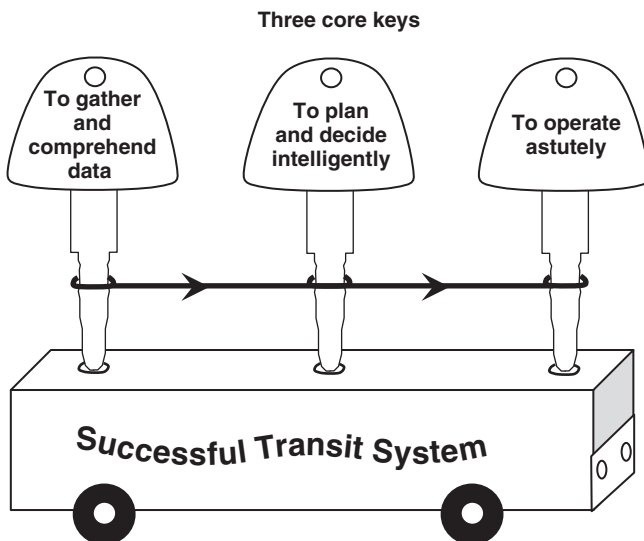


Figure 2.1 *The three important keys for a successful transit system*

2.2 Data-collection techniques

Transit-data-collection techniques required for operations planning can be divided into three categories: (i) manual-based methods, (ii) automated-based methods, and (iii) AVL-based methods (AVL = automated vehicle location). Fundamentally, there are only two types of methods, manual and automated. However, AVL or AVM (automated vehicle monitoring) systems give more accurate information, especially in time and space, than do other item-specific automated methods and, therefore, can be looked at as a separate category. The AVL-based technique will be covered in Chapter 17 in connection with linking the technique to remedies for reliability problems.

Five primary techniques for collecting transit data may be identified: point check, ride check, deadhead check, passenger survey and population surveys.

2.2.1 Point check

Point check is usually described as counts and measurements performed by a checker stationed at a transit stop. The stop selected is virtually the maximum (peak) load point, at which the transit vehicle departing this stop has, on average, the maximum on-board load across all route segments. A route segment is defined as a section of the route between two adjacent stops. For each vehicle passing the stop, the point check usually contains load counts, arrival and departure times, and vehicle and route identifications. Other point-check locations than the peak stop are more (multiple) peak points, end points and strategic points. Multiple point checks accommodate situations in which there are simultaneously several peak points and a situation of long routes and branching routes, in which a branching route is one that stretches along the base route and adds a certain branch. End point checks accommodate running-time measurements and, if applicable, record farebox readings. Strategic point checks are useful for item-specific checks, such as at major transfer points (measuring transfer time and successful vehicle meetings), major activity centres (observing passenger behaviour in selecting a competitor mode), and new neighbourhoods (measuring changes in passenger demand).

2.2.2 Ride check

Ride check refers to counts and measurements performed by either a checker riding the transit vehicle along the entire route or an automated instrument (hence, replacing the human checker). The ride check contains mainly on and off passenger counts, from which one can derive the on-board passenger load for each route segment, arrival and departure times for each stop, and sometimes item-specific surveys or measurements (vehicle running speed, boarding by fare category, gender of passengers and baggage size), and record farebox readings. The common automated instrument for ride checks, called APC (automated passenger counter), can perform the main ride-check tasks. It cannot, however, replace the checker in counting boarding by fare category and in surveying passengers. A special ride check is one performed by the operator (driver). It usually involves the interaction of the driver and a farebox, with the driver inserting into the machine information related to fare category and O-D per passenger. The action, however, increases the time spent at the stop.

Figure 2.2 depicts passenger demand as interpreted by hourly passenger load, using both point-check and ride-check methods. This is a vital data element affecting operations planning. From point check, only the load for the peak segment can be obtained; with ride check, the entire load profile by either time or space can be shown. Figure 2.2 will be used as space-based (passenger-km) illustrations in the next chapter, which is concerned with a determination of vehicle frequencies and headways. The time-based (passenger-hour) load profile will be used

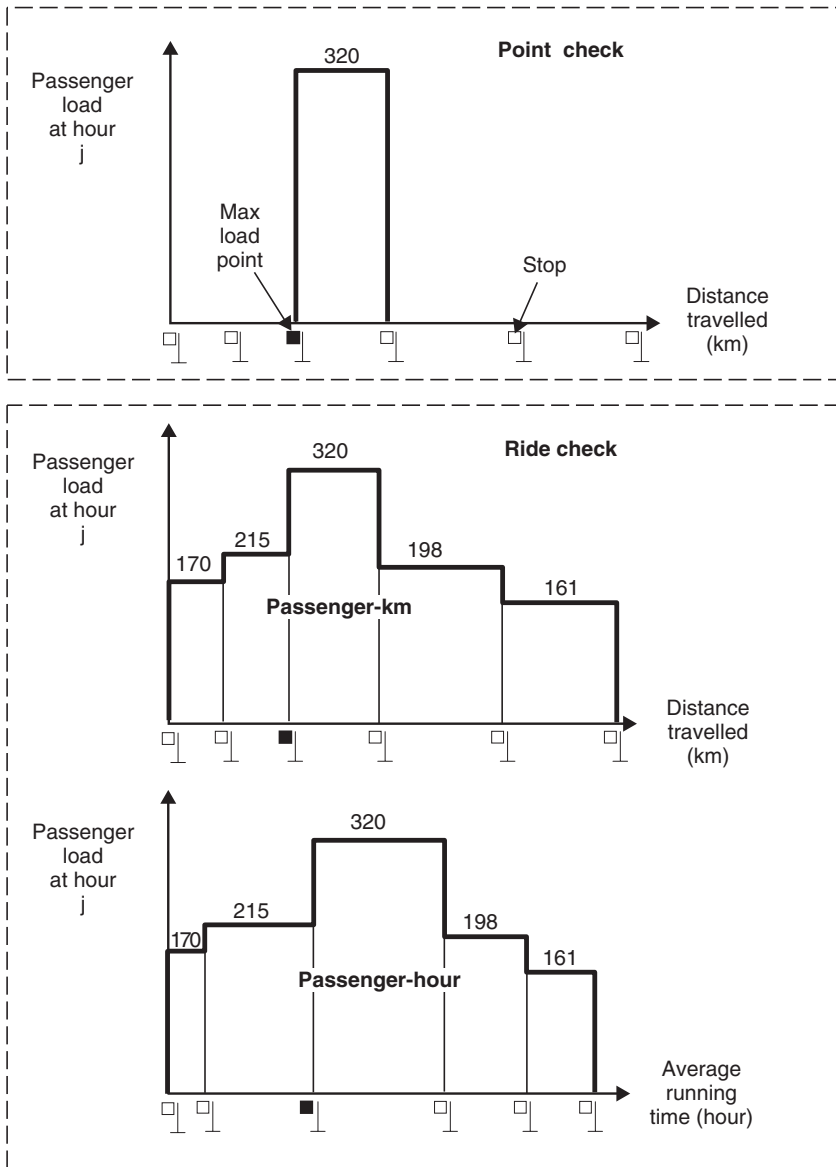


Figure 2.2 Main information obtained for operations planning through point check and ride check

in Chapters 13 and 14, which deal with route-design service and network. Figure 2.2 shows the location of the maximum (max) load point (stop), from which 320 passengers are transported during hour j along the maximum (peak) load segment. In addition, Figure 2.2 demonstrates the different visual appearance of the space and time-load profiles. Whereas the x-axis of the space-based profile is fixed for a given route direction, the x-axis of a time-based profile is set according to average values.

2.2.3 *Deadhead check*

Deadhead check refers to the average vehicle running time between an arrival point on one route and a departure point on another route. This deadheading time is required in a transit system with interlining routes. It is measured mainly by agency cars travelling along the shortest path (in time) between the two route end points. This shortest path varies by time of day, day of week and type of day.

2.2.4 *Passenger survey*

Transit passenger surveys are conducted in essence while directly confronting the passengers. The known survey methods of this type are on-board, at stop, at terminal, and mail-back (postage-free forms). The most common is the on-board survey. All the surveys are carried out by agency checkers or drivers or by especially trained staff who either distribute forms to fill out or ask questions in person. One way to increase the response incentive is to hand out a symbolic gift; e.g. a good pen or key-holder. Such a token may open up more opportunities for cooperation as the gesture is appreciated. Basically, there are general surveys and special purpose surveys. In a general survey, multi-type information is obtained, such as O-D, access and egress modes and distances, trip purpose, routes selected on a trip, fare paid, type of payment, frequency of use by time of day, and socio-economic and attitude elements. Special purpose surveys aim at eliciting only *one type* of information, such as O-D, opinion of service changes, transfer activities, pass-holder usage, attitude toward possible fare changes, or the proportion of different fare types (adult, student, free passes, transfer and special fare).

2.2.5 *Population survey*

Generally the population surveys are conducted at the regional level at home, shop or workplace. These surveys are usually interview-based, involving both transit users and non-users, in order to capture public attitudes and opinions about transit service changes (including the impact on household location decisions), fare changes, and transportation, traffic and land-use projects. The population interviews of users and non-users also address the vital issues of potential ridership, market segmentation, market opportunities and suggestions for new transit initiatives.

Further description of some of the data-collection techniques can be found in the transit-data-collection manual (UMTA, 1985). With this background of transit-data-collection techniques and methods, two important questions can be answered: (1) What data are required for enhancing specific service components? (2) How can we best exploit and analyse the data? The next section covers these issues.

2.3 Data requirements

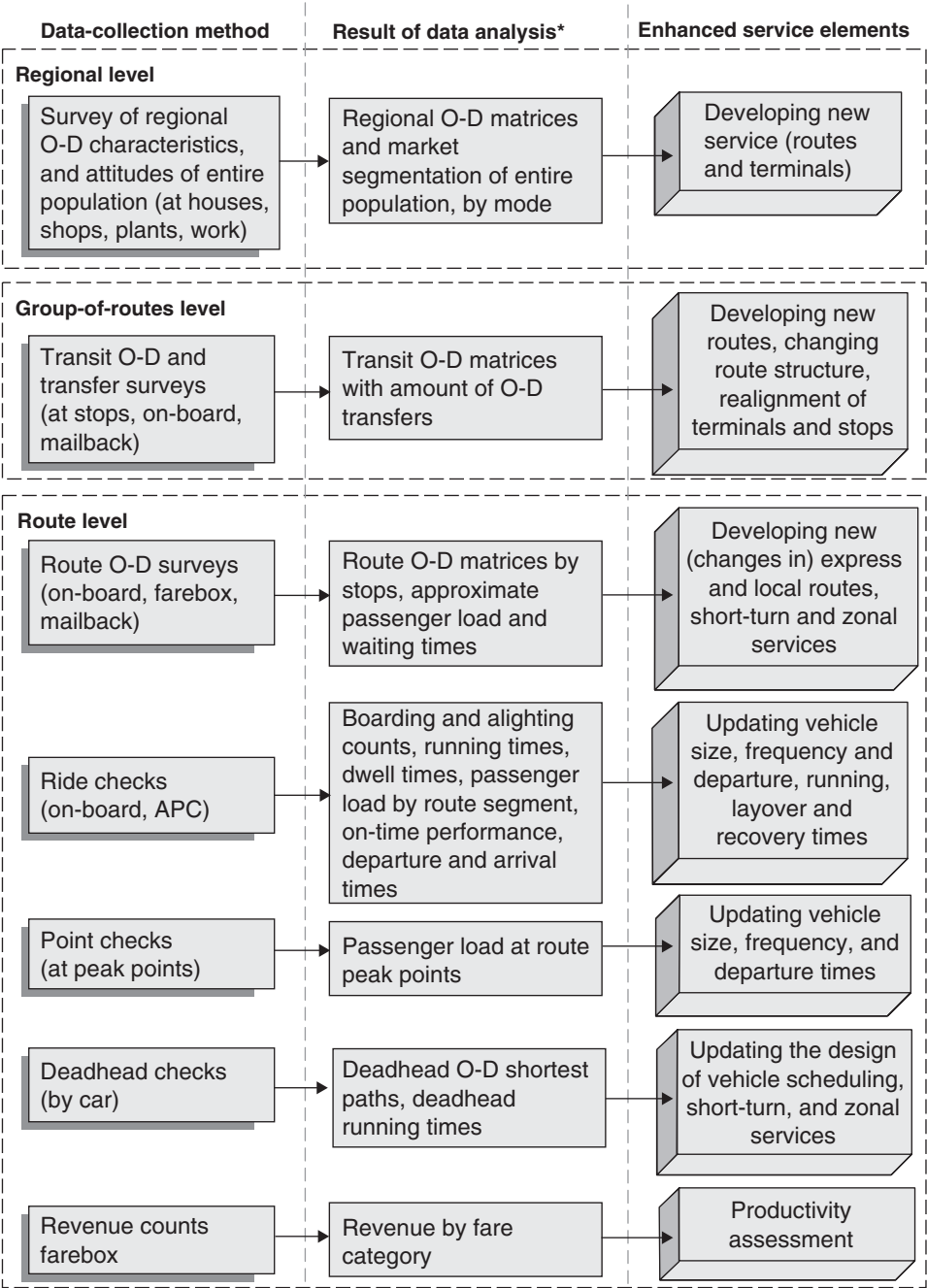
This section focuses on the need for certain data elements in order to provide an opportunity for achieving service enhancements. Generally, detailed information on passenger demand and service characteristics is not available at the route level. Without this information, however, the efficient deployment of transit service commensurate with demand is impossible. This is an example of why I emphasize that important data elements be acquired so as to: (1) obtain badly needed data and drop insignificant data, and (2) gain an overview of the entire data requirements for attaining a reduced-cost data-collection system. Since a cost appraisal of the data-collection effort is ostensibly high, let us emphasize Einstein's saying: "Everything should be made as simple as possible, but not simpler".

There is a common thread of data needs across all transit agencies. This thread exists as long as agencies share these objectives: (a) improving service and operations, (b) improving productivity and efficiency by better matching supply and demand, (c) improving levels of service through increased reliability as a result of better control and response, and (d) reducing data-gathering, processing, and reporting costs. Broadly speaking, the data items gathered by the techniques described above are useful for one or more aspects of route and service design, scheduling, information system, marketing, deficit allocation, monitoring management and external reporting. This section will show, however, only the data elements pertaining to operations planning.

Figure 2.3 portrays in flowchart form three operations-planning categories: (i) an adequate data-collection method, (ii) the results required from a data analysis, and (iii) relevant, enhanced service elements. Clarification of data-collection methods appear in the previous section. Figure 2.3 refers to three planning levels: regional, group-of-routes and route. At the regional level, we focus on developing new or improved transit service by means of routes and terminals. To accomplish this, the entire regional population O-D matrices and market segmentation by mode of travel is needed. Chapters 11–14 cover some aspects of this planning undertaking. At the group-of-routes level, we focus solely on transit service with the objective of creating new and improved routes, stops and terminals. This differs from the regional level by the treatment of a specific group of routes characterized by their own O-D matrices and transfer activities.

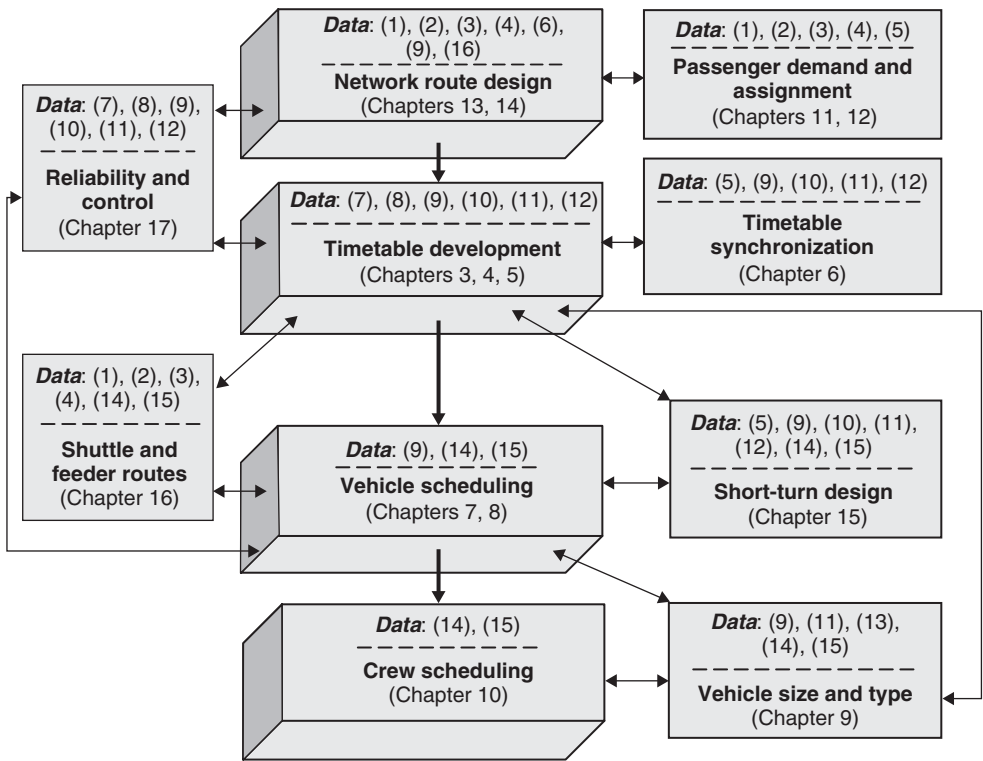
At the route level in Figure 2.3, there is the bulk of enhanced service elements. We start with route O-D matrices, approximate passenger loads, and waiting times in order to develop or improve express and local routes, as well as short-turn and zonal services. The analysis of ride check data yields the next set of enhanced elements, which includes the evaluation of vehicle size, determination of frequencies and headways, and the construction of timetables with their accompanying parameters. The third set of enhanced elements is based on point-check data analysis for updating vehicle sizes; e.g. mini, standard or articulated buses; frequencies and headways based on peak point counts; and their derived timetables. The fourth set of elements relies on O-D average running-time information pertaining to deadheading trips, together with a description of O-D shortest paths. The last set of enhanced elements, which is revenue-based, assesses productivity measures. The procedures and modelling to attain these route-level enhanced service elements appear in Chapters 3–9 and 12–17; that is to say, in the majority of the chapters in this book.

All together, there are 16 crucial data elements required for transit operations planning. The lower part of Figure 2.4 lists these 16 data elements. Figure 2.4 presents the abstract



*By time of day, day of week, type of day

Figure 2.3 Data collection methods and resultant analysis and service elements

**Key to data elements:**

(1) Regional O-D matrices	(9) Running times
(2) Market segmentation by travel mode	(10) On-time performance
(3) Transit O-D matrices	(11) Vehicle loads by route segment (between stops)
(4) Transfer counts	(12) Actual departure and arrival times
(5) Route O-D matrices by stop	(13) Vehicle loads at route peak points
(6) Approximate vehicle loads	(14) Deadhead O-D shortest paths
(7) Approximate passenger waiting times	(15) Deadhead running times
(8) Passenger boarding and alighting counts	(16) Revenue by fare category

Figure 2.4 Data requirements by book chapter themes

relationships among the four main planning activities depicted in the centre of the figure and the remaining themes of this book, citing the relevant data elements. This figure virtually follows the illustration in Figure 1.9, but with an emphasis on the data elements.

Two vital issues must accompany any data-collection system. The first is the *sample size* required; the second is *how often* the data should be collected. It is obvious that not all transit trips or transit passengers can be observed in the data-collection process. Since transit agencies usually collect only a fraction of the data, it is uncertain how well the sample data represents the entire population. The next section provides the fundamental statistical tools that can broadly cover the sample-size issue.

The second issue is the matter of how often should data be collected. We recognize that transit service exists in a dynamic environment, in which changes in passenger demand occur regularly. Transit agency road inspectors and supervisors are supposed to deliver information on noticeable changes that would warrant a new data-collection effort. Unfortunately we cannot rely on the drivers to do so, since by nature they try to maximize their comfort and will not report changes that may put more of a burden on their shoulders (e.g. reduced passenger load – lots of empty seats – that justify trip cancellations). Essentially there are data items that are recommended to be collected quarterly, and others annually or every few years. The data elements numbered 8–13 in Figure 2.4 are those that usually need to be collected quarterly. It should be noted that the statistical properties and characteristics of each data element – e.g. its variability – can shed light on the ‘how often’ issue as will be seen in the next section.

2.4 Basic statistical tools

Statistical techniques play an important role in achieving the objectives, of proper transit-data acquisition, handling and interpretation. *Merriam-Webster’s Collegiate Dictionary* (2000) defines ‘statistics’ as follows: “(1) a branch of mathematics dealing with a collection, analysis, interpretation of masses of numerical data; (2) a collection of quantitative data”. Basically the two definitions possess common elements implying a collection of data with inference or ‘inference-making’ as the objective. A large body of data is called *population*, and the subset selected from it is a *sample*. In any transit-data collection system, the objective of the statistics is to infer (estimate or decide) something about the population, based on the information contained in a sample.

2.4.1 Histogram and distribution

Consider a study the aim of which is to determine important variables affecting headways between transit vehicles moving on the same route and in the same direction. Among these variables are existing passenger demand, potential or hidden passenger demand, and a minimum level of service regardless of demand (see standard 12 in Figures 1.4 and 1.5). The transit planner will wish to measure these variables with the objective of utilizing the information in the sample to infer the approximate relationship of the variables with headways and to measure the strength of this relationship. Certainly the objective will be to determine optimum conditions for setting the best headways from different perspectives; that of the passenger, the agency or the government (community).

Any set of measurements can be characterized by a frequency histogram – a graph with an x-axis subdivided into intervals of equal width, with rectangles over each interval. The height of these rectangles is proportional to the fraction of the total number of measurements in each interval. For example, in order to characterize ten measurements of transit vehicle running times (in minutes) – say, 36, 52, 38, 44, 48, 39, 63, 28, 36 and 55 minutes – we could divide the x-axis into 10-minute intervals as is shown in the upper part of Figure 2.5. We wish in this example to obtain information on the form of the distribution of the data, which is usually mound-shaped. The larger the amount of data, the greater will be the number of intervals that can be included and still present a satisfactory picture of the data.

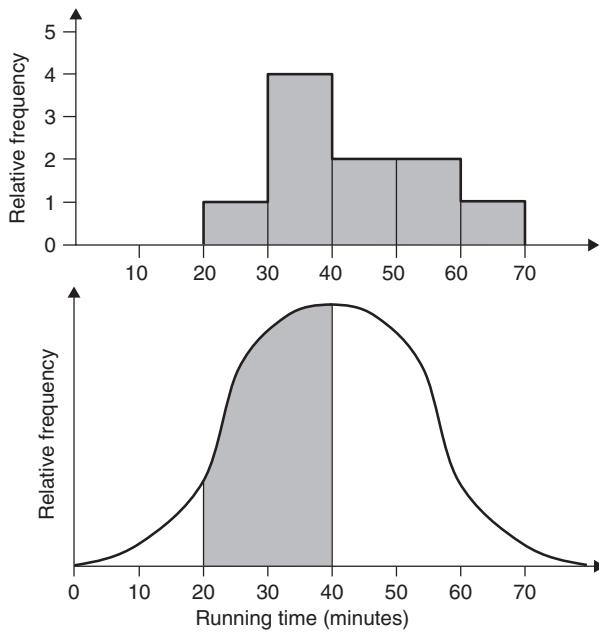


Figure 2.5 Relative frequency histogram and distribution of the running times in the example

In addition, we can decide that if a measurement falls on a point of division, it will belong to the interval on its left on the x-axis.

If it can be assumed that the measurements were selected at random from the original set, we can move one step further to a probabilistic interpretation. The probability that a measurement would fall in a given interval is proportional to the area under the histogram that lies over the interval. For example, the probability of selecting a running-time measurement over the 20 to 40-minute interval in Figure 2.5 is 0.5. This interpretation applies to the distribution of any set of measurements, including populations. If the lower part of Figure 2.5 gives the frequency distribution of running times (of a given transit route, direction of travel and time period), the probability that the running time will lie in the 20 to 40 interval is proportional to the shaded area under the distribution curve. In statistical terms, the distribution function for a random variable running time is denoted $F(t')$ and given by $F(t') = P(T' \leq t')$ for $-\infty < t' < \infty$, where $P(T' \leq t')$ is the probability of T' being less than or equal to the value of t' , and T' is continuous. The function $F(t')$ is called the distribution function (or cumulative distribution function), and its derivative (if existing) is $f(t') = dF(t')/dt'$. The function $f(t')$ is called a probability density function (pdf) for the random variable T' . The distribution that appears in the lower part of Figure 2.5 is $f(t')$.

Since many similar histograms could be formed from the same set of measurements, we need to rigorously define quantities with which to measure the sample data, n . The most common measure of a central tendency used in statistics is the *arithmetic mean* or *average*, the latter being the term used in this book. The most common measure of variability used in statistics is the *variance*, which is the function of deviations of the sample measurements from their average. The square root of the variance is called *standard deviation*. These three

measures – variance, deviation and standard deviation – appear in equations (2.1) and (2.2) for a set of n measured responses $x_1, x_2, \dots, x_n$, where $\bar{x}$, s^2 and s are the average, variance and standard deviation, respectively. The corresponding three measures of population are denoted μ , σ^2 , and σ .

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (2.1)$$

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (2.2)$$

For the example in Figure 2.5, we obtain: $\bar{x} = 43.9$ minutes, $s^2 = 98.578$ minutes-square, and $s = 9.929$ minutes. It may be noted that many distributions of transit data (as most other data in real life) are mound-shaped, which can be approximated by a bell-shaped frequency distribution. This is known as a *normal* curve, with a normal density function, and may be expressed as follows:

$$f(x) = \frac{e^{-(x-\mu)^2 / 2\sigma^2}}{\sigma\sqrt{2\pi}}, \sigma > 0, -\infty < \mu < \infty, -\infty < x < \infty \quad (2.3)$$

where e is the base of a natural logarithm ($= 2.7182818 \dots$), π is the ratio of the length of the circle to its diameter ($= 3.14159 \dots$), μ and σ are the average and standard deviation, respectively, and hence locate the centre of the distribution and measure its spread. Data possessing the normal type of distribution have the following empirical rule: $\mu \pm i\sigma$ contains approximately 68%, 95% and 99.7% of the measurements for $i = 1, 2, 3$, respectively. In practical terms, we find that most random variables observed in nature lie within one, two, or three standard deviations of their average, with probabilities in the vicinity of 0.68, 0.95 and 0.997, respectively.

2.4.2 Estimation

Let us now return to the objective of statistics – making an inference about the population, based on information contained in a sample. Usually this inference is about one or more population parameters. For example, we might wish to estimate the fraction of trips that would have running times of less than or equal to 20 minutes. Let T' equal the running time; the parameter of interest is $P(T' \leq 20)$, which is the area under the probability density function (lower part of Figure 2.5) over the interval $t' \leq 20$. Suppose we wish to estimate the average of T' and use the sample in Figure 2.5, with its average estimate of 43.9 minutes. One way to measure the goodness of this estimation is in terms of the distance between the estimate and its real (unknown) population value. This distance, which varies in a random manner in repeated sampling, is called the *error of estimation*, and we certainly would like it to be as small as possible. If this error is given in percentages, it is often called precision ($\pm$ given %). If it is in actual values (minutes for T'), it is often called tolerance ($\pm$ given minutes).

Assuming that we want this average running time of 43.9 minutes to have a precision of $\pm 10\%$, we then ask: What is the probability that the average of T' will fall in the interval between 39.51 and 48.29 minutes. This probability is called *confidence level* or *confidence coefficient*, and the interval is called *confidence interval*. The objective here is to find the tolerance that generates a narrow interval that encloses the average of T' with a high confidence level (probability) of usually 0.95, or 95%. A standard normal random variable with a normal distribution as expressed in Equation (2.3) is usually denoted Z . If we want to find a confidence interval that possesses a confidence coefficient equal to $(1 - \alpha)$ for Z , we can use the following probability statement:

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha \quad (2.4)$$

where P is the probability, and $z_{\alpha/2}$ and $-z_{\alpha/2}$ are two-tail end values of the standard normal distribution (the probability in between the tails is $1 - \alpha$). For instance, Z has a 0.95 probability ($\alpha = 0.5$) of being between -1.96 and $+1.96$ (based on normal distribution tables). In this case, Equation (2.4) becomes $P(-1.96 < Z < 1.96) = 0.95$. That is, $Z_{1-\alpha/2} = Z_{1-0.5/2} = Z_{0.975} = 1.96$. More interpretations and description of this subject can be found in numerous books on statistics, such as Washington *et al.* (2003).

The transit-data-collection manual (UMTA, 1985) contains a reference to the US national database specification. In this specification, the precision is $\pm 10\%$ and the confidence level is at 95% for reporting annual passenger-boarding counts and passenger-miles. This is a system-wide requirement by the Urban Mass Transit Administration (UMTA) for an average weekday and for Saturday and Sunday.

Other useful measures of the error of estimation described in the transit-data-collection manual (UMTA, 1985) are absolute tolerance (AT) and absolute equivalent tolerance (AET). AT and AET represent the error of estimation of a *proportion of observations* rather than the average of the data. The UMTA manual states as an example that schedule adherence is the proportion of trips lying in the categories Early, On-time, and Late. For instance, an estimated proportion of on-time performance may be 0.8 with $AT = \pm 0.1$. However, when the estimated proportion approaches 0 or 1.0, the AT resulting from a given sample size is small compared to the AT value for an estimated proportion in the vicinity of 0.5. Thus, the second measure, AET, is established as the absolute (equivalent) tolerance that can be attained with the following formula (UMTA, 1985) if the estimated proportion is 0.5:

$$AET = \frac{0.5AT}{\sqrt{p(1-p)}} \quad (2.5)$$

where p is the expected proportion.

2.4.3 Selecting sample size

The sampling procedure affects the quantity of information per measurement. This procedure, together with the sample size, n , controls the total amount of relevant information in the sample. Start with the simplest sampling situation, in which random sampling is conducted among a relatively large population. For an independent, continuous variable X , we define the average or the arithmetic mean as the expected value, $E(X)$, with a variance of $V(X)$.

It is known (see, for example, Washington *et al.*, 2003) that for $x_1, x_2, \dots, x_n$ independent random variables (or trials of an experiment), with $E(x_i) = \mu$ and $V(x_i) = \sigma^2$, we will obtain $E(\bar{X}) = \mu$ and $V(\bar{X}) = \sigma^2/n$, where $\bar{X}$ is distributed with an average of μ and a standard deviation of $\sigma_{\bar{X}} = \sigma^2/\sqrt{n}$.

Suppose we wish to estimate the average vehicle load at peak hour μ' , and we wish the tolerance (error of estimation) to be fewer than 8 passengers, with a confidence level (probability) of 95%. This is for a given route and travel direction during weekdays. Since approximately 95% of the sample averages, in terms of the number of trips to be sampled, will fall within $2\sigma_{\bar{X}}$ of μ' in repeated sampling, we can ask that $2\sigma_{\bar{X}}$ equal 8 passengers, or $2\sigma_{\bar{X}} = 2\sigma/\sqrt{n} = 8$, from which we obtain the sample size $n = \sigma^2/16$. Certainly we do not know the population standard deviation, σ ; hence, we must use an estimate from a previous sample or knowledge of the range in which the measurement will fall. An empirical rule is that this range is approximately 4σ and, in this way, we will find σ . For our example, suppose the range is 50 passengers (average load during peak hours lies between 20 and 70 passengers). Thus, $n = (50/4)^2/16 = 9.77$. Accordingly, we will sample ten trips during the peak hours of a given route and direction of travel spread out over several weekdays. We would then be reasonably certain (with a probability of 95%) that our estimate will lie within $2\sigma_{\bar{X}} = 8$ passengers of the true average of the vehicle load. The general sample size formula for this case is as follows:

$$n = Z_{1-\frac{\alpha}{2}}^2 \frac{\sigma^2}{TO^2} \quad (2.6)$$

where $Z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ percentile normal deviate, and TO is the tolerance. Based on the empirical rule cited above, we can use the formula in a practical situation:

$$n = \frac{4s^2}{(TO)^2} \quad (2.7)$$

where s is the estimate of σ .

In effect, we would expect in the example used that the tolerance would be fewer than 8 passengers. According to the empirical rule, there is a probability of 68% that the tolerance is less than $\sigma_{\bar{X}} = 4$ passengers. Note that these probabilities of 95% and 68% are inexact, because σ was approximated.

2.4.4 Practical sample size for origin-destination (O-D) survey

The importance of O-D data in improving transit service is apparent from Figure 2.3 for all operations-planning levels. O-D matrices ought to be constructed for developing new transit routes, terminals and stops, changing route structure and strategies, and amending timetables and vehicle and crew schedules. Therefore, this section will be devoted to a practical method of finding an adequate sample size for O-D surveys.

While sampling *from where* people want to go and *to where* (by time of day, purpose of trip, etc.), we enter a process of interaction between events from a statistical perspective. We want to survey N people, with the outcome of each falling into one of the O-D cells (each

cell is a specific origin-destination). The experiment of such N independent trials is called a *multinomial experiment* (generalization of the binomial experiment) and has a useful probability (multinomial) distribution for discrete random variables. The proper O-D matrix is, in fact, a contingency table based on a multinomial distribution, in which contingency means the relationship between O-D classifications. That is, there are $B_1, B_2, \dots, B_k$ possible O-D relationships with the corresponding $\theta_1, \theta_2, \dots, \theta_k$ probabilities, such that $\sum_i \theta_i = 1$.

We are interested in knowing the probability that, in N trials, B_1 will be observed x_1 times, B_2 will be observed x_2 times, $\dots$, and B_k will be observed x_k times where $\sum_i x_i = N$. Because of independency, the following is obtained: $P(B_1) = \theta_1^{x_1}$, $P(B_2) = \theta_2^{x_2}$, $\dots$, $P(B_k) = \theta_k^{x_k}$ and

$$P(x_1, x_2, \dots, x_k) = \frac{N! \theta_1^{x_1} \cdot \theta_2^{x_2} \cdot \dots \cdot \theta_k^{x_k}}{x_1! x_2! \cdot \dots \cdot x_k!} \text{ (multinomial distribution)} \quad (2.8)$$

where P represents probability.

Consider now x_i as a binomial variable with average $N\theta_i$ and variance $N\theta_i(1 - \theta_i)(1 - (n/N))$, where n is the sample size. For $n \ll N$, the fraction n/N approaches zero. Further, for large n , say over 50, x_i can be considered normally distributed as in Equation (2.3): $x_i \sim \text{Normal} [\mu = n\theta_i, \sigma^2 = n\theta_i(1 - \theta_i)]$.

Let $p_i = x_i/n$ represent a group i proportion within the population. Thus, $p_i \sim \text{Normal} [\theta_i, \theta_i(1 - \theta_i)/n]$. The sample size in Equations (2.6) and (2.7) yields: $n = Z_{1-(\alpha/2)}^2 (s^2/TO_i^2)$, where s , the estimate of σ , can be represented by $\sqrt{p_i(1-p_i)}$. Since p_i is the group proportion, $p_i(1 - p_i)$ can never go beyond 0.25 (the case in which $p_i = 0.5$). Hence, it can be used as an upper bound for the sample size. The required sample size, therefore, is:

$$n = Z_{1-\frac{\alpha}{2}}^2 \frac{p_i(1-p_i)}{TO_i^2}, i = 1, 2, \dots, k \quad (2.9)$$

In cases in which we cannot consider $n \ll N$, we can, as a practical step, use the following expression for n_0 sample size: $n_0 = n/1 - n/N$.

We can now establish a stepwise procedure for choosing the sample size:

- (i) Ratio estimation in each cell of the O-D matrix. That is, what is the ratio, R , between the number of passengers travelling from a given origin to a given destination, and the population (which ratio can be estimated by performing a small pretest)?
- (ii) Determination of the required precision in each cell of the O-D matrix.
- (iii) Determination of the confidence interval.
- (iv) Determination of the base ratio for the sample-size selection procedure. It should be noted that the selection of the smallest ratio in the O-D matrix will result in the largest sample size, n_m . However, an average ratio will result in a smaller sample size

- than n_m , but with reduced confidence intervals for cells having lower ratios than the average.
- (v) Selection of the sample size based on given precision tables (see example below).

The end of this chapter present the precision table for commonly used sample sizes for an O-D survey: 500, 1,000, 2,000, 5,000, and 10,000. For example, there are in Table 2.1 three origins and three destinations, **a**, **b** and **c**, and the estimated ratios, R , between the number of passengers for each O-D cell and the true population. We then assume for all the cells a precision of $TO_i/p_i = \pm 10\%$ and a confidence interval of 95%. The base ratio $\bar{R}$ is selected to be the average, or close to the average, of all the ratios in Table 2.1.

Table 2.1 *Example of ratios between each O-D cell and the true population*

From	a	b	c
To			
a	–	0.05	0.20
b	0.15	–	0.10
c	0.15	0.05	–

$\bar{R} = 0.12$ ($\approx 0.7/6$). Table 2.2 extracts the relevant portion of the entire precision table at the end of the chapter and emphasizes the precisions for $\bar{R} = 0.12$ for all commonly used n . That is,

- for $n = 500$, the precision is $\pm 23.74\%$,
- for $n = 1,000$, the precision is $\pm 16.78\%$,
- for $n = 2,000$, the precision is $\pm 11.87\%$,
- for $n = 5,000$, the precision is $\pm 7.51\%$, and
- for $n = 10,000$, the precision is $\pm 5.31\%$.

It may be seen that none of the precision levels is $\pm 10\%$. In order to attain greater precision, we can decide on $\pm 7.51\%$ (the closet to $\pm 10\%$ from below) and a sample of 5,000 people. Consequently, for $n = 5,000$, and based on Table 2.2, the precision is ± 12.08 for the cell with $R = 0.05$,

- for 0.10, the precision is $\pm 8.32\%$,
- for 0.15, the precision is $\pm 6.60\%$,
- for 0.20, the precision is $\pm 5.54\%$.

These cell-based precisions are also emphasized in Table 2.2. Finally more transit-related sample size formulae appear in the *Transit Data Collection Design Manual* (UMTA, 1985).

Table 2.2 Precision table for the example

n = sample size	95% Confidence Interval				
	500	1,000	2,000	5,000	10,000
R = ratio					
0.01	87.21	61.67	43.61	23.58	19.50
0.02	61.36	43.39	30.68	19.40	13.72
0.03	49.84	35.24	24.92	15.76	11.14
0.04	42.94	30.36	21.4	13.58	9.60
0.05	38.21	27.02	19.10	12.08	8.54
0.06	34.69	24.53	17.35	10.97	7.76
0.07	31.95	22.59	15.97	10.10	7.14
0.08	29.72	21.02	14.86	9.40	6.65
0.09	27.87	19.71	13.94	8.81	6.23
0.10	26.30	18.59	13.15	8.32	5.88
0.11	24.93	17.63	12.47	7.88	5.58
0.12	23.74	16.78	11.87	7.51	5.31
0.13	22.68	16.03	11.34	7.17	5.07
0.14	21.72	15.36	10.86	6.87	4.86
0.15	20.87	14.75	10.43	6.60	4.67
0.16	20.08	14.20	10.04	6.35	4.49
0.17	19.37	13.70	9.68	6.12	4.33
0.18	18.71	13.23	9.35	5.92	4.18
0.19	18.10	12.80	9.05	5.72	4.05
0.20	17.53	12.40	8.77	5.54	3.92

2.5 Literature review and further reading

This section reviews papers that deal with the collection of transit data, focusing on data on service performance and ridership. These data are necessary for operations planning in the short run, and for network design in the long run. The collection of general travel-behaviour data, passenger-trip diaries, or data concerning passenger preferences is not covered by this review.

According to the *Transit Data Collection Design Manual* (UMTA, 1985), the route-specific data needed for effective basic operation planning include total boardings, loads at key points, running time, schedule adherence, revenues and some passenger characteristics. At the design level, data include passenger origins and destinations, fare-category distribution, boardings and alightings by stop, transfers between routes, passenger attitudes and passengers' travel behaviour. The manual divides the techniques for collecting transit data into four groups, according to the identity and location of the surveyor: (i) ride checks, which are taken on board; (ii) point checks, taken by an observer standing at the road side; (iii) driver checks, with the data collected by the driver; and (iv) automated checks, in which the information is collected by machines. Eight different types of counts and readings are described: on/off passenger counts, counts of boardings only, passenger-load counts farebox readings, revenue counts, transfer counts, origin-destination stop counts and passenger surveys. The manual also presents sampling methods, techniques for sample-size determination, calculation of conversion factors to enable survey results to represent a whole population, and data-accuracy determination.

The review of transit-data-collection techniques by Booz Allen and Hamilton (1985) documents the most common collection techniques. Nine leading, manual collection techniques are described: ride checks, point checks, boarding counts, farebox readings, revenue counts, speed and delay studies, running-time checks, transfer counts and stop studies. The review describes each technique and procedures for its application, as well as the data items collected by the technique. Instructions are given on how to aggregate and manipulate the data.

Bamford *et al.* (1984) assert that on-board passenger information can be gathered either by using self-completed questionnaires or by an interviewer. They examine the benefits of a data-collection method based on a very short questionnaire that can be self-completed by means of 'rub-off', within no more than one minute and without the use of a pen. They claim that this method is cheaper than most other on-board surveying methods and yields higher response rates.

A modular approach to on-board automatic data-collection systems (NCTRD report, 1986) defines standard requirements for such systems. The authors argue that these systems offer advantages over manual systems in many transit applications: improved data-processing time, lower collection costs and better quality data. They describe a standard microprocessor-based system that provides an electronic means of gathering passenger, fare and schedule data. Step-by-step guidelines for selecting and implementing the data-collection system are presented, as is a discussion of the cost of such a system. The paper also enumerates the requirements for equipping a bus with an automatic data-collection system and details a technique for determining the necessary number of equipped buses, which works out to almost 10% of the fleet. Perhaps the most important conclusion derived from equipping a bus fleet with automatic data-collection devices is the need for coordination with the existing dispatching procedures. The cooperation of dispatchers and drivers, particularly in regard to modifications of dispatching practices, is crucial.

Furth and McCollom (1987) focus on the calculation of conversion factors, which are used for converting survey results (number of boardings, alightings, peak load, revenue, etc.) into representative values. The authors discuss statistical aspects of the conversion-factor

approach, including sample-size estimation and a determination of accuracy measures. They outline several scenarios concerning available input data, since preliminary information about mean values is not always available. An optimal sampling plan to minimize costs is offered.

Macbriar (1989) discusses data retrieval from electronic ticket machines. These machines constitute a potential source of information not only on fares and revenues, but also on passenger characteristics that change with fare, origins and destinations, and riding times. The paper discusses in detail the importance of keeping a simple, well-arranged enumeration system for trips, stops, etc., which is necessary for documenting the data. The paper also discusses data needs for reimbursing concession-fare schemes; i.e. the way operators receive compensation for being forced to provide reduced prices for specific sectors.

Weinstein and Albom (1998) discuss techniques for collecting data on qualitative features of transit performance that are difficult to measure, such as station/vehicle cleanliness. They present a methodology, including rating criteria, for estimating these features.

Richardson (1998) describes a survey technique that has been used to measure ticket-usage rates. This information should assist in the allocation of farebox revenue to operators of a private transit system. Because of the unrepresentative sampling method, a principal component of the methodology presented is a weighting procedure, based on data partially collected beforehand. A combination of questionnaires and observational surveys is used in order to obtain accurate bias estimates.

Naverrete (1999) describes a methodology for using digital assistants instead of paper records when inspectors collect field data. Criteria for the efficiency of the digital data-collection device are developed, including ease of data entry and retrieval, utilization of existing technologies, and effective handwriting recognition. A suitable device is chosen and tested. According to field tests, the most difficult problem encountered is the learning curve for the handwriting-recognition programme.

Barua *et al.* (2001) discuss the use of geographical information systems (GIS) for transit origin-destination data analysis following an on-board transit survey. GIS is used not just for geocoding origin and destination locations, but also as a tool for checking survey-data quality and for validating the data analysis. A suggested methodology for weighting survey results and converting them into a representative trip matrix is based on spatial criteria. The GIS-based approach for validating the quality of the results includes a comparison of transit-link volumes with data from passenger counts and a parallel validation of sub-area transit-demand characteristics.

Some of the literature in the field of data collection presents the features of specific software or hardware systems, without detailing methodologies. For example, Barnes and Urbanik (1990) describe a software for a computerized data-collection process. Rossetti (1996) reports on a transit-monitoring system based on radio-frequency identification that integrates automatic vehicle locationing (AVL) with automatic passenger counting. Other papers on AVL and AVM are discussed in Chapter 17.

Table 2.3 summarizes the main features and characteristics of the literature reviewed by categories. Having presented the overview of the first key in Figure 2.1, I will proceed in the next 12 chapters will deal with the second key: how to plan and decide intelligently. The third key will be dealt with in Chapters 16 and 17.

Table 2.3 Summary of the main features of the literature reviewed, by category

Source	Type of data collection	Data collection method	Contribution of the paper	Includes methodology for weighting survey results?	Includes discussion of data accuracy?
Booz <i>et al.</i> (1985)	Various	Various	Review	Yes	Yes
Booz <i>et al.</i> (1985)	Various	Various	Review	Yes	Yes
Bamford <i>et al.</i> (1984)	Various	On-board questionnaires	Technical features of the questionnaire	No	No
NCTRD (1986)	Various	On-board automatic	Definition of system requirements	No	No
Furth and McCollom (1987)	Various	Any method	Analytical method	Yes	Yes
Macbriar (1989)	Various	On-board automatic (through ticket machine)	Review/discussion	No	No
Weinstein and Albom (1998)	Qualitative data, such as cleanliness	Observation (on-board or at station)	Analytical method	No	Yes
Richardson (1998)	Ticket-usage rates	On-board, combination of questionnaires and observations	Analytical method	Yes	Yes
Naverrete (1999)	Various	Digital assistants	New technology	No	No
Barua <i>et al.</i> (2001)	Passenger origin-destination	Any (focus is on analysis method)	Analytical method	Yes	Yes
Barnes and Urbanik (1990)	Various	On-board automatic	New technology	No	No

(Continued)

Table 2.3 Summary of the main features of the literature reviewed, by category (continued)

Source	Type of data collection	Data collection method	Contribution of the paper	Includes methodology for weighting survey results?	Includes discussion of data accuracy?
Rossetti (1996)	Various	Combination of AVL and on-board automatic counters	New technology	No	No

References

- Bamford, C. G., Carrick, R. J. and MacDonald, R. (1984). Public transport surveys: A new effective technique of data collection. *Traffic Engineering and Control*, **25** (6), 318–319.
- Barnes, K. E. and Urbanik, T. (1990). Automated transit ridership data collection software development and user's manual, Texas Transportation Institute, Report UMTA-TX-08-1087-91-1.
- Barua, B., Boberg, J., Hsia, J. S. and Zhang, X. (2001). Integrating geographic information systems with transit survey methodology. *Transportation Research Record*, **1753**, 29–34.
- Booz, Allen and Hamilton (1985). *Review of Transit Data Collection Techniques*. Philadelphia, PA.
- Furth, P. G. and McCollom, B. (1987). Using conversion factors to lower transit data collection costs. *Transportation Research Record*, **1144**, 1–6.
- Macbriar, I. D. (1989). Current issues in public transport data collection. *Proceedings of Seminar C held at the 17th PTRC Transport and Planning Summer Annual Meeting*, P318, **99**, 219–229.
- Merriam-Webster's Collegiate Dictionary* (2000). Tenth Edition. Merriam-Webster, Springfield, MA.
- NCTRD (National Cooperative Transit Research & Development Program) *Modular approach to on-board automatic data collection systems* (1986). Report No. 9, Transportation Research Board, Washington, DC.
- Naverrete, G. (1999). In the palm of your hand: Digital assistant's aid in data collection. *Journal of Management in Engineering*, **15** (4), 43–45.
- Richardson, T. (1998). Public transport ticket usage surveys: A methodological design. *Australian Transport Research Forum*, **22**, 697–712.
- Rossetti, M. D. (1996). Automatic Data Collection on Transit Users Via Radio Frequency Identification, Report of investigation, Transit-IDEA program, No. 10.

- Urban Mass Transportation Administration (UMTA) (1985). *Transit Data Collection Design Manual*. Washington, DC.
- Washington, S. P., Karlaftis, M. G. and Mannering, F. L. (2003). *Statistical and Econometric Methods for Transportation Data Analysis*. CRC Press LLC, Florida, USA.
- Weinstein, A. and Albom, R. (1998). Securing objective data on the quality of the passenger environment for transit riders: Redesign of the passenger environment measurement system for the Bay Area Rapid Transit District. *Transportation Research Record*, **1618**, 213–219.

Precision table for five survey sample sizes

Precision table															
R\n	90% Confidence interval					95% Confidence interval					99% Confidence interval				
	500	1,000	2,000	5,000	10,000	500	1,000	2,000	5,000	1,0000	500	1,000	2,000	5,000	10,000
0.01	73.19	51.75	36.60	23.15	16.37	87.21	61.67	43.61	27.58	19.50	114.62	81.05	57.31	36.25	25.63
0.02	51.49	36.41	25.75	16.28	11.51	61.36	43.39	30.68	19.40	13.72	80.64	57.02	40.32	25.50	18.03
0.03	41.83	29.58	20.91	13.23	9.35	49.84	35.24	24.92	15.76	11.14	65.50	46.32	32.75	20.71	14.65
0.04	36.04	25.48	18.02	11.40	8.06	42.94	30.36	21.47	13.58	9.60	56.43	39.90	28.22	17.85	12.62
0.05	32.06	22.67	16.03	10.14	7.17	38.21	27.02	19.10	12.08	8.54	50.21	35.51	25.11	15.88	11.23
0.06	29.12	20.59	14.56	9.21	6.51	34.69	24.53	17.35	10.97	7.76	45.60	32.24	22.80	14.42	10.20
0.07	26.81	18.96	13.41	8.48	6.00	31.95	22.59	15.97	10.10	7.14	41.99	29.69	20.99	13.28	9.39
0.08	24.95	17.64	12.47	7.89	5.58	29.72	21.02	14.86	9.40	6.65	39.06	27.62	19.53	12.35	8.74
0.09	23.39	16.54	11.70	7.40	5.23	27.87	19.71	13.94	8.81	6.23	36.63	25.90	18.31	11.58	8.19
0.10	22.07	15.60	11.03	6.98	4.93	26.30	18.59	13.15	8.32	5.88	34.56	24.44	17.28	10.93	7.73
0.11	20.92	14.80	10.46	6.62	4.68	24.93	17.63	12.47	7.88	5.58	32.77	23.17	16.38	10.36	7.33
0.12	19.92	14.09	9.96	6.30	4.45	23.74	16.78	11.87	7.51	5.31	31.19	22.06	15.60	9.86	6.98
0.13	19.03	13.46	9.51	6.02	4.26	22.68	16.03	11.34	7.17	5.07	29.80	21.07	14.90	9.42	6.66
0.14	18.23	12.89	9.12	5.77	4.08	21.72	15.36	10.86	6.87	4.86	28.55	20.19	14.28	9.03	6.38
0.15	17.51	12.38	8.76	5.54	3.92	20.87	14.75	10.43	6.60	4.67	27.42	19.39	13.71	8.67	6.13

(Continued)

Precision table for five survey sample sizes (continued)

Precision table															
R\n	90% Confidence interval					95% Confidence interval					99% Confidence interval				
	500	1,000	2,000	5,000	10,000	500	1,000	2,000	5,000	10,000	500	1,000	2,000	5,000	10,000
0.16	16.85	11.92	8.43	5.33	3.77	20.08	14.20	10.04	6.35	4.49	26.39	18.66	13.20	8.35	5.90
0.17	16.25	11.49	8.13	5.14	3.63	19.37	13.70	9.68	6.12	4.33	25.45	18.00	12.73	8.05	5.69
0.18	15.70	11.10	7.85	4.96	3.51	18.71	13.23	9.35	5.92	4.18	24.59	17.39	12.29	7.78	5.50
0.19	15.19	10.74	7.59	4.80	3.40	18.10	12.80	9.05	5.72	4.05	23.78	16.82	11.89	7.52	5.32
0.20	14.71	10.40	7.36	4.65	3.29	17.53	12.40	8.77	5.54	3.92	23.04	16.29	11.52	7.29	5.15
0.21	14.27	10.09	7.13	4.51	3.19	17.00	12.02	8.50	5.38	3.80	22.34	15.80	11.17	7.07	5.00
0.22	13.85	9.79	6.93	4.38	3.10	16.50	11.67	8.25	5.22	3.69	21.69	15.34	10.85	6.86	4.85
0.23	13.46	9.52	6.73	4.26	3.01	16.04	11.34	8.02	5.07	3.59	21.08	14.90	10.54	6.67	4.71
0.24	13.09	9.26	6.55	4.14	2.93	15.60	11.03	7.80	4.93	3.49	20.50	14.50	10.25	6.48	4.58
0.25	12.74	9.01	6.37	4.03	2.85	15.18	10.74	7.59	4.80	3.39	19.95	14.11	9.98	6.31	4.46
0.26	12.41	8.78	6.20	3.92	2.77	14.79	10.46	7.39	4.68	3.31	19.43	13.74	9.72	6.15	4.35
0.27	12.10	8.55	6.05	3.82	2.70	14.41	10.19	7.21	4.56	3.22	18.94	13.39	9.47	5.99	4.24
0.28	11.80	8.34	5.90	3.73	2.64	14.06	9.94	7.03	4.44	3.14	18.47	13.06	9.24	5.84	4.13
0.29	11.51	8.14	5.75	3.64	2.57	13.71	9.70	6.86	4.34	3.07	18.02	12.75	9.01	5.70	4.03
0.30	11.24	7.95	5.62	3.55	2.51	13.39	9.47	6.69	4.23	2.99	17.60	12.44	8.80	5.56	3.93
0.31	10.97	7.76	5.49	3.47	2.45	13.08	9.25	6.54	4.14	2.92	17.19	12.15	8.59	5.43	3.84

0.32	10.72	7.58	5.36	3.39	2.40	12.78	9.03	6.39	4.04	2.86	16.79	11.87	8.40	5.31	3.75
0.33	10.48	7.41	5.24	3.31	2.34	12.49	8.83	6.24	3.95	2.79	16.41	11.61	8.21	5.19	3.67
0.34	10.25	7.25	5.12	3.24	2.29	12.21	8.64	6.11	3.86	2.73	16.05	11.35	8.02	5.08	3.59
0.35	10.02	7.09	5.01	3.17	2.24	11.94	8.45	5.97	3.78	2.67	15.70	11.10	7.85	4.96	3.51
0.36	9.81	6.94	4.90	3.10	2.19	11.69	8.26	5.84	3.70	2.61	15.36	10.86	7.68	4.86	3.43
0.37	9.60	6.79	4.80	3.04	2.15	11.44	8.09	5.72	3.62	2.56	15.03	10.63	7.52	4.75	3.36
0.38	9.40	6.64	4.70	2.97	2.10	11.20	7.92	5.60	3.54	2.50	14.71	10.40	7.36	4.65	3.29
0.39	9.20	6.51	4.60	2.91	2.06	10.96	7.75	5.48	3.47	2.45	14.41	10.19	7.20	4.56	3.22
0.40	9.01	6.37	4.50	2.85	2.01	10.74	7.59	5.37	3.39	2.40	14.11	9.98	7.05	4.46	3.15
0.41	8.82	6.24	4.41	2.79	1.97	10.51	7.44	5.26	3.33	2.35	13.82	9.77	6.91	4.37	3.09
0.42	8.64	6.11	4.32	2.73	1.93	10.30	7.28	5.15	3.26	2.30	13.54	9.57	6.77	4.28	3.03
0.43	8.47	5.99	4.23	2.68	1.89	10.09	7.14	5.05	3.19	2.26	13.26	9.38	6.63	4.19	2.97
0.44	8.30	5.87	4.15	2.62	1.86	9.89	6.99	4.94	3.13	2.21	13.00	9.19	6.50	4.11	2.91
0.45	8.13	5.75	4.07	2.57	1.82	9.69	6.85	4.85	3.06	2.17	12.74	9.01	6.37	4.03	2.85
0.46	7.97	5.64	3.99	2.52	1.78	9.50	6.72	4.75	3.00	2.12	12.48	8.83	6.24	3.95	2.79
0.47	7.81	5.52	3.91	2.47	1.75	9.31	6.58	4.65	2.94	2.08	12.23	8.65	6.12	3.87	2.74
0.48	7.66	5.41	3.83	2.42	1.71	9.12	6.45	4.56	2.88	2.04	11.99	8.48	5.99	3.79	2.68
0.49	7.50	5.31	3.75	2.37	1.68	8.94	6.32	4.47	2.83	2.00	11.75	8.31	5.88	3.72	2.63
0.50	7.36	5.20	3.68	2.33	1.64	8.77	6.20	4.38	2.77	1.96	11.52	8.15	5.76	3.64	2.58

(Continued)

Precision table for five survey sample sizes (continued)

Precision table															
R\n	90% Confidence interval					95% Confidence interval					99% Confidence interval				
	500	1,000	2,000	5,000	10,000	500	1,000	2,000	5,000	10,000	500	1,000	2,000	5,000	10,000
0.51	7.21	5.10	3.61	2.28	1.61	8.59	6.08	4.30	2.72	1.92	11.29	7.98	5.65	3.57	2.52
0.52	7.07	5.00	3.53	2.23	1.58	8.42	5.95	4.21	2.66	1.88	11.07	7.83	5.53	3.50	2.47
0.53	6.93	4.90	3.46	2.19	1.55	8.25	5.84	4.13	2.61	1.85	10.85	7.67	5.42	3.43	2.43
0.54	6.79	4.80	3.39	2.15	1.52	8.09	5.72	4.04	2.56	1.81	10.63	7.52	5.32	3.36	2.38
0.55	6.65	4.70	3.33	2.10	1.49	7.93	5.61	3.96	2.51	1.77	10.42	7.37	5.21	3.30	2.33
0.56	6.52	4.61	3.26	2.06	1.46	7.77	5.49	3.88	2.46	1.74	10.21	7.22	5.11	3.23	2.28
0.57	6.39	4.52	3.19	2.02	1.43	7.61	5.38	3.81	2.41	1.70	10.01	7.07	5.00	3.16	2.24
0.58	6.26	4.43	3.13	1.98	1.40	7.46	5.27	3.73	2.36	1.67	9.80	6.93	4.90	3.10	2.19
0.59	6.13	4.34	3.07	1.94	1.37	7.31	5.17	3.65	2.31	1.63	9.60	6.79	4.80	3.04	2.15
0.60	6.01	4.25	3.00	1.90	1.34	7.16	5.06	3.58	2.26	1.60	9.41	6.65	4.70	2.97	2.10
0.61	5.88	4.16	2.94	1.86	1.32	7.01	4.96	3.50	2.22	1.57	9.21	6.51	4.61	2.91	2.06
0.62	5.76	4.07	2.88	1.82	1.29	6.86	4.85	3.43	2.17	1.53	9.02	6.38	4.51	2.85	2.02
0.63	5.64	3.99	2.82	1.78	1.26	6.72	4.75	3.36	2.12	1.50	8.83	6.24	4.41	2.79	1.97
0.64	5.52	3.90	2.76	1.74	1.23	6.57	4.65	3.29	2.08	1.47	8.64	6.11	4.32	2.73	1.93
0.65	5.40	3.82	2.70	1.71	1.21	6.43	4.55	3.22	2.03	1.44	8.45	5.98	4.23	2.67	1.89

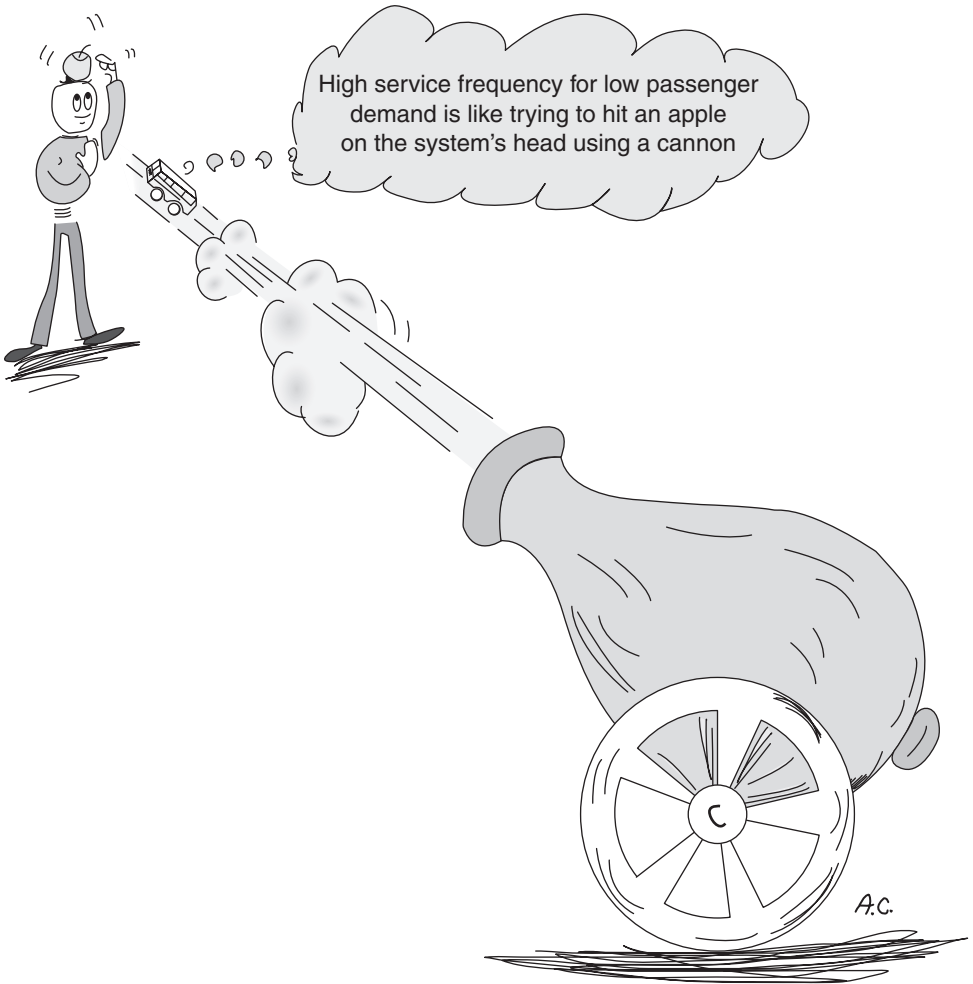
0.66	5.28	3.73	2.64	1.67	1.18	6.29	4.45	3.15	1.99	1.41	8.27	5.85	4.13	2.61	1.85
0.67	5.16	3.65	2.58	1.63	1.15	6.15	4.35	3.08	1.95	1.38	8.08	5.72	4.04	2.56	1.81
0.68	5.05	3.57	2.52	1.60	1.13	6.01	4.25	3.01	1.90	1.34	7.90	5.59	3.95	2.50	1.77
0.69	4.93	3.49	2.47	1.56	1.10	5.88	4.15	2.94	1.86	1.31	7.72	5.46	3.86	2.44	1.73
0.70	4.82	3.41	2.41	1.52	1.08	5.74	4.06	2.87	1.81	1.28	7.54	5.33	3.77	2.38	1.69
0.71	4.70	3.32	2.35	1.49	1.05	5.60	3.96	2.80	1.77	1.25	7.36	5.21	3.68	2.33	1.65
0.72	4.59	3.24	2.29	1.45	1.03	5.47	3.87	2.73	1.73	1.22	7.18	5.08	3.59	2.27	1.61
0.73	4.47	3.16	2.24	1.41	1.00	5.33	3.77	2.67	1.69	1.19	7.01	4.95	3.50	2.22	1.57
0.74	4.36	3.08	2.18	1.38	0.97	5.20	3.67	2.60	1.64	1.16	6.83	4.83	3.41	2.16	1.53
0.75	4.25	3.00	2.12	1.34	0.95	5.06	3.58	2.53	1.60	1.13	6.65	4.70	3.33	2.10	1.49
0.76	4.13	2.92	2.07	1.31	0.92	4.93	3.48	2.46	1.56	1.10	6.47	4.58	3.24	2.05	1.45
0.77	4.02	2.84	2.01	1.27	0.90	4.79	3.39	2.40	1.51	1.07	6.30	4.45	3.15	1.99	1.41
0.78	3.91	2.76	1.95	1.24	0.87	4.66	3.29	2.33	1.47	1.04	6.12	4.33	3.06	1.93	1.37
0.79	3.79	2.68	1.90	1.20	0.85	4.52	3.20	2.26	1.43	1.01	5.94	4.20	2.97	1.88	1.33
0.80	3.68	2.60	1.84	1.16	0.82	4.38	3.10	2.19	1.39	0.98	5.76	4.07	2.88	1.82	1.29
0.81	3.56	2.52	1.78	1.13	0.80	4.25	3.00	2.12	1.34	0.95	5.58	3.95	2.79	1.76	1.25
0.82	3.45	2.44	1.72	1.09	0.77	4.11	2.90	2.05	1.30	0.92	5.40	3.82	2.70	1.71	1.21
0.83	3.33	2.35	1.66	1.05	0.74	3.97	2.81	1.98	1.25	0.89	5.21	3.69	2.61	1.65	1.17
0.84	3.21	2.27	1.61	1.02	0.72	3.83	2.71	1.91	1.21	0.86	5.03	3.55	2.51	1.59	1.12

(Continued)

Precision table for five survey sample sizes (continued)

Precision table															
R\n	90% Confidence interval					95% Confidence interval					99% Confidence interval				
	500	1,000	2,000	5,000	10,000	500	1,000	2,000	5,000	10,000	500	1,000	2,000	5,000	10,000
0.85	3.09	2.19	1.55	0.98	0.69	3.68	2.60	1.84	1.16	0.82	4.84	3.42	2.42	1.53	1.08
0.86	2.97	2.10	1.48	0.94	0.66	3.54	2.50	1.77	1.12	0.79	4.65	3.29	2.32	1.47	1.04
0.87	2.84	2.01	1.42	0.90	0.64	3.39	2.40	1.69	1.07	0.76	4.45	3.15	2.23	1.41	1.00
0.88	2.72	1.92	1.36	0.86	0.61	3.24	2.29	1.62	1.02	0.72	4.25	3.01	2.13	1.35	0.95
0.89	2.59	1.83	1.29	0.82	0.58	3.08	2.18	1.54	0.97	0.69	4.05	2.86	2.02	1.28	0.91
0.90	2.45	1.73	1.23	0.78	0.55	2.92	2.07	1.46	0.92	0.65	3.84	2.72	1.92	1.21	0.86
0.91	2.31	1.64	1.16	0.73	0.52	2.76	1.95	1.38	0.87	0.62	3.62	2.56	1.81	1.15	0.81
0.92	2.17	1.53	1.08	0.69	0.49	2.58	1.83	1.29	0.82	0.58	3.40	2.40	1.70	1.07	0.76
0.93	2.02	1.43	1.01	0.64	0.45	2.40	1.70	1.20	0.76	0.54	3.16	2.23	1.58	1.00	0.71
0.94	1.86	1.31	0.93	0.59	0.42	2.21	1.57	1.11	0.70	0.50	2.91	2.06	1.46	0.92	0.65
0.95	1.69	1.19	0.84	0.53	0.38	2.01	1.42	1.01	0.64	0.45	2.64	1.87	1.32	0.84	0.59
0.96	1.50	1.06	0.75	0.47	0.34	1.79	1.27	0.89	0.57	0.40	2.35	1.66	1.18	0.74	0.53
0.97	1.29	0.91	0.65	0.41	0.29	1.54	1.09	0.77	0.49	0.34	2.03	1.43	1.01	0.64	0.45
0.98	1.05	0.74	0.53	0.33	0.23	1.25	0.89	0.63	0.40	0.28	1.65	1.16	0.82	0.52	0.37
0.99	0.74	0.52	0.37	0.23	0.17	0.88	0.62	0.44	0.28	0.20	1.16	0.82	0.58	0.37	0.26

3 Frequency and Headway Determination



Chapter 3 Frequency and Headway Determination

Chapter outline

- 3.1 Introduction
 - 3.2 Max load (point check) methods
 - 3.3 Load profile (ride check) methods
 - 3.4 Criterion for selecting point check or ride check
 - 3.5 Conclusion (two examples)
 - 3.6 Literature review and further reading
 - Exercises
 - References
-

Practitioner's Corner

This chapter addresses the topic of determining frequencies on transit routes, a problem that must receive attention, either explicitly or implicitly, several times a year. The chapter provides insights and a solution for intelligently integrating resource saving and an effective level of service. The basic premise here is that such an integration is achievable.

The importance of ridership information has led transit agencies to introduce automated surveillance techniques or, alternatively, to increase the amount of manually collected data. Naturally, the transit agencies are expected to gain useful information for operations planning by obtaining more accurate passenger counts. This chapter describes and analyses several appropriate data-collection approaches in order to set frequencies and headways efficiently. Four different methods are presented for deriving frequency, two based on point-check (max load) data and two on ride-check (load profile) data. A ride check provides more complete information than does a point check, but at greater cost; there is a question whether the additional information gained justifies the expense. The four methods, all of which are based on available old profiles, provide the planner with adequate guidance in selecting the type of data-collection procedure. In addition, the planner or scheduler can evaluate the minimum expected vehicle runs when the load standard is eased and avoid overcrowding (in an average sense) at the same time.

The chapter starts by describing the significance of proper frequency calculation, followed by a presentation and interpretation of the four methods. The fourth section established a criterion, the profile-based load, for selecting either the point-check or the ride-check data-collection technique. In the fifth section, two examples demonstrate the four methods, one of the examples employing real-life, heavy bus-route data. The chapter ends with a literature review and exercises for practicing the methods. The majority of the chapter is suitable for practitioners; however, the end of Sections 3.2 and 3.4 may be skipped to some extent. They can be replaced by the procedure depicted, in flowchart form, in Figure 3.10 in Section 3.5.

Dealing with data collection to affect the number of vehicle runs brings to mind the following real-life anecdote. A chief planner in a bus agency once observed that there

was approved extra but unnecessary runs on one route. He then found out that this run was performed at the end of a duty day in order to bring the driver close to his home. The chief planner immediately ordered the local planner to eliminate this run. Six months later, the chief planner learned, to his surprised, that this run still existed. Angrily he then called the local planner and asked for an explanation, to which the latter replied: “Oh, I told this driver that on the day the checker is there, not to make this run”.

We can wrap up this corner with some pithy advice and admonition: the chains of traditional methods are too weak to be felt until they are too strong to be broken. In transit-operations planning, therefore, there should always be a window for flexibility, new initiatives and changes.

3.1 Introduction

One of the major foci in determining transit service is the selection of the most suitable frequency (vehicles/hour) for each route in the system, by time-of-day, day-of-week, and day-type. This chapter provides an initial focus on the subject of the second key in Figure 2.1, intelligent decision-making when planning transit service. This second key will accompany us through Chapter 16.

Figure 1.2 showed that frequency determination was in essence of paramount importance for creating transit timetables. Furth and Wilson (1981) wrote that transit agencies typically used service standards as the basis for setting frequencies, while combining this action with experience, judgement and passenger counts. The service standards in normal use appear in Figure 1.4: crowding level, allowed maximum standees, and upper (policy) and lower limits on headways. Figure 1.5 emphasizes that these standards cannot be rules of thumb; they must be based on determined criteria. This is especially true because of the high cost involved in providing higher frequencies, an act that may not always be necessary. Prudent transit management requires a balance between increasing frequency and the cost of its implementation. Often we observe that some transit agencies, following the belief that last week’s frequency is good for this week, do not adjust their service to fluctuating demand. To those with this philosophy, one can only say: it is better to wear out by changing than to rust out by believing that what you have is good enough.

This chapter presents different methods that allow not only for efficient frequency-setting but also for a sensitivity analysis of possible changes. Basically, the objectives of this chapter are two-fold: (i) to set vehicle frequencies in order to maintain adequate service quality and minimize the number of required vehicle runs; and (ii) to construct an evaluation tool that will efficiently allocate the cost of gathering appropriate passenger-load data at the route level.

The main input data for this chapter will be point-check and ride-check data. These types of passenger counts are explicitly described in Section 2.2 and Figure 2.2; they will be combined in the analysis with crowding level and minimum frequency (inverse of policy headway) standards. It is common for load-profile data to be gathered annually or every few years along the entire length of the transit route (ride check). Usually the most recent passenger-load information will come from one or more selected stops along that part of the route where the vehicle carries its heaviest loads (point check). Point-check information is routinely surveyed

several times a year for the purpose of possible schedule revisions, which can range from completely new timetables for new or revised routes to daily adjustments to accommodate changes in central business district (CBD) and industrial area working hours and school-dismissal times. A ride check provides more complete information than a point check, but is more expensive because of the need for either additional checkers to provide the required data or an automated surveillance system (e.g. APC, AVL). The question is whether the additional information justifies the expense of gathering it. This chapter explores ways in which a transit agency can use the old profile to determine which method, ride check or point check, is more appropriate and less costly for collecting the new data.

Following Furth and Wilson (1981), transit agencies use methods to set headways that are commonly based on existing service standards of crowding level and policy headway. These standards are based on two requirements: (i) adequate space will be provided to meet passenger demand; and (ii) an upper-bound value is placed on the headways to assure a minimum frequency of service. The first requirement is appropriate for heavily travelled route hours (e.g. peak period), and the second for lightly travelled hours. The first requirement is usually met by a widely used *peak-load factor* method (point check), which is similar to the max load procedure – both will be explained below. The second requirement is met by *policy headway*, which usually does not exceed 60 minutes and is restricted in some cases to 30 minutes. Occasionally, a lower-bound value is set on the headway, based on productivity or revenue/cost measures. There are also mathematical programming techniques to approach the problems of route design and service frequency simultaneously. Such a technique was adopted by Furth and Wilson (1981) to find the appropriate headway that would maximize social benefit, subject to the constraints of total subsidy, fleet size and bus-occupancy levels. Nevertheless, mathematical programming models are hardly employed, since they cannot incorporate practical operational considerations in the optimization analysis. A further review of the literature appears at the end of the chapter.

This chapter consists of three major parts. First, two methods are presented to derive vehicle frequencies based on point-check (max load) data. Second, two further methods are proposed, using ride-check (load profile) data. Third, a criterion is established for determining the appropriateness of each data-collection method. All four methods are applied – and analysed – on a route basis, following Ceder (1984). These three parts are succeeded by two examples (one real-life), a literature review and exercises.

3.2 Max load (point check) methods

One of the basic objectives in the provision of transit service is to ensure adequate space to accommodate the maximum number of on-board passengers along the entire route over a given time period. Let us denote the time period (usually an hour) as j . Based on the peak-load factor concept, the number of vehicles required for period j is:

$$F_j = \frac{\bar{P}_{mj}}{\gamma_j \cdot c} \quad (3.1)$$

where $\bar{P}_{mj}$ is the average maximum number of passengers (max load) observed on-board in period j , c represents the capacity of a vehicle (number of seats plus the maximum allowable standees), and γ_j is the load factor during period j , $0 < \gamma_j \leq 1.0$. For convenience, let us refer to

the product $\gamma_j \cdot c$ as d_{oj} , the desired occupancy on the vehicle at period j . The standard γ_j can be set so that d_{oj} is equal to a desired fraction of the capacity (e.g. d_{oj} = number of seats). It should be noted here that if $\bar{P}_{mj}$ is based on a series of measurements, one can take its variability into account. This can be done by replacing the average value in Equation (3.1) with $\bar{P}_{mj} + b \cdot S_{pj}$, where b is a predetermined constant and S_{pj} is the standard deviation associated with $\bar{P}_{mj}$.

The max load data are usually collected by a trained checker, who stands and counts at the transit stop located at the beginning of the max load section(s). This stop is usually determined from old ride-check data or from information given by a mobile supervisor. Often, the checkers are told to count at only one stop for the entire day instead of switching among different max load points, depending on period j . Certainly it is less costly to position a checker at one stop than to have several checkers switching among stops. Given that a checker is assigned to one stop, that which apparently is the heaviest *daily* load point along the route, we can establish the so-called *Method 1* for determining the frequency associated with this single stop:

$$F_{1j} = \max \left(\frac{P_{mdj}}{d_{oj}}, F_{mj} \right), \quad j = 1, 2, \dots, q \quad (3.2)$$

$$P_{md} = \max_{i \in S} \sum_{j=1}^q P_{ij} = \sum_{j=1}^q P_{i^*j}$$

$$P_{mdj} = P_{i^*j}$$

where F_{mj} is the minimum required frequency (reciprocal of policy headway) for period j , there are q time periods; S represents the set of all route stops i excluding the last stop, i^* is the *daily max load point*, and P_{ij} is a defined statistical measure (simple average or average plus standard deviation) of the total number of passengers on-board all the vehicles departing stop i during period j . The terms P_{mdj} and P_{md} are used for the (average) observed load at the daily max load point at time j and the total load observed at this point, respectively.

Figure 3.1 exhibits an example of passenger counts (Example 1) along a 10-km route with six stops between 6:00 a.m. and 11:00 a.m. The second column in the table in this figure presents the distances, in km, between each stop. The desired occupancy and minimum frequency are the same for all hours, and hence their time period subscript is dropped: $d_o = 50$ passengers and $F_m = 3$ vehicles, respectively. The set of stops S includes 5 i 's, $j = 1, 2, \dots, 5$, each period of one hour being associated with a given column. The last column in the table represents $\sum_{j=1}^5 P_{ij}$ in which each entry in the table is P_{ij} (an average value across several checks). Thus, i^* is the 3rd stop with $P_{md} = 1740$, and P_{mdj} in Equation (3.2) refers only to those entries in the 3rd row.

The second point-check method, or *Method 2*, is based on the max load observed in each time period. That is,

$$F_{2j} = \max \left(\frac{P_{mj}}{d_{oj}}, F_{mj} \right), \quad j = 1, 2, \dots, q \quad (3.3)$$

where $P_{mj} = \max_{i \in S} P_{ij}$, which stands for the maximum observed load (across all stops) in each period j .

In the table in Figure 3.1, the values of P_{mj} are circled, and a rectangle is placed around P_{md} . Figure 3.1 also illustrates passenger counts in three dimensions (load, distance and period), from

Stop no	Distance (km) to next stop	Average observed load (passengers), by hour					Total load (passengers)
		6:00–7:00	7:00–8:00	8:00–9:00	9:00–10:00	10:00–11:00	
1	2	50	136	245	250	95	776
2	1	100	510	310	208	122	1250
3	1.5	400	420	400	320	200	1740
4	3	135	335	350	166	220	1206
5	2.5	32	210	300	78	105	725

Notes: (1) Route length is 10 km, and stop no. 6 is the last stop
(2) For all hours, $d_o = 50$, $c = 90$ passenger, $F_m = 3$ veh/hr

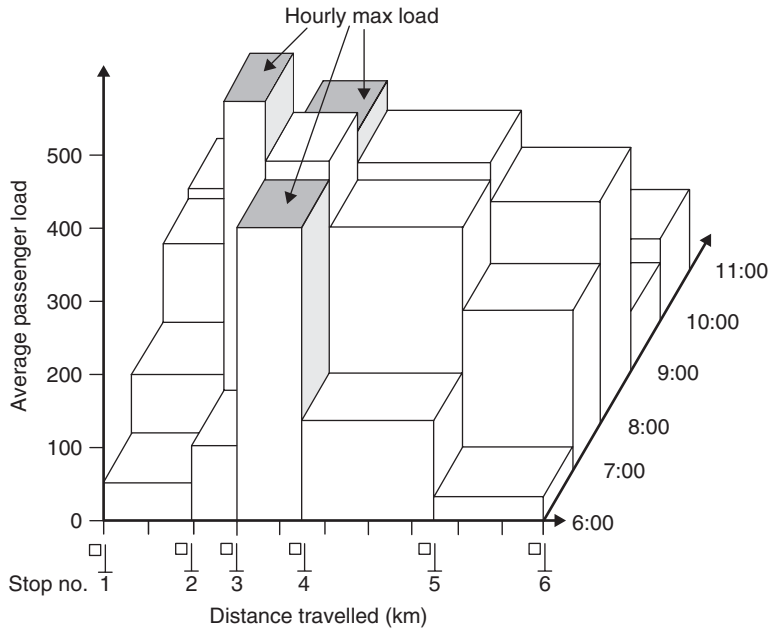


Figure 3.1 Five-hour load profiles, with indications of hourly and daily max load points

which the hourly max load is observed for the first three hours. The results of Equations (3.1) and (3.2) applied to Example 1 appear in Table 3.1 for both frequency (F_{kj}) and headways (H_{kj}) rounded to the nearest integer, where $k = 1, 2$. The only non-rounded headway is $H_{kj} = 7.5$ minutes, since it fits the so-called *clock headways*: these have the feature of creating timetables that repeat themselves every hour, starting on the hour. Practically speaking, $H_{kj} = 7.5$ can be implemented in an even-headway timetable by alternating between $H_{kj} = 7$ and $H_{kj} = 8$.

We will also retain non-rounded F_{kj} 's and show in the next chapter how to use these determined values for constructing timetables with and without even headways.

Although Method 1 has an advantage over Method 2 through cost saving in data gathering, it cannot be traded-off with accuracy of the results. That is, it is less costly and more

Table 3.1 Frequency and headway results for Example 1 as shown in Figure 3.1, according to Methods 1 and 2

Period j	Method 1 (Daily max load point)		Method 2 (Hourly max load point)	
	F _{1j} (veh/hr)	H _{1j} (minutes)	F _{2j} (veh/hr)	H _{2j} (minutes)
6:00–7:00	8.0	7.5	8.0	7.5
7:00–8:00	8.4	7	10.2	6
8:00–9:00	8.0	7.5	8.0	7.5
9:00–10:00	6.4	9	6.4	9
10:00–11:00	4.0	15	4.4	14

convenient to retain a checker at one transit stop throughout the entire working day than to assign the same checker or others to a different stop at every period j . Consequently, we will now concentrate on a comparison of the two methods: if the difference is statistically not significant, then the routine use of Method 1 will be preferred. This comparison may be carried out by the chi-square statistic, χ^2 , for testing the hypothesis concerning the population variances of the data supplied for each method. We assume that the random sample of passenger counts comes from a normal distribution (see Section 2.4 in this book for basic statistics), with averages μ_1, μ_2 and variances σ_1^2, σ_2^2 for Methods 1 and 2, respectively. The hypothesis that $\sigma_1^2 = \sigma_2^2$ will be examined, using the χ^2 test while knowing that σ_2^2 is the theoretical (expected) variance with which it is to be compared. The statistic $\chi^2, \chi^2 = (n - 1) s^2 / \sigma_2^2$ has a chi-square probability distribution with $(n - 1)$ degrees of freedom, where s^2 is computed from the sample. If $\sigma_1^2 > \sigma_2^2$, both s^2 and χ^2 will be larger than expected. The hypothesis will be rejected if $\chi^2 > \chi_\alpha^2$, where χ_α^2 is chosen so that $P(\chi^2 > \chi_\alpha^2) = \alpha$ (see Section 2.4); the χ_α^2 tables can be found, for example, in the following website: <http://www.stat.lsu.edu/exstweb/statlab/Tables/TABLES98-Chi.html>.

It should be noted that there is often a misjudgement of statistical data when chi-square tests are incorrectly utilized (Ceder and Dressler, 1980). For example, it is incorrect to compare the frequency or headway results given by the two methods. Instead, the actual observations should be compared, because any transformation of the actual units distorts the χ^2 test. If the observed (O) and the expected (E) data are multiplied by k , then the calculated χ^2 is equal to k times the corrected χ^2 value: $\sum_i [(kO_i - kE_i)^2 / kE_i] = k \sum_i [(O_i - E_i)^2 / E_i]$. In our case, the observed items are the passenger counts of Method 1 and the expected items are the counts of Method 2, using the following calculation:

$$\chi^2 = \sum_{j=1}^q \left[\frac{(P_{i^*j} - P_{mj})^2}{P_{mj}} \right] \quad (3.4)$$

where i^* is the daily max load point.

A comparison of the two methods can now be applied to Example 1 using Equation (3.4). This yields: $\chi^2 = (420 - 510)^2 / 510 + (200 - 220)^2 / 220 = 17.7$; for $\alpha = 0.05$, we obtain $\chi_\alpha^2 = 9.45$ from the chi-square table, with a degree of freedom of four. Hence the possibility

of assigning a checker only to stop 3 of Example 1 must be rejected, and we can skip over the Method 1 results.

3.3 Load profile (ride check) methods

The data collected by ride check enables the planner to observe the load variability among the transit stops, or what is termed the *load profile*. Usually, a recurrent, unsatisfactory distribution of loads will suggest the need for possible improvements in route design. The most common operational strategy resulting from observing the various loads is short turning (shortlining). A start-ahead and/or turn-back point(s) after the start and/or before the end of the route may be chosen, creating a new route that overlaps the existing route. This short-turn design problem is covered in Chapter 15. Other route-design-related actions using load data are route splitting and route shortening, both of which are dealt with in Chapter 13. This section will use the ride-check data for creating more alternatives to derive adequate frequencies, while assuming that the route remains the same. Nevertheless, we know that in practice the redesign of an existing route is not an activity often undertaken by transit agencies.

Two examples of load profiles are illustrated in Figure 3.2. These profiles are extracted from Example 1 in Figure 3.1 for the first and third hours. It may be noted that in most available transit-scheduling software (see, for example, Section 1.2), these load profiles are plotted with respect to each stop without relating the x-axis to any scale. A more appropriate way to plot the loads is to establish a passenger-load profile with respect to the distance travelled from the departure stop to the end of the route. It is also possible to replace the (deterministic) distance by the average running time; in the latter case, however, it is desirable for the running time to be characterized by low and persistent variations. These plots furnish the important evaluation measures of passenger-km and passenger-hour, as is also shown in Figure 3.2.

Let us observe the area marked by dashed lines in Figure 3.2. If a straight line is drawn across the load profile where the number of passengers is equal to the observed average hourly max load, then the area below this line but above the load profile is a measure of needless productivity. When Method 2 is used to derive the headways, this area represents empty space-kilometres. Furthermore, if d_{oj} in Method 2 is equal to the number of seats – often this is the desired occupancy or load factor used – then this measure is empty seat-kilometres. In light of this measure of unproductive service, we can see in Figure 3.2 that the 8:00–9:00 load profile is more than twice as productive as the 6:00–7:00 profile, although both have the same (max load-point based) frequency. We can now use the additional information supplied by the load profile to overcome the problem exhibited in Figure 3.2 when using Method 2. This can be done by introducing frequency-determination methods based on passenger-km rather than on a max load measure. The first load-profile method considers a lower-bound level on the frequency or an upper bound on the headway, given the same vehicle-capacity constraint. We call this *Method 3*, and it is expressed as follows:

$$F_{3j} = \max \left[\frac{A_j}{d_{oj} \cdot L}, \frac{P_{mj}}{c}, F_{mj} \right] \quad (3.5)$$

$$A_j = \sum_{i \in S} P_{ij} \cdot \ell_i, \quad L = \sum_{i \in S} \ell_i$$

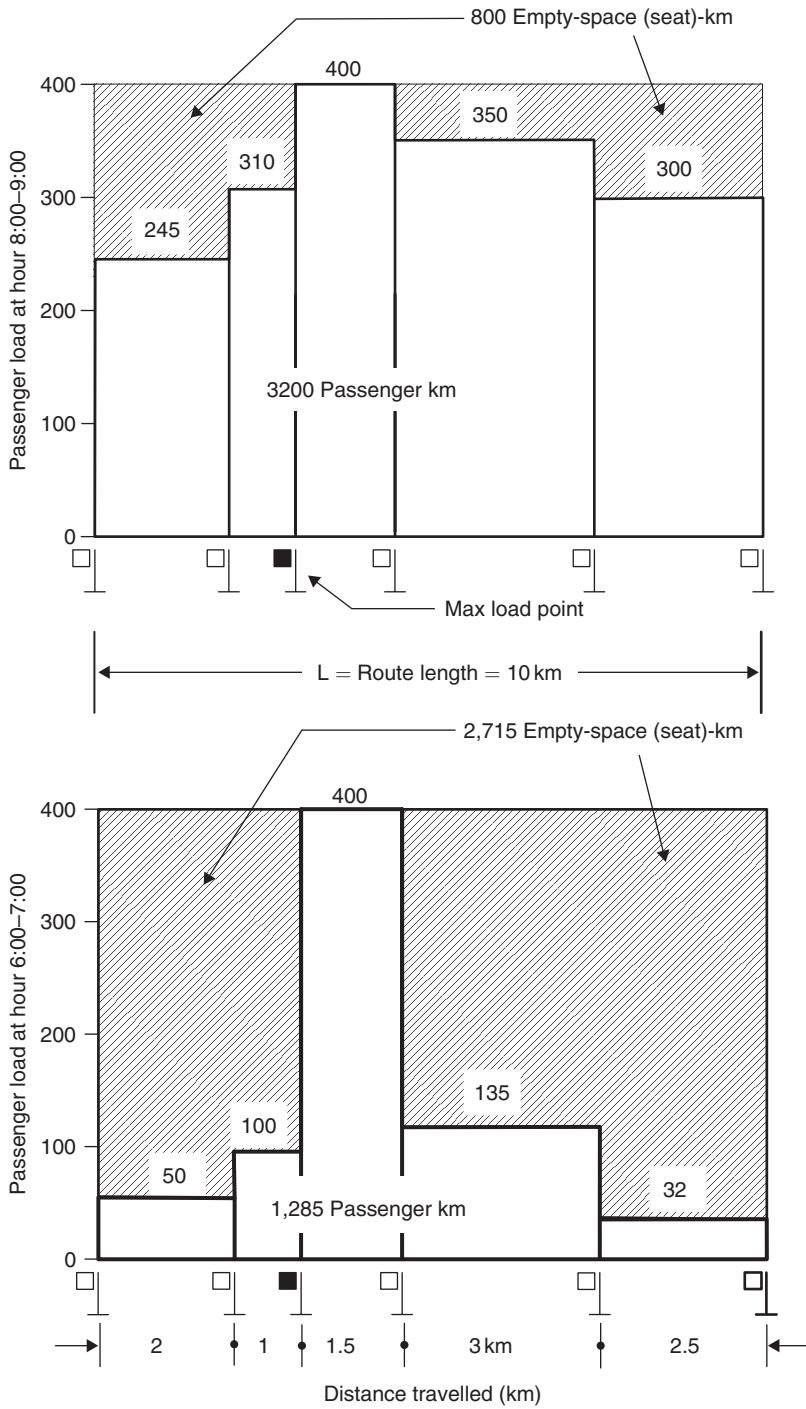


Figure 3.2 Two load profiles from Example 1 with the same 8-vehicle frequency, but with different passenger- and empty-space-km

Where l_i is the distance between stop i and the next stop ($i + 1$), A_j is the area in passenger-km under the load profile during time period j and L is the route length. The other notations were previously defined in Equations (3.1), (3.2) and (3.3).

One way to look at Method 3 is to view the ratio A_j/L as an average representative of the load P_{ij} (regardless of its statistical definition), as opposed to the max load (P_{mj}) in Method 2. Method 3 guarantees, on the average basis of P_{ij} , that the on-board passengers at the max load route segment will not experience crowding above the given vehicle capacity c . This method is appropriate for cases in which the planner wishes to know the number of vehicle runs (frequency) expected, while relaxing the desired occupancy standard constraint and, at the same time, avoiding situations in which passengers are unable to board the vehicle in an average sense. Using the results of Method 3 allows planners to handle: (i) demand changes without increasing the available number of vehicles; (ii) situations in which some vehicles are needed elsewhere (e.g. breakdown and maintenance problems or emergencies); and (iii) occasions when there are fewer drivers than usual (e.g. owing to budget cuts or problems with the drivers' union). On the other hand, Method 3 can result in unpleasant travel for an extended distance in which the load (occupancy) is above d_{oj} .

To eliminate or control the possibility of such an undesirable phenomenon, we introduce another method, called *Method 4*. This method establishes a level-of-service consideration by restricting the total portion of the route length having loads greater than the desired occupancy. Method 4 takes the explicit form:

$$F_{4j} = \max \left[\frac{A_j}{d_{oj} \cdot L}, \frac{P_{mj}}{c}, F_{mj} \right] \quad (3.6)$$

$$\text{subject to (s.t.)} \quad \sum_{i \in I_j} \ell_i \leq \beta_j \cdot L,$$

where mathematically $I_j = \left\{ i: \frac{P_{ij}}{F_j} > d_{oj} \right\}$; in other words, I_j is the set of all stops i in time

period j , such that the load P_{ij} exceeds the product of d_{oj} times the frequency F_{4j} , and β_j is the allowable portion of the route length at period j in which P_{ij} can exceed the product $F_{4j} \cdot d_{oj}$. The other notations in Equation (3.6) were previously defined. By controlling parameter β_j , it is possible to establish a level-of-service criterion. We should note that for $\beta_j = 0$ and $\beta_j = 1.0$, Method 4 converges to Methods 2 and 3, respectively.

The load profile of Example 1 (see Figure 3.1) is presented in Figure 3.3 pertaining to the hour 9:00–10:00. The *considered load level* associated with Method 3 is simply the area under the load profile, divided by $L = 10$ km, or 188.1 passengers in this case. This is an average load profile. However, all loads between stops 1 and 4 (4.5 km), which stretch out for 45% of the route length, exceed this average. To avoid the load exceeding the desired profile for more than a predetermined % of the route length, Method 4 can be introduced. If this percentage is set at 40% of the route length, the considered load will be 208 passengers, thus allowing only the stretch between stops 1 and 2 (2 km) and that between stops 3 and 4 (1.5 km) to have this excess load. We term this situation Method 4 (40%). Setting the percentage to 20% results in an average of 250 passengers; in the case of 10% (1 km), the considered load level converges to the Method 2 average of 320 passengers.

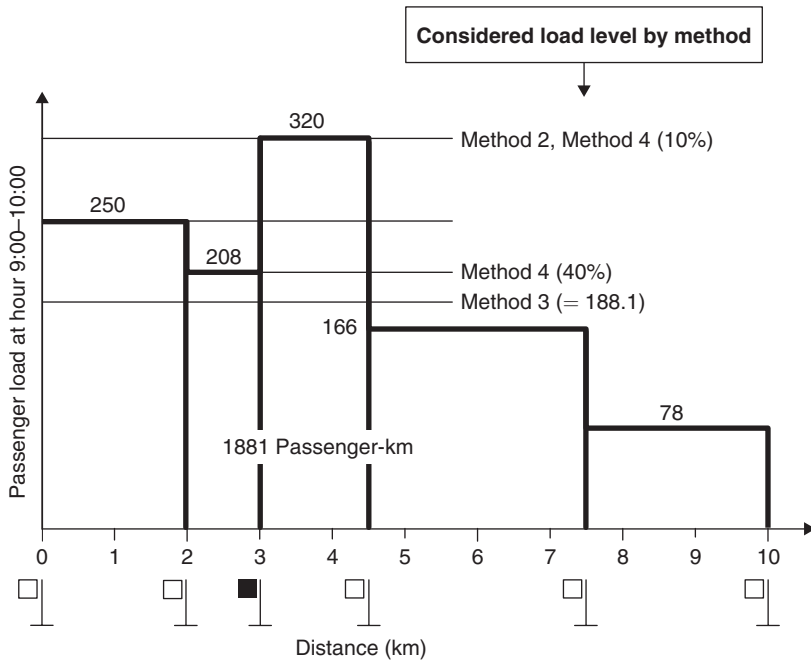


Figure 3.3 Load profile from Example 1 between 9:00 and 10:00 with considered load levels for three methods and Method 4 standing for 10%, 20% and 40% of the route length

Figure 3.4 illustrates the fundamental trade-off between the load profile and max load concepts. We will show it for the case of Method 4 (20%) and the 9:00–10:00 hour. Based on Figure 3.1 data and Equation (3.6), with $\beta_j = 0.2$, we attain $F_4 = \max(250/50, 320/90, 3) = 5$ veh/hr from raising the considered load level from 188.1 to 250 passengers.

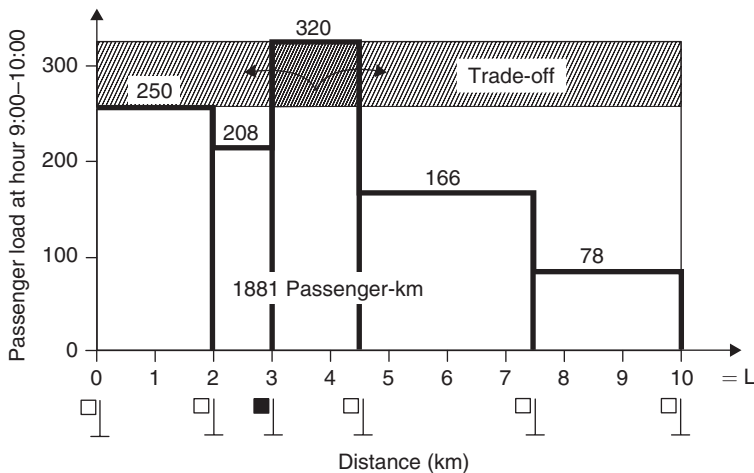


Figure 3.4 Load profile from Example 1 (9:00–10:00) using Method 4 (20%), with an indication of trade-off between this method and Method 2 (more crowding in return for less empty space-km)

The 5 assigned departures will, in an average sense, carry $320/5 = 64$ passengers between stops 3 and 4 (14 more than the desired load of 50). Nonetheless, this excess load for 1.5 is traded off with $(320 - 250) 8.5 = 595$ empty-space km as is shown in Figure 3.4. This trade-off can be interpreted economically and perhaps affect the ticket tariff.

In case the calculated frequency in Equations (3.5) and (3.6) is the result of (P_{mj}/c) , then the considered load will be determined by the product $(P_{mj}/c) \cdot d_{oj}$. Table 3.2 shows the results for Methods 3 and 4, in which the percentage of route length allowed to have an excess load in Method 4 is set at 10%, 20% and 30%.

Table 3.2 Frequency and headway results for Example 1 in Figure 3.1 for Methods 3 and 4

Period j	Method 3		Method 4					
			10%		20%		30%	
	F _{3j} (veh/hr)	H _{3j} (min)	F _{4j} (veh/hr)	H _{4j} (min)	F _{4j} (veh/hr)	H _{4j} (min)	F _{4j} (veh/hr)	H _{4j} (min)
6:00–7:00	4.44	14	8.00	7.5	4.44	14	4.44	14
7:00–8:00	5.88	10	8.40	7.0	8.40	7	6.70	9
8:00–9:00	6.40	9	8.00	7.5	7.00	9	7.00	9
9:00–10:00	3.72	16	6.40	9.0	5.00	12	5.00	12
10:00–11:00	3.07	20	4.40	14	4.40	14	4.00	15

We may observe that in the Method 3 results in Table 3.2, the first hour relies on P_{mj}/c or, specifically, $400/90 = 4.44$ veh/hr. When we turn to Method 4 for this first hour, the results of Method 2 for 10% are attained, since the max load stretches along more than 10% of the route length. For 20% and 30%, the vehicle-capacity constraint still governs. In the second hour, 7:00–8:00, the following is obtained for Method 3: $F_3 = \max(2942/50 \cdot 10, 510/90, 3) = 5.88$ veh/hr. Continuing in the second hour for Method 4 (10%), the average max load of 294.2 passengers rises to 420, resulting in $F_4 = 8.40$ veh/hr. Table 3.2 continues to be fulfilled in the same manner. It should be noted that, as in Table 3.1, all the headways are rounded to their nearest integer.

Although we aim at a resource saving using Methods 3 and 4, there is a question as to whether this saving justifies the additional expense involved in using ride check as opposed to point check. The next section attempts to answer this question by constructing a criterion suggesting when to use the point check or, otherwise, the ride-check data-collection technique.

3.4 Criterion for selecting point check or ride check

This section will test an assumption that particular load-profile characteristics suggest the data-collection technique to be used. The basic idea is to provide the transit agency with adequate preliminary guidance in selecting the type of technique based on old load profiles. The assumption will be investigated whether a relatively flat profile suggests the use of a

point-check procedure (Method 1 or 2) or whether a ride-check procedure (Method 3 or 4) would be more appropriate.

One property of the load profile is its density, ρ . This is the observed measure of total passenger-km (i.e. total ridership over the route), divided by the product of the length of the route and its maximum load. The product, in the denominator, is the passenger-km that would be observed if the max load existed across all stops. Thus, the load-profile density for hour j , ρ_j , is

$$\rho_j = \frac{A_j}{P_{mj} \cdot L} \quad (3.7)$$

The load-profile density is used to examine profile characteristics. High values of ρ indicate a relatively flat profile, whereas low values of ρ indicate load variability among the route stops.

One way to explore load-profile density is to approximate the observed shapes of profile curves through a mathematical model. The log-normal model will be selected for this purpose, since it provides a family of curves that can be controlled by varying the two parameters, μ and σ . The log-normal model takes the form:

$$f(x) = \frac{1}{x \cdot \sigma \sqrt{2\pi}} e^{-(\ln x - \mu)^2 / 2\sigma^2}; \quad x > 0 \quad (3.8)$$

The equation satisfying the optimum conditions, $df(x)/dx = 0$, is $x_0 = e^{\mu - \sigma^2}$, in which Equation (3.8) reaches maximum. In our case, the maximum coincides with the max load. This continuous model can only approximate some of the observed load profiles, since it has only one peak and represents monotonically increasing and decreasing functions before and after this peak, respectively. Nonetheless, this model is useful for observing some general differences between the ride-check and point-check methods.

In order to compare the methods, $f(x)$ will be used as a normalized load (the load divided by the max load) and x as a normalized distance (the distance from the initial route stop divided by the length of the route). At a given time interval of one hour, j , the considered max load is $P_{mj} = 650$ passengers. Given that $d_{oj} = 65$, $c = 100$ and $F_m = 2$, the determined frequency and headway for both Methods 1 and 2 are $F_j = 10$ and $H_j = 6$. By applying this information to Methods 3 and 4, and using a variety of log-normal curves, the frequencies and headways shown in Table 3.3 will be obtained. The results in this table are arranged in order of increasing density. For Method 3, the capacity constraint determines the values of F and H up to and including $\rho = 0.64$, and up to different ρ values (if any) for Method 4. Examples of the log-normal normalized curves are shown in Figures 3.5 and 3.6, the latter showing the determination of the frequency for one curve ($\mu = -1.5$, $\sigma = 1.0$). Given that the normalized route length is 1.0, then the different percentages (for this length) in Figure 3.6 are set for the Method 4 criterion, with an explicit calculation example for the 10% case.

From Table 3.3, it appears that, for Method 3, the ride-check data result in the same rounded headway as the point-check data for $\rho > 0.84$. For Method 4, the ride-check and point-check information tend to yield the same headways for $\rho > 0.34$ (10% case), $\rho > 0.50$ (20% case) and $\rho > 0.64$ (30% case).

In practice, the transit agency wishes to save vehicle runs and, eventually, to perform the matching between demand and supply with fewer vehicles. As we will show in Chapters 4 and 6, different headway values do not necessarily save vehicle runs or reduce the required fleet size. However, the analysis of the profile-density measure can be used by the transit agency as

Table 3.3 Frequencies (*F*), rounded headways (*H*), and density (ρ) for different load-profile configurations (log-normal-model based), using Methods 3 and 4

Log-normal model parameters		Profile density	Method 4							
			Method 3		10%		20%		30%	
μ	σ	ρ	F	H	F	H	F	H	F	H
−1.0	0.2	0.18	6.50*	9	7.60	7	6.50*	9	6.50*	9
−1.5	0.5	0.25	6.50*	9	8.46	7	6.50*	9	6.50*	9
−1.5	1.0	0.32	6.50*	9	8.36	7	6.50*	9	6.50*	9
−1.0	1.5	0.34	6.50*	9	7.55	7	6.50*	9	6.50*	9
−1.0	1.2	0.43	6.50*	9	9.00	6^	7.05	8	6.50*	9
−1.0	0.6	0.44	6.50*	9	9.50	6^	8.10	7	6.50*	9
−0.8	0.5	0.47	6.50*	9	9.55	6^	8.40	7	6.50*	9
−1.0	0.9	0.48	6.50*	9	9.45	6^	8.05	7	7.05	8
−0.4	1.5	0.50	6.50*	9	9.05	6^	7.35	8	6.50	9
0.0	0.5	0.56	6.50*	9	9.92	6^	9.67	6^	9.27	9
−0.5	0.5	0.57	6.50*	9	9.76	6^	9.11	6^	8.16	6
−0.4	0.5	0.59	6.50*	9	9.81	6^	9.31	6^	8.46	7
−0.4	1.2	0.62	6.50*	9	9.65	6^	8.85	6^	7.80	7
−0.5	1.0	0.64	6.50*	9	9.79	6^	9.04	6^	8.19	7
−0.4	0.8	0.68	6.77	8	9.87	6^	9.42	6^	8.72	7
0.5	1.5	0.75	7.46	8	9.86	6^	9.36	6^	8.76	6^
0.0	1.0	0.76	7.63	7	9.93	6^	9.68	6^	9.23	6^
0.5	1.0	0.78	7.77	7	9.97	6^	9.87	6^	9.72	6^
1.0	1.5	0.84	8.41	7	9.96	6^	9.76	6^	9.46	6^
1.5	1.5	0.87	8.72	6	9.97	6^	9.92	6^	9.82	6^

Notes: For Methods 1 and 2: $F = 10$, $H = 6$, where $d_0 = 65$, $c = 100$
* Whenever $F = 6.5$, $H = 9$, the capacity constraint is met
^ Whenever $H = 6$, ride check and point check yield the same headway

a preliminary check before embarking on a more comprehensive analysis. The following are practical observations: (i) for densities below 0.5, savings are likely to result by gathering the load-profile information and using ride-check methods (alternatively, the profile can be examined with such low ρ values for short-turn strategies, to be discussed in Chapter 14);

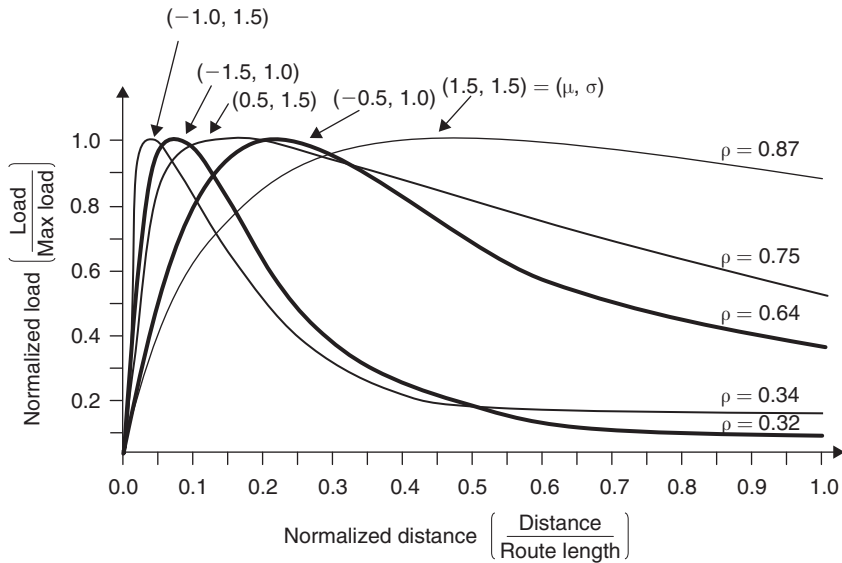


Figure 3.5 Five approximated load profiles based on the log-normal model with normalized scales

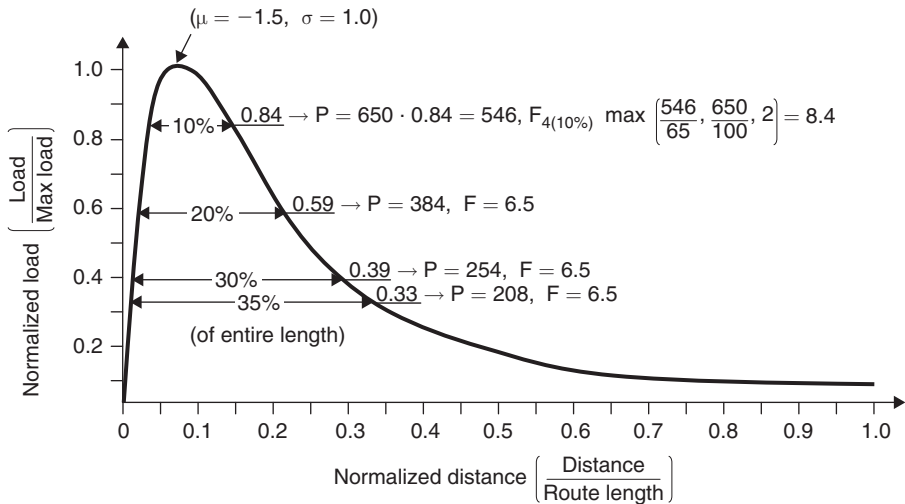


Figure 3.6 An approximated load profile based on the log-normal model, with indications of different (route) length percentages associated with Method 4

(ii) for densities between 0.5 and 0.85, actual comparisons can be made between the point-check and ride-check methods, along with further saving considerations (for constructing timetables and vehicle schedules); and (iii) for densities above 0.85, it is likely that the majority of the required information for the headway calculation can be obtained from a point-check procedure (either Methods 1 or 2). A further simplified and explicit practical criterion is to use Methods 3 and 4 for $\rho \leq 0.5$, and Methods 1 and 2 otherwise. It may be argued that for the range $0.5 < \rho \leq 0.85$, the use of load-profile methods cannot produce significant gains over the max

load methods although theoretically it may be justified. This argument is supported by the relatively small amount of passenger-km in the trade-off situation exhibited in Figure 3.4.

3.5 Conclusion (two examples)

Another simple example, which we call Example 2, will now ascertain the procedures employed. This example will be followed by a real-life example that includes the results of the four methods. This section will end by portraying the methodology proposed in flowchart form, with an orientation towards practice.

Example 2 which is presented and illustrated in Figure 3.7, will also be used throughout Chapters 4 and 5. The basic required input consists of: (1) distance between stops; (2) desired

Stop no.	Distance (km) to next stop	Average observed load (passenger), by scheduled departure time					Average load, by hour		Total hourly load (passengers)
		6:15	6:50	7:15	7:35	7:50	6:00–7:00	7:00–8:00	
1	2	22	25	50	60	45	67	135	202
2	1.5	52	40	90	87	75	128	(216)	[344]
3	4.5	35	65	85	44	83	(134)	178	312
Number of observed scheduled vehicles							2	2	Calculated
Desired occupancy							50	60	
Minimum frequency							1	2	
Vehicle capacity (seats + allowed standees)							90	90	
Area under the load profile (passenger-km)							929	1395	

See calculation in Figure 3.8.

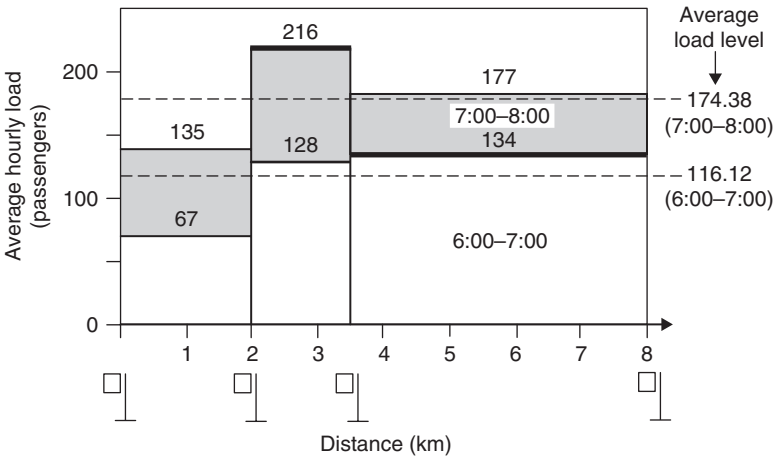


Figure 3.7 Example 2: Three-hour data for frequency and headway determination by each of the four methods

occupancy per hour (for every planned vehicle type); (3) minimum frequency per hour (could be the inverse of policy headway); (4) number of observed (scheduled usually) vehicles each hour; (5) observed load (an average value or a consideration of its variability) between both adjacent stops per hour; and (6) vehicle capacity (for every planned vehicle type). A two-hour operation period (06:00–08:00) is chosen for simplicity in this example. Furthermore, as a matter of decision, all vehicles depart on the hour pertaining to this next hour.

Let us recall the assumption that the observed loads in each hour are based on a uniform passenger-arrival rate (demand). That is, the number of passengers carried by the first vehicle in each hour is divided proportionally to reflect the demand in two time intervals: (i) the start of the hour and the departure time of the first vehicle that hour; and (ii) last departure time in the previous hour and the start of the considered hour. Hence, vehicles departing on the hour carry a demand from the previous hour. Figure 3.8 interprets this proportionality for the 7:15 departure; in this figure, five load profiles are shown for each observed vehicle. We assume that the load on the 7:15 vehicle contains 2/5 of the 6:00–7:00 demand (10 minutes from 6:50 to 7:00, and 15 minutes from 7:00 to 7:15). This proportion is marked in Figure 3.8 and inserted in the table of Figure 3.7 under average load by hour.

We can see from the table in Figure 3.7 that the daily max load point is stop 2, with stops 3 and 2 being the hourly max points for the 6:00–7:00 and 7:00–8:00 hours, respectively. The load profiles are depicted on the same scales in Figure 3.7. The last column, last row, and ‘average load, by hour’ column in Figure 3.7 are calculated items. The difference between Methods 1 and 2 is simply between the set of counts (128, 216) and (134, 216), respectively. Using Equation (3.4), we obtain $\chi^2 = 0.269$, where $\chi^2 = 3.84$ for $\alpha = 0.05$ with a single d.f. Hence, Method 1 can substitute for Method 2 for future passenger counts. Moreover, by using Equation (3.7) and Figure 3.7, we obtain $\rho_{6-7} = 0.87$ and $\rho_{7-8} = 0.81$, meaning that the max load point data-collection procedure is acceptable.

The results of Example 2 for all four methods are shown in Table 3.4 along with their calculations. Equations (3.2) and (3.3) may be applied for Methods 1 and 2, respectively, and Equations (3.5) and (3.6) for Methods 3 and 4, respectively. Table 3.4 considers, for Method 4, the constraint that no more than 20% of the route length is allowed to have an excess load,

Table 3.4 Frequency (F) and rounded headway (H) results for Example 2, by each of the four methods

Hour	Method 1		Method 2		Method 3		Method 4 (20%)	
	F	H	F	H	F	H	F	H
6:00–7:00	2.56	23	2.68	22	2.32	26	2.68	22
7:00–8:00	3.60	17	3.60	17	2.91	21	2.95	20
Calculation	6:00–7:00		6:00–7:00		6:00–7:00		6:00–7:00	
	$\max\left(\frac{128}{50}, 1\right) = 2.56$		$\max\left(\frac{134}{50}, 1\right) = 2.68$		$\max\left(\frac{929}{50.8}, \frac{134}{90}, 1\right) = 2.32$		$\max\left(\frac{134}{50}, \frac{134}{90}, 1\right) = 2.68$	
Calculation	7:00–8:00		7:00–8:00		7:00–8:00		7:00–8:00	
	$\max\left(\frac{216}{60}, 2\right) = 3.60$		$\max\left(\frac{216}{60}, 2\right) = 3.60$		$\max\left(\frac{1395}{60.8}, \frac{216}{90}, 2\right) = 2.91$		$\max\left(\frac{177}{60}, \frac{216}{90}, 2\right) = 2.95$	

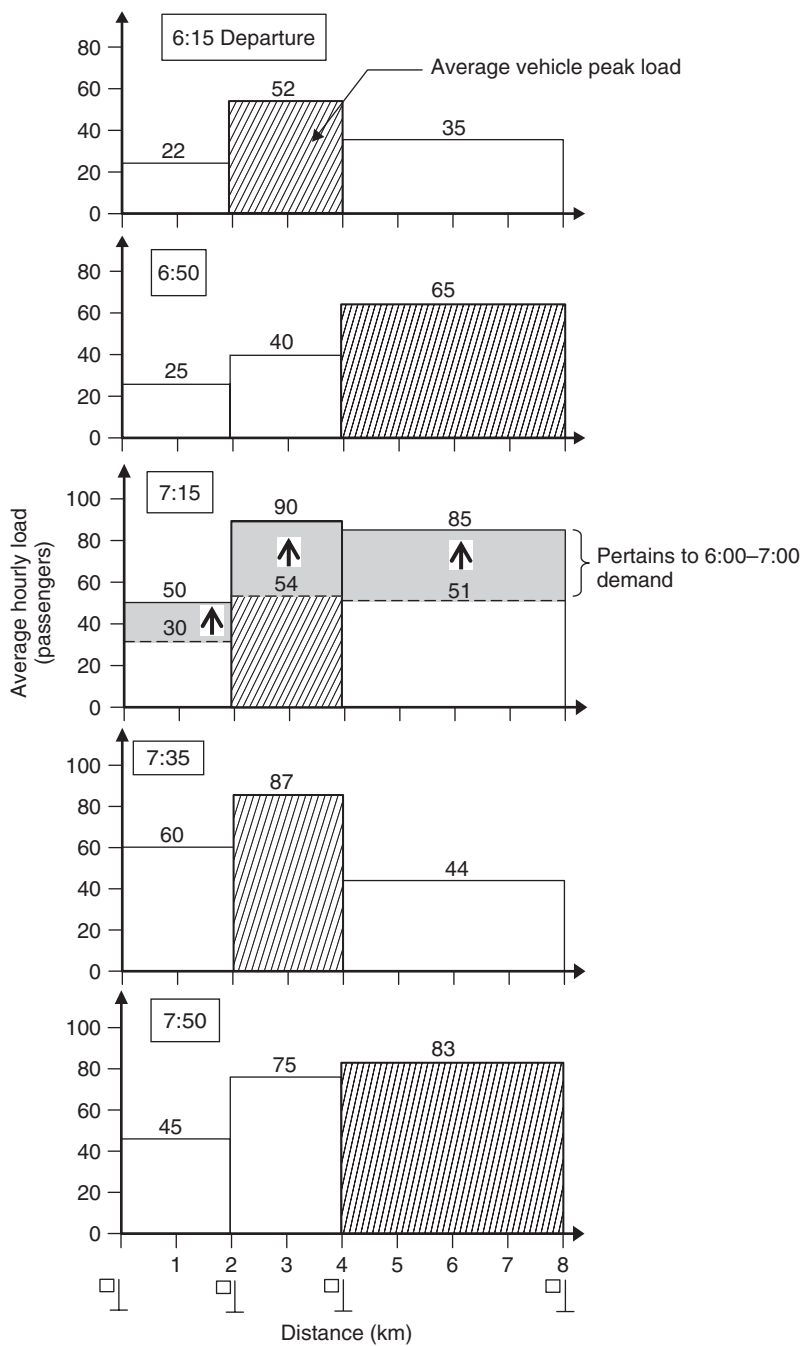


Figure 3.8 Example 2 data by vehicle load, showing the demand (passenger-km) that pertains to the 6:00–7:00 period, assuming uniform passenger-arrival rates

provided that the considered load level for this excess load is equal to or below the hourly max load divided by vehicle capacity. The headways in Table 3.4 are rounded off to their nearest integer. The frequencies in this table will be used in the next chapter (using the same example) as non-integer numbers while attempting to construct alternative (and automated) timetables.

For a real-life situation, we will demonstrate the procedures developed when using ride-check data from an old Los Angeles Metro (previously called SCRTD) bus route. This route, 217, was considered a heavy route and was characterized by 60 stops. All the trips on this route cross the daily max load point. The complete northbound ride-check input for route 217 appears in Table 3.5. The content of this table is comprehensive: it includes the distances (in km) to the next stop and the stop name in the first and second columns, respectively; the last column represents the total load for the whole day. The entries in Table 3.5 are based on rounded average values following several checks. Vehicle capacity, required for Methods 3 and 4, is 80 passengers.

An automated program is constructed for the four methods to be commensurate with Equations (3.2), (3.3), (3.5), and (3.6). The intermediate result of this program, which concerns max load information, is illustrated in Table 3.6. The daily max load point is the Fairfax/Rosewood stop, with a total of 4,413 observed passengers over the entire day. The hourly max load points for each hour appear in Table 3.6. Occasionally there are multiple-stop results; e.g. at 10:00–11:00 and 22:00–23:00. The results of the load-profile density measure reveal that all $\rho < 0.5$; hence, the ride-check count is the data-collection technique appropriate for this route. The frequency and headway results for route 217 are summarized in Table 3.7 for all four methods. The chi-square comparison between Methods 1 and 2 for every hour over all hours, using $\alpha = 0.05$, shows that the hypothesis concerning equal methods is rejected; therefore, the hourly max load data-collection technique should be applied for this data set. Nonetheless, as we mentioned before, the ride check is the appropriate technique. The full results, in Table 3.7, are presented solely for comparison purposes; they show that from 6:00–7:00 and after 19:00, route 217 relies on its minimum frequency. Methods 2 and 3 serve as upper and lower bounds on the frequency, respectively; and Method 4 ranges between the two, depending on the percentage of route length allowed to have excess load.

A graphical comparison of the frequency results of Methods 2, 3 and 4 (20%) for both directions of route 217 is presented in Figure 3.9. This figure also contains the actual observed (scheduled) frequency. We can easily see that the observed frequencies in both directions represent an excessive amount of bus runs. The transit agency provided the input, including the desired occupancy, set forth by the planners. The method used by the agency's planning department overlapped with Method 2. The observed frequency, therefore, poses an apparent enigma: why does the scheduled frequency not overlap, or at least come close to, the results of Method 2? More in-depth investigation can provide an answer, which has to do with the objective function of the drivers who attempt to maximize their comfort. When a bus run is overloaded repeatedly, the driver requests reinforcement; that is, an added run. When a bus run is almost empty, silence becomes golden. With time, the number of bus runs only increases, and this is the main reason behind the phenomenon exhibited in Figure 3.9. Incidentally, the agency's schedulers fixed their schedule following the presentation of the Figure 3.9 results. Nonetheless, the actual data support the use of Method 4 (20%) from a resource-saving perspective. It is interesting to note that Method 4 (20%) results in a much lower frequency than does Method 2, particularly in the southbound direction. The lower bound on the frequency needed to accommodate the passenger load while

Table 3.5 Ride-check data for LA bus route 217 northbound

Time period			6:00– 7:00	7:00– 8:00	8:00– 9:00	9:00– 10:00	10:00– 11:00	11:00– 12:00	12:00– 13:00	13:00– 14:00
No. of buses observed			6	10	8	8	8	7	8	8
Minimum frequency (veh/hr)			2	2	2	2	2	2	2	2
Desired occupancy (passengers)			60	70	70	60	50	50	50	50
Dist. (km)	Stop name									
0.16	Adams	/Washingt	13	27	115	24	11	11	6	4
0.26	Fairfax	/Adams	21	68	148	39	16	27	26	8
0.27	Fairfax	/Washingt	22	89	160	48	22	32	29	15
0.26	Fairfax	/Apple	25	101	163	51	26	34	29	16
0.27	Fairfax	/Venice	29	112	183	53	37	43	37	22
0.27	Fairfax	/Venice F	37	124	217	84	52	44	51	38
0.24	Fairfax	/18TH	40	119	188	83	51	45	52	39
0.24	Fairfax	/Airdrome	40	131	192	89	52	49	54	42
0.24	Fairfax	/Pickford	45	151	195	94	48	50	54	42
0.24	Fairfax	/Saturn	48	167	211	97	58	50	55	45
0.24	Fairfax	/Pico	56	217	246	116	83	78	84	95
0.26	Fairfax	/Packard	59	228	252	120	91	80	87	98
0.26	Fairfax	/Whitwort	63	250	257	125	98	87	90	100
0.00	Fairfax	/Olympic	59	244	275	130	108	98	103	120
0.24	Olympic	/Ogden	59	332	278	152	123	99	116	142
0.22	Fairfax	/San Vice	70	355	325	171	144	124	141	181
0.24	Fairfax	/8th St	66	357	330	175	150	128	145	188
0.29	Fairfax	/Wilshire	54	363	349	236	288	250	354	356
0.29	Fairfax	/6th St	55	369	351	238	291	257	363	365
0.29	Fairfax	/Drexel	54	376	355	234	291	258	381	378
0.29	Fairfax	/3rd St	48	401	370	234	261	258	338	381
0.29	Fairfax	/1st St	48	400	366	232	265	256	339	385
0.34	Fairfax	/Beverly	44	392	354	232	249	252	314	362

Table 3.5 Ride-check data for LA bus route 217 northbound (Continued)

Time period			6:00– 7:00	7:00– 8:00	8:00– 9:00	9:00– 10:00	10:00– 11:00	11:00– 12:00	12:00– 13:00	13:00– 14:00
No. of buses observed			6	10	8	8	8	7	8	8
Minimum frequency (veh/hr)			2	2	2	2	2	2	2	2
Desired occupancy (passengers)			60	70	70	60	50	50	50	50
Dist. (km)	Stop name									
0.35	Fairfax	/Oakwood	52	391	351	237	271	282	315	366
0.34	Fairfax	/Rosewood	49	370	326	246	288	292	327	376
0.35	Fairfax	/Melrose	46	113	175	165	245	265	308	378
0.34	Fairfax	/Willowgh	44	112	173	155	238	246	291	358
0.35	Fairfax	/Santa Mo	33	90	140	148	154	174	216	292
0.35	Fairfax	/Fountain	35	92	134	153	149	172	208	281
0.29	Fairfax	/Sunset	32	70	95	132	127	144	172	249
0.29	Sunset	/Genesee	31	69	90	133	125	143	170	248
0.29	Sunset	/Stanley	30	71	92	135	127	135	171	248
0.29	Sunset	/Gardner	31	73	85	132	117	124	165	240
0.29	Sunset	/Martel	32	75	82	131	110	116	164	228
0.29	Sunset	/Poinsett	33	73	77	127	118	115	171	220
0.30	La Brea	/Sunset	36	65	70	122	113	106	163	211
0.18	La Brea	/Hollywood	36	62	68	118	116	98	156	199
0.18	Hollywood	/Sycamore	36	65	67	118	116	104	155	192
0.18	Hollywood	/Orange	32	62	66	112	112	107	152	176
0.18	Hollywood	/Highland	19	33	49	84	90	94	130	147
0.18	Hollywood	/Las Palm	16	26	45	77	83	89	120	126
0.18	Hollywood	/Whitley	15	22	42	62	67	76	111	92
0.18	Hollywood	/Wilcox	14	21	40	54	62	66	105	84
0.19	Hollywood	/Cahuenga	11	20	36	44	61	66	90	80
0.19	Hollywood	/Ivar	9	16	33	33	50	57	73	61

14:00– 15:00	15:00– 16:00	16:00– 17:00	17:00– 18:00	18:00– 19:00	19:00– 20:00	20:00– 21:00	21:00– 22:00	22:00– 23:00	23:00– 24:00	24:00– 25:00	
8	10	8	9	6	5	3	3	2	2	2	
2	2	2	2	2	2	2	2	2	2	2	
50	60	70	70	60	60	50	50	50	50	50	
											Total
352	416	467	370	220	106	57	48	51	22	9	4383
367	418	481	371	214	102	57	50	51	20	8	4413
365	427	439	370	201	101	51	47	50	20	8	3774
344	406	411	353	191	91	51	45	48	19	8	3584
273	322	353	299	171	97	45	50	51	19	4	2931
244	300	321	283	157	89	42	49	49	20	4	2782
210	270	280	264	144	80	44	48	48	20	4	2433
206	269	287	259	146	80	42	48	47	20	4	2417
206	260	288	253	145	77	43	46	48	18	4	2397
210	249	274	241	138	83	43	46	45	17	4	2317
208	236	263	228	128	82	40	47	45	17	4	2236
193	232	260	222	129	81	38	55	46	18	4	2212
177	222	252	218	132	86	39	54	45	17	2	2130
172	213	242	206	127	81	36	56	42	16	1	2045
165	208	225	194	122	78	38	56	45	17	1	2002
166	197	204	179	113	75	36	54	43	17	1	1904
154	185	146	146	88	65	33	55	35	16	2	1571
140	168	141	139	85	58	31	55	34	15	2	1450
116	144	129	116	68	43	27	53	31	13	2	1229
108	131	119	98	58	38	18	44	27	13	2	1102
96	123	107	85	50	40	16	41	25	13	1	1005
84	111	98	84	47	35	14	29	24	10	1	869

(Continued)

Table 3.5 Ride-check data for LA bus route 217 northbound (Continued)

Time period			6:00– 7:00	7:00– 8:00	8:00– 9:00	9:00– 10:00	10:00– 11:00	11:00– 12:00	12:00– 13:00	13:00– 14:00
No. of buses observed			6	10	8	8	8	7	8	8
Minimum frequency (veh/hr)			2	2	2	2	2	2	2	2
Desired occupancy (passengers)			60	70	70	60	50	50	50	50
Dist. (km)	Stop name									
0.19	Hollywood	/Vine	3	14	19	17	31	39	35	33
0.19	Argyle	/Hollywood	4	20	17	20	31	35	32	44
0.19	Argyle	/Yucca	3	20	17	19	30	28	32	44
0.27	Franklin	/Argyle	2	16	17	16	21	22	29	34
0.11	Gower	/Franklin	2	8	13	11	10	16	21	21
0.16	Beachwood	/Franklin	2	11	11	12	8	9	15	12
0.18	Beachwood	/Midway	2	11	11	11	8	6	14	12
0.18	Beachwood	/Scenic	2	10	11	10	7	5	13	11
0.16	Beachwood	/Temple H	2	9	10	9	7	4	10	9
0.18	Beachwood	/Winans	1	7	10	9	7	4	5	6
0.18	Beachwood	/Cheremoy	1	7	9	6	7	3	5	4
0.18	Beachwood	/Glen Ald	1	7	7	6	6	3	4	4
0.19	Beachwood	/Glen Dak	1	7	7	6	6	3	2	4
0.21	Beachwood	/Westshir	0	7	6	6	2	3	2	3
0.00	Beachwood	/Westshir	–	–	–	–	–	–	–	–

neglecting the load factor (desired occupancy) is exhibited by the results of Method 3. In most hours, it is as much as half the observed frequency.

Finally, the procedures described and interpreted in this chapter will be demonstrated, using a flowchart (Figure 3.10), which provides a pragmatic overview leading to the use of either max load-point method (Method 1 or 2) or load-profile Method 4. The method chosen has a direct impact on the technique required for data acquisition. From a practitioner’s perspective, it may be noted that the noticeable importance of ridership information has led transit agencies to introduce automated surveillance techniques mixed with manually collected data.

Many transit agencies worldwide are considering implementing APC (automatic passenger counters). Such systems are based on: (i) infrared beam interruption; (ii) pressure-sensitive stairwell mats; (iii) counts by weighing the load on the vehicle and (iv) ultrasonic beam

14:00– 15:00	15:00– 16:00	16:00– 17:00	17:00– 18:00	18:00– 19:00	19:00– 20:00	20:00– 21:00	21:00– 22:00	22:00– 23:00	23:00– 24:00	24:00– 25:00	
8	10	8	9	6	5	3	3	2	2	2	
2	2	2	2	2	2	2	2	2	2	2	
50	60	70	70	60	60	50	50	50	50	50	
											Total
68	87	73	67	47	25	14	24	17	9	1	623
69	91	83	66	47	20	16	24	15	10	0	644
68	81	79	64	44	15	13	21	11	9	0	598
47	64	59	47	25	11	10	17	8	6	0	451
27	53	38	30	10	8	8	6	0	0	0	282
25	49	21	24	7	4	6	3	0	0	0	219
25	48	19	19	5	2	5	3	0	0	0	201
16	44	17	18	4	2	5	3	0	0	0	178
14	44	14	13	3	2	3	3	0	0	0	156
13	39	13	12	1	1	3	3	0	0	0	134
10	34	11	7	1	1	3	3	0	0	0	112
8	30	10	5	1	1	2	1	0	0	0	96
8	26	8	4	1	1	2	1	0	0	0	87
0	5	5	2	1	0	0	0	0	0	0	42
–	–	–	–	–	–	–	–	0	0	0	–

interruption. Naturally, the transit agencies expected to gain useful information for operations planning by obtaining more accurate passenger counts. However, as mentioned above, there is always a point at which the additional accuracy is not worth the accompanying costs of data collection and analysis. Finally, in light of the methods proposed, as opposed to existing methods in use, the following rule may apply: forgetting is as important as remembering in the practical use of what is right to do.

3.6 Literature review and further reading

Only papers that focus on the determination of frequencies or headways with a linkage to the creation of timetables are reviewed in this section. Models in which headways or

Table 3.6 Max load information for LA route 217 northbound, using Methods 1 and 2

Time period	Max load point		Max load (passengers)
06:00–07:00	Fairfax	/San Vice	70
07:00–08:00	Fairfax	/3rd St	401
08:00–0900	Fairfax	/3rd St	370
09:00–10:00	Fairfax	/Rosewood	246
10:00–1100	Fairfax	/6th St	291
	Fairfax	/Drexel	
11:00–12:00	Fairfax	/Rosewood	292
12:00–13:00	Fairfax	/Drexel	381
13:00–14:00	Fairfax	/1st St	385
14:00–15:00	Fairfax	/1st St	373
15:00–16:00	Fairfax	/Melrose	427
16:00–17:00	Fairfax	/Rosewood	481
17:00–18:00	Fairfax	/Beverly	377
18:00–19:00	Fairfax	/Oakwood	220
19:00–20:00	Fairfax	/Oakwood	106
20:00–21:00	Fairfax	/Beverly	59
21:00–22:00	La Brea Hollywood	/Hollywood /Sycamore	56
22:00–23:00	Fairfax	/Oakwood	51
	Fairfax	/Rosewood	
	Fairfax	/Santa Monica	
23:00–24:00	Fairfax	/Oakwood	22
24:00–25:00	Fairfax	/Drexel	9
	Fairfax	/3rd St	
	Fairfax	/1st St	
	Fairfax	/Beverly	
	Fairfax	/Oakwood	
All day	Fairfax	/Rosewood	4413

Table 3.7 Frequency and headway results for LA route 217 northbound, by each of the four methods

Time period	Method 1		Method 2		Method 3		Method 4					
	F (veh/hr)	H (minutes)					10%		20%		30%	
			F	H	F	H	F	H	F	H		
06:00–07:00	2.00	30	2.00	30	2.00	30	2.00	30	2.00	30	2.00	30
07:00–08:00	5.28	11	5.72	10	5.01	12	5.41	11	5.11	12	5.01	12
08:00–09:00	4.65	13	5.28	11	4.62	13	5.02	12	4.72	13	4.62	13
09:00–10:00	4.09	15	4.09	15	3.07	20	3.97	15	3.07	20	3.07	20
10:00–11:00	5.75	10	5.82	10	3.63	17	5.43	11	4.93	12	3.63	17
11:00–12:00	5.83	10	5.83	10	3.65	16	5.25	11	5.05	12	3.65	16
12:00–13:00	6.53	9	7.61	8	4.76	13	6.76	9	6.16	10	4.76	13
13:00–14:00	7.51	8	7.69	8	4.81	12	7.61	8	7.21	8	5.01	12
14:00–15:00	7.33	8	7.46	8	4.66	13	7.06	8	6.96	9	4.66	13
15:00–16:00	6.96	9	7.11	8	5.33	11	6.93	9	6.73	9	5.33	11
16:00–17:00	6.87	9	6.87	9	6.01	10	6.31	10	6.01	10	6.01	10
17:00–18:00	5.30	11	5.38	11	4.71	13	5.31	11	4.91	12	4.71	13
18:00–19:00	3.56	17	3.66	16	2.75	22	3.45	17	3.25	18	2.75	22
19:00–20:00	2.00	30	2.00	30	2.00	30	2.00	30	2.00	30	2.00	30
20:00–21:00	2.00	30	2.00	30	2.00	30	2.00	30	2.00	30	2.00	30
21:00–22:00	2.00	30	2.00	30	2.00	30	2.00	30	2.00	30	2.00	30
22:00–23:00	2.00	30	2.00	30	2.00	30	2.00	30	2.00	30	2.00	30
23:00–24:00	2.00	30	2.00	30	2.00	30	2.00	30	2.00	30	2.00	30
24:00–25:00	2.00	30	2.00	30	2.00	30	2.00	30	2.00	30	2.00	30

frequencies are set as part of a comprehensive network-design process are reviewed in Chapter 14. Methods for setting headways and frequencies for feeder routes are discussed in Chapter 16.

Furth and Wilson (1981) mention four common approaches for determining the headways of transit routes: policy headways, which are not directly derived from passenger demand; headways determined according to peak passenger load and vehicle capacity; headways designed such that the revenue/cost ratio will not exceed a preset value; and headways designed to achieve a desired ratio of passenger miles or vehicle miles per hour. An optimization (non-linear programming) model and a solution algorithm are developed for setting frequencies in a given route network. The model assumes that the demand on each line is elastic, i.e. sensitive

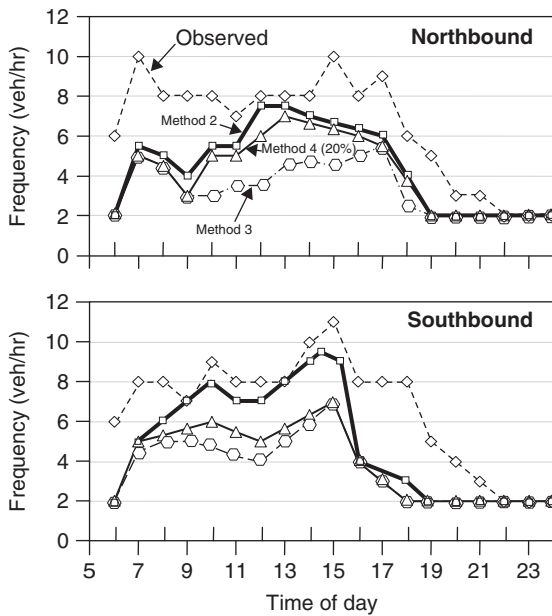


Figure 3.9 Comparison of frequency results of Methods 2, 3 and 4 (20%) and the observed frequency of route 217 in Los Angeles

to frequency changes. However, there are no demand relationships between different routes. The model allocates available buses to routes, subject to subsidy, fleet size and vehicle-loading-level constraints. Multiple day periods are taken into account. The frequencies determined depend on fares, subsidies, fleet size and the value of waiting time.

Koutsopoulos *et al.* (1985) develop a programming problem for determining frequencies in a transit network with a demand that varies along daily periods. Operating costs and travel times are assumed to be time varying, as well. The optimization is subject to subsidy, fleet size and vehicle-capacity constraints. The model is not based on the common assumption of a half-headway average waiting time, but on a more sophisticated waiting-time sub-model; for example, the situation in which some passengers cannot board the first arriving bus because of crowding is considered. Another sub-model measures passenger inconvenience, which is included in the total cost that the model aims to minimize. Since the solution procedure of the nonlinear problem presented may be complicated, the authors propose a simplified version of the initial formulation, in which the daily periods are divided into sub-periods, during which headways are constant. The simplified problem can be solved by, for example, using linear programming.

LeBlanc (1988) introduced a model for determining frequencies, using a modal-split assignment programming model with distinct transit routes. The author shows how to refine conventional modal-split assignment models to include features of the proposed methodology. Multimodal considerations, such as the proportion of transit riders and the influence of the determined frequencies on road congestion, are taken into account.

Banks (1990) presents a model for setting headways in a transit-route system. A case in which route demands vary with frequency is compared to a fixed-demand case; in the

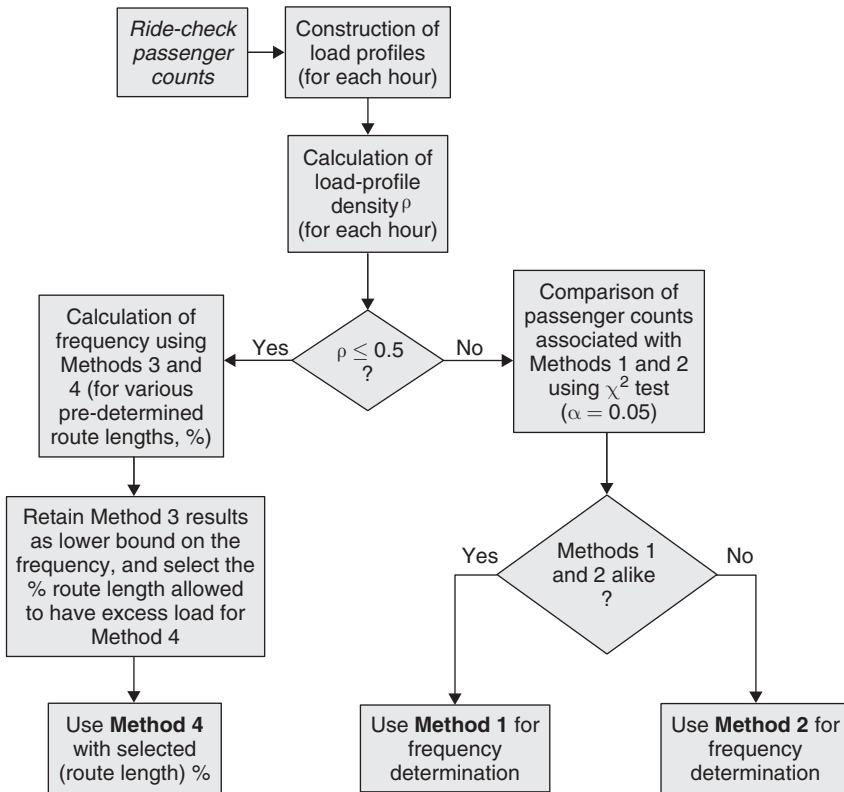


Figure 3.10 Summary in flowchart form of procedures for determining frequency

frequency-dependent demand case, there is no inter-line dependency, and demand functions are assumed to be unknown. An optimal solution is discussed in an unconstrained case and in several cases with subsidy, fleet-size and capacity constraints. In the unconstrained case, a complete analytical solution is presented, showing that the optimal headway is proportional to the square root of operating cost and travel time. It is also inversely proportional to the square root of waiting time and cost and the number of passengers. No analytical expressions for optimal headway are developed for the constrained cases.

Although Khasnabis and Rudraraju (1997) do not propose a methodology for frequency determination, they use simulation experiments to show that pre-emption strategies that are used in signalized intersections along the bus route are an important factor that should be taken into consideration when setting service headway.

Wirasinghe (2003) examines the validity of a traditional method, formulated by Newell, for determining frequencies. According to this method, frequency is proportional to the square root of the arrival rate of passengers if the vehicles are sufficiently large; and to the arrival rate, otherwise. According to a modification of this method, suggested by Newell as well, optimal frequency is also proportional to the square root of the ratio of the value of the passengers' waiting time to the cost of dispatching a vehicle. The validity of Newell's formulas is examined in several situations, based on assumptions that are different from the original: uniform headway

during off-peak periods; policy headways; alternative waiting-time assumptions; stochastic demand; many-to-many demand. Wirasinghe concludes that the original ‘square root policy’ proposed by Newell is applicable, with some modifications, under most conditions.

Table 3.8 summarizes, by different categories, the main features and characteristics of the literature reviewed. Although we are finished with the four methods described in this chapter, we should not count our chickens before they are hatched. That is, we still need to check whether our results fit the other parts of the operational planning process. Consequently, following the calculation of frequencies and headways, we will proceed to the next chapter, which deals with the construction of alternative public timetables.

Table 3.8 *Characteristics of the literature covered in Section 3.6*

Source	Demand features	Constraints	Other features	Variables in the frequency formula derived
Furth and Wilson (1981)	Elastic frequency but without inter-line influence; multi-period	Subsidy, fleet size, vehicle-loading level		Fares, subsidy, fleet size, value of waiting time
Koutsopoulos <i>et al.</i> (1985)	Multi-period	Subsidy, fleet size, vehicle capacity	Waiting-time sub-model, passenger inconvenience sub-model	
LeBlanc (1988)			Multimodal considerations	
Banks (1990)	Elastic/fixed cases examined	Subsidy, fleet size, vehicle capacity (several cases examined, including unconstrained cases)		Operating cost, travel time, value of waiting time, demand level (in the unconstrained case)
Khasnabis and Rudraraju (1997)			Emphasis on the need to consider signal pre-emption strategies	
Wirasinghe (2003)	Among the cases discussed: stochastic demand, one-to-many demand, many-to-many demand		Alternative waiting-time assumptions	Operating cost, value of waiting time, passenger-arrival rate

Exercises

- 3.1 Given a bus route with 6 stops and all buses on the route having 50 seats and the same total capacity of 80 passengers, the table below provides average passenger-load counts in each of the five hours between 06:00 and 11:00, the number of buses observed, desired occupancy (passengers), and minimum frequency (bus/hr). Construct a load profile for each hour and calculate its area; determine the future data-collection technique proposed for each hour. Explain.

Stop no.	Distance to next stop (km)	Time period				
		06:00– 07:00	07:00– 08:00	08:00– 09:00	09:00– 10:00	10:00– 11:00
1	2.2	72	161	65	182	138
2	0.8	90	328	199	318	193
3	1.4	85	468	365	300	222
4	0.6	68	397	388	212	166
5	1	54	286	140	147	84
Number of observed buses	2	6	5	4	3	
Desired occupancy (passengers)	50	60	60	50	50	
Minimum frequency	2	2	2	2	2	

- Calculate the frequency and headway for each hour with each of the four methods, Method 4 being associated with a maximum of 20% of the route length having an excess load.
 - Compare the results of the four methods and draw conclusions.
 - What is the trade-off between excess load (in passenger-km) using Methods 3 and 4 (20%) and the reduced empty space-km for each hour?
 - If the cost of a single empty space-km is 3¢/km (operations cost), what is the range of a reduced single tariff per km for route segments suffering excess load that yield a total saving (cost associated with reduced empty space-km) larger than the total reduced (farebox) income?
- 3.2 Given a transit route and two-hour data with a bus capacity of 80 passengers, the basic ride-check data is shown below.
- Determine the hourly max loads and max load points.
 - Determine the frequency for each hour, using Method 2.

Stop no.	Distance to next stop (km)	Period and departure times			
		06:00–07:00		07:00–08:00	
		06:15	06:45	07:10	07:40
1	2	22	25	50	60
2	3	52	40	60	75
3	3	35	65	80	45
Desired occupancy (passengers)		40		50	
Minimum frequency (veh/hr)		2		3	

(c) Determine the frequency using Method 4 (40%), in which a maximum of 40% of the route length has excess load.

References

Banks, J. H. (1990). Optimal headways for multi-route transit systems. *Journal of Advanced Transportation*, **24**, 127–154.

Ceder, A. (1984). Bus frequency determination using passenger count data. *Transportation Research*, **18A** (5/6), 439–453.

Ceder, A. and Dressler, O. (1980). A note on the χ^2 test with applications and results of road accidents in construction zones. *Accident Analysis and Prevention*, **12**, 7–10.

Furth, P. G. and Wilson, W. H. M. (1981). Setting frequencies on bus routes: Theory and practice. *Transportation Research Record*, **818**, 1–7.

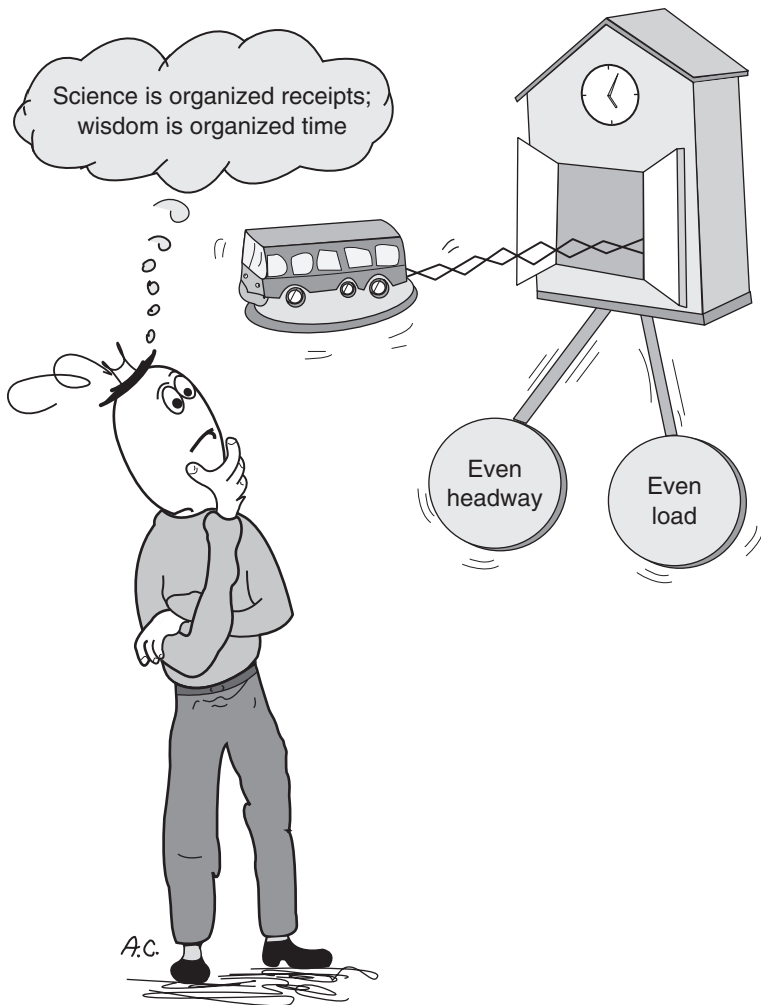
Khasnabis, S. and Rudraraju, R. K. (1997). Optimum bus headway for preemption: A simulation approach. *Transportation Research Record*, **1603**, 128–136.

Koutsopoulos, H. N., Amedeo, R. O. and Wilson, N. H. M. (1985). Determination of headways as a function of time varying characteristics on a transit network. In *Computer Scheduling of Public Transport 2* (J. M. Rousseau, ed.), 391–413. Elsevier Science, Amsterdam, the Netherlands.

LeBlanc, L. J. (1988). Transit system network design. *Transportation Research*, **22B**, 383–390.

Wirasinghe, S. C. (2003). Initial planning for urban transit systems. In *Advanced Modeling for Transit Operations and Service Planning* (W. H. K. Lam and M. G. H. Bell, eds), 1–29. Elsevier Science, Oxford, UK.

4 Timetable Development



Chapter 4 Timetable Development

Chapter outline

- 4.1 Introduction
 - 4.2 Objectives, optional timetables and comparison measures
 - 4.3 Even headways with smooth transitions
 - 4.4 Headways with even average loads
 - 4.5 Automation, test runs and conclusion
 - 4.6 Literature review and further reading
 - Exercises
 - References
-

Practitioner's Corner

The aim of this chapter is to create and present approaches and procedures for deriving prudent public timetables. These timetables and their compliance mirror the quality of the transit service provided. Hence, vehicles departing too early or ahead of schedule need to be restrained, just as those leaving late must be scheduled or rescheduled to be on time. Because of existing problems of transit reliability, there is need to improve the correspondence of vehicle-departure times with passenger demand instead of assuming that passengers will adjust themselves to given time-tables (excluding situations characterized by short headways). The following riddle may shed some light on the substance of adequate, and accurate, timetables.

Two identical shopping precincts, M and K, are joined by a road connecting. A bus departs every 10 minutes (headway) from each precinct to the other (M to K, and K to M); the travel time in each direction is the same. A bus stop is located at exactly the mid-point between the two precincts. Passengers arrive randomly at this bus stop and take the first bus to arrive, regardless of the precinct to which it is headed. Observations at this bus stop reveal that 9 out of every 10 passengers board the bus that goes from M to K. Is this possible? If so, how? (The answer appears at the end of this Practitioner's Corner.)

The chapter starts with an explication of the advantage of having optional or alternative public timetables. This feature gives rise to improved service quality and resource saving, using derived comparison measures that constitute a basis for comparing the alternative timetables. A procedure for calculating departure times is provided, based on even headways and a smooth transition between time periods (usually hours), which allows for even average loads on the last and first vehicles in each period. Thereafter, another procedure is presented for calculating departure times based on even average passenger loads (as required by the standard), rather than on even headways, at the hourly max load point. The chapter considers four frequency-determination methods and gives examples (one of them real life) used in Chapter 3, along with possible practical decisions on headways and number of departures. The chapter ends with a literature review and practical exercises.

Practitioners are encouraged to read the entire chapter, but to concentrate on the examples and their figures while studying the two procedures, rather than attempting to fully comprehend the methodological arguments. The overall analysis of the optional timetables also contains a direct planner/scheduler intervention in setting frequencies. Such an intervention is required in situations known only to the planner and does not rely on passenger loads.

We can now return to the riddle. Indeed, the situation described can be realized if the timetables from M to K and K to M are set to differ by a 1-minute difference. Specifically, the bus departs from M to K at 7:00, 7:10, 7:20 and so on; whereas the bus leaves from K to M at 7:01, 7:11, 7:21, etc. Ostensibly, the average waiting time for the M to K bus is nine times that for the opposite direction; proportionally, therefore, more passengers will be on hand for the M to K bus. This example shows how important it is to set timetables adequately. Situations similar to that of the riddle, but for a network of routes, undoubtedly imply an unreliable transit service. Improvement in creating timetables can offer a remedy.

4.1 Introduction

Public transit timetables constitute the most profound bridge between the transit agency (and/or the community) and the passenger seeking a reliable service. Inadequate and/or inaccurate timetables not only confuse the passengers but also reinforce the bad image of public transit as a whole. There is a saying: “A stitch in time would have confused Einstein”. Along these lines, one can say that many stitches in a transit timetable would confuse those (the passengers) who want to use the service.

The assumption that passengers will adjust themselves to a given timetable (with headways of, say, longer than 10 minutes), instead of planners’ adjusting the timetable to passenger demand, constitutes a major source of unreliable service. When passenger demand is not met, transit vehicles slow down (i.e. increase dwell time), travelling behind schedule and entering the inevitable process of further slowing down. This situation will eventually lead to the known bunching phenomenon with the vehicles that follow. In contrast, a situation of overestimating demand may result in transit vehicles running ahead of time. Neither situation is observed when the service is frequent and characterized by a low variance in headway distribution.

The products of the derived frequencies and headways given in Chapter 3 yield the timetables for the public, for the drivers, as well as for the supervisors and inspectors. Once the timetables are constructed, as shown in Figure 1.2, it is possible to initiate the task of scheduling vehicles and crews for the previously determined trips. Naturally, the transit agency wishes to utilize its resources more efficiently by minimizing the number of required vehicles and crew costs. To accomplish this, the planner (or an automated procedure) examines optional timetables during the vehicle and crew-assignment processes. Optional timetables are derived by shifting departure times and/or reducing the number of departures without prudent consideration of the load profile. Often, because of the uncertainty involved, some planners/schedulers prefer to shift departure times in small increments instead of adjusting them to the data-based demand–supply point. This can follow

the rule that if you put aside a small amount for savings every day, you will be surprised to learn how little was accumulated in a year. All in all, it is desirable to extend the analysis of deriving appropriate headways to an evaluation of optional timetables in conjunction with the required resources.

This chapter, and the next one, as well, will continue with an explication of the second planning activity in Figure 1.2, timetable development. The basic challenge is how to improve the correspondence of vehicle-departure times with fluctuating passenger demand. It is known that passenger demand varies even within the space of 1 hour. This dynamic behaviour can be detected through passenger-load counts and information provided by road inspectors. A more balanced load timetable is achievable by adjusting departure times. The two procedures presented in this chapter, and one in the next, can be applied to both single and interlining transit routes, and they can be carried out in an automated manner. While the first procedure yields departure times (a timetable) for vehicles with even scheduled headways, the second procedure produces departure times for vehicles having even average loads at the hourly max load point. A third procedure, discussed in Chapter 5, refers to balanced (even) loads on individual vehicles at their (individual) peak-load segments. The key point here is to be able to control the loading, instead of repeatedly exposing passengers to unreliable service resulting from an imbalance in loading situations.

This chapter consists of three major parts. First, a spectrum of optional timetables is exhibited, along with utilization measures, thus enabling a comparison of various timetables. Second, a procedure is presented for constructing departure times in which the headways are evenly spaced. In this context, smoothing techniques are developed in the transition segments between adjacent time periods, usually hourly segments. Third, another procedure is offered, this one for a case in which the headways are allowed to be unevenly spaced; here, the departure times are shifted so as to obtain uniform average loads at the hourly max load point, instead of even headways. All the procedures are applied on a route basis; they are similar to, corrected and improved over those presented earlier by Ceder (1986, 2003). A literature review and exercises follow these three parts.

4.2 Objectives, optional timetables and comparison measures

A cost-effective and efficient transit timetable embodies a compromise between passenger comfort and service cost. A good match between vehicle supply and passenger demand occurs when transit schedules are constructed so that the observed passenger demand is accommodated while the number of vehicles used is minimized. This approach helps in minimizing the agency cost in terms of drivers' wages and capital costs required to purchase vehicles. This cost-effective concept led to the establishment of five objectives in creating public transit timetables:

1. Evaluate optional timetables in terms of required resources.
2. Improve the correspondence of vehicle departure times with passenger demand, while minimizing resources.
3. Permit, in the timetable construction procedure, direct vehicle-frequency changes for possible exceptions (known to the planner/scheduler) that do not rely on passenger-demand data.

4. Allow the construction of timetables with headway-smoothing techniques (similar to those performed manually) in the transition segments between adjacent time periods.
5. Integrate different frequency-setting methods and different timetable-construction procedures.

This chapter attempts to show how to fulfil these five objectives. The next chapter continues this attempt.

4.2.1 Elements in practice

Different transit agencies use different scheduling strategies based on their own experience. Consequently, it is unlikely that two independent transit agencies will use exactly the same scheduling procedures, at least at the level of detail. In addition, even within the same transit agency, planners may use different scheduling procedures for different groups of routes. Therefore, I believe that when developing computerized procedures, there is need to supply the schedulers with schedule options, along with an interpretation and explanation of each option. Undoubtedly, it is desirable that one of the options should coincide with the scheduler's manual/semi-manual procedure. In this way, the scheduler will be in a position not only to expedite manual/semi-manual tasks, but also to compare procedures and methods in regard to the trade-off between the passengers' comfort and the operating cost.

The number of vehicle runs determined by the timetable and, eventually, the number of vehicles required are sensitive to the procedure that the scheduler uses to construct the departure times. Some transit agencies routinely round off the frequency F_j to the next highest integer and then calculate the appropriate headways for the time period. By doing so, they increase the number of daily departures beyond what is needed to appropriately match demand with supply. Such a procedure may result in non-productive runs (many empty seat-km). For example, Table 3.7, in Chapter 3, lists the number of required daily departures $\sum_{j=1}^{19} F_j$ for LA Metro (previously SCRTD) route 217: 86.52 and 78.11 for Methods 2 and 4 (20% case), respectively. When the quantity F_j is 'rounded up', we obtain, respectively, 92 and 85 daily departures for these two methods. Obviously, by rounding off F_j to the next highest integer, the level of passenger comfort is increased, but at the expense of unnecessary operating costs. However, in some cases, the 'rounding-off' procedure may be justified if P_{ij} (see Equation 3.2) is used as an average load when the load variance is high. In this case (provided that additional runs are made by rounding off F_j upward), possible overcrowding situations may be reduced as opposed to increasing average empty seat-km. Nonetheless, to overcome the problem of highly variable loads, a statistical load measure can be used that considers its variance as input to a frequency method: this was explained in Chapter 3 under Equation (3.2).

4.2.2 Optional timetables

The five objectives enumerated above, and current timetable-construction procedures, provide the basis to establish a spectrum of optional timetables. Three categories of options may be identified: (i) type of timetable, (ii) method or combination of methods for setting frequencies, and (iii) special requests. These three groups of options are illustrated in Figure 4.1. A selected path in this figure provides a single timetable. Hence, there are a variety of timetable options.

The first category in Figure 4.1 concerns alternative types of timetables. The even-headway type simply means constant time intervals between adjacent departures in each time period, or the case of evenly spaced headways. Even average load refers to unevenly spaced headways in each time period, but the observed passenger loads at the hourly max load point are similar on all vehicles. A second type of timetable entails situations in which even headways will result in significantly uneven loads. Such uneven-load circumstances occur, for example around work and school dismissal times, but they may in fact occur on many other occasions. Figure 4.1 shows that the average even load can be managed either at the hourly max load point (even loads on all vehicles at that point) or at each individual vehicle’s max load point. The average even load at the hourly max load point type is dealt with in this chapter, and the other type in the next chapter.

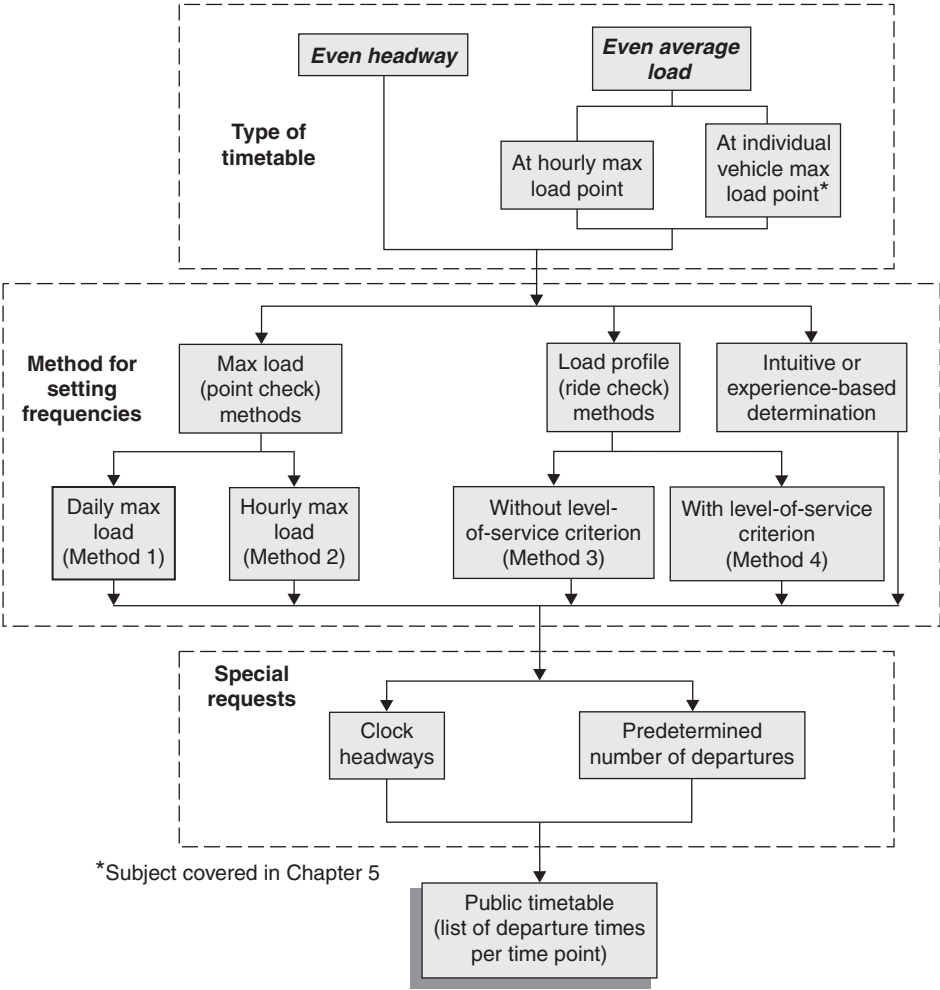


Figure 4.1 Optional public timetables

In the second category of options, it is possible to select different frequency or headway-setting methods. This category allows for the selection of one method or for combinations of methods for different time periods. The methods considered, and indicated in Figure 4.1, are the two-point check and the two-ride check, both described in Chapter 3. In addition, there might be procedures used by the planner/scheduler that are not based on data, but on observations made by road supervisors and inspectors or other sources of information.

In the third category of selections, special scheduling requests are allowed for. One characteristic of existing timetables is the repetition of departure times, usually every hour. These easy-to-memorize departure times are based on so-called *clock headways*: 6, 7.5, 10, 12, 15, 20, 30, 40, 45 and 60 minutes. Ostensibly, headways less than or equal to 5 minutes are not thought to influence the timing of passenger arrivals at a transit stop. The clock headway is obtained by rounding the derived headway down to the nearest of these *clock* values. Consequently, and similar to the ‘rounding off upward’ of frequencies, clock headways require a higher number of departures than what is actually necessary to meet the demand.

A second possible special request is to allow the scheduler to predetermine the total number of vehicle departures during any time period. This request is most useful in crises, when the agency needs to supply a working timetable for an operation based on tightly limited resources (vehicles and/or crews). By controlling the total number of departures while complying with other requests, the scheduler achieves better results than by simply dropping departures without any systematic procedure. Furthermore, there might be cases in which the agency would like to increase the level of service by allowing more departures in the belief that passenger demand can be increased by providing improved (more frequent) service. Certainly, this special request can also be approached through varying the desired occupancy (load factor) standard; however, this option can be a compulsory standard.

Finally, it should be emphasized that not all the paths concerning clock headways in Figure 4.1 are meaningful. The selection of the even average load type of timetable cannot be performed if there is a clock-headway constraint. Moreover, the number of departures cannot be predetermined for clock headways because of the specific time restrictions on those headways.

4.2.3 Comparison measures

With computerized timetable construction, the transit agency can assess a variety of optional timetables rather than being limited to examining one or a few. Two interrelated measures may be useful for the agency to compare optional timetables: (i) number of required runs (departures); and (ii) required single-route fleet size.

The first comparison measure, total number of departures, can serve as an indicator of the number of vehicles required and also whether or not it is possible to save vehicle runs.

The second comparison measure refers to each route separately and provides an estimate of the required fleet size at the route level. In a large transit agency, an efficient arrangement of vehicle blocks includes interlining (switching a vehicle from one route to another) and deadheading trips. Hence, fleet size is not determined at the route level but at the network level (see Chapters 7 and 8). The second comparison measure, however, is based on a simple formula derived by Salzborn (1972) for a continuous time function and explicitly shown by Ceder (1984) for discrete time points. This formula states that if T is the round-trip time,

including layover and turn-around time, then the minimum fleet size is the largest number of vehicles departing at any time interval during T . This value is adequate for a single route with a coinciding departure and arrival location. Consequently, the second comparison measure can be used for each direction separately, as well as for both directions when selecting the maximum of two derived values. This single-route fleet-size formula will be elaborated in Chapter 7.

4.2.4 Anchoring the timetable to a single timepoint

A public timetable usually consists of lists of vehicle departure times at all route stops or at selected stops called timepoints. Occasionally, this public timetable is given at just a single point – the route departure point or a major boarding stop. The running time between adjacent timepoints may vary from one time period to another, based on ride-check information. In essence, the timetable can initially be constructed at only one point, and be referred to as such, and then extended forward and backward using the average running-time information. That is, in order to ensure an appropriate transit service to meet the variations in passenger demand, it suffices to construct the timetable at one point. This observation is stated in the following proposition:

Proposition

For a route-based transit timetable consisting of more than one timepoint, the association of max observed load in each time period with only a single timepoint ensures that the average vehicle load on each route segment is less than or equal to the desired occupancy.

Proof and interpretation

For simplicity sake, refer to the single timepoint as the route-departure point. The derived frequency is greater than or equal to the maximum required frequency (across all route segments) in each time period. The proof for frequencies determined by Method 2 was provided in Chapter 3; however, this can easily be converted to, and shown by, any of the four methods. As a note, when only the daily max load point is considered, this does not necessarily imply that the observed max load in each period occurs at that point. The problem can be treated similarly to the three-dimensional Time–Load–Distance representation in Figure 4.2, based on Example 2 (which appears in Figure 3.7 in Chapter 3). That is, the shaded two-part area in Figure 4.2 describes the max loads of Example 2 in each hour: 134 (between 06:00–07:00) and 216 passengers (07:00–08:00). Calculation of the frequency at Distance = 0 in Figure 4.2 ensures compliance with Method 2.

Since Example 2 is used throughout this chapter, Table 4.1 presents its entire input data required for constructing optional timetables. Table 4.1 also contains information on: (a) T , the time required for the same vehicle, at the same departure point, to pick up another departure; and (b) the calculated data on an individual vehicle basis. For the sake of clarity, Figure 4.3 shows five time–space vehicle trajectories of Example 2. Because of possible different average running times in each period, the headways at each stop do not necessarily coincide as seen in Figure 4.3. However, since the hourly max load is used for calculating the frequency, the timetable, set only at one timepoint, reflects the maximum number of required vehicles at the observed hourly max load point. There is, then, a question of

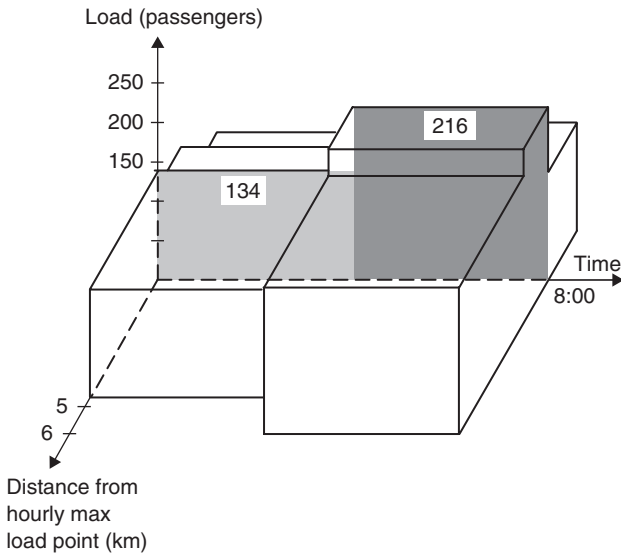


Figure 4.2 Three-dimensional illustration of Example 2 load profiles positioned relative to the distance from the hourly max load point

Table 4.1 Part of the Example 2 (Figure 3.7) data and calculated complementary data required for illustrating the optional timetable procedure

Time period	Departure time	Average load (passengers) at the max load point, by method number		Area under the load profile (passenger-km)	Frequency (veh/hr) and headway (minutes, in parenthesis), by method number		
		1	2		1	2	4 (20%)
6:00–7:00	06:00*	—	—	929	2.56	2.68	2.68
	06:15	52	35		(23)	(22)	(22)
	06:50	40	65				
7:00–8:00	07:15	90	90	1395	3.60	3.60	2.95
	07:35	87	87		(17)	(17)	(20)
	07:50	75	75				

Added data: **T** (round trip time + layover and turn around time) = 60 minutes

*First predetermined trip

whether the determined frequency at the observed hourly max load point complies with the desired occupancy constraint.

It is important to note that the running-time information must rely on the fact that in an average sense vehicles do not overtake each other. Average running times should be

determined not only from ride-check data, but also from the requirement that the first vehicle to depart cannot be the second to arrive at any timepoint. Thus, the time–space trajectories in Figure 4.3 cannot cross each other in an average (deterministic) context.

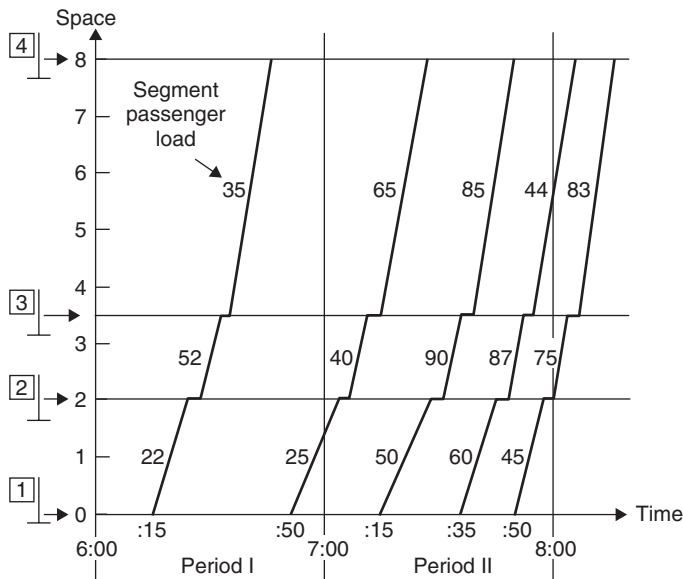


Figure 4.3 Time–space trajectories of all vehicles in Example 2, including segment loads

Based on possible differences in actual average headways at each stop, the associated time span at the observed hourly max load point, covering all trips across the time period at the route’s departure point, can be shorter or longer than the time span of the route’s departure point. For example, we see in Figure 4.3 that the time span of the last three departures, 35 minutes (7:15–7:50), is larger than the time span of these trips at the hourly max load point of period II, stop 3. Fortunately in either case, what governs the frequency calculation is the observed max load. Hence, by construction, the result is that as long as the time–space trajectories do not cross one another, the requested frequency at each stop must be less than or equal to the frequency determined at the route’s departure point.

Finally, it may be noted that the above proposition can also be applied to Methods 1, 3 and 4 – each with its own loading standards and constraints.

4.3 Even headways with smooth transitions

One characteristic of existing timetables is the repetition of the same headway in each time period. However, a problem facing the scheduler in creating these timetables is how to set departure times in the transition segments between adjacent time periods. This section addresses the issue.

4.3.1 Underlying principle

A common headway smoothing rule in the transition between time periods is to use an average headway. Many transit agencies employ this simple rule, but it may be shown that it can result in either undesirable overcrowding or underutilization. For example, consider two time periods, 06:00–07:00 and 07:00–08:00, in which the first vehicle is predetermined to depart at 06:00. In the first time period, the desired occupancy is 50 passengers, and in the second 70 passengers. The observed maximum demand to be considered in these periods is 120 and 840 passengers, respectively. These observed loads at a single point are based on the uniform passenger-arrival-rate assumption. The determined frequencies are $120/50 = 2.4$ vehicles and $840/70 = 12$ vehicles for the two respective periods, and their associated headways are 25 and 5 minutes, respectively. If one uses the common average headway rule, the transition headway is $(25 + 5)/2 = 15$ minutes; hence, the timetable is set to 06:00, 06:25, 06:50, 07:05, 07:10, 07:15, . . . , 07:55, 08:00. By assuming a uniform passenger arrival rate, the first period contributes to the vehicle departing at 07:05 the average amount of $(10/25) \cdot 50 = 20$ passengers at the max load point; the second period contributes $(5/5) \cdot 70 = 70$ passengers. Consequently, the expected load at the max load point is $20 + 70 = 90$, a figure representing average overcrowding over the desired 70 passengers after 7:00. Certainly, the uniform arrival-rate assumption does not hold in reality. However, in some real-life situations (e.g. after work and school dismissals), the observed demand in 5 minutes can be more than three times the observed demand during the previous 10 minutes, as is the case in this example. In order to overcome this undesirable situation, the following principle, suggested by Ceder (2003), may be employed.

Principle 1: Establish a curve representing the cumulative (non-integer) frequency determined versus time. Move horizontally for each departure until intersecting the cumulative curve and then vertically; this will result in the required departure time.

Proposition 1: Principle 1 provides the required evenly spaced headways with a transition load approaching the *average desired occupancies* of d_{0j} and $d_{0(j+1)}$ for two consecutive time periods, j and $j+1$.

Proof: Figure 4.4 illustrates Principle 1 using the frequency results from Example 2 (see Figure 3.7) appearing in Table 3.4 (Chapter 3). Since the slopes of the lines are 2.68 and 3.60 for $j = 1$ and $j = 2$, respectively, the resultant headways are those required. The transition load is associated with the 7:05 departure, which consists of arriving passengers during 16 minutes for $j = 1$, and arriving passengers during 5 minutes for $j = 2$. Therefore, $(16/22) \cdot 50 + (5/17) \cdot 60 = 54$ approximately. This transition load is not the exact average between $d_{01} = 50$ and $d_{02} = 60$, since departures are made in integer minutes. That is, the exact determined departure after 7:00 is $(3 - 2.68) \cdot 60 / 3.60 = 5.33$ minutes. Inserting this value, instead of the 5 minutes mentioned above, yields a value that is closer to the exact average. Basically, the proportions considered satisfy the proof-by-construction of Proposition 1.

4.3.2 Interpretation using Example 2

The upper part of Figure 4.4 exhibits the resulting six departures for the even-headway procedure, using frequency setting of Method 2, whereby the determined frequencies are kept

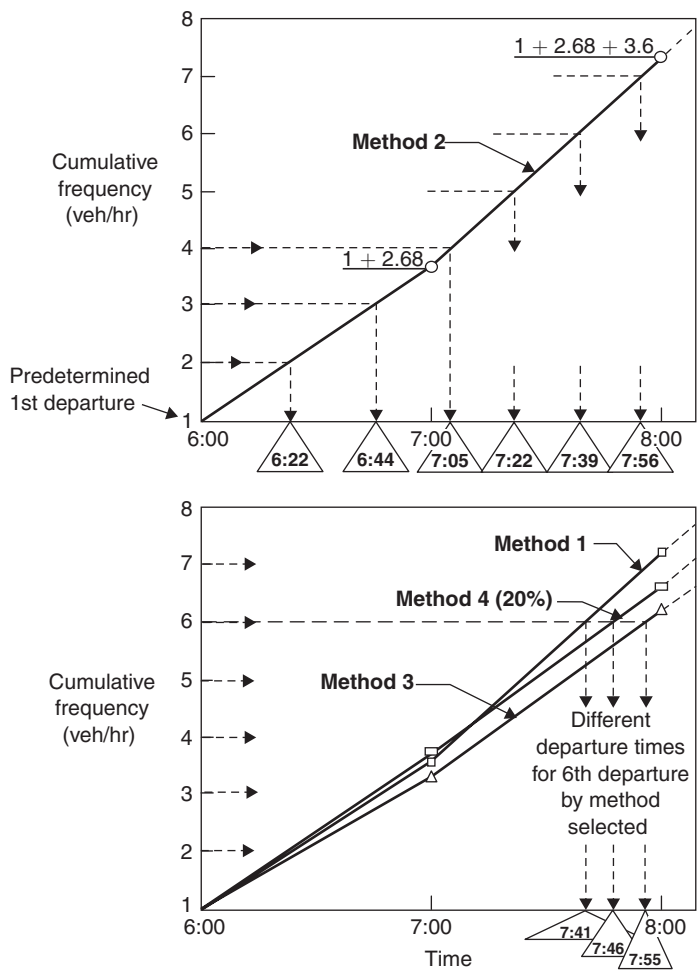


Figure 4.4 Determination of departure times for evenly spaced headways and for all methods in Example 2, including a smoothing process between periods

non-integer as in Table 3.4. The lower part of Figure 4.4 shows the cumulative frequency results for Methods 1, 3 and 4 (20%). For the sake of clarity, the sixth departure is determined to be 7:41, 7:46 and 7:56 according to Methods 1, 4 (20%) and 3, respectively. Overall, Principle 1 allows for saving some unnecessary vehicle runs and also stabilizes the average load during the transition segment between time periods.

The resultant departure times at the route-departure point – it can be any other timepoint – appear in Table 4.2 using the even-headway procedure. This table presents five optional timetables (at a single timepoint), using three frequency setting methods, one of which, Method 2, is also combined with special requests. These five timetables are accompanied by the two comparison measures: total departures and minimum single-route fleet size. The first two timetables in Table 4.2 are based on Figure 4.4; the three others deserve explication.

Table 4.2 Five derived sets of departure times using Example 2 for even headways, by method or combination of methods selected

Timetable characteristics (even headways)	Method 2	Method 4 (20%)	Method 1 for 1st hour, Method 4 (20%) for 2nd hour	Method 2 with clock headway	Method 2 with predetermined number (= 5) of departures
Departure times at anchored timepoint (6:00–8:00)	6:00* 6:22 6:44 7:05 7:22 7:39 7:56	6:00* 6:22 6:44 7:06 7:26 7:46	6:00* 6:23 6:46 7:08 7:28 7:48	6:00* 6:20 6:40 7:00 7:15 7:30 7:45 8:00	6:00* 6:35 7:08 7:34 8:00
Total no. of departures	7	6	6	8	5 (as requested)
Minimum single-route fleet size (T = 60 minutes)	4	3	3	4	3

*First predetermined trip

One option in selecting optional timetables (see Figure 4.1) is to use different frequency setting methods for different time periods. In this way, the transit agency can examine the effect of these settings on vehicle (resource) requirements during peak and off-peak periods; for example, examining the use of Method 4 during peak periods, in which the need for more vehicles is highest, and Method 2 for off-peak periods. In the third option in Table 4.2, Method 1 is applied to the first time period, and Method 4 (20% case) to the second (peak) period. The timetable-construction procedure combines, in the transition between hours, the leftover vehicle demand utilizing Method 1 with the frequency required by Method 4. That is, using the Table 3.4 results, $F = 2.56$ according to Method 1, from 6:00–7:00 and $F = 2.95$ according to Method 4, from 7:00–8:00. The last vehicle according to Method 1, therefore, departs at 6:46, and the leftover vehicle demand is $14 \cdot (2.56/60) = 0.597$. Hence, $1 - 0.597 = 0.403$ is the applicable vehicle demand for Method 4, and this yields $0.403 \cdot (60/2.95) = 8.2$ minutes, which is rounded off to 8 minutes past 7:00. The remaining departures after 7:08 follow $H = 20$ of Method 4 (20%).

The fourth timetable option in Table 4.2 is based on Method 2 with clock headways. In this case, we simply round down the headway determined by Method 2 to its nearest clock headway. Thus, from 6:00–7:00, $H = 22$ (see Table 3.4), which is rounded to 20; and $H = 17$, from 7:00–8:00, is rounded down to 15.

The last timetable option is based on five predetermined departures (including the first 6:00 departure) while using Method 2. The total number of required vehicles using Method 2 (excluding 6:00) is $2.68 + 3.60 = 6.28$. Since only four departures have to

be constructed, the frequencies are modified proportionally by the ratio $4/6.28 = 0.637$. The procedure continues the same way as that without the special request. That is, $F = 1.71$ ($= 2.68 \cdot 0.637$) with $H = 35$ for 6:00–7:00, and $F = 2.29$ with $H = 26$ for 7:00–8:00. Certainly, if the demand remains the same, the scheduler should recognize the potential risk of overcrowding when restricting the total number of departures. Nevertheless, the purpose of this special request is to have a systematic computerized procedure to manage both crisis situations (limited resources) and situations in which additional passenger demand can possibly be attracted (i.e. requesting more departures than calculated).

This explication of Table 4.2 ends with the derivation of the single-route fleet size using the round-trip time information of $T = 60$ minutes. The derivation of this comparison measure will be illustrated for the first option in Table 4.2 (Method 2). Adding 60 minutes to the first 6:00 departure reveals that the vehicle associated with this departure can perform the 7:05 departure. Hence, three vehicles are required for this step (to perform 6:00, 6:22 and 6:44 departures). By continuing to add 60 to 6:22, this will enable the vehicle to pick up the 7:22 departure (again, three vehicles are required). Adding 60 to 6:44 shows that the next pick up is executed by the 7:56 departure, resulting in four independent departures (6:44, 7:05, 7:22 and 7:39). Continuing to count the number of departures in each 60-minute window leads to finding the maximum number, which is the minimum single-route fleet size required. For the first option in Table 4.2, this number ostensibly is four.

4.4 Headways with even average loads

This section opens with the following premise: transit managers/planners/schedulers, who believe that problems related to attracting more transit users and reliability problems are drowned in the ‘ocean’ of even-headway timetables, should be told that these problems know . . . how to swim. In other words, even-headway timetables do not necessarily deliver the merchandise (satisfactory transit service) to the customer (passengers).

As already noted, passenger demand varies even within a single time period, hence resulting in even headways in an imbalanced load on individual vehicles at the hourly max load point. On heavy-load routes and short headways, the even-headway timetable suffices. However, in the course of reducing reliability problems, we may occasionally prefer to use the even-load instead of the even-headway procedure. Moreover, the availability of APCs provides a framework in which to investigate systematically the variation in passenger demand. With the anticipated vast amount of passenger-load data, it is easier to match vehicle departure times with variable demands. Two procedures carry out this endeavour. The first, addressed in this section, deals with average even load on individual vehicles at the hourly (or daily) max load point. The second procedure, addressed in the next chapter, ensures an average even load at each individual vehicle max load point. In this section, the procedures described by Ceder (1986, 2003) will be corrected and improved.

4.4.1 Underlying principle

A simple example is presented here to illustrate the underlying load-balancing problem. Consider an evenly spaced headway timetable in which vehicles depart every 20 minutes between 07:00 and 08:00; i.e. at 07:20, 07:40 and 08:00. The observed load data consistently

show that the second vehicle, which departs at 07:40, has significantly more passengers than the third vehicle. The observed (average) max load during this 60-minute period is 150 passengers, and the desired occupancy is 50 passengers. Hence, using Method 2, three vehicles are required to serve the demand as in the evenly spaced headways timetable. The average observed loads at the hourly max load point on the three vehicles are 50, 70 and 30 passengers, respectively. Given that these average loads are consistent, then the transit agency can adjust the departure times so that each vehicle has a balanced load of 50 passengers on the average at the hourly max load point. The assumption of a uniform passenger-arrival rate results in $70/20 = 3.5$ passengers/minute between 07:20 and 07:40, and $30/20 = 1.5$ passengers/minute between 07:40 and 08:00. If the departure time of the second vehicle is shifted by X minutes backward (i.e. an early departure), then the equation $3.5X = 70 - 50$ yields the balanced schedule, with $X = 5.7 \approx 6$ minutes, or departures at 07:20, 07:34 and 08:00. The third departure will add this difference of 20 passengers at the hourly max load point. The even-headway setting assures enough vehicles to accommodate the hourly demand, but it cannot guarantee balanced loads for each vehicle at the peak point. In order to avoid this imbalanced situation, the following principle should be exploited.

Principle 2: Construct a curve representing the cumulative loads observed on individual vehicles at the hourly max load points. Move horizontally per each d_{oj} for all j , until intersecting the cumulative-load curve, and then vertically; this results in the required departure times.

Proposition 2: Principle 2 results in departure times such that the average max load on individual vehicles at the hourly j th max load point approaches the desired occupancy d_{oj} .

Proof: Figure 4.5 illustrates Principle 2 for the Example 2 problem appearing in Figure 3.7 and Table 4.1. Method 2 will be used in the upper part of Figure 4.5 in which the derived departure times are unevenly spaced to obtain even loads at stop 3 for $j = 1$ and at stop 2 for $j = 2$. These even loads are constructed on the cumulative curve to approach $d_{o1} = 50$ and $d_{o2} = 60$. If we assume a uniform passenger-arrival rate between two observed departures, it can be shown that the load (at stop 3) of the first derived departure (6:23) consists of the arrival rate between 6:00 and 6:15 ($35/15 = 2.33$) and the rate between 6:15 and 6:50 ($65/35 = 1.86$). Thus, $2.33 \cdot 15 + 1.86 \cdot 8 \approx 50$. In the transition between $j = 1$ and $j = 2$ (in the upper part of Figure 4.5), the value of $d_2 = 60$ is considered, since the resultant departure comes after 7:00. The load of the vehicle departing at 7:07 at its hourly max load point, stop 2, is simply $17 \cdot (90/25) = 61.2$ from rounding off the departure time to the nearest integer. That is, $(10 + y) \cdot (90/25) = 60$ results in $y = 6.67$ minutes. This completes the proof-by-construction of Proposition 2.

4.4.2 Interpretation of even load using Example 2

In its upper part, Figure 4.5 shows the resulting seven departures for the even-load procedure, using the loads observed at the hourly (Method 2) max load points (as in Table 4.1). A cumulative curve of straight lines can be drawn from observed departure times and max loads ($35 + 65 + 90 + \dots$). The slope of each line will represent the uniform arrival rate. For example, the arrival rate between the 6:50 and 7:15 departures is $90/25 = 3.6$ passengers/minute. The time on the curve associated with the first desired occupancy, $d_{o1} = 50$ passengers, is 06:23, and the second is 6:50; this means that the second departure is unchanged, since $35 + 65 = 50 + 50$. We then can check the cumulative load level of 150

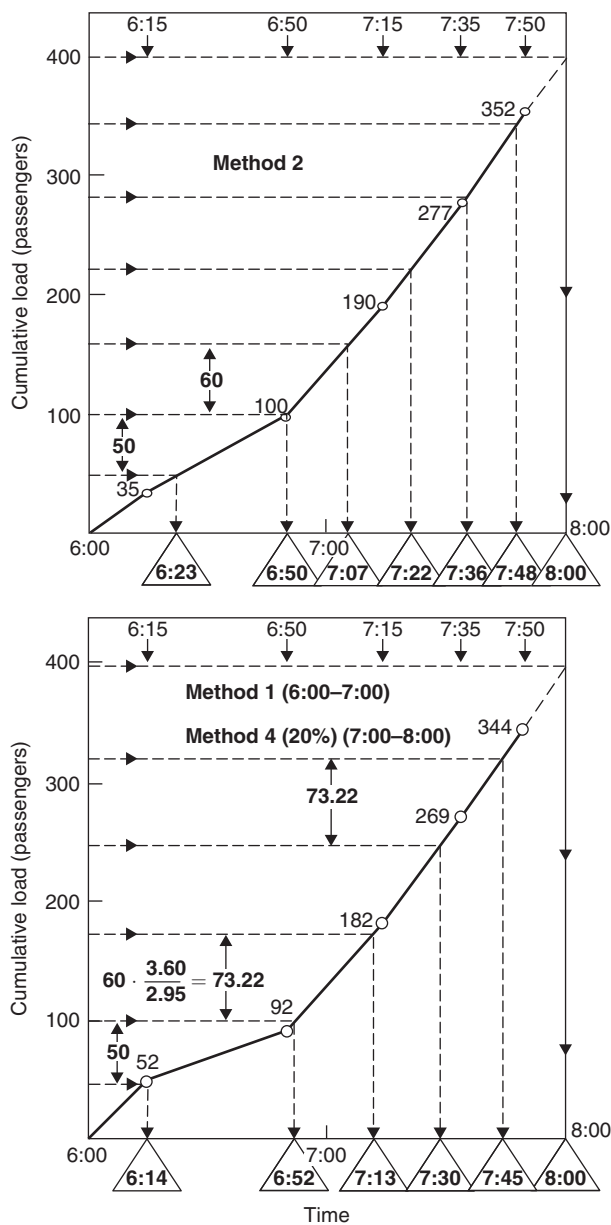


Figure 4.5 Determination of Example 2 departure times with even load at the hourly max load point using methods 2 and at the daily max load point using a combination of Methods 1 and 4 (20%)

and learn that its associated time on the curve is after 7:00. Hence $d_{o2} = 60$ will be applied for the rest of the example. The cumulative-load curve also ensures that the first vehicle to depart in any time period will accommodate the desired occupancy assigned to that period. The only exception in this process occurs when the determined frequency is based on the

minimum frequency standard of Equation (3.3) in Chapter 3. In such a case, disregard even-loads for a given time period j , and switch to even-headways, since virtually no overcrowding is expected. However, to maintain the even-load situation for the minimum-frequency case, then d_{oj} must be replaced according to the formula shown in the following section.

The lower part of Figure 4.5 shows the derivation of an average even-load timetable at the daily max load point as opposed to the hourly max load point described for Method 2. In this example, combine the results of Method 1 (6:00–7:00) with Method 4 using the 20% criterion (7:00–8:00). The cumulative-load curve is based on the loads observed only at stop 2 (daily max load point) and appearing in Table 4.1 and Figure 3.7. While the desired occupancy for Method 1 is the same as for Method 2 ($d_{o1} = 50$), the value of d_{o2} needs to be commensurate with the frequency result given by Method 4 (20%). Thus, the substitution for d_{o2} can be calculated simply in proportional terms:

$$d_{ij} = d_{oj} \cdot \frac{F_{ij}}{P_{mj} / d_{oj}}, \quad i = 3, 4 \quad \text{and} \quad j = 1, 2, \dots, q \quad (4.1)$$

where d_{ij} is the adjusted ‘desired occupancy’; i is the frequency-determination method number; j is the time period; and F_{ij} , P_{mj} and d_{oj} are defined in Chapter 3 as the frequency, maximum observed load, and desired occupancy, respectively. Equation (4.1) also holds for situations in which the determined frequency is the minimum frequency standard for all methods. This standard is the result of Equations (3.3) and (3.5); that is, in the case of Method 2, $F_{2j} = \max(P_{mj}/d_{oj}, F_{mj}) = F_{mj}$, where F_{mj} is the minimum frequency in time period j .

Based on Equation (4.1), and using Example 2, Method 4 (20%) yields the following: $d_{4,2} = 60(3.60/2.95) = 73.22$. This adjusted quantity is used in Figure 4.5 for constructing the timetable from the route’s departure point. The results of the two timetables of Figure 4.5 and of other optional timetables for Example 2 appear in four columns in Table 4.3.

When applying the load profile-based (frequency) methods, expect higher than the desired loads at the max load points as a trade-off for less empty space-km. The second column of Table 4.3 presents the resulting timetable at the hourly max load points according to Method 4 (20%). This timetable is derived according to Principle 2, with the same cumulative-load curve as in the upper part of Figure 4.5, but an adjusted d_{ij} based on Equation (4.1). Since the frequencies given by Methods 2 and 4 (20%) for the first 6:00–7:00 period coincide, $d_{i1} = 50$ for both. For the 7:00–8:00 period, $d_{i2} = 73.22$, with both d_{i1} and d_{i2} used for the hourly max loads. That is, the first departure after 7:00 will be y minutes from 6:50 with a load of 73.22; hence, $y = 73.22 \cdot 25/90 = 20.33 \approx 20$, thereby determining the 6:50 + 20 = 7:10 departure.

The last timetable option in Table 4.3 is based on five predetermined departures (including the first 6:00 departure) while using Method 2. The desired occupancy is modified proportionally in the same manner as was done in Section 4.3.2, by the ratio $4/6.28 = 0.637$. Consequently, the two periods in Example 2 will use the load quantities $50/0.637 = 78.5$ and $60/0.637 = 94.2$, respectively, instead of $d_{o1} = 50$ and $d_{o2} = 60$. With these quantities, Principle 2 will be applied on the upper cumulative curve of Figure 4.5 to arrive at the results. It should be understood that the special request of a clock-headway cannot be

Table 4.3 Four derived sets of departure times using Example 2 for an even average load (at a single timepoint), by method or combination of methods selected

Timetable characteristics (even average load)	Method 2	Method 4 (20%)	Method 1 (6:00–7:00), Method 4 (20%) (7:00–8:00)	Method 2 with predetermined number (= 5) of departures
Point where average even load is expected	Hourly max load	Hourly max load	Daily max load	Hourly max load
Departure times at anchored timepoint (6:00–8:00)	6:00*	6:00*	6:00*	6:00*
	6:23	6:23	6:14	6:38
	6:50	6:50	6:52	7:10
	7:07	7:10	7:13	7:33
	7:22	7:28	7:30	7:52
	7:36	7:44	7:45	
	7:48	7:58	8:00	
	8:00			
Total no. of departures	8	7	7	5 (as requested)
Minimum single-route fleet size (T = 60 minutes)	5	4	4	3

* First predetermined trip

applied to even-load procedures. Ostensibly, timetables with clock headways cannot incorporate uneven headways.

The comparison measures in Table 4.3 are derived in the same manner as for the even-headway procedure as explained in Section 4.3.2. The comparison measures of number of departures and single-route fleet size should be virtually similar for the even-headway and even-load procedures. Nonetheless, because of different usages of the end of the schedule (8:00), this expected similarity almost disappears. The next section, using SCRTD real-life data, will show that this similarity indeed exists, as expected, since both types of procedures rely on the same frequencies.

4.5 Automation, test runs and conclusion

The outcome of Chapters 3 and 4 can be jointly automated. Practically speaking, a set of computer programs was created that perform: (i) conversion from transit agency files to adequate input files, (ii) setting of frequencies and headways by four methods, and (iii) construction of a public timetable at all route timepoints. In practice, it is necessary to actually

test optional timetables prior to drawing a definite conclusion. Often, the following saying is true: logic is a systematic method of coming to the wrong conclusion with confidence. Hence, only real testing, using a before and after study, can reveal the magnitude – if at all – of the improvement.

Figure 4.6 exhibits a summary of the procedures that can take place in any possible automation. This flowchart is fed by the data required before embarking on frequency-setting and timetable-construction procedures. Compliance with the third objective of this chapter (see Section 4.2) is assured, and the possibility is furnished of direct frequency insertion and the manual construction of timetables.

For a real-life example and test runs, the ride-check data from the old LA Metro route 217 (in Los Angeles), which has already been described and employed in Chapter 3,

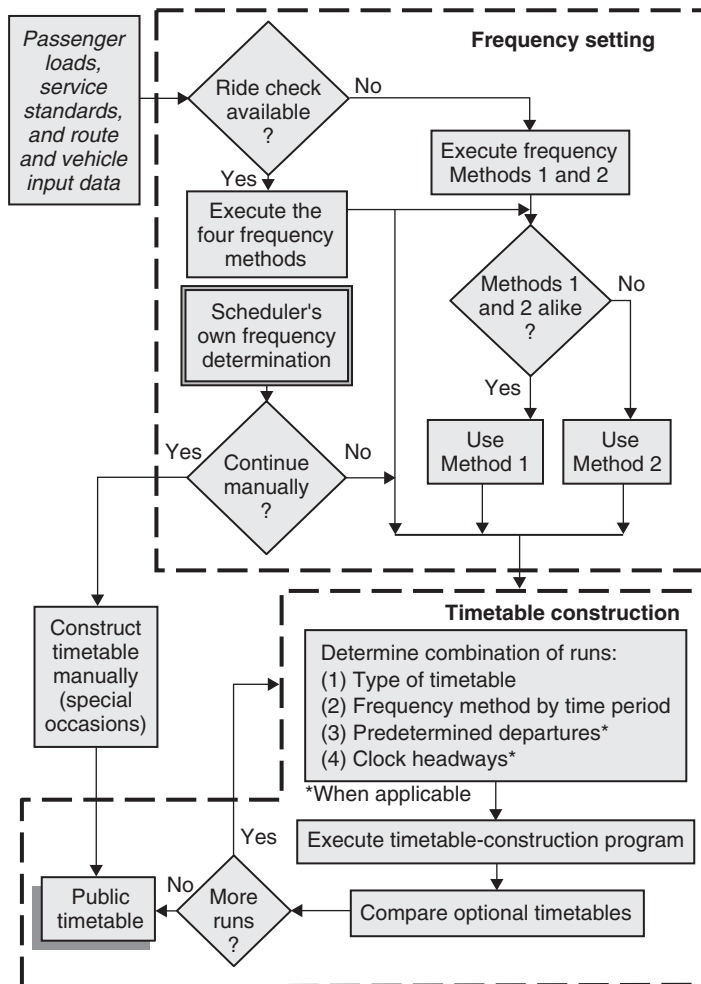


Figure 4.6 Overall automated procedure, in flowchart form, for optional timetable development

- (a) Type of timetable – ‘1’ for evenly spaced headways;
– ‘2’ for even average max load (at hourly max load point).
- (b) ‘Number’ of method to be used (among the inserted frequency-setting methods).
- (c) For each used method, the user specifies:
 - method ‘number’;
 - the time-period ‘number’ in which to start using the method;
 - the last time-period ‘number’ in which to use the method in the combination being considered (i.e. the same method can be used several times for different time periods, but each combination must be specified).
- (d) Clock headway – ‘0’ for not required;
– ‘1’ for required
- (e) Predetermined number of departures
– ‘0’ for no need;
– ‘given number’ of departures for using the constraint.

As was noted earlier, the comparison measures across all the real-life results in Tables 4.5, 4.6 and 4.7 are the same for the two types of timetables (even-headway, even-load). Where they differ is in the frequency method selected and the special requests (clock headway, predetermined number of departures). See, for example, the comparison measures for the lower and upper bounds of route 217 northbound. Specifically, Method 3 delivers a lower bound of 67 and 11 for the number of departures and the single-route fleet size, respectively. This contrasts with the upper bound, given by Method 2 and a clock headway, of 89 departures and 16 vehicles required. Finally, Table 4.8 presents a complete timetable (across all eight timepoints), this time for route 217 southbound, using Method 4 (20%) and an even-headway procedure.

In conclusion, the consequence of this chapter can be described generally in the light of the five study objectives set forth in Section 4.2. Using passenger-load data, the procedures developed provide optional timetables in terms of vehicle departure times at all specified timepoints. Each timetable is accompanied by two comparison measures. These evaluation

Table 4.4 *Combination of requests for test runs using LA route 217 ride-check data*

Run number	Direction of travel S = South N = North	Type of timetable 1 = Even headway 2 = Even load	Method number	Clock headway 1 = With 0 = Without	Predetermined number of departures	Results of comparison measures	
						Number of departures	Minimum fleet size
1	N	1	1	0	0	84	15
2	N	1	2	0	0	87	15
3	N	1	3	0	0	67	11
4	N	1	4	0	0	78	14
5	N	1	2,4*	0	0	85	15
6	N	1	2	1	0	80	16
7	N	1	4	1	0	78	14
8	N	1	2	0	70	70	12
9	N	1	2	0	140	140	24
10	N	2	1	0	0	84	15
11	N	2	2	0	0	87	15
12	N	2	3	0	0	67	11
13	N	2	4	0	0	78	14
14	N	2	2,4*	0	0	85	15
15	N	2	2	0	70	70	12
16	N	2	2	0	140	140	24
17	S	1	1	0	0	88	19
18	S	1	2	0	0	91	19

(Continued)

Table 4.4 *Combination of requests for test runs using LA route 217 ride-check data (continued)*

Run number	Direction of travel S = South N = North	Type of timetable 1 = Even headway 2 = Even load	Method number	Clock headway 1 = With 0 = Without	Predetermined number of departures	Results of comparison measures	
						Number of departures	Minimum fleet size
19	S	1	3	0	0	69	13
20	S	1	4	0	0	73	13
21	S	1	2,4*	0	0	89	19
22	S	1	2	1	0	91	20
23	S	1	4	1	0	74	14
24	S	1	2	0	70	70	15
25	S	1	2	0	140	140	30
26	S	2	1	0	0	88	19
27	S	2	2	0	0	91	20
28	S	2	3	0	0	69	13
29	S	2	4	0	0	73	13
30	S	2	2,4*	0	0	89	20
31	S	2	2	0	70	70	15
32	S	2	2	0	140	140	30

*Method 2 is used for off-peak periods no. (1), (4, 5, 6, 7, 8, 9, 10, 11), (14, 15, 16, 17, 18, 19); Method 4 is used for peak periods no. (2, 3), (12, 13) – a total of five combinations

Table 4.5 Comparison of Methods 2 and 4 (20% criterion) using the two types of timetables for LA route 217 northbound

Characteristics		Even headways, Method 2					Even headways, Method 4 (20%)					
Departure times at route departure point	06:00	09:31	12:11	14:10	16:15	19:28	06:00	10:04	12:54	15:04	17:36	23:59
	06:30	09:45	12:19	14:18	16:24	19:58	06:30	10:16	13:03	15:13	17:48	00:30
	07:00	10:00	12:27	14:26	16:33	20:28	07:00	10:28	13:12	15:22	18:01	01:00
	07:11	10:10	12:35	14:34	16:42	20:58	07:12	10:41	13:20	15:31	18:19	
	07:21	10:21	12:43	14:42	16:50	21:29	07:24	10:53	13:29	15:40	18:38	
	07:32	10:31	12:51	14:50	16:59	21:59	07:36	11:05	13:37	15:49	18:57	
	07:42	10:42	12:59	14:58	17:10	22:29	07:48	11:17	13:46	15:58	19:25	
	07:53	10:52	13:07	15:07	17:21	22:59	08:00	11:30	13:54	16:08	19:56	
	08:04	11:02	13:15	15:15	17:33	23:29	08:13	11:42	14:03	16:18	20:26	
	08:15	11:13	13:22	15:24	17:44	24:00	08:26	11:54	14:11	16:28	20:57	
	08:27	11:23	13:30	15:32	17:55	00:30	08:39	12:05	14:20	16:38	21:27	
	08:38	11:33	13:38	15:41	18:09	01:00	08:51	12:15	14:29	16:49	21:57	
	08:49	11:44	13:46	15:49	18:26		09:07	12:25	14:38	16:59	22:28	
	09:01	11:54	13:54	15:58	18:42		09:26	12:34	14:46	17:11	22:58	
	09:16	12:03	14:02	16:06	18:59		09:46	12:44	14:55	17:23	23:29	
Total no. of departures		87					78					
Minimum single-route fleet size		15					14					

(Continued)

Table 4.5 Comparison of Methods 2 and 4 (20% criterion) using the two types of timetables for LA route 217 northbound (continued)

Characteristics		Even headways, Method 2						Even headways, Method 4 (20%)					
Departure times at route departure point		06:00	09:25	12:07	14:08	16:14	19:28	06:00	10:03	12:51	15:04	17:32	23:59
		06:38	09:39	12:14	14:15	16:26	19:58	06:36	10:17	13:03	15:15	17:51	00:12
		07:03	10:00	12:25	14:29	16:33	20:23	07:03	10:31	13:07	15:20	18:02	01:00
		07:23	10:12	12:33	14:35	16:43	20:59	07:21	10:41	13:15	15:31	18:17	
		07:35	10:24	12:40	14:44	16:49	21:34	07:35	10:54	13:24	15:37	18:36	
		07:42	10:35	12:49	14:53	16:59	21:58	07:44	11:06	13:33	15:46	18:55	
		07:50	10:44	12:58	14:59	17:09	22:26	07:52	11:16	13:46	15:58	19:26	
		07:56	10:54	13:05	15:08	17:19	22:59	08:00	11:30	13:54	16:07	19:56	
		08:02	11:04	13:10	15:16	17:27	23:17	08:07	11:39	14:02	16:20	20:19	
		08:10	11:13	13:17	15:23	17:46	23:59	08:17	11:53	14:08	16:30	20:57	
		08:21	11:20	13:25	15:34	17:56	00:13	08:33	12:04	14:16	16:42	21:26	
		08:36	11:32	13:34	15:40	18:12	01:00	08:44	12:12	14:30	16:49	21:57	
		08:44	11:43	13:44	15:48	18:25		09:03	12:24	14:37	16:59	22:26	
		09:01	11:54	13:53	15:58	18:41		09:20	12:34	14:48	17:10	22:58	
		09:11	12:03	14:02	16:05	18:57		09:37	12:41	14:55	17:20	23:16	
Total no. of departures		87						78					
Minimum single-route fleet size		15						14					

Table 4.6 Comparison of Methods 2 and 4 (20% criterion) using the two types of timetables and of the two types of timetables for a combination of these methods for LA route 217 northbound

Characteristics		Even clock headways, Method 2					Even clock headways, Method 4 (20%)					
Departure times at route departure point	06:00	09:30	12:07	14:00	15:52	18:45	06:00	10:00	12:50	15:00	17:36	00:00
	06:30	09:45	12:15	14:07	16:00	19:00	06:30	10:12	13:00	15:10	17:48	00:30
	07:00	10:00	12:22	14:15	16:10	19:30	07:00	10:24	13:07	15:20	18:00	01:00
	07:10	10:10	12:30	14:22	16:20	20:00	07:12	10:36	13:15	15:30	18:20	
	07:20	10:20	12:37	14:30	16:30	20:30	07:24	10:48	13:22	15:40	18:40	
	07:30	10:30	12:45	14:37	16:40	21:00	07:36	11:00	13:30	15:50	19:00	
	07:40	10:40	12:52	14:45	16:50	21:30	07:48	11:12	13:37	16:00	19:30	
	07:50	10:50	13:00	14:52	17:00	22:00	08:00	11:24	13:45	16:10	20:00	
	08:00	11:00	13:07	15:00	17:12	22:30	08:12	11:36	13:52	16:20	20:30	
	08:12	11:10	13:15	15:07	17:24	23:00	08:24	11:48	14:00	16:30	21:00	
	08:24	11:20	13:22	15:15	17:36	23:30	08:36	12:00	14:10	16:40	21:30	
	08:36	11:30	13:30	15:22	17:48	00:00	08:48	12:10	14:20	16:50	22:00	
	08:48	11:40	13:37	15:30	18:00	00:30	09:00	12:20	14:30	17:00	22:30	
	09:00	11:50	13:45	15:37	18:15	01:00	09:20	12:30	14:40	17:12	23:00	
	09:15	12:00	13:52	15:45	18:30		09:40	12:40	14:50	17:24	23:30	
Total no. of departures		89					78					
Minimum single-route fleet size		16					14					

(Continued)

Table 4.6 Comparison of Methods 2 and 4 (20% criterion) using the two types of timetables and of the two types of timetables for a combination of these methods for LA route 217 northbound (continued)

Characteristics		Even headways, combination of Methods 2 and 4 (20%)					Even average load, combination of Methods 2 and 4 (20%)					
Departure times at route departure point	06:00	09:48	12:20	14:19	16:25	20:28	06:00	09:43	12:16	14:17	16:27	20:23
	06:30	10:02	12:28	14:27	16:34	20:59	06:38	10:02	12:26	14:30	16:33	20:59
	07:00	10:12	12:36	14:35	16:43	21:29	07:01	10:15	12:34	14:36	16:44	21:35
	07:12	10:22	12:44	14:43	16:51	21:59	07:21	10:27	12:41	14:46	16:50	21:58
	07:24	10:33	12:52	14:51	17:00	22:29	07:35	10:36	12:50	14:54	17:00	22:27
	07:36	10:43	13:00	15:00	17:13	22:59	07:43	10:46	13:00	15:00	17:11	22:59
	07:47	10:54	13:08	15:08	17:25	23:29	07:52	10:55	13:05	15:11	17:22	23:17
	07:59	11:04	13:16	15:16	17:37	00:00	08:00	11:07	13:11	15:16	17:35	00:00
	08:12	11:14	13:24	15:25	17:50	00:30	08:07	11:14	13:18	15:24	17:53	00:13
	08:25	11:25	13:31	15:33	18:03	01:00	08:16	11:21	13:26	15:35	18:07	01:00
	08:38	11:35	13:39	15:42	18:22		08:33	11:34	13:35	15:41	18:20	
	08:50	11:45	13:47	15:50	18:40		08:43	11:45	13:46	15:50	18:38	
	09:04	11:56	13:55	15:59	18:59		09:02	11:56	13:54	15:59	18:58	
	09:18	12:05	14:03	16:08	19:28		09:14	12:03	14:02	16:06	19:29	
	09:33	12:12	14:11	16:16	19:58		09:27	12:08	14:09	16:16	19:58	
Total no. of departures		85					85					
Minimum single-route fleet size		15					15					

Table 4.7 Comparison of timetables associated with a predetermined number of departures and of timetables using Methods 1 and 3 for LA route 217 northbound

Characteristics		Even average load, 70 departures, Method 2				Even average load, 140 departures, Method 2						
Departure times at route departure point	06:00	10:57	14:07	17:21	06:00	08:37	10:56	12:35	14:05	15:38	17:18	20:57
	06:49	11:11	14:16	17:36	06:32	08:42	11:02	12:40	14:09	15:42	17:22	21:20
	07:13	11:19	14:31	17:54	06:49	08:49	11:10	12:46	14:14	15:47	17:28	21:39
	07:32	11:34	14:39	18:12	06:55	09:01	11:14	12:51	14:23	15:54	17:39	21:52
	07:42	11:48	14:52	18:30	07:12	09:06	11:18	12:57	14:29	15:59	17:48	22:11
	07:51	12:01	15:00	18:49	07:24	09:15	11:22	13:02	14:33	16:04	17:55	22:28
	07:59	12:07	15:15	19:21	07:31	09:23	11:32	13:05	14:36	16:08	18:03	22:45
	08:08	12:16	15:19	19:59	07:38	09:32	11:39	13:08	14:41	16:14	18:13	23:04
	08:21	12:28	15:33	20:30	07:42	09:41	11:46	13:12	14:49	16:23	18:20	23:14
	08:38	12:37	15:41	21:18	07:47	09:55	11:53	13:16	14:54	16:28	18:31	23:39
	08:50	12:48	15:53	21:51	07:51	10:03	11:59	13:21	14:57	16:32	18:41	00:02
	09:06	13:00	16:02	22:27	07:55	10:11	12:03	13:27	15:01	16:37	18:51	00:09
	09:24	13:06	16:12	23:04	07:58	10:19	12:06	13:32	15:08	16:44	19:06	00:32
	09:44	13:14	16:27	23:38	08:03	10:27	12:09	13:38	15:15	16:47	19:24	01:00
	10:04	13:24	16:35	00:09	08:07	10:33	12:14	13:45	15:16	16:53	19:44	
	10:20	13:35	16:46	01:00	08:12	10:38	12:21	13:50	15:21	16:59	20:01	
	10:34	13:48	16:58		08:19	10:44	12:27	13:55	15:29	17:04	20:15	
	10:46	13:58	17:10		08:31	10:51	12:31	14:01	15:34	17:12	20:32	
Total no. of departures		70 (as requested)				140 (as requested)						
Minimum single-route fleet size		12				24						

(Continued)

Table 4.7 Comparison of timetables associated with a predetermined number of departures and of timetables using Methods 1 and 3 for LA route 217 northbound (continued)

Characteristics		Even headways, Method 1						Even headways, Method 3					
Departure times at route departure point	06:00	09:32	12:06	14:06	16:04	18:41	06:00	09:50	13:21	16:11	19:55		
	06:30	09:47	12:15	14:14	16:13	18:58	06:30	10:09	13:34	16:21	20:26		
	07:00	10:01	12:24	14:22	16:22	19:27	07:00	10:25	13:46	16:31	20:56		
	07:12	10:12	12:33	14:30	16:31	19:57	07:13	10:42	13:59	16:41	21:27		
	07:23	10:22	12:43	14:39	16:40	20:28	07:25	10:59	14:12	16:51	21:57		
	07:35	10:33	12:52	14:47	16:48	20:58	07:37	11:16	14:25	17:02	22:28		
	07:46	10:43	13:01	14:55	16:57	21:28	07:49	11:32	14:38	17:15	22:58		
	07:57	10:54	13:09	15:03	17:08	21:58	08:01	11:49	14:51	17:28	23:29		
	08:10	11:04	13:17	15:12	17:19	22:29	08:14	12:04	15:04	17:41	23:59		
	08:23	11:14	13:25	15:21	17:31	22:59	08:28	12:17	15:15	17:54	00:30		
	08:36	11:25	13:33	15:29	17:42	23:29	08:41	12:30	15:27	18:12	01:00		
	08:49	11:35	13:41	15:38	17:53	00:00	08:54	12:43	15:38	18:34			
	09:02	11:45	13:49	15:47	18:07	00:30	09:11	12:56	15:49	18:56			
	09:17	11:56	13:57	15:56	18:24	01:00	09:31	13:08	16:01	19:25			
Total no. of departures		84						67					
Minimum single-route fleet size		15						11					

Table 4.8 Complete timetable with even-headway, Method 4 (20%), at all timepoints of route 217 southbound in Los Angeles

Departure no.	Beachwood Westshire	Gower Franklin	Hollywood Vine	La Brea Sunset	Fairfax Sta Moni	Fairfax Beverly	Fairfax Olympic	Adams Washington
1	05:37	05:41	05:44	05:50	05:56	06:00	06:05	06:11
2	06:07	06:11	06:14	06:20	06:26	06:30	06:36	06:42
3	06:33	06:38	06:41	06:47	06:55	07:00	07:06	07:13
4	06:46	06:51	06:54	07:01	07:08	07:13	07:19	07:26
5	06:59	07:04	07:07	07:14	07:21	07:26	07:32	07:39
6	07:12	07:17	07:20	07:27	07:35	07:40	07:46	07:52
7	07:25	07:30	07:33	07:40	07:48	07:53	07:59	08:05
8	07:35	07:40	07:43	07:51	08:00	08:05	08:11	08:19
9	07:47	07:52	07:55	08:03	08:12	08:17	08:23	08:31
10	07:59	08:04	08:07	08:15	08:24	08:29	08:35	08:43
11	08:11	08:16	08:19	08:27	08:36	08:41	08:47	08:55
12	08:23	08:28	08:31	08:39	08:48	08:53	08:59	09:07
13	08:34	08:39	08:42	08:50	08:59	09:04	09:10	09:18
14	08:45	08:50	08:53	09:01	09:10	09:15	09:21	09:29
15	08:57	09:02	09:05	09:13	09:22	09:27	09:33	09:41
16	09:08	09:13	09:16	09:24	09:33	09:38	09:44	09:52
17	09:19	09:24	09:27	09:35	09:44	09:49	09:55	10:03

(Continued)

Table 4.8 Complete timetable with even-headway, Method 4 (20%), at all timepoints of route 217 southbound in Los Angeles (continued)

Departure no.	Beachwood Westshire	Gower Franklin	Hollywood Vine	La Brea Sunset	Fairfax Sta Moni	Fairfax Beverly	Fairfax Olympic	Adams Washington
18	09:30	09:35	09:38	09:46	09:55	10:00	10:07	10:15
19	09:40	09:45	09:48	09:56	10:05	10:10	10:17	10:25
20	09:50	09:55	09:58	10:06	10:15	10:21	10:27	10:36
21	10:00	10:05	10:08	10:16	10:25	10:31	10:37	10:46
22	10:11	10:16	10:19	10:27	10:36	10:41	10:48	10:56
23	10:21	10:26	10:29	10:37	10:46	10:51	10:58	11:06
24	10:31	10:36	10:39	10:47	10:56	11:02	11:09	11:18
25	10:42	10:47	10:50	10:58	11:07	11:13	11:20	11:29
26	10:53	10:58	11:01	11:09	11:18	11:24	11:31	11:40
27	11:04	11:09	11:12	11:20	11:29	11:35	11:42	11:51
28	11:15	11:20	11:23	11:31	11:40	11:46	11:53	12:02
29	11:26	11:31	11:34	11:42	11:51	11:57	12:04	12:13
30	11:38	11:43	11:46	11:54	12:03	12:09	12:16	12:27
31	11:50	11:55	11:58	12:06	12:15	12:21	12:28	12:39
32	12:02	12:07	12:10	12:18	12:27	12:33	12:40	12:52
33	12:15	12:20	12:23	12:31	12:40	12:46	12:53	13:04
34	12:27	12:32	12:35	12:43	12:52	12:58	13:05	13:16
35	12:37	12:43	12:46	12:54	13:03	13:09	13:16	13:25

36	12:48	12:54	12:56	13:05	13:14	13:20	13:27	13:36
37	12:59	13:04	13:07	13:16	13:25	13:31	13:38	13:47
38	13:10	13:15	13:18	13:26	13:35	13:41	13:48	13:57
39	13:21	13:26	13:29	13:37	13:46	13:52	13:59	14:08
40	13:29	13:34	13:37	13:47	13:57	14:03	14:11	14:20
41	13:38	13:43	13:46	13:56	14:06	14:12	14:20	14:29
42	13:47	13:52	13:55	14:05	14:15	14:21	14:29	14:38
43	13:57	14:02	14:05	14:15	14:25	14:31	14:39	14:48
44	14:06	14:11	14:14	14:24	14:34	14:40	14:48	14:57
45	14:16	14:21	14:24	14:34	14:44	14:50	14:58	15:07
46	14:25	14:30	14:33	14:43	14:53	14:59	15:07	15:16
47	14:34	14:39	14:42	14:52	15:02	15:08	15:16	15:25
48	14:42	14:47	14:50	15:00	15:10	15:16	15:24	15:33
49	14:51	14:56	14:59	15:09	15:19	15:25	15:33	15:42
50	14:59	15:04	15:07	15:17	15:27	15:33	15:41	15:50
51	15:08	15:13	15:16	15:26	15:36	15:42	15:50	15:59
52	15:16	15:21	15:24	15:34	15:44	15:50	15:58	16:07
53	15:25	15:30	15:33	15:43	15:53	15:59	16:07	16:16
54	15:41	15:46	15:49	15:59	16:09	16:15	16:23	16:32
55	15:58	16:03	16:06	16:16	16:26	16:32	16:40	16:49

(Continued)

Table 4.8 Complete timetable with even-headway, Method 4 (20%), at all timepoints of route 217 southbound in Los Angeles (continued)

Departure no.	Beachwood Westshire	Gower Franklin	Hollywood Vine	La Brea Sunset	Fairfax Sta Moni	Fairfax Beverly	Fairfax Olympic	Adams Washington
56	16:15	16:20	16:23	16:33	16:43	16:49	16:57	17:06
57	16:35	16:40	16:43	16:53	17:03	17:09	17:17	17:26
58	17:00	17:04	17:07	17:17	17:27	17:33	17:41	17:50
59	17:24	17:29	17:32	17:42	17:52	17:58	18:06	18:15
60	17:57	18:01	18:04	18:13	18:22	18:27	18:34	18:43
61	18:27	18:31	18:34	18:43	18:52	18:57	19:04	19:13
62	18:59	19:04	19:07	19:15	19:23	19:28	19:34	19:42
63	19:29	19:34	19:37	19:45	19:53	19:58	20:04	20:12
64	20:00	20:05	20:08	20:16	20:23	20:28	20:34	20:42
65	20:30	20:35	20:38	20:46	20:53	20:58	21:04	21:12
66	21:01	21:06	21:09	21:17	21:24	21:29	21:35	21:43
67	21:31	21:36	21:39	21:47	21:54	21:59	22:05	22:13
68	22:01	22:06	22:09	22:17	22:24	22:29	22:35	02:43
69	22:31	22:36	22:39	22:47	22:54	22:59	23:05	23:13
70	23:01	23:06	23:09	23:17	23:24	23:29	23:35	23:43
71	23:32	23:37	23:40	23:48	23:55	00:00	00:06	00:14
72	00:02	00:07	00:10	00:18	00:25	00:30	00:36	00:44
73	00:32	00:37	00:40	00:48	00:55	01:00	01:06	01:14

measures fulfil the first objective. One set of options when selecting the procedure to construct the timetables is referred to as balancing passenger loads on individual vehicles while allowing unevenly spaced headways. The underlying approach in that set of options is to shift the departure times of individual vehicles so as to obtain even average loads while relaxing the even-headways requirement. The latter accomplishes the second objective.

The third objective of the study is to allow for direct vehicle-frequency changes. This is achieved in the process described in Figure 4.6, whereby the schedulers can either interject their own set of intuitive or experience-based frequencies, or they can substitute some of the derived frequencies. In this way, possible scheduling exceptions (e.g. special passenger demand because of a sports event) could be inserted. The common manual process in creating timetables often encounters a problem in smoothing the headways in the transition segments between adjacent time periods. In both types of timetables that have been described (even-headway and even-load), the procedures developed ensure, in an average sense, the fulfilment of the desired occupancy standard. This basically fulfils the smoothing necessity expressed in the fourth objective. The fifth and last objective is attained by allowing the transit agency to request a selection of different frequency-setting methods for different time periods. The scheduler/planner can then select for peak periods those methods that are more sensitive to resource saving (e.g. Method 4), and for off-peak periods methods that are more sensitive to passenger comfort (e.g. Method 2).

4.6 Literature review and further reading

The problem of finding the best dispatching policy for transit vehicles on fixed routes has a direct impact on constructing timetables. This dispatching-policy problem, which has been dealt with quite extensively in the literature, can be categorized into four groups: (1) models for an idealized transit system, (2) simulation models, (3) mathematical programming models, and (4) data-based models.

The first group, idealized transit systems, was investigated by, for example, Newell (1971), Osuna and Newell (1972), Hurdle (1973), Wirasinghe (1990, 2003), and De Palma and Lindsey (2001). Newell (1971) assumed a given passenger-arrival rate as a smooth function of time, with the objective of minimizing total passenger waiting time. He showed analytically that the frequency of transit vehicles with large capacities (in order to serve all waiting passengers) and the number of passengers served per vehicle varied with time approximately as the square root of the arrival rate of passengers. Osuna and Newell (1972) developed control strategies for either holding back a transit vehicle or dispatching it immediately, based on a given number of vehicles, random round-trip travel times with known distribution functions, and uniform passenger-arrival rates with a minimum waiting-time objective. Using dynamic programming, they found that the optimal strategy for two vehicles and a small coefficient of variation of trip time retained nearly equally spaced dispatch times. Hurdle (1973), investigating a similar problem, used a continuum fluid-flow model to derive an optimal dispatching policy while attempting to minimize the total cost of passenger waiting time and vehicle operation.

Wirasinghe (1990, 2003) examined and extended Newell's dispatching policy while considering the cost components initially used by Newell (1973). Wirasinghe considered the average value of a unit waiting time per passenger (C_1) and the cost of dispatching a vehicle (C_2) to show that the passenger-arrival rate in Newell's square root formula is multiplied by

($C_1/2C_2$). Wirasinghe also showed how to derive the equations of total mean cost per unit of time by using both uniform headway policy and Newell's variable-dispatching policy.

De Palma and Lindsey (2001) developed a method for designing an optimal timetable for a single line with only two stations. The method is suitable for a situation in which each rider has a precise time in which the person wants to travel; travelling earlier or later than desired increases the total cost. The objective is to minimize riders' total schedule-delay costs. Two cases are analysed with respect to passenger preferences. In the first case, all passengers treat a unit of delay equally. The second case assumes several rider groups, with different levels of delay costs ascribed to riders from the different groups. In addition, the researchers compared two models: a 'line' model, in which preferred travel times are uniformly distributed over part of the day and trips cannot be rescheduled between days; and a 'circle' model, in which preferred travel times are uniformly distributed over the full 24-hour day and trips can be rescheduled between days. Optimal timetables are derived for each of the models.

In the second group, simulation models were studied by, for example, Marlin *et al.* (1988), Adamski (1998) and Dessouky *et al.* (1999). Marlin *et al.* (1988) developed a simulation model for dispatching transit vehicles every day. They checked the feasibility of the results and used mathematical programming for vehicle assignments in an interactive computer-support system. Adamski (1998) employed a simulation model for real-time dispatching control of transit vehicles while attempting to increase the reliability of service in terms of on-time performance. His simulation implemented optimal stochastic control with linear feedback. The use of intelligent transportation systems was applied by Dessouky *et al.* (1999) in a study of bus dispatching at timed transfer points. The researchers used a simulation analysis to show that the benefit of knowing the location of the bus was most significant when the bus was experiencing a significant delay, especially when there was a small number of connecting buses at transfer points.

Mathematical programming methods, the third group for determining frequencies and timetables, have been proposed by Furth and Wilson (1981), Koutsopoulos *et al.* (1985), Ceder and Stern (1984), Eberlein *et al.* (1998), Gallo and Di-Miele (2001), and Peeters and Kroon (2001). Furth and Wilson sought to maximize the net social benefit, consisting of ridership benefit and waiting-time saving, subject to constraints on total subsidy, fleet size, and passenger-load levels. Koutsopoulos *et al.* extended this formulation by incorporating crowding-discomfort costs into the objective function and treating the time-dependent character of transit demand and performance. Their initial problem consisted of a nonlinear optimization program relaxed by linear approximations. Ceder and Stern addressed the problem with an integer programming formulation and a heuristic person-computer interactive procedure. Their approach focuses on reconstructing timetables when the available vehicle fleet size is restricted. Eberlein *et al.* (1998) studied a special dispatching problem for the purpose of introducing deadheading trips in high-frequency transit operations. They solved their dispatching strategy optimally; they also determined the number of stops that could be skipped in order to minimize total passenger cost in the system.

Gallo and Di-Miele (2001) produced a model for the special case of dispatching buses from parking depots. Their model is based on the decomposition of generalized assignments and design, non-crossing and matching sub-problems. It can be extended to a case in which there is an overlap between arrival and departure-vehicle flows. Peeters and Kroon (2001) present a procedure for planning an optimal cyclic rail timetable; i.e. a timetable in

which trains leave at the same minute every hour. The problem is represented through a constraint graph, in which each node is an event that needs to be scheduled; cycles are examined according to a calculation of tensions and potentials. The model is formulated as a mixed-integer nonlinear program with the objectives of minimizing passenger time, maximizing timetable robustness and minimizing the number of required trains. A solution procedure is suggested, by which the nonlinear part of the formulation is transformed into a mixed-integer linear problem that is an approximation of the original problem; further actions are taken in order to reduce the number of constraints.

In the fourth and last group, the data-based models described in this chapter are based on Ceder (1986, 2003). Advanced models pertaining to this group will be presented in the next chapter.

Exercises

- 4.1 During a morning time period of 90 minutes (07:00–08:30), ride-check data were collected on a given bus route (average across several days). The route is 7-km long, and there is only one stop (B) between the first (departure) stop (A) and last (arrival) stop (C), the distance from A to B being 3 km. The average of the data gathered and information on three trips are as follows:
Given also a predetermined departure time of 07:00:

Departure time	Average load (passengers) when departing the stop		Desired occupancy (passengers)	Minimum frequency (veh/hr) (07:00–08:30)	Bus capacity (passengers)
	A	B			
07:20	50	25			
07:55	70	49	50	2	80
08:30	40	30			

- (a) Construct an even-load timetable at point A using: (i) Method 3, and (ii) Method 4 (30%) for frequency determination, with Method 4 having an excess load along a maximum of 30% of its route length.
- (b) Repeat (a) above, but with a bus capacity of 60 (instead of 80) for all trips.
- 4.2 Based on the data and results of Exercise 3.1 (Chapter 3):
- (a) Construct evenly spaced headway timetables using Methods 2 and 4 (20%) at the route departure point.
- (b) Construct clock-headway timetables using Methods 2 and 4 (20%) at the route departure point.

- (c) Find the comparison measures for each of the four timetables in (a) and (b) (total no. of departures and minimum required single-route fleet size).
- 4.3 Based on the data and results of Exercise 3.2 (Chapter 3):
 - (a) Construct an even average load timetable according to Method 2.
 - (b) Use the results of (a) to answer the following: (i) Are the loads at each stop, on each vehicle, in each hour, the same? (ii) If the individual loads are not the same, at which stop and for what bus will the average load exceed the desired occupancy?
- 4.4 Using the following data

Time	Average max load (passengers)	Desired occupancy (passengers)	Passengers' arrival-rate pattern	Minimum required frequency (veh/hr)
07:00–08:00	240	64	uniform	2
08:00–09:00	180	24	uniform	4

and a requirement that the average max load on the last vehicle between 07:00 and 08:00 be the same as on the first vehicle between 08:00 and 09:00, or $44 = (64 + 24)/2$ passengers:

- (a) Construct an even-headway timetable at the max load point (the same for both hours), assuming that passenger demand starts at 07:00; include the first departure after 09:00, assuming that demand after 09:00 is the same as that from 08:00–09:00.
- (b) What departure times and average loads are derived by Method 2 using an even-headway procedure in the transition between the two hours? Explain differences from (a) and draw conclusions.

References

Adamski, A. (1998). Simulation support tool for real-time dispatching control in public transport. *Transportation Research*, **32A**(2), 73–87.

Ceder, A. (1984). Bus frequency determination using passenger count data. *Transportation Research*, **18A**(5/6), 439–453.

Ceder, A. (1986). Methods for creating bus timetables. *Transportation Research*, **21A**(1), 59–83.

Ceder, A. (2003). Public transport timetabling and vehicle scheduling. In *Advanced Modeling for Transit Operations and Service Planning* (W. Lam and M. Bell, eds), pp. 31–57, Elsevier Ltd.

Ceder, A. and Stern, H. I. (1984). Optimal transit timetables for a fixed vehicle fleet. In *Proceedings of the 10th International Symposium on Transportation and Traffic Theory* (J. Volmuller and R. Hammerslag, eds), pp. 331–355, UNU Science Press.

- De Palma, A. and Lindsey, R. (2001). Optimal timetables for public transportation. *Transportation Research*, **35B**, 789–813.
- Dessouky, M., Hall, R., Nowroozi, A. and Mourikas, K. (1999). Bus dispatching at timed transfer transit stations using bus tracking technology. *Transportation Research*, **7C**(4), 187–208.
- Eberlein, X. J., Wilson, N. H. M., and Barnhart, C. (1998). The real-time deadheading problem in transit operations control. *Transportation Research*, **32B**(2), 77–100.
- Furth, P. G. and Wilson, N. H. M. (1981). Setting frequencies on bus routes: theory and practice. *Transportation Research Board*, **818**, 1–7.
- Gallo, G. and Di-Miele, F. (2001). Dispatching buses in parking depots. *Transportation Science*, **35**(3), 322–330.
- Hurdle, V. F. (1973). Minimum cost schedules for a public transportation route. *Transportation Science*, **7**(2), 109–157.
- Koutsopoulos, H. N., Odoni, A., and Wilson, N. H. M. (1985). Determination of headways as function of time varying characteristics on a transit network. In *Computer Scheduling of Public Transport 2* (J. M. Rousseau, ed.), pp. 391–414, North-Holland Publishing Co.
- Marlin, P. G., Nauess, R. M., Smith, L. D. and Rhoades, M. (1988). Computer support for operator assignment and dispatching in an urban transit system. *Transportation Research*, **22A**(1), 13–26.
- Newell, G. F. (1971). Dispatching policies for a transportation route. *Transportation Science*, **5**, 91–105.
- Newell, G. F. (1973). Scheduling, location, transportation and continuous mechanics: some simple approximations to optimization problems. *SIAM Journal of Applied Mathematics*, **25**(3), 346–360.
- Osuna, E. E. and Newell, G. F. (1972). Control strategies for an idealized public transportation system. *Transportation Science*, **6**(1), 52–72.
- Peeters, L. and Kroon, L. (2001). A cycle based optimization model for the cyclic railway timetabling problem. In *Computer-aided Scheduling of Public Transport*. Lecture Notes in Economics and Mathematical Systems, **505** (S. Voss and J. R. Daduna, eds), pp. 275–296, Springer-Verlag.
- Salzborn, F. J. M. (1972). Optimum bus scheduling. *Transportation Science*, **6**, 137–148.
- Wirasinghe, S. C. (1990). Re-examination of Newell's dispatching policy and extension to a public transportation route with many to many time varying demand. In *Transportation and Traffic Theory* (M. Koshi, ed.), pp. 363–378, Elsevier Ltd.
- Wirasinghe, S. C. (2003). Initial planning for an urban transit system. In *Advanced Modeling for Transit Operations and Service Planning* (W. Lam and M. Bell, eds), pp. 1–29, Elsevier Ltd.

This page intentionally left blank

5

Advanced Timetables I: Maximum Passenger Load



Chapter 5 Advanced Timetables I: Maximum Passenger Load

Chapter outline

- 5.1 Introduction
 - 5.2 Even max load on individual vehicles
 - 5.3 Optimization, operations research and complexity
 - 5.4 Minimum passenger-crowding timetables for a fixed vehicle fleet
 - Exercise
 - References
-

Practitioner's Corner

The term ‘advanced’ in the title of this chapter has already surrendered its level of computation. Indeed not all the sections may interest practitioners. Nonetheless, this chapter and the next present a practical framework and tools for improving and optimizing the construction and setting of transit timetables. Practitioners can, almost effortlessly, study the procedures with the aid of graphical schemes, although the process of constructing the timetables themselves requires some mathematical treatment.

This present chapter contains two main, independent parts. First, it picks up where Chapter 4 left off, with the construction of timetables, while assuring an even average max load at each vehicle’s max load point. This contrasts with Chapter 4’s even average max load at a specific timepoint, usually the hourly max load point. Second, optimal timetables in which passenger-crowding situations are minimized are derived for a reduced fleet size. In other words, an answer is given to the question: What are the best timetable frequencies that can be assigned for a given fleet size when trying to minimize expected overcrowding situations? Background on optimization techniques is given in Section 5.3 to aid the second part of this chapter and the next chapter.

After reading Section 5.1, practitioners are encouraged to follow carefully the implementation of Example 2, used initially in Chapters 3 and 4 in the case of even loads at different vehicle max load points. This implementation takes place in section 5.2.3, which includes an interpretation and explanation of the procedure developed. Practitioners may be especially interested in Sections 5.4.1 and 5.4.3, and possibly 5.4.4, for this special problem of optimal timetables for a fixed fleet size. The exercise at the end is recommended for practising the procedure of the first part of the chapter.

The following is a short story about people who don’t know to keep up with a timetable. Three people arrive at a train station and ask for the next train to a certain destination. The conductor in the booth replies that the train has just left and the next one will leave in one hour. The three look at one another and say: “Let’s go and have a beer, we have time”. When they return, the conductor tells them that they missed the train by two minutes and they will have to wait one hour again. The three go off again to have

a beer. For the third time, they miss the train, by one minute. They are then told that the next train will be the last one of the day and will be leaving in 90 minutes, so they had better be on time. The three nod and go to have more beer. Arriving back at the station, they see the train starting to pull out. All three run after it, two succeeding to hang on while the third person remains standing and laughing on the platform. The conductor asks him: “Why are you laughing?” He replies: “You see the two people hanging onto the train, they just wanted to escort me onto the train”. We hope in this chapter to offer some planning tools to help passengers . . . to be on time, and also to put to rest the following saying: “The only way to catch a bus/train is to miss the one before it”.

5.1 Introduction

Although the even-load procedure described in Chapter 4 ensures even average loads of d_{0j} at the j -th hourly max load point, it does not guarantee that the average load for individual vehicles at other stops will not exceed d_{0j} ; therefore, the result may be overcrowding. In other words, the hourly max load point represents an average peak point at j , whereas the max load point for an individual vehicle can come at another stop and exceed d_{0j} . For a given time period, then, each vehicle may have a different max load point and a different observed average load across the entire vehicle route. The purpose of the first part of this chapter is to derive a timetable such that, on average, all vehicles will have even loads (equal to the desired occupancy) at the max load stop for each vehicle. The adjustments in the timetable are not intended for highly frequent urban services, in which the headway may be less than, say, 10 minutes or an hourly frequency of about 6 vehicles or more. The objective of the first part of this chapter (based on Ceder, 2001) is to construct a timetable so as to avoid passenger-overcrowding situations (i.e. loads greater than d_{0j}).

The second part of the chapter will examine optimal timetables for a network of transit routes with reduced fleets. In a practical context, this problem can arise when the fleet size is reduced because of vehicle-age attrition and budget policy decisions. Although the next two chapters cover vehicle scheduling and minimization of fleet size, they assume that the frequencies of departure times are given. The purpose of the second part of this chapter is to construct timetables while assuring that: (i) the difference between the observed max load at hour j and d_{0j} is minimized for all routes, and (ii) the given (available) fleet size is sufficient. This is a problem of optimal multi-terminal timetable construction for different fixed sizes of a vehicle fleet. The formulation used in this part follows Ceder and Stern (1984).

5.2 Even max load on individual vehicles

This section continues Sections 4.3 and 4.4 in Chapter 4, but employs a procedure to assure average even load at each vehicle's max load point. This is in contrast to Principle 2 and Proposition 2 (Section 4.4.1), in which even loads are practicable only at the hourly max load point. To have an even desired load at the (individual) vehicle's max load is an additional

notion in the attempt to further reduce overcrowding; the procedure can be implemented in certain transit-operation scenarios by route and time period. We will continue to use Example 2 as it appeared in Figure 3.7 and Table 4.1.

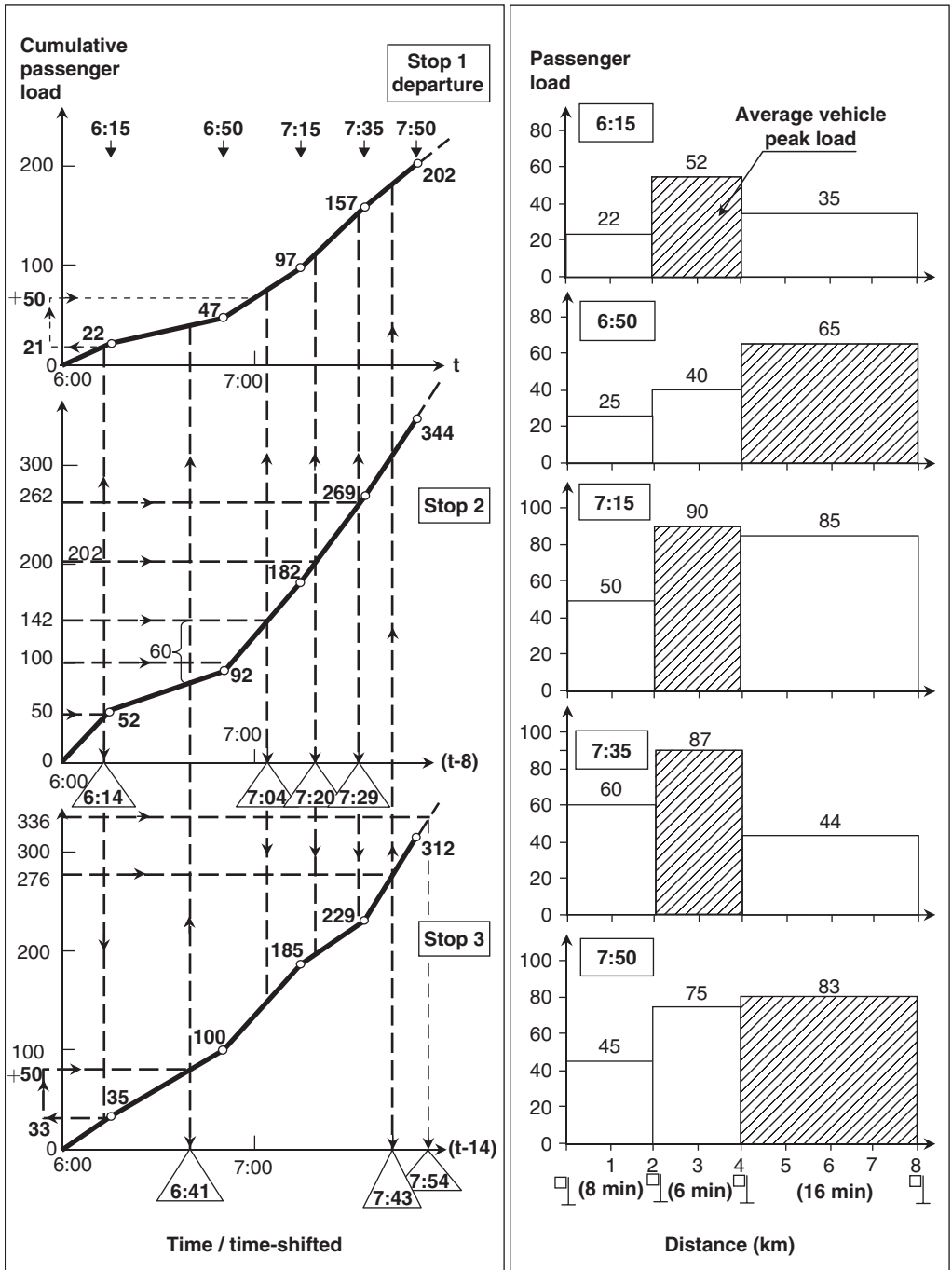
5.2.1 The underlying principle

The results of applying Proposition 2 on Example 2 are shown in Figure 4.5 and Table 4.3. The upper part of Figure 4.5 and the first column of Table 4.3 show the derived departure times, based on even load at stop 3 (6:00–7:00) and stop 2 (7:00–8:00), using Method 2 for frequency determination. Take, for example, the resultant departure at 7:48 (Figure 4.5 and Table 4.3, using Method 2), with 60 average passengers on-board at stop 2 (hourly max load point). This (desired) load of 60 is derived by construction. However, if we check the average load of this 7:48 departure at stop 3, we will find a larger load than 60. That is, from Figure 3.7 (Chapter 3), the average passenger-arrival rate between 7:35 and 7:50 is $83/15 = 5.53$ passengers per minute. Hence, between 7:36 and 7:48 (the results based on Proposition 2), this rate holds and will yield $12 \cdot 5.53 \approx 66$ passengers on the 7:48 bus at stop 3. In order to overcome such undesirable overcrowding (above the desired load of 60), the following principle is employed:

Principle 3: Construct a cumulative passenger-load curve at each stop (except for the arrival point), moving horizontally per each d_{oj} (desired occupancy at time period j) for all j on each curve until each of the cumulative-load curves is intersected, and then vertically to establish a departure time for each curve. The required departure time is the *minimum* time across all curves. Using the last determined departure time, set the corresponding loads across all curves; add to these loads d_{oj} or the next d_{oj} (in the transition between time periods), and move horizontally and vertically, as in the first step, to derive the next departure time. Repeat until the end of the time span.

Proposition 3: Principle 3 results in departure times such that the observed average max load on individual vehicles approaches the desired occupancy d_{oj} .

Proof: Proposition 3 is proved, as in Chapter 4, by construction. Fortunately, Principle 3, similar to Principles 2 and 3 in Chapter 4, can be graphically interpreted to ease the proof. Figure 5.1 illustrates Principle 3 for Example 2 in Figure 3.7. Figure 5.2 is an enlarged part of Figure 5.1 for the start times at stops 2 and 3. Figure 5.2 shows the load profiles of the five departures and three cumulative-load curves at three stops. The curves at stops 2 and 3 are shifted by 8 and 14 minutes, respectively, to allow for an equal time basis at the route's departure point. At the initialization, the value of $d_{o1} = 50$ is coordinated with the three cumulative-load curves to obtain: 6:51.5 at stop 1, 6:14 at stop 2, and 6:41 at stop 3. Although this is clearly seen in Figure 5.3 for stops 2 and 3, the determination at stop 1 is based on an arrival rate of 2 passengers per minute ($= 50/25$) and an additional load of 3 beyond the observed load of 47 at 6:50. According to Principle 3, the minimum time among the three is selected: the departure at 6:14 (emphasized in the figures). This means that the scheduled 6:15 departure is shifted backward by one minute to have an average of 50 instead of 52 passengers at stop 2. Then, $d_{o1} = 50$ is added to the value of the cumulative-load curves at 6:14 to attain three more candidate departure times: 7:02 (see following note), 6:52 and 6:41 for stops 1, 2 and 3, respectively. Hence, 6:41 is selected. If the candidate time is beyond the boundary of the time period (7:00), another d_{oj} , $d_{o2} = 60$, should be applied; therefore, 7:02



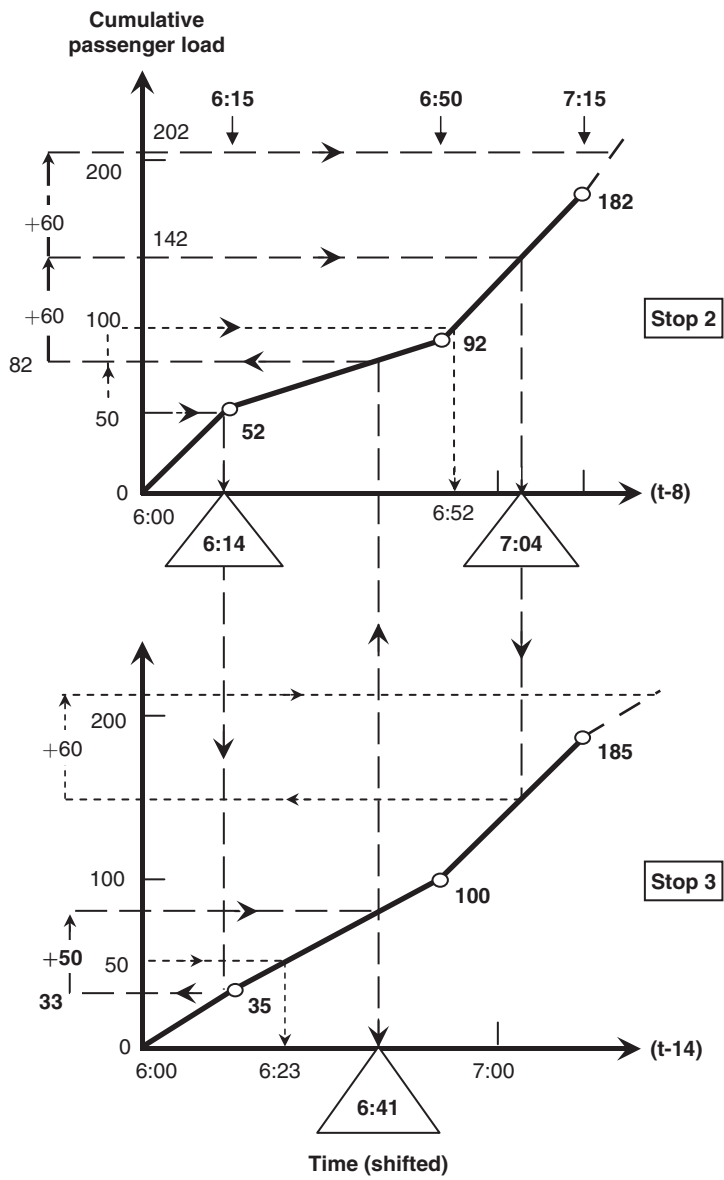


Figure 5.2 Enlarged part for stops 2, 3 of the graphical interpretation of the Algorithm Balance shown in Figure 5.1

becomes 7:07 for stop 1. The pro-cedure continues with $d_{o2} = 60$ and results in 7:04 as the third departure having an average even 60-passenger load at stop 2 and four more departures as illustrated in Figure 5.1. This completes the proof-by-construction of Proposition 3.

The basic assumptions of Proposition 3 are as follows: (i) the change of departure times will not affect the arrival pattern of passengers; (ii) the change of departure times (with the same frequency) will not affect passenger demand. These assumptions are not critical to the

procedure described, since changes in passenger-arrival patterns and demand will be captured frequently using updated data.

5.2.2 Minimum frequency standard

The procedure described by Principle 3, similar to Principle 2, does not guarantee that the minimum frequency standard (the inverse of policy headway), F_{mj} , for each time interval j , will always be met. Usually F_{mj} is necessary at the beginning and end of the day. Therefore, during the process of Principle 3, F_{mj} needs to be checked in the transition between time intervals – i.e. whenever a new derived departure is in advance of the time boundary. Using N_j , which is the number of departures derived by Principle 3 during time interval $j - 1$, F_{mj} is checked in the following manner:

- (a) check for each j , $N_j > F_{mj}$?; if yes – END, otherwise continue $j = 1, 2, \dots, q$;
- (b) calculate the alternative desired occupancy, d_{om} , for the minimum frequency situation:

$$d_{om} = \frac{\max_{i=1,2,\dots,n} [L_i(t_j) - L_i(t_{j-1})]}{F_{mj} + 1}, \text{ and return to the procedure based on}$$

- (c) Principle 3 at the $(j - 1)$ th transition time, where there are n stops, t_j is the j th transition time (between time periods), and $L_i(t_j)$ is the cumulative-load curve at stop i at t_j ; change d_{oj} to d_{om} and go to (a).

The procedure described by steps (a) and (b) ensures that the minimum frequency criterion will always be met. If the procedure described by Principle 3 results in time interval j with a frequency of less than F_{mj} , then the maximum difference in passenger loads on the cumulative-load curve between t_j and t_{j-1} determines a new desired occupancy, d_{om} . This d_{om} then replaces d_{oj} until the derivation of the first departure time at interval $j + 1$.

There are two more comments that should be mentioned about the procedure described by Principle 3 and the minimum frequency case:

- (i) If a different type of vehicle (than the one for which d_{oj} is set) is considered for the determination of a given departure, d_{oj} can be changed in Principle 3. This may be the case for an excessive load, which may result in too short a headway, or for the opposite case of a large amount of empty seat-km, resulting in F_{mj} . Both cases can be observed from the cumulative-load curves or load profiles.
- (ii) If use of a headway criterion policy (or the inverse of F_{mj} for each j) is preferred, then there will be an extra check in the procedure for F_{mj} in (b) for each derived departure time k at j : if $h_{kj} > \frac{1}{F_{mi}}$, set $h_{kj} = \frac{1}{F_{mj}}$ and continue with the procedure described by Principle 3, where h_{kj} is the headway between the k th and $(k - 1)$ th departures, otherwise – END.

5.2.3 Comparison

The comparison between the observed data of Example 2 and the results of the procedures using Principles 1, 2 (Chapter 4) and 3 is summarized in Table 5.1 and illustrated in

Figure 5.3. Table 5.1 shows the associated *individual* average max load under each departure; its corresponding stop appears under max load.

Table 5.1 *Departure times for Example 2 route-departure points (three even headways and even-load procedures), accompanied by their associated max load and its starting stop, using frequencies in Method 2*

Trip characteristics associated with route- departure point		Procedures			
		Observed data	Evenly spaced headways	Even average load at hourly max load point	Even average load at individual max load point
1st* departure	Depart. time	6:15	6:22	6:23	6:14
	Max load	52	60	61	50
	Max load point	Stop 2	Stop 2	Stop 2	Stop 2
2nd		6:50 65 Stop 3	6:44 41 Stop 3	6:50 50 Stop 3	6:41 50 Stop 3
3rd		7:15 90 Stop 2	7:05 61 Stop 2	7:07 61 Stop 2	7:04 60 Stop 2
4th		7:35 87 Stop 2	7:22 66 Stop 2	7:22 59 Stop 2	7:20 60 Stop 2
5th		7:50 83 Stop 3	7:39 77 Stop 2	7:36 62 Stop 2	7:29 60 Stop 2
6th		—	7:56 94 Stop 3	7:48 66 Stop 3	7:43 60 Stop 3
7th		—	—	8:00 66 Stop 3	7:54 60 Stop 3

* Subsequent to a predetermined 6:00 departure.

Figure 5.3 shows the diversity of the individual max loads across all the procedures and the observed ones. Certainly, this comparison applies only to the specific Example 2, as it will vary from one situation to another. In situations in which the hourly max load point usually coincides with the individual max load, the results of Principle 2 will be close to those of Principle 3, as is the case of Example 2. In the examples that were presented by Ceder (2001), the differences are more significant.

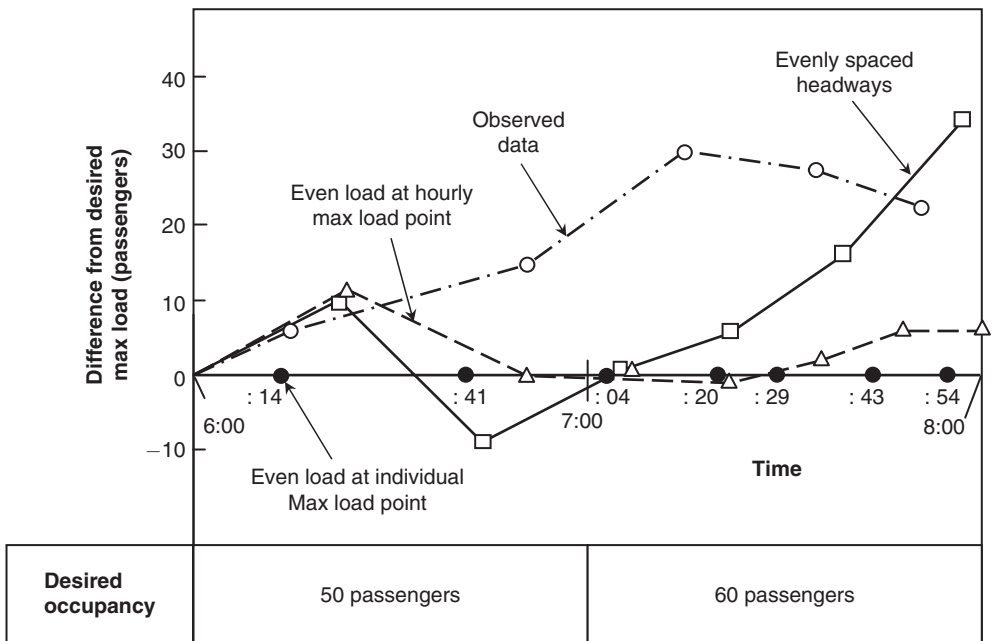


Figure 5.3 Differences between individual vehicle's max load and the desired occupancy according to the procedure selected in Example 2

This section ends with a brief overview of what is done in practice. Different transit agencies employ different scheduling strategies based primarily on their own planners'/schedulers' experience, and secondarily on their scheduling software (if any). As a result, it is unlikely that two independent transit agencies will use exactly the same scheduling procedures, at least on the detailed level. In addition, even in the same transit agency, schedulers may use different procedures for different groups of routes. Consequently, there is a need when developing computerized procedures to supply the planners/schedulers with alternative schedule options along with an interpretation and explanation of each alternative. Two such alternatives were presented in Chapter 4 and one in this section. Undoubtedly, it is desirable that one of the alternatives should coincide with the scheduler's manual procedure. In this way, the scheduler will be in a position not only to expedite manual tasks, but also to compare the different procedures with regard to the trade-off between passenger comfort and operating cost. The next section will focus on how to set certain parts of the transit timetables so as to obtain, at the planning stage, the simultaneous arrivals of specific trips. This synchronization challenge, if it ends successfully, supplies an important improvement to transit level-of-service.

5.3 Optimization, operations research and complexity

This section introduces the notion of optimization, using operations research (OR) principles/methods. Optimization usually means finding the best solution to some problem

from a set of alternatives. Following Ceder (1999), OR means a scientific approach to decision-making, or formulating problems so they can be solved quantitatively. Hence, optimization involves dealing with mathematical models of situations or phenomena that exist in the real world. The development of powerful computers has had a very strong impact on the motivation to develop complex, sophisticated OR models.

The basic elements in OR models include the following: (1) *Variables* (decision, policy, independent or dependent variables); these are quantities that can be manipulated to achieve some desired objective or outcome. (2) *Objective function* (profit, revenue or cost function); this is a measure of effectiveness (efficiency, productivity or utility value) associated with some particular combination of the variables and is a function that is to be optimized (minimized or maximized). (3) *Constraints* (feasibility conditions); these are equations or inequalities that the variables must satisfy in addition to providing a minimum or maximum value of the objective function.

Mathematically speaking, the variables are often represented as $x_1, x_2, \dots, x_{1n}$ or the vector of the variables $\mathbf{x} = (x_1, x_2, \dots, x_{1n})$. The objective function, as a single-value function, is represented as $z = f(x_1, x_2, \dots, x_{1n})$ or $z = f(\mathbf{x})$. The constraints are also represented as a function of $\mathbf{x}$, $g_v(\mathbf{x})$, with v as a constraint number. The general variable optimization problem, with m constraints, is to

$$\begin{aligned} &\text{minimize/maximize} \quad z = f(\mathbf{x}) \\ &\text{subject to (s.t.)} \quad g_v(\mathbf{x}) \{ \geq, =, \leq \} 0, v = 1, 2, \dots, m \end{aligned} \quad (5.1)$$

It is apparent that $\min[f(\mathbf{x})] = -\max[-f(\mathbf{x})]$, and hence each minimization problem can be treated as a maximization problem and vice versa. Notations that correspond to OR and are used in the formulations in this book contain the symbols $\forall$ (for all), $\in$ (belong to), $\notin$ (not belong to), and $\emptyset$ (empty set).

Occasionally, we cannot prove that optimality is attained; nor can we prove that optimality cannot be attained in all those cases. In these situations, the OR model is termed a *heuristic procedure*. In certain cases, a mathematical formulation may possibly be too complex to be handled even by advanced computers, thereby leading to the intentional construction of a heuristic procedure. This is actually the case with timetable synchronization, to be dealt with in due course.

A framework for developing a mathematical OR model of a system's problem is shown in Figure 5.4 for a deterministic (average-based) study. It relies on an eight-step procedure. OR analysts should first define the system's (organization's) problem while searching for its actual objectives, then look for the entire system's limitations and constraints. The third step, in Figure 5.4, translates the first two steps into a formulation similar to (5.1). Once the formulation is correctly established, a check can be made of the existence of feasible solutions, or solutions that comply fully with all the constraints. If feasibility does not exist, return to the first two steps for possible changes and/or relaxation of some constraints. Otherwise, select a software package or develop an algorithm to solve the OR formulation automatically. The sixth step will involve examining the important issue of the complexity (described below) of the algorithm to verify whether indeed the system's problem can be solved by computer within a reasonable amount of time. If the algorithm is too complex, thus requiring too long a running time, a simplified heuristic procedure may be constructed to solve the problem. Otherwise, a check will have to be conducted of the

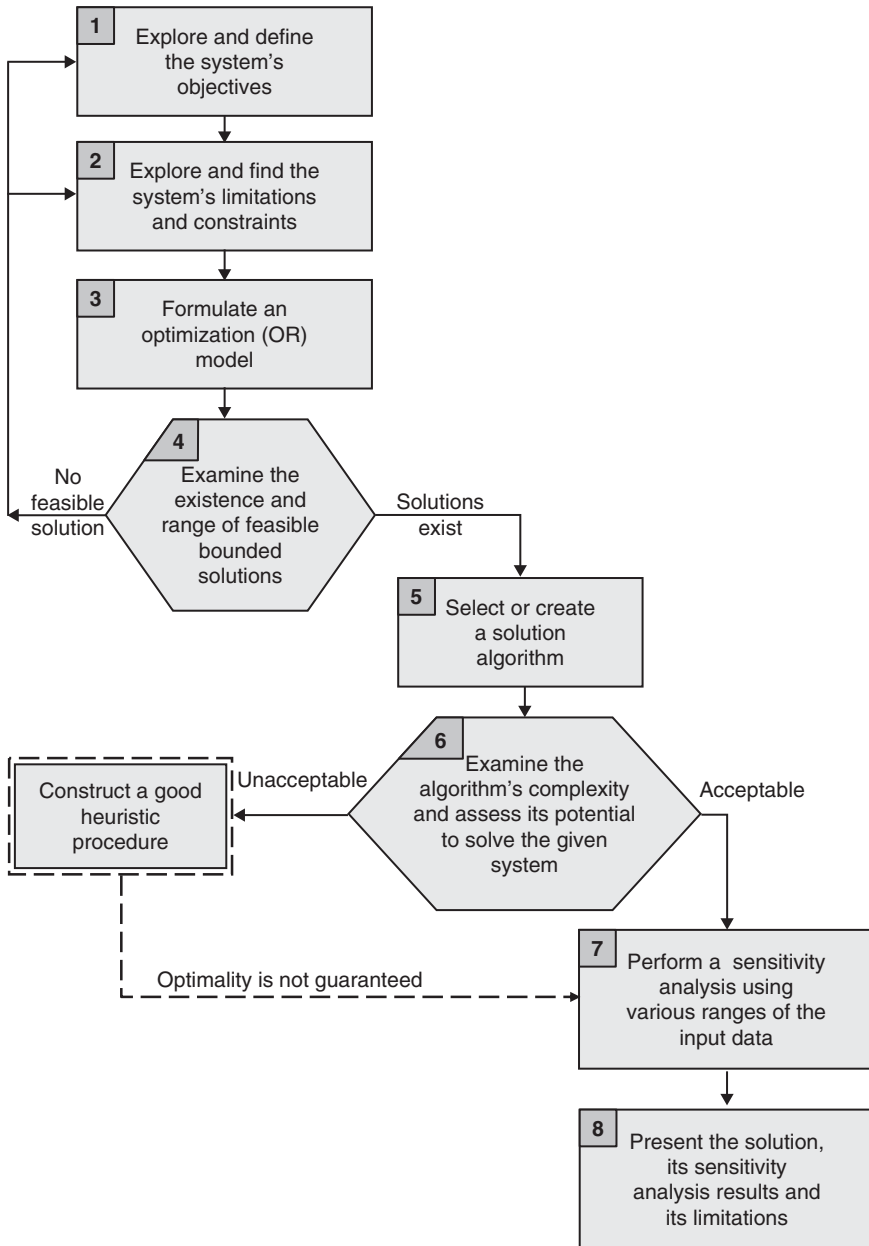


Figure 5.4 Schematic framework of an OR process to solve a deterministic optimization problem analytically

sensitivity of the problem's objective value to different input data. The final step appropriately presents the results along with the sensitivity analysis and the problem's limitations. In the course of a heuristic procedure, it becomes apparent that optimality cannot be guaranteed.

One important characteristic of an OR model/procedure/algorithm is the efficiency with which the formulation may be run by all the various computing resources required. The solution time is usually checked for the worst-case scenario or worst-case complexity, which is well described by Ahuja *et al.* (1993). It becomes commonplace to use the notation $O(f(k,q))$, an operation of $f(k,q)$, for expressing the running-time complexity function, which is dependent on the input parameter size k and q , where $f(k,q)$ is the function of those two parameters. For example, if the running time is $f(k,q) = (10k + 0.01k^3) \cdot q^5$, then for all $k \geq 100$ (as we always look for the worst case), the second term in the parenthesis dominates, and the complexity of this function is $O(k^3 \cdot q^5)$. The $O(\cdot)$ notation provides, for sufficiently large input values, an indication of the number of elementary mathematical operations required. Following Ahuja *et al.* (1993), the complexity of OR optimization algorithms are often interpreted as *efficient* if its worst case is bounded by a polynomial function of the input parameters; for examples, algorithms with $O(k^3 + q^5)$, $O(kq + q^3 \log(k^2))$. Nonetheless, the algorithm found in many combinatorial optimization problems is of the type called *exponential-time algorithm* with, for instance, $O(k!)$, $O(3^q)$, $O(k^{\log(q)})$. Some of these hard-to-solve problems belong to a class known as *NP-complete* (non-polynomial complete). Certainly it is a challenge to develop a polynomial-time algorithm for a problem in this class. In these NP-complete problems, we may depart from a completely proven optimization solution and begin developing a heuristic procedure.

The next two sections, as well as several subsequent chapters, will use the OR optimization concept as a framework for achieving the best results.

5.4 Minimum passenger-crowding timetables for a fixed vehicle fleet

The problem addressed in this section is that of optimal multi-route timetable construction for a transit agency. The motivation for this task lies in the challenge facing schedulers/planners in reconstructing timetables when the available vehicle fleet is reduced. The purpose of this section is to present the formulation to solve the problem optimally by using OR tools (i.e. integer programming). Because of the complexity involved in solving integer programming formulation, a heuristic approach designed for a person-computer interactive procedure will also be proffered in the next chapter.

5.4.1 Problem description

More concretely, the problem is as follows: given a fixed fleet size (number of vehicles) and passenger load at the max load segment between a set of terminals (predefined routes), what is the frequency of vehicle departures that will maximize some measure of passenger satisfaction? A fixed set of terminals and a fixed planning horizon are given.

The time horizon for each terminal is partitioned into a fixed number of q time periods (usually hours but not necessarily equal). For each time period j , $j = 1, 2, \dots, q$ and route, the max load is given. The number of vehicle departures, F_{jk} , for a given terminal k and for each time period j must lie within a bounded range, the two standards of the lower and upper bounds of the frequency (see Figure 1.4 in Chapter 1). The lower bound is based on a predetermined policy headway, and the upper bound on a minimum passenger-loading factor. The objective is to minimize the sum of all service measures that might be

interpreted as a minimum passenger-disservice cost, or crowding, at the max load segments.

5.4.2 Formulation

Let $d(k, t)$ represent the total number of departures less the total number of trip arrivals at terminal k up to and including time t . The maximum value of $d(k, t)$ over the schedule horizon $[T_1, T_2]$ is designated $D(k)$. If we denote the set of all terminals as T , the sum of $D(k) \forall k \in T$ is equal to the minimum number of vehicles required to service the set T . This is known as the Fleet Size Formula. It was independently derived by Bartlett (1957), Gertsbach and Gurevich (1977), and Salzborn (1972). Mathematically, for a given fixed schedule:

$$N = \sum_{k \in T} D(k) = \sum_{k \in T} \max_{t \in [T_1, T_2]} d(k, t) \quad (5.2)$$

where N is the minimum number of vehicles to service the set T . Further interpretation and analysis using Equation (5.2) appear in Chapter 7.

Let T_k be the set of all departure times from terminal k . Let $(j, k_1, k_2) \equiv (\cdot)$ represent a possible trip bundle departing during period j from terminal k_1 to terminal k_2 .

Define a 0-1 variable as:

$$x^F(\cdot) = \begin{cases} 1, & \text{if } F \text{ departures are selected during period } j \text{ from terminal } k_1 \text{ to } k_2 \\ 0, & \text{otherwise} \end{cases}$$

where F is an index running from $L(\cdot)$ to $U(\cdot)$; i.e. $F = L(\cdot), L(\cdot) + 1, L(\cdot) + 2, \dots, U(\cdot) - 1, U(\cdot)$. Using the definition of Equation (3.2), let $P_m(\cdot)$ be the max load, and $d_0(\cdot)$ the desired occupancy – both definitions are associated with period j and the route from terminal k_1 to k_2 . The crowding measure (i.e. cost) associated with $x^F(\cdot)$ is defined as

$$c^F(\cdot) = \max [P_m(\cdot) - F d_0(\cdot), 0] \quad (5.3)$$

and the total crowding defining the objective function, Z :

$$Z = \sum_{\forall(\cdot)} \sum_{F=L(\cdot)}^{U(\cdot)} c^F(\cdot) \cdot x^F(\cdot)$$

The mathematical programming formulation, which is inspired from Salzborn's note (1972) on Fleet Routing Models for transportation systems, is stated below:

$$\text{minimize} \quad \sum_{\forall(j, k_1, k_2)} \sum_{F=L(j, k_1, k_2)}^{U(j, k_1, k_2)} c^F(j, k_1, k_2) \cdot x^F(j, k_1, k_2) \quad (5.4)$$

s.t.

Bundle departure constraints

$$\sum_{F=L(\cdot)}^{U(\cdot)} x^F(\cdot) = 1, \quad \forall(\cdot) \quad (5.5)$$

Assigned vehicle bounds

$$\left\{ \begin{array}{l} \text{the net number of departures less arrivals} \\ \text{that occur before or at } t \text{ at terminal } k \text{ as} \\ \text{determined by the value of } x^F(\cdot) \end{array} \right\} \leq D(k), t \in T_k, k \in T \quad (5.6)$$

where $D(k)$ is the maximum number of vehicles assigned to terminal k .

Resource constraint (total fleet size):

$$\sum_{k \in T} D(k) \leq N_o, \quad (5.7)$$

where N_o is the total (fixed and given) fleet size.

Variable constraints

$$\begin{aligned} x^F(\cdot) &= 0, 1 \quad \forall F, (\cdot) \\ D(k) &\geq 0 \quad \forall k \in T \end{aligned} \quad (5.8)$$

In the overall formulation, $D(k)$ will be an integer in any feasible solution.

Constraint (5.5) ensures that only one bundle of departures is selected for a given terminal pair (route) and time period. Constraint (5.6) ensures that the number of vehicles using a given terminal k up to time t does not exceed the number of vehicles, $D(k)$, assigned to terminal k . The left-hand side of constraint (5.6) can be represented as a linear function of the $x^F(\cdot)$ variable (as will be demonstrated in the example in the next section). Not all departure times need be considered, as some lead to redundant equations; this can occur when there is a sequence of departure times unbroken by intervening arrivals. Constraint (5.7) indicates that the sum of vehicles assigned to all terminals is no greater than the given fleet size N_o . Finally, constraint (5.8) defines $x^F(\cdot)$ as a binary, and the boundary of $D(k)$.

The optimal solution $\{x_o^F(\cdot), D_o(k)\}$ yields both the assignment of vehicles to terminals, $D_o(k)$ and the optimal number of departures within each time period. Thus an optimal timetable may be constructed from $x_o^F(\cdot)$.

5.4.3 Example

In order to gain further understanding of the underlying structure of the mathematical formulation, an example consisting of two terminals and two time periods will now be presented. It follows Ceder and Stern (1984). The basic input data of the example problem, along with the 0–1 variables and their associated cost, $c^F(\cdot)$, are indicated in Table 5.2. For each trip bundle $(\cdot)$, the upper and lower bounds on the number of departures are shown in Table 5.2. For example: $L(1, b, a) = 2$ and $U(1, b, a) = 3$. The indicated cost for each $(\cdot)$ is calculated in accordance with Equation (5.3); that is, $\max(P_m - F' d_o, 0)$. There are three 0–1 variables for the first trip bundle $(1, a, b)$ and two variables for every other trip bundle. Altogether there are 11 variables – 9 x 's, $D(a)$, and $D(b)$.

The construction of the bounds for the constraint (5.6) is based on a determination of the arrival and departure times for each of the 9 x 's. Table 5.3 contains the information, with the headways equally spaced for each set of departures. This assumes an arrival rate like that shown in Table 5.2.

Table 5.2 Basic input data for the example

Terminals (k_1, k_2)	Travel time	Period j	Time span	No. of passengers P_m	Desired occupancy d_o	Number of departures					
						F=1		F=2		F=3	
						c^F	variable	c^F	variable	c^F	variable
(a,b)	4	1	0–4	145	65	80	x_1	15	x_2	0	x_3
		2	4–8	75	47	28	x_4	0	x_5	–	–
(b,a)	3	1	0–4	160	65	–	–	30	x_6	0	x_7
		2	4–8	70	47	23	x_8	0	x_9	–	–

Table 5.3 Departure and arrival times for each set of departures

Departure terminal										<i>a</i>										<i>b</i>									
Arrival terminal										<i>b</i>										<i>a</i>									
Variable	x_1			x_2			x_3			x_4			x_5			x_6			x_7			x_8			x_9				
Departure time	4	2	4	1½	2½	4	8	6	8	2	4	1½	2½	4	8	6	8												
Arrival time	8	6	8	5½	6½	8	12	10	12	5	7	4½	5½	7	11	9	11												

The example problem can now be constructed in terms of integer programming:

$$\text{minimize } \{Z = 80x_1 + 15x_2 + 28x_4 + 30x_6 + 23x_8\}$$

s.t.

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_4 + x_5 &= 1 \\ x_6 + x_7 &= 1 \\ x_8 + x_9 &= 1 \end{aligned} \tag{i}$$

$$\begin{aligned} x_1 + 2x_2 + 3x_3 &\leq D(a) \\ x_1 + 2x_2 + 3x_3 + x_5 - x_6 - 2x_7 &\leq D(a) \\ x_1 + 2x_2 + 3x_3 + x_4 + 2x_5 - 2x_6 - 3x_7 &\leq D(a) \\ 2x_6 + 3x_7 &\leq D(b) \\ -x_2 - x_3 + 2x_6 + 3x_7 + x_9 &\leq D(b) \\ -x_1 - 2x_2 - 3x_3 + 2x_6 + 3x_7 + x_8 + 2x_9 &\leq D(b) \end{aligned} \tag{ii}$$

$$D(a) + D(b) \leq N_o \tag{iii}$$

$$\begin{aligned} x_i &= 0, 1; i = 1, 2, \dots, 9 \\ D(a), D(b) &\geq 0 \end{aligned} \tag{iv}$$

The constraints in (i) to (iv) are based on Equations (5.5) to (5.6), respectively, and on the information given in Tables 5.2 and 5.3. In constraints (ii), each possible combination of the net number of departures for a given terminal is restricted so as not to exceed the number of vehicles assigned to that terminal. For example, the first constraint in (ii) refers to $0 \leq t \leq 4$, and the second-to $0 \leq t \leq 6$, in regard to the net number of departures in terminal a . For $0 \leq t \leq 6$ in terminal a , we take three possible departures of x_3 , two departures of x_2 , and one departure of x_1 and x_5 , as opposed to two arrivals of x_7 and one arrival of x_6 .

The known OR package MPSX has been used to solve this simple example for $N_0 = 7, 6, 5, 4, 3, 2$. It is possible to obtain the solutions by relaxing the integrality constraint on the x 's and, if necessary, round off any fractions to the nearest integer. The results are presented in Table 5.4. Note that the right-hand side on N_0 may be used to directly obtain a curve of fleet size versus minimum cost in a single computer run. This trade-off is shown as the solid line in Figure 5.5. In the next section, the problem is resolved by allowing small departure times between vehicles.

Table 5.4 Optimal results for different fleet sizes

Fleet size no.	Sets of departures in solutions $x_i=1$	D(a)	D(b)	Minimum cost Z
7	x_3, x_5, x_7, x_9	3	3	0
6	x_3, x_5, x_7, x_9	3	3	0
5	x_2, x_5, x_7, x_9	2	3	15
4	x_2, x_5, x_6, x_9	2	2	45
3	x_1, x_5, x_6, x_8	1	2	133
2	Infeasible solution			–

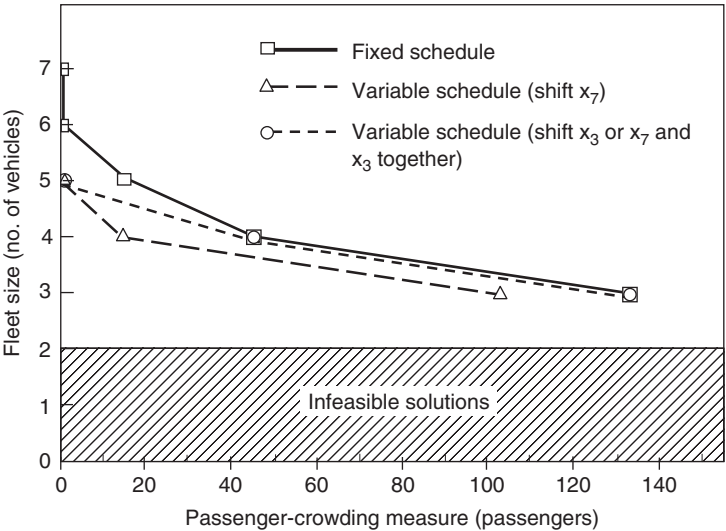


Figure 5.5 Optimal results of the example problem for a minimum passenger-crowding timetable

5.4.4 Variable scheduling

Practical vehicle scheduling often involves shifting departure times in order to better match vehicle assignment with a given set of trips. This practice is mentioned in Chapter 1 and described in Chapter 8. In this case, the departure times are allowed to vary over pre-specified limits.

A pre-computation analysis can be carried out of the bounds on the number of departures. The steps of this analysis are shown in flowchart form in Figure 5.6. The functions $d(k, t)$ are first constructed for the minimum and maximum number of departures, $L(\cdot)$ and $U(\cdot)$, for

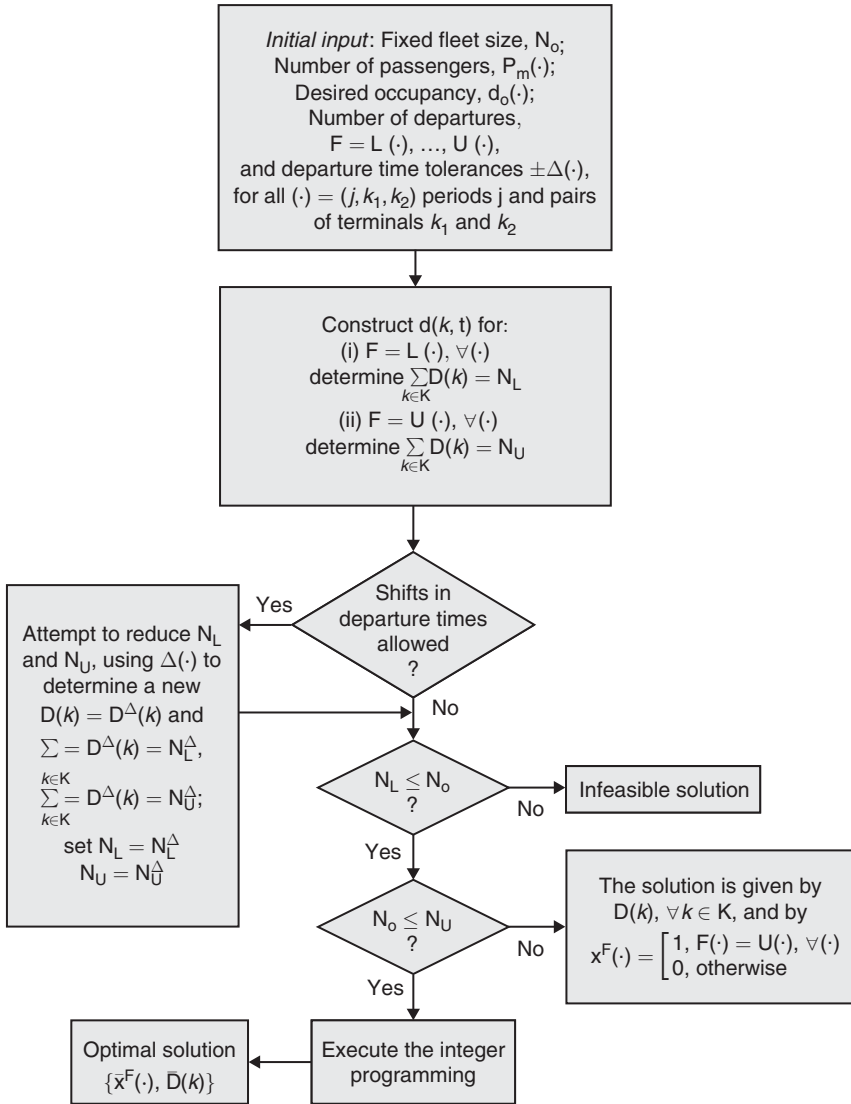


Figure 5.6 A pre-computation analysis

all the trip bundles $(\cdot)$. The lower and upper bounds on the fleet size are then determined as N_L and N_U , respectively. Before executing the integer program, two tests should be performed: (1) feasibility, and (2) appropriateness of the $U(\cdot)$ solution. These two tests are shown in Figure 5.6. Applying this pre-computation stage (without shifting departure times) to the example problem obviates the need to execute the integer programming for three fleet size values $N_o = 7, 6$ and 2 , since $N_L = 3$ and $N_U = 6$.

Assume that the departure-time tolerance for the example problem, described in Tables 5.2 and 5.3, is $\Delta(\cdot) = \pm\frac{1}{3}$ time units for all trip bundles. The analysis of $d(k,t)$, carried out in Chapter 8, reveals that N_U can be reduced by one by varying individual departure times, whereas N_L remains unchanged. The value of N_U can be reduced from 6 to 5 through three alternative shift procedures:

- (I) decrease the first departure time of x_7 , from $1\frac{1}{3}$ to 1.0 ;
- (II) increase the third departure time of x_3 from 4.0 to $4\frac{1}{3}$;
- (III) shift both departure times in (I) and (II) in opposite directions by $\frac{1}{6}$ (i.e. from $1\frac{1}{3}$ to $1\frac{1}{6}$ and from 4.0 to $4\frac{1}{6}$).

Note that in all three cases, it is possible to maintain even headways by shifting all departures in each bundle by the same amount, but this is not true in general.

The integer programming for these possibilities is similar to that explained in Section 5.4.3, except for the following:

For case (I), the first constraint in (ii) is changed into:

$$3x_3 + 2x_2 + x_1 - x_7 \leq D(a)$$

For cases (II) and (III), the first and last constraints in (ii) are changed into:

$$\begin{aligned} 2x_3 + 2x_2 + x_1 &\leq D(a) \\ 3x_7 + 2x_6 - 2x_3 - 2x_2 - x_1 + x_8 + 2x_9 &\leq D(b) \end{aligned}$$

and an additional constraint is added to (ii):

$$3x_3 + 2x_2 + x_1 - x_7 \leq D(a).$$

The results of the modified problem are presented in Table 5.5.

Table 5.5 Optimal results for the example variable scheduling problem

Case	Fleet size no.	Sets of departures in the solution, $x_i = 1$	D(a)	D(b)	Minimum cost
(I)	≥ 5	x_3, x_5, x_7, x_9	2	3	0
	4	x_2, x_5, x_7, x_9	1	3	15
	3	x_1, x_5, x_7, x_8	0	3	103
(II) and	≥ 5	x_3, x_5, x_7, x_9	2	3	0
	4	x_2, x_5, x_6, x_9	2	2	45
(III)	3	x_1, x_5, x_6, x_8	1	2	133

The pre-computation analysis shown in Figure 5.6 has been applied to this modified problem. Consequently, the MPSX package was used (while relaxing the integrality constraints) only for $N_0 = 4,3$ in all shifting cases. The optimal results exhibited in Tables 5.4 (without shifting) and 5.5 (after shifting) are compared in Figure 5.5. This figure demonstrates the trade-off between passenger-crowding (disservice) cost and fleet size. Such a graphical representation can be used as an evaluation tool by the transit planner/scheduler.

Exercise

Ride-check data, shown in the table below, were collected on a given bus route (average across several days). There are two stops (B and C) between the first (route departure) stop (A) and the last (route arrival) stop (D). The data cover a two-hour morning period (6:00–8:00). Average travel times between stops are as follows: eight minutes (A to B), 14 minutes (B to C) and 22 minutes (C to D). Minimum frequency is 2 veh/hr for both hours.

Using Method 2 for frequency determination, *find* and *compare* the departure times obtained by: (a) the existing (given) data, (b) the even-headway procedure, (c) the even-load procedure at the hourly max load point, and (d) the even-load procedure at the individual vehicle max load point. In this comparison, *provide* the average max load and its associated max load point for each departure and procedure used. Note: only one departure having a desired occupancy of 50 passengers will be derived after 8:00.

Average load (passengers) when departing stop:				
Departure time (at A)	A	B	C	Desired occupancy (passengers)
6:00	25	18	12	40
6:35	65	72	48	
7:10	67	82	35	60
7:45	84	33	47	
8:00	75	41	57	50

References

- Ahuja, R. K., Magnanti, T. L. and Orlin, J. B. (1993). *Network Flows*. Prentice Hall.
- Bartlett, T. E. (1957). An algorithm for the minimum number of transport units to maintain a fixed schedule. *Naval Research Logistics Quarterly*, **4**, 139–149.
- Ceder, A. (1999). *Systems Analysis as an Introduction to Operations Research*. Michlol Publication – Technion (Israel).

- Ceder, A. (2001). Bus timetables with even passenger loads as opposed to even headways. *Transportation Research Record*, **1760**, 28–33.
- Ceder, A. and Stern, H. I. (1984). Optimal transit timetables for a fixed vehicle fleet. In *Transportation and Traffic Theory* (J. Volmuller and R. Hammerslag, eds) pp. 331–355, UNU Science Press.
- Gertsbach, I. and Gurevich, Y. (1977). Constructing an optimal fleet for a transportation schedule. *Transportation Science*, **11**, 20–36.
- Salzborn, F. J. M. (1972). Optimum bus scheduling. *Transportation Science*, **6**, 137–148.

6

Advanced Timetables II: Maximum Synchronization



Chapter 6 Advanced Timetables II: Maximum Synchronization

Chapter outline

- 6.1 Introduction
 - 6.2 Formulating an OR model for synchronization
 - 6.3 The Synchro-1 Procedure
 - 6.4 The Synchro-2 Procedure
 - 6.5 Examples
 - 6.6 Literature review and further reading
 - Exercises
 - References
-

Practitioner's Corner

The second part of constructing advanced timetables continues with the objective, as in Chapter 5, to improve and optimize timetables through the use of both mathematical formulations and step-wise procedures. Practitioners can study the procedures, although it necessitates some mathematical notation.

This chapter contains four main parts. First, I address and formulate the problem of generating headway-based timetables, in which the number of simultaneous vehicle arrivals at connection (transfer) points is maximized. This is a maximum synchronization problem in time and space that is meant to enable the transfer of passengers from one route to another with minimum waiting time at the transfer points. Second, a heuristic procedure (i.e. a procedure or algorithm that does not result for sure in an optimum solution) is proffered, in which the setting of departure times follows efficient rules. Third, I improve the results (i.e. increasing the number of meetings between vehicles on different routes) by allowing a shift in departure time within given boundaries. This improvement uses defined efficient rules imbedded in the second heuristic procedure. Fourth, I provide detailed examples to illustrate the two proposed procedures explicitly. These examples follow the two procedures step-by-step and can help practitioners gain a better grasp of the analysis and results. Overall, this chapter asserts that, basically, a non-coordinated transit network is simply the unfolding of a miscalculation.

After reading Section 6.1 (scope), practitioners are encouraged to read Sections 6.5 (examples) and to glance at Sections 6.3 and 6.4 (description of the two procedures). It is also useful to solve the exercises at the end of the chapter by means of the procedures described.

As is customary in Practitioner's Corner, following is a brief story that may initially be seen as a transit misconnection case: A bus starts to move downhill, when a passenger in the back notices someone running extremely fast after the bus. "Stop running, you won't catch it," he yells to the runner, who shouts back: "I'd better, I'm the driver".

The chapter ends with a literature review on transit coordination and synchronization, followed by exercises. We hope that the planned-synchronization ideas presented will resolve another passenger's grievance: "Look at all the buses that now want exact change. I figure if I give them exact change, they should transfer me from one bus to another without my having to wait".

6.1 Introduction

This chapter focuses on the problem of how to construct a timetable with a maximum number of simultaneous vehicle arrivals at connection (transfer) points. From the user perspective, a global approach to establishing these timetables usually considers the minimization of travel and waiting (and possibly walking) times. This, then, becomes a transit-network scheduling problem, which utilizes O-D (Origin-Destination) data. Assume, as is done in practice, that there is an existing transit network of routes with a certain passenger demand according to the time of day. Maximum synchronization is a rather important objective from both the transit-network's and the users' perspectives. Ceder and Wilson (1986), in a study of transit-route design at the network level, emphasized the importance of eliminating a large number of transfer points because of their adverse effect on the user. From the operating-cost perspective, a trade-off no doubt exists between this elimination and the efficiency of the transit-route network. This trade-off is depicted schematically in Figure 6.1.

In order to allow for an adequate or compromised level-of-service, the planners/schedulers face a synchronization task to ensure maximum smooth transfers involving the switching of passengers from one route to another without waiting time. This task is extensive, but minimizes the waiting time for passengers who require connections. The result is that the planner/scheduler creates a more attractive transit system, one that generates the opportunity for increasing the number of riders. Figure 6.1 illustrates how the concept of synchronized transfers mitigates any adverse impact on the level-of-service while achieving a more cost-effective transit network.

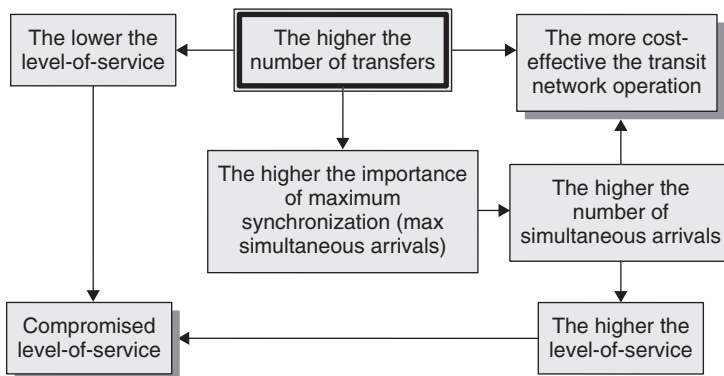


Figure 6.1 Trade-offs dependent on transit-passenger transfers

Actually, synchronization is one of the most difficult but vital scheduling tasks. A poor transfer can cause the user to stop using the transit service. It is analogous to the Chinese proverb: “Two leaps [in our case, two trips with a transfer] per chasm are fatal”. Nevertheless, synchronization is addressed almost intuitively. Generally speaking, the planner/scheduler attempts to fix the departure times in the timetable while conforming to three elements: (i) required frequency, (ii) efficient assignment of trips to a single vehicle chain, and (iii) maximum synchronization of arrivals. The second element will be dealt with in Chapters 7 and 8. This chapter will present an efficient mathematical procedure to help achieve maximum synchronization; the procedure can be employed as a useful tool for creating timetables. The process described in this chapter follows Ceder *et al.* (2001) and Ceder and Tal (2001).

6.2 Formulating an OR model for synchronization

Chapters 4 and 5 established optional timetables in conjunction with adequate frequency and satisfactory loads for a single transit route. When two or more routes interact, particularly at connection (transfer) points, the issue of coordination becomes critical. In order to gain a better approach to coordination at connection points, there is need to relax some of the rigorously defined timetable parameters. The parameter that can offer a measure of flexibility is the headway. Fortunately, only certain routes, at certain time periods, are able to be included in this optimal search for reducing waiting time at transfer points. This section follows the meaning of optimization as a framework for mathematical treatment, as described in Chapter 5, and formulates the problem of maximizing simultaneous vehicle arrivals at transfer points.

The following OR model, based on Ceder *et al.* (2001), is presented in five parts: (i) Notation of known (given) data, (ii) Decision variables, (iii) Objective function, (iv) Constraints, and (v) Assumptions and notes.

Notation of known (given) data

The given transit network is presented by $G = \{A, \bar{N}\}$,
where:

A = a set of directed arcs representing the travelling path of transit routes

$\bar{N}$ = a set of transfer nodes in network G .

The problem data are the following:

T = planning horizon (departure times can be set in the interval $[0, T]$, which is a discrete interval)

M = number of transit routes in the network

N = number of transfer nodes in the network

$H_{\min_k}$ = minimum headway (agency's requirements) between two adjacent departures on route k ($1 \leq k \leq M$)

$H_{\max_k}$ = maximum headway (policy headway) permitted between two adjacent bus departures on route k ($1 \leq k \leq M$)

F_k = number of departures to be scheduled for route k during the interval $[0, T]$ ($1 \leq k \leq M$)

T_{kj} = travel time from the starting point of route k to node j ($1 \leq k \leq M, 1 \leq j \leq N$); travel times are considered to be deterministic and can be referred to as mean travel times.

Decision variables

- (a) X_{ik} represents the i -th departure time on route k ($1 \leq i \leq F_k$);
 (b) Z_{ikjqn} is a binary variable that yields the value 1 if the vehicle of the i -th departure on route k meets the vehicle of the j -th departure of route q at node n ; otherwise, it yields the value 0;

Let $A_{kq} = \{n: 1 \leq n \leq N, T_{kn} \geq 0, T_{qn} \geq 0\}$.

Objective function

$$\text{Max} \sum_{k=1}^{M-1} \sum_{i=1}^{F_k} \sum_{q=k+1}^M \sum_{j=1}^{F_q} Z_{ikjqn} \quad (6.1)$$

Constraints

$$X_{1k} \leq H\max_k, \quad 1 \leq k \leq M \quad (6.2)$$

$$X_{F_k k} \leq T, \quad 1 \leq k \leq M \quad (6.3)$$

$$H\min_k \leq X_{(i+1)k} - X_{ik} \leq H\max_k, \quad 1 \leq k \leq M, 1 \leq i \leq F_k - 1 \quad (6.4)$$

$$Z_{ikjqn} = \text{Max}[1 - |(X_{ik} + T_{kn}) - (X_{jq} - T_{qn})|, 0] \quad (6.5)$$

Constraint (6.2) ensures that the first departure time will not begin beyond the maximum headway from the start of the time horizon, while constraint (6.3) ensures that the last departure is executed within the planning horizon. Constraint (6.4) indicates the headway limits and constraint (6.5) defines the binary variable of the objective function.

Assumptions and notes

- (a) It is assumed that the first departure on each route k must take place in the interval $[0, H\max_k]$;
 (b) The problem is impractical unless the following constraints hold for each k :

$$1. H\max_k \geq H\min_k \quad (6.6)$$

$$2. T \geq (F_k - 1) \cdot H\min_k \quad (6.7)$$

$$3. T \leq F_k \cdot H\max_k \quad (6.8)$$

- (c) The case of a route k that does not pass through a node j is represented by $T_{kj} = -1$.

The above OR model can be simplified by defining a variable, Y_{kq} , representing the overall number of simultaneous arrivals of vehicles on route k with vehicles on route q . The model is changed to

$$\text{Max} \sum_{k=1}^{M-1} \sum_{q=k+1}^M Y_{kq} \quad (6.9)$$

s.t.

$$Y_{kq} = \sum_{n \in A_{kq}} \sum_{i=1}^{F_k} \sum_{j=1}^{F_q} \text{Max}[1 - |(X_{ik} + T_{kn}) - (X_{jq} + T_{qn})|, 0] \quad (6.10)$$

Constraints (6.2) to (6.4) remain unchanged.

The last formulation represents a nonlinear programming problem. It can be reformulated as a mixed integer linear programming (MIP) problem, which can be solved (up to certain sizes), by several software packages. The nonlinear constraint is (6.10). Let D_{nijkq} denote a binary variable (defined over the same domain as Z_{ikjqn}), and B a large number ($B = T + \text{Max}_{i,j} T_{ij}$). The constraint in (6.10) is exchanged with the following constraints:

$$B \cdot D_{nijkq} \geq X_{ik} + T_{kn} - (X_{jq} + T_{qn}) \quad (6.11)$$

$$B \cdot D_{nijkq} \geq X_{jq} + T_{qn} - (X_{ik} + T_{kn}) \quad (6.12)$$

$$Y_{kq} < \sum_{n \in A_{kq}} \sum_{i=1}^{F_k} \sum_{j=1}^{F_q} (1 - D_{nijkq}) \quad (6.13)$$

If $X_{ik} + T_{kn} = X_{jq} + T_{qn}$, there is a simultaneous arrival of the vehicle of the i -th departure on route k with the vehicle of the j -th departure of route q at node n . The variable D_{nijkq} can yield the value 0, and Y_{kq} is increased by one, according to (6.13).

If $X_{ik} + T_{kn} \neq X_{jq} + T_{qn}$, the arrivals do not coincide, and D_{nijkq} must yield the value in order to satisfy constraints (6.11) and (6.12). The number of simultaneous arrivals between vehicles of routes k and q (Y_{kq}) is not increased in (6.13).

The following is an upper bound on the number of possible simultaneous arrivals in a given transit network:

$$Z^* = \sum_{k=1}^{M-1} \sum_{q=k+1}^M \sum_{n \in A_{kq}} \text{minimum}(F_k, F_q) \quad (6.14)$$

The number of integer variables in an MIP problem is generally a good index of its complexity (represented by the computerized processing time required). The variable D_{nijkq} represents the simultaneous arrival of the vehicle of the i -th departure of route k and the vehicle of the j -th departure of route q at node n . This means an integer variable for every combination of two trips on different routes that intersect at node n . Let F be $\text{Max}(F_k)$; the number of integer variables in the worst case is $O(NM^2F^2)$, which is a very large number. However, in a more realistic setting, the number is $O(M^2F^2)$, where N can be replaced by the average number of nodes commonly shared by any two routes.

The problem formulated by (6.9) to (6.13) is a large, mixed-integer, linear-programming problem. Running small network examples (five routes, five nodes), using General Algebraic Modeling System (GAMS) software on a PC, requires hours, even days. This time-consuming process provided the motivation to develop heuristic procedures that would solve such problems within a reasonable time. Two heuristic procedures were implemented in Turbo-Pascal, and many examples were checked and compared to the optimal solutions obtained by using the GAMS. The first procedure is based on the selection of nodes within the network. In each step, the next node is selected, provided that not all the departure times have been determined for that node. After the departure time is resolved, all its corresponding arrival times are set. The second procedure attempts to improve the solution through possible shifts in departure times. The following two sections present these two heuristic procedures.

6.3 The Synchro-1 Procedure

Definition 1: A node is defined as ‘possible’ if:

- (a) at least one transit route that passes through the node, and not all the departure times for that route are set;
- (b) more simultaneous arrivals can be created at the node.

Definition 2: A node is defined as ‘new’ if no arrival times have been set for it.

The Synchro-1 Procedure uses several components, as shown in the flowchart in Figure 6.2. Its steps are as follows.

Step 1: Initialization; check whether the problem is feasible and set the data structure. Mark all nodes as possible.

Step 2: Select the next node, NO, from the possible nodes.

Step 3: If NO is new, perform component FIRST, otherwise perform component MIDDLE.

Step 4: If there is any possible node, go to *Step 2*; if there are any more routes, perform component CHOOSE and go to *Step 2*, otherwise stop.

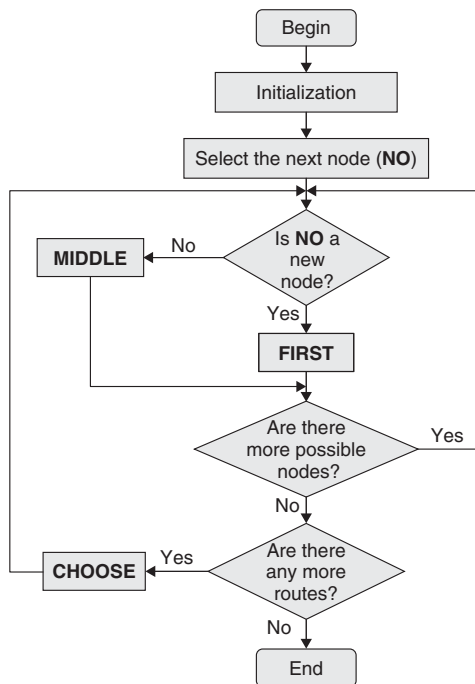


Figure 6.2 Synchro-1 Procedure flowchart

Step 1 contains a check of whether the problem is feasible; if so, towards that end, two data structures are built:

- (a) A structure called *route* for each transit route i , which includes $H_{min,i}$, $H_{max,i}$, F_i , the number of nodes the route passes through, and the departure times that have already been set.

- (b) A structure called *node* for each node n , which includes two elements: (i) the number of routes passing through the node, and (ii) the route with the maximum travelling time to the node. It also includes the number of simultaneous arrivals at the node at each timepoint in interval $[0, T + \text{Max}T_{ij}]$, where the max is over all i, j .

All the nodes are marked possible.

In Step 2, the next node, NO, is selected from among the possible nodes. Node NO must satisfy the following conditions:

- (a) The number of different vehicle-arrival times at the node is at its maximum; in such a node, there is a greater probability that another vehicle departure can be set so that it will arrive at NO by any one of the (already set) arrival times.
- (b) Among the nodes satisfying the preceding condition, NO is that through which a maximum number of routes pass; in such a node, there is a greater potential for simultaneous arrivals.
- (c) Among the nodes satisfying the preceding condition, NO minimizes the maximum travel time of all routes from their origin to the node (after the departure times of vehicles are set in order to meet at NO, the potential for simultaneous arrivals at distant nodes still exists).

A distinction is made in Step 3 between a new node and another node. For a new node, the component FIRST attempts to set the departure times of vehicles that pass through it, such that the vehicles will arrive at the node at the earliest time possible and simultaneous arrivals will continue to be created at the node according to the $H_{\min}$, $H_{\max}$ of the routes. For example, transit vehicles on routes 2 and 3 arrive simultaneously at a certain node at time t_0 ; if the next departure time for each of these routes can be set at a fixed difference, d minutes, from the last departure time of the route [for all $i = 1, 2$ and 3 , $\max(H_{\min_i}) \leq d \leq \text{minimum}(H_{\max_i})$], there will be additional simultaneous arrivals at the node at time $t_0 + d$. Component FIRST finds the minimum d possible. If parameter d cannot be set ($H_{\min_i} > H_{\max_j}$), the next departure times of vehicles on these routes will not be resolved in this step.

For a node that is not 'new', an attempt is made to set the vehicle-departure times on routes passing through it, such that the vehicles will arrive at the node the earliest of all arrival times already set in that node. If no more simultaneous arrivals are available at the node, the node is marked as 'not possible'. This component is called MIDDLE.

Step 4 tests whether any more nodes are possible. If not, there may be routes on which not all the vehicle departure times were set. In such cases, the route that passes through the maximum number of nodes is chosen, and its next departure time is set by using the difference $H_{\min_i}$ from the last departure. In this way, the procedure sets additional vehicle arrivals for the maximum number of nodes possible. All the nodes through which route i passes are marked possible, and the procedure returns to Step 2. This component is called CHOOSE.

The complexity of the Synchro-1 Procedure is, in the worst case, $O(NTFM^2)$, where N is the number of transfer nodes, M is the number of routes, T is the planning horizon, and F is the maximum number of scheduled departure times across all routes.

6.4 The Synchro-2 Procedure

The results of Synchro-1 could be improved by allowing a shifting of departure times in the timetable obtained by the procedure. That is, for each node and for each two timepoints, t_1 and

t_2 ($t_1 < t_2$), at which vehicles arrive at the node, an attempt can be made to introduce a shift in all the vehicle departure times for vehicles arriving at the node at t_1 so that they will arrive at time t_2 . If this succeeds, the timetable is changed accordingly, and the number of simultaneous arrivals increases. It should be noted that in order to shift a single departure time, the following must be checked:

- (a) After the shifting is done, the constraints on H_{min} and H_{max} must still hold. Otherwise, additional departure times on the route should be shifted.
- (b) As a result of shifting the departure time of vehicle i , its arrival time for all nodes through which it passes is changed. Therefore, the departure time of each vehicle that arrives simultaneously with vehicle i at any node must also be shifted. For each shift, the constraints on H_{min} and H_{max} must be checked, and so on. This process is recursive, and an attempt to shift a single departure time may cause the shifting of all network departure times.
- (c) A shift of Δt minutes in the departure time of vehicle i may result in changing other departure times, which ultimately will lead to changing the departure time of vehicle i by more than Δt . A check must eliminate this situation.

The shifting procedure can be added to the Synchro-1 Procedure with two components, MERGE and MOVE, and a process TRY-MOVE, as shown in Figure 6.3.

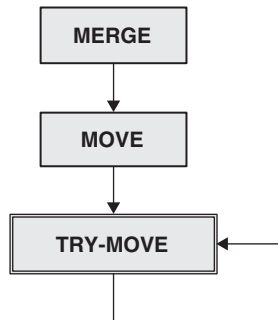


Figure 6.3 Synchro-2 Procedure flowchart

Component MERGE

Component MERGE identifies two departure times, t_1 and t_2 ($t_1 < t_2$), for a given node NO and delivers three elements to component MOVE: the transit routes arriving at t_1 and t_2 , and the required shift $t_2 - t_1$.

Component MOVE

The MOVE component attempts to shift t_1 by $\Delta t = t_2 - t_1$, where t_2 is changed by route number. This component also contains the number of vehicles that need to be shifted, and a vehicle array with all possible vehicles that need to be checked for shifting. Each vehicle is identified by route number and vehicle code. Component MOVE uses the process TRY-MOVE, which allows for individual shifts to be checked.

TRY-MOVE is a recursive process that attempts to shift the departure time of a given vehicle and indicates whether MOVE is 'successful' or 'false'. This TRY-MOVE process

checks whether the required new departure time is within $[O,T]$ and whether or not the resultant headway exceeds H_{max} .

Synchro-1 and Synchro-2 combined

Synchro-1 and Synchro-2 can be combined to shift Synchro-1 results. Further, the first selected node NO can change N times: $NO = 1, 2, \dots, N$; that is, in order to run Synchro-1 N times. Such N time-runs may be required when the criterion of Synchro-1 for selecting NO cannot be justified; e.g. in a case in which the node with the maximum routes passing through it is far from the origin. These N runs over different NOs can also be combined with Synchro-2. Example 3 below examines four variations: (i) Synchro-1, (ii) N runs of Synchro-1 (first node varies), (iii) Synchro-2, and (iv) Shifts in the results of N runs of Synchro-1.

The complexity (running time as a function of network size parameters in the worst case) of each variation is certainly not the same. Synchro-1 has the lowest complexity, $O(NTFM^2)$, followed by N runs of Synchro-1, $O(N^2TFM^2)$, which is less than the complexity of Synchro-2, $O(N^2T^3F^2M^2)$. The most complex run is the one with N changes in the first node while running Synchro-2, $O(N^3T^3F^2M^2)$.

6.5 Examples

Three examples of transit networks will now be described, following Ceder *et al.* (2001) and Ceder and Tal (2001). Two of the examples are presented in detail for the sake of clarity. An optimal procedure with the Linear Programming (LP) software GAMS was applied to each example for comparison purposes.

Detailed example 1

Figure 6.4 presents a simple network that combines two transfer points with two routes. The numbers on the arcs are average planned travel times (in minutes). A demonstration of Synchro-1 follows.

Step 1: Building the route data structure:

i	Hmin _i	Hmax _i	F _i	No. of nodes
I	5	15	4	2
II	8	20	3	2

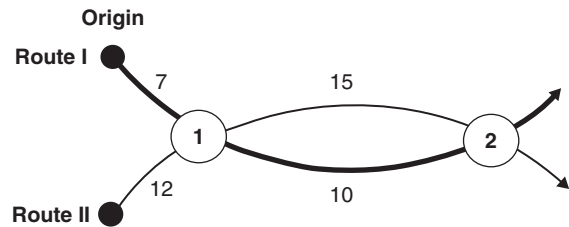


Figure 6.4 Example 1, basic network

and the *node* data structure:

Node n	No. of routes	Route with Max T_{in}
1	2	II
2	2	II

The data comply with the three constraints (6.6), (6.7) and (6.8).

Step 2: Three criteria must be selected: NO with maximum arrival time, maximum routes crossing, and minimum of all routes' maximum travel time (min-max criterion) from origin to NO. Thus, the number of departure (arrival) times equals 0 for both nodes. The number of crossing routes equals 2 for both nodes. The maximum travel time for node 2 is $\max(17, 27) = 27$; for node 1, it is $\max(7, 12) = 12$; the minimum of the two is 12. The selected NO is, therefore, node 1.

Step 3: The earliest time possible is set for node 1. Continuing in this node, the synchronization is based on $\max(5, 8) \leq d \leq \text{minimum}(15, 20) \rightarrow d_{\min} = 8$. This is the FIRST component, which provides the following results:

Departure time		
Route I	Route II	Meeting time
5	0	12
13	8	20
21	16	28

where meeting refers to a simultaneous arrival. The number of departures for route II complies with $F_2 = 3$.

Step 4: Because node 2 has yet to be examined, Step 2 is selected.

Step 3: Because node 2 is not new, component MIDDLE is applied. $F_1 = 4 > 3$ (currently created). Thus, one more departure time is set for route I based on the already created departure times for route II (at 0, 8, 16), so that the synchronization is made at node 2. The result is the additional departure time for route I (at 26, which meets the route II vehicle's departure at 16).

Step 4: No more routes are left to examine and the algorithm ends with the final results, as follow:

Departure time		Meeting time		Total meetings
Route I	Route II	At node 1	At node 2	
5	0	12		
13	8	20		4
21	16	28		
26			43	

The optimal and heuristic procedures coincide in this simple example.

Detailed example 2

Figure 6.5 presents the second example with a case of $Hmin_i > Hmax_j$ for a two-route, four-node network. The numbers on the arcs are average planned travel times (in minutes). Both Synchro-1 and Synchro-2 may be used for this example.

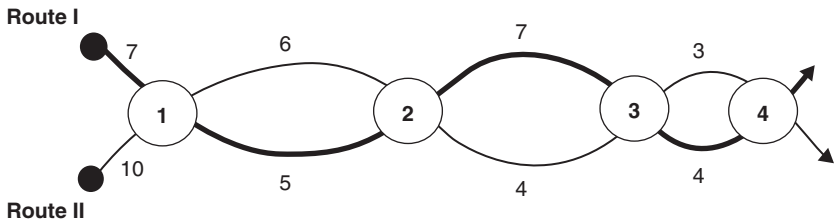


Figure 6.5 Example 2 basic network

The data for Example 2 are as follow:

i	Hmin _i	Hmax _i	F _i	T
Route I	6	10	4	30
Route II	3	5	6	

in which $Hmin_1 > Hmax_2$.

Using Synchro-1

Step 1: In the route structure data, the number of nodes through which each route passes is four.

Node data structure:

Node n	No. of routes	Route with Max T _{in}
1	2	II
2	2	II
3	2	II
4	2	I, II

Step 2: NO is selected such that: (i) the number of departure (arrival) times equals 0 for all nodes, (ii) the number of crossing routes equals 2 for all nodes, and (iii) the route's maximum travel times for the nodes are 10, 16, 20 and 23, respectively.

The minimum is 10, so $NO = 1$.

Step 3: Component FIRST is applied. The first meeting time possible at node 1 is 10. The component cannot set the parameter d . Therefore, procedure FIRST results in only the first departure time for each route. That is, departure times 3 and 0 for routes I and II, respectively, to meet at 10.

Step 4: With more possible nodes, Synchro-1 returns to Step 2.

Step 2: Number of arrival times is 1, 2, 2 and 2, respectively, for nodes 1, 2, 3 and 4. $NO = 2$ (min-max travel time is 16).

Step 3: Component MIDDLE sets a new meeting at node 2, at time 27, by setting the third departure time for route I to 15, and the fourth departure time for route II to 11.

Steps 2, 3: $NO = 3$. Component MIDDLE sets new meetings at node 3, at times 28 and 34, by setting the second departure time for route I to 9, the third departure time for route II to 8, and the fifth departure for route II to 14.

Steps 2, 3: $NO = 4$. No meetings are possible.

Synchro-1 continues to perform component MIDDLE until it ends with the following results:

Departure time		Meeting time at node 1	Meeting time at node 2	Meeting time at node 3	Total meetings
Route I	Route II				
3	0	10			
9	3			28	
15	8		27	34	5
21	11	28			
	14				
	18				

The meeting time is shown in the same row as its associated departure for route I.

Using Synchro-2

Component MERGE and MOVE check with the possible shifting of TRY-MOVE. The only successful shifting is the second departure for route II from 3 to 5. The new results are as follows:

Departure time		Meeting time at node 1	Meeting time at node 2	Meeting time at node 3	Total meetings
Route I	Route II				
3	0	10			
9	5		21	28	
15	8		27	34	6

(Continued)

Table (Continued)

Departure time		Meeting time at node 1	Meeting time at node 2	Meeting time at node 3	Total meetings
Route I	Route II				
21	11	28			
	14				
	18				

The optimal solution is 6 meetings, coinciding with the solution given by Synchro-2.

Example 3

Figure 6.6 exhibits the third example, in which the numbers on the arcs are average planned travel times (in minutes).

The following data apply for Figure 6.6:

i	Hmin _i	Hmax _i	F _i	T
Route I	3	8	20	
Route II	4	7	Combinations	20
Route III	5	12	(See Table 6.1)	

This third example was examined with four variations of procedures: Synchro-1, Synchro-1 with N changes of first node, Synchro-2, and Synchro-2 with N changes of first node. The results of these four variations appear in Table 6.1, including a comparison of the optimal results. Table 6.1 shows 20 combinations of frequencies F_1 , F_2 and F_3 for routes I, II and III, respectively.

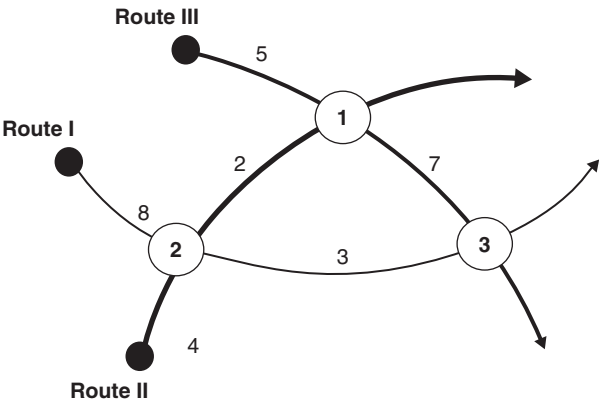


Figure 6.6 Example 3 basic network

Table 6.1 Maximum number of meetings produced by four variations of Synchro-1 (S1) and Synchro-2 (S2) with 20 combinations of vehicle frequencies

Combination no.	F_1	F_2	F_3	S1	Procedure variation			Optimal
					S1 + N runs	S2	S2 + N runs (with shifts)	
1	1	1	1	2	2	2	2	2
2	1	1	2	2	2	2	2	3
3	1	2	1	3	3	3	3	3
4	1	2	2	3	3	4	4	4
5	2	1	1	2	3	2	3	3
6	2	1	2	2	3	3	4	4
7	2	2	1	3	3	3	4	4
8	2	2	2	4	4	5	5	5
9	3	1	1	2	3	2	3	3
10	3	1	2	3	3	3	4	4
11	3	2	1	3	3	3	4	4
12	3	2	2	4	5	5	6	6
13	3	3	1	4	4	4	5	5
14	3	3	2	5	5	6	6	7
15	3	1	3	3	4	3	4	4
16	3	2	3	4	5	6	7	7
17	3	4	4	7	7	10	10	10
18	4	4	4	8	8	11	11	11
19	5	5	5	8	8	8	8	8
20	6	6	5	4	7	4	7	7

The results of Table 6.1 are as follows: (1) for the most complex variation, with Synchro-2 and N changes at first node (of Synchro-1), there are 18 (out of 20) optimal results; (2) Synchro-1 and Synchro-1 with N runs result in 13 identical results (i.e. the decision about the first node in Synchro-1 was the best selection in 13 of the 20 cases); (3) based on example 3, it is not obvious that Synchro-2 provides better results than Synchro-1 + N runs.

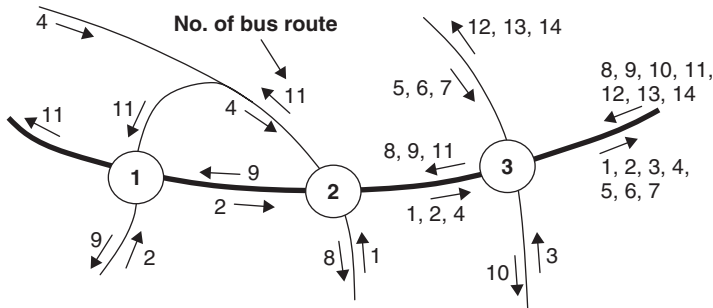


Figure 6.7 A real-life example of bus routes that need synchronization at three points

Real-life example

The study by Ceder *et al.* (2001) selected a real-life synchronization problem in Israel for testing Synchro-1 and Synchro-2. The real-life network, shown in Figure 6.7, consists of a main road that has three major transfer nodes. There are seven bus routes travelling in each direction on this network, and a total of 14 routes that meet at three nodes. (Route numbers appear in Figure 6.7.) The time span for the synchronization problem was three hours during morning rush hour. The best results (using all four variations) were 240 meetings using Synchro-1 with N changes for the first selected node. The real-life example could not be checked against the optimal solution because of its large size.

The approach presented in this section can serve as a useful tool to assist the transit scheduler in one of the most difficult scheduling tasks. The problem of a successful connection is reflected in the following definition of a missed transfer: the same thing happened today that happened yesterday but usually to different passengers. Thus, the more successful connections that can be created, the better and more reliable will be the service. This planning tool needs to be complemented with the on-time performance of the transit vehicles. The latter is expected to be realized when using new AVL and communications technologies. All together, the procedures described in this section may offer a step forward in changing the image of transit services and, ultimately, in view of increasing traffic congestion, lead to increased ridership. The next section provides a literature review of the research conducted in this area.

6.6 Literature review and further reading

The need to transfer between routes generates a major cause of discomfort for transit users. Designing routes and schedules with a minimum amount of waiting time during a transfer may decrease the level of inconvenience. This section reviews methodologies for the design of such synchronized transit services.

Rapp and Gehner (1976) describe a graphic, interactive tool for minimizing transfer delays in a transit network. Optimal system-wide timetable coordination, under operating-cost constraints, is achieved through determining deviations from the original departure times. The diversity of headways along different routes and the interdependence of layover times

in terminals are taken into account. Seeking an optimal solution involves an iterative assignment of passenger demand on the route network.

Salzborn (1980) presents a methodology for scheduling a system that includes several urban feeder lines and one inter-urban line. The latter passes by a few terminals, and its timetable is coordinated with the arrival times of passengers in all terminals. Two separate scheduling processes are described, one for the feeder lines and one for the inter-urban line. The objective is to minimize both passenger delays and the number of required vehicles. The decision variables consist of time slots that are assigned to line departures from the terminals.

Although Nelson *et al.* (1981) do not focus on the development of a design methodology, their paper mentions four different transfer-design strategies: simple timed-transfer, pulse scheduling, line-up, and neighbourhood pulse. The simple timed-transfer strategy means adjusting timetables so that two buses will arrive at a station at the same time. The pulse-scheduling strategy requires network-wide actions that include route adjustments, headways, layover times and various control strategies. The two other strategies are similar to pulse-scheduling, but their implementation is limited either to off-peak hours (line-up strategy) or to a small geographical area (neighbourhood pulse strategy).

Kyte *et al.* (1982) present the process of building a route network in which a main trunk line passes through a series of transit centres. The process includes a determination of clock headways; i.e. headways that form the same schedule every hour, with the objective of coordinating transfer times at the transit centres.

Schneider *et al.* (1984) provide a detailed list of criteria for choosing a proper site for the location of a time-transfer transit centre.

Hall (1985) develops a model for schedule coordination at a single transit terminal between a set of feeder routes and the line that they feed. The travel time on each of the routes is assumed to be subject to random delay. The optimized variable is the slack time between feeder arrivals and the main-line departure.

Abkowitz *et al.* (1987) relate to the simple case of two bus routes that meet at one spot. Their analysis aims at determining the conditions under which timed-transfer provides an improved service level compared with unscheduled transferring. A computer program is developed that simulates four transfer strategies: (1) unscheduled transfers; (2) scheduled transfers without vehicle waiting; (3) scheduled transfers in which the lower-frequency bus is held until the higher-frequency vehicle arrives; (4) scheduled transfers, in which both buses are held until a transfer event occurs. Mathematical expressions for waiting times are developed for each of the four cases. Simulation results of the four cases are analysed to arrive at a generalization of situations for which each of the transfer strategies is suitable.

Klemm and Stemme (1988) formulate an integer-programming problem that attempts to minimize transfer times in a network, represented as a graph. The authors describe rules for the construction of a route-graph on which the solution procedure is based.

Voss (1990) discusses two cases of a schedule-synchronization problem, formulated as a programming problem. The first case is similar to the one presented by Klemm and Stemme; namely, a relaxed version of a quadratic assignment problem. The transit routes in this case meet at a given set of transfer points. The second case, a modification of the first one, represents a transit system in which different routes partly use the same streets or tracks. Solution approaches are discussed, and initial feasible solutions generated; they are then improved by using a Tabu search algorithm.

Desilets and Rousseau (1990) propose a model that is basically equivalent to that of Klemmt and Stemme, but they also include an extended discussion of the definition of the cost that the model aims at minimizing. Several alternative cost definitions are suggested, such as a definition that focuses on service reliability and a definition based on choosing a group of connections for which synchronization is especially desirable. The paper also describes a heuristic solution based on an initial random solution and on a local search technique for solution improvement.

Lee and Schonfeld (1991) introduce two models for optimizing the slack time between a train line and a bus line that interchange passengers at adjoining terminals. The models are analytical but are solved numerically or by using simulation. The first model assumes that train-arrival times are always the same, whereas bus arrivals are according to a given probability function. In the second model, both train and bus arrivals are probabilistic. Transfer-cost functions are developed, showing relationships between slack times and headways, transfer volumes, passenger-time values, bus operating costs, and the variability of bus and train arrivals. Schedule coordination between the two routes is found to be not worth attempting when the standard deviation of arrivals exceeds certain levels.

Adamski (1993) develops several transfer-optimization models. The first is a static model for a single transfer point, under the assumption of random vehicle travel time. Optimal timetable offsets are calculated with the objective of minimizing passenger disutility, which is a function of waiting time. Four alternative forms of the disutility function are examined. Cases with and without a holding control are analysed. In the second model, optimal offset times are determined in the case of different lines that use a common road segment. The third model that Adamski develops is a dynamic model for real-time transfer optimization. A measure of off-schedule deviations is used to determine the value of control variables, such as holding times at control points.

Daduna and Voss (1993) discuss the optimization of transfers, using a quadratic semi-assignment problem formulated as a programming problem. The authors focus on heuristic solution procedures rather than on a new synchronization methodology. A regret heuristic is proposed for finding an initial solution. Tabu search and simulated annealing strategies for improving the solution are described.

Adamski and Chmiel (1997) formulate two schedule-synchronization problems, which they solve by using genetic algorithms, the first for the synchronization problem of transfers at transfer points and the second for synchronization of different transit lines that travel a common road segment. Deterministic and stochastic cases are compared. Decision variables are offsets entered into timetables in order to minimize waiting times.

Clever (1997) discusses the differences between systems with a single-mode transfer centre and systems with multimodal, multi-centre transfer timing. Characteristics of both types of networks are presented and compared. Among the aspects discussed are service symmetry, relations between service and passenger demand, minimal headway and economic aspects.

Shih *et al.* (1998) present a heuristic model for the design of a coordinated network with transfer centres. The routes are not predetermined, but an output of the model. In addition, this model determines the appropriate vehicle size for each route. The proposed design process is composed of four major procedures. The first procedure generates a set of routes. In the second procedure, the network is analysed and frequencies are set; this procedure incorporates a trip-assignment model. In the third procedure, transfer-centre locations are

determined. The last procedure involves network improvements. Algorithms are described qualitatively, using flowcharts for each of the four procedures.

Becker and Spielberg (1999) present principles for designing and implementing a multi-centre-based timed-transfer network. To construct such a network, a prime headway is chosen, and all headways in the network are set to be multiples of this headway. Transfer centres are located in such a manner that the distances between them are also multiples of the prime headway. Other adjustments are made in order to improve network coordination, such as connecting routes at common ends and connecting feeder routes to the transfer centres.

Maxwell (1999) discusses four transfer strategies: (1) close headways with short waiting times, so that synchronization is not needed; (2) main line trunk with timed-transfer branch lines; (3) a single hub with spoke lines, so that all transfers occur at the same spot; (4) several transfer hubs with fixed-interval timed-transfers. Advantages and disadvantages of each strategy are discussed and compared. A methodology for using the fourth strategy is presented in detail. Transfer points on a symmetrical route are shown to be located in spots that divide the route into an integer amount of equal segments. It is also shown that other locations can be chosen if speeds are adjusted accordingly. The use of a schedule map and symmetrical train graphs as analysis tools is demonstrated.

It should be noted that several guidelines for the design or the implementation of synchronized transit systems are mentioned repeatedly in most of the papers reviewed. Some of these guidelines are as follows:

- For an efficient synchronized system, route design and schedule design should be effected simultaneously.
- The success of timed-transfer services relies heavily on the use of real-time control. Schedule adherence is crucial.
- The effort in introducing synchronized services may not be worthwhile when frequencies are relatively high. Most papers mention a 30-minute headway as an example of a headway that justifies transfer timing.

The main characteristics of the models reviewed are described in Table 6.2.

Table 6.2 *Characteristics of works covered in Section 6.6, Literature review and further reading*

Source	Network structure/ transferring strategies	Location of synchronized transfer	Special features
Adamski (1993)	1st model: several routes that meet at a terminal. 2nd model: several routes that traverse a common road segment. 3rd model: real-time control strategies	1st model: terminals. 2nd model: a road segment along which several routes pass. 3rd model: control points	Random vehicle travel time; several objective-function alternatives in the 1st model
Adamski and Chmiel (1997)	1st model: several routes that meet at a terminal. 2nd model: several routes	1st model: terminals. 2nd model: a road segment along which	Deterministic and stochastic cases

(Continued)

Table 6.2 *Characteristics of works covered in Section 6.6, Literature review and further reading (continued)*

Source	Network structure/ transferring strategies	Location of synchronized transfer	Special features
	that traverse a common road segment	several routes pass	
Abkowitz <i>et al.</i> (1987)	Two bus routes that meet at one spot. Four transfer strategies: unscheduled; scheduled without waiting; scheduled when the less-frequent bus waits; scheduled when both buses wait	A single meeting point	
Becker and Spielberg (1999)	Any route structure	Multiple centres	
Ceder <i>et al.</i> (2001); Ceder and Tal (2001)	Any route structure	All meeting points	Several headway- design options and several objective- function alternatives
Clever (1997)	Any route structure	1st option: single transfer centre. 2nd option: multiple centres	
Desilets and Rousseau (1990)	Any route structure	All meeting points	Several objective function alternatives
Daduna and Voss (1993)	Any route structure	All meeting points	
Hall (1985)	Several feeder lines and one main line	A single terminal	Random delay
Klemt and Stemme (1988)	Any route structure	All meeting points	
Kyte <i>et al.</i> (1982)	A main trunk line with complementary services	Several centres through which the main line passes	
Lee and Schonfeld (1991)	A train line and a bus line that meet at one spot	A single meeting point	Two models with different assumptions regarding delay variability

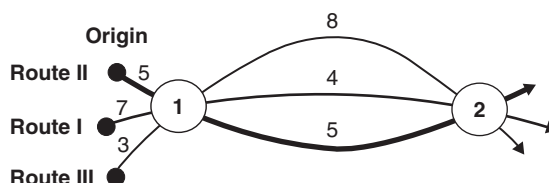
(Continued)

Table 6.2 Characteristics of works covered in Section 6.6, Literature review and further reading (continued)

Source	Network structure/ transferring strategies	Location of synchronized transfer	Special features
Maxwell (1999)	Four transfer strategies: close headways without synchronization; main-line trunk with timed-transfer branch lines; a single hub- and-spoke; several transfer hubs with fixed-interval timed transfers		
Nelson <i>et al.</i> (1981)	Four transfer strategies: simple timed transfer; pulse scheduling; line-up; neighbourhood pulse		
Rapp and Gehner (1976)	Any route structure	All meeting points	
Salzborn (1980)	Several feeder lines and one main line	Several centres through which the main line passes	Objective includes minimizing the number of required vehicles
Schneider <i>et al.</i> (1984)		Detailed discussion of transfer-centre location	
Shih <i>et al.</i> (1998)	Any route structure (routes are generated by the model)	Multiple centres	Transfer-centre locations and optimal vehicle sizes are determined
Voss (1990)	Any route structure	1st case: meeting points. 2nd case: road segments along which several routes pass	

Exercises

- 6.1 Given the simple network below, which combines two transfer points with three routes. The numbers on the arcs are travel times (average in minutes).

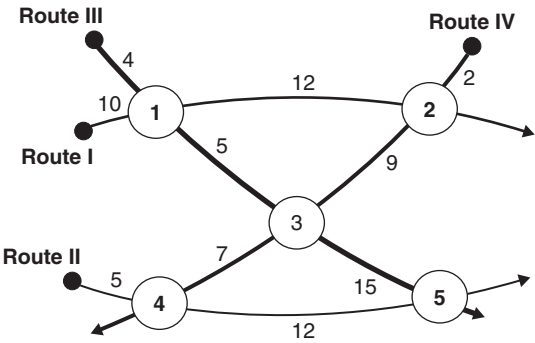


The data for the network:

i	Hmin _i	Hmax _i	F _i	T
Route I	7	13	2	
Route II	7	15	2	20
Route III	–	20	1	

Note that there is no Hmin for route III, since only one vehicle is assigned. Find the number of meetings, their upper-bound Z^* , and their associated departure and meeting times, using Synchro-1 and Synchro-2. *Suggestion:* run an optimization software for optimal (maximum number of meetings) results.

6.2 Given the network below, which combines five transfer points with four routes. The numbers on the arcs are travel times (average in minutes).



The data for the network:

i	Hmin _i	Hmax _i	F _i	T
Route I	5	10	3	
Route II	6	10	3	30
Route III	7	8	3	
Route IV	10	15	3	

Find the number of meetings, their upper-bound Z^* , and their associated departure and meeting times, using Synchro-1 and Synchro-2. *Suggestion:* run an optimization software for optimal (maximum number of meetings) results.

6.3 Synchro-1 and Synchro-2 are intended to maximize the number of simultaneous arrivals (meetings). Given the average number of passengers that need to be transferred between routes, construct a heuristic procedure that, instead of maximizing

the number of meetings, will maximize the number of passenger-meetings (each passenger transfer, without waiting time, counts as 1). The new objective function weighs each passenger transfer the same.

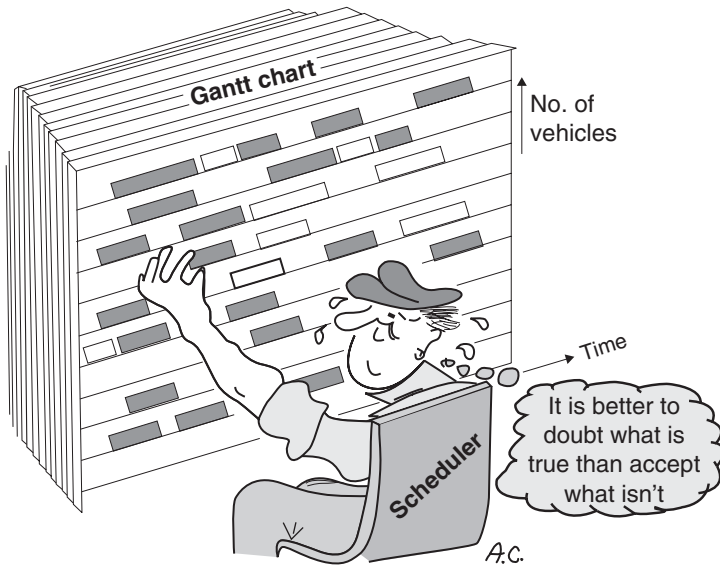
References

- Abkowitz, M., Josef, R., Tozzi, J. and Driscoll, M. K. (1987). Operational feasibility of timed transfer in transit systems. *Journal of Transportation Engineering*, **113**, 2, 168–177.
- Adamski, A. (1993). Transfer optimization in public transport. In *Computer-Aided Transit Scheduling* (J. R. Daduna, I. Branco and J. M. P. Paixao, eds), pp. 23–38, Lecture Notes in Economics and Mathematical Systems, **430**, Springer-Verlag.
- Adamski, A. and Chmiel, W. (1997). Optimal service synchronization in public transport. In *Transportation Systems* (M. Papageorgiou and A. Pouliezios, eds), **3**, 1283–1287.
- Becker, A. J. and Spielberg, F. (1999). Implementation of a timed transfer network at Norfolk, Virginia. *Transportation Research Record*, **1666**, 3–13.
- Ceder, A. (1999). *Systems Analysis as an Introduction to Operations Research*. Michlol Publication – Technion (Israel).
- Ceder, A. (2001). Bus timetables with even passenger loads as opposed to even headways. *Transportation Research Record*, **1760**, 28–33.
- Ceder, A., Golany, B., and Tal, O. (2001). Creating bus timetables with maximal synchronization. *Transportation Research*, **35A**(10), 913–928.
- Ceder, A. and Stern, H. I. (1984). Optimal transit timetables for a fixed vehicle fleet. In *Transportation and Traffic Theory* (J. Volmuller and R. Hammerslag, eds) pp. 331–355, UNU Science Press (the Netherlands).
- Ceder, A. and Tal, O. (2001). Designing synchronization into bus timetables. *Transportation Research Record*, **1760**, 3–9.
- Ceder, A. and Wilson, N. H. M. (1986). Bus network design. *Transportation Research*, **20B**(4), 331–344.
- Clever, R. (1997). Integrated timed transfer: a European perspective. *Transportation Research Record*, **1571**, 109–115.
- Daduna, J. R. and Voss, S. (1993). Practical experiences in schedule synchronization. In *Computer-Aided Transit Scheduling* (J. R. Daduna, I. Branco and J. M. P. Paixao, eds) pp. 39–55, Lecture Notes in Economics and Mathematical Systems, **430**, Springer-Verlag.
- Desilets, A. and Rousseau, J. M. (1990). SYNCHRO: a computer-assisted tool for the synchronization of transfers in public transit networks. In *Computer-Aided Transit Scheduling* (M. Desrochers and J.-M. Rousseau, eds), pp.153–166, Lecture Notes in Economics and Mathematical Systems, **386**, Springer-Verlag.
- Hall, R. W. (1985). Vehicle scheduling at a transportation terminal with random delay en route. *Transportation Science*, **19**, 308–320.
- Klemmt, W. D. and Stemme, W. (1988). Schedule synchronization for public transit networks. In *Computer-Aided Transit Scheduling* (J. R. Daduna and A. Wren, eds), pp. 327–335, Lecture Notes in Economics and Mathematical Systems, **308**, Springer-Verlag.
- Kyte, M., Stanley, K. and Gleason, E. (1982). Planning, implementing, and evaluating a timed-transfer system in Portland, Oregon. *Transportation Research Record*, **877**, 23–29.

- Lee, K. K. T. and Schonfeld, P. (1991). Optimal slack time for timed transfers at a transit terminal. *Journal of Advanced Transportation*, **25**(3), 281–308.
- Maxwell, R. R. (1999). Intercity rail fixed-interval, timed-transfer, multihub system: applicability of the 'Integraler Taktfahrplan' strategy to North America. *Transportation Research Record*, **1691**, 1–11.
- Nelson, M., Brand, D. and Mandel, M. (1981). Use and consequences of timed-transfers on US transit properties. *Transportation Research Record*, **197**, 50–55.
- Rapp, M. H. and Gehner, C. D. (1976). Transfer optimization in an interactive graphic system for transit planning. *Transportation Research Record*, **619**, 27–33.
- Salzborn, F. J. M. (1980). Scheduling bus systems with interchanges. *Transportation Science*, **14**, 211–220.
- Schneider, J. B., Deffebach, C. and Cushman, K. (1984). The timed-transfer/transit center concept as applied in Tacoma/Pierce County, Washington. *Transportation Quarterly*, **38**(3), 393–402.
- Shih, M. C., Mahmassani, H. S. and Baaj, M. H. (1998). Planning and design model for transit route networks with coordinated operations. *Transportation Research Record*, **1623**, 16–23.
- Voss, S. (1990). Network design formulations in schedule synchronization. In *Computer-Aided Transit Scheduling* (M. Desrochers and J.-M. Rousseau, eds), pp.137–152, Lecture Notes in Economics and Mathematical Systems, **386**, Springer-Verlag.

7

Vehicle Scheduling I: Fixed Schedules



Chapter 7 Vehicle scheduling I: Fixed Schedules

Chapter outline

- 7.1 Introduction
 - 7.2 Fleet size required for a single route
 - 7.3 Example of an exact solution for multi-route vehicle scheduling
 - 7.4 Max-flow technique for fixed vehicle scheduling
 - 7.5 Deficit-function model with deadheading trip insertion
 - 7.6 Depot-constrained vehicle scheduling
 - 7.7 Literature review and further reading
 - Exercises
 - References
 - Appendix 7.A: The maximum-flow (max-flow) problem
-

Practitioner's Corner

Given demand and coordinated timetables, the next phase is to establish chains of daily trips or vehicle blocs. Each chain or bloc constitutes a vehicle's schedule for a day. The problem of minimizing the number of chains and, at the same time, fulfilling agency requirements (refuelling, maintenance, etc.) is complex and cumbersome for medium and large-scale transit agencies. This chapter provides an overview of the problem, outlines specific solution procedures and tools, and describes some of the experience with a single transit agency.

The chapter contains five main parts following an introductory section. Section 7.2 exhibits a simple method for ascertaining the minimum number of vehicles required for a given single route without interlining. Section 7.3 presents exact mathematical-programming solutions for the case of multi-route vehicle scheduling. Practitioners can skip the math formulation here, but should concentrate on the experience of a 4,000-bus agency that led to the adoption of different approaches (described in the fourth part). In Section 7.4, the problem is converted to a known network-flow problem for enriching the quantitative interpretation of the underlying required analysis. Practitioners can skip this third part. Section 7.5 is the core of the chapter, proffering a graphical person-computer interactive approach based on a step function called deficit function. This approach provides a highly informative graphical technique that is simple to interact with and use. Practical suggestions can be interjected by the scheduler, followed immediately by describing the effects of the suggestions on the vehicle's schedule. Section 7.6 suggests a treatment for an often imperative postulate – maintaining a balance in the number of vehicles starting and ending at a depot. This part shows the formulation of both balancing depots and complying with limitations on the number of overnight spaces at depots or parking facilities. The chapter ends with literature review and exercises.

Practitioners may skip all of Section 7.4 and the Appendix, as well as the maths treatment (formulation) in Sections 7.3, 7.5 and 7.6. They should, though, pay especial attention to the examples and procedures depicted in Figures 7.6–7.12.

The subjects of vehicle and crew scheduling, the latter to be dealt with in Chapter 10, do receive attention in available software. This is not to say that those agencies using the software are fully happy with the results. Nonetheless, the existing competition among the various software companies is healthy, and brings to mind the following story. A large branch of Chase Manhattan Bank in New York displayed a sign with the company's then slogan: "You have a friend at Chase". Bank Leumi Le-Israel (BLL) decided to open one of its large branches adjacent to this branch of Chase. To compete with it, BLL also put up a huge sign: "With BLL you have a FAMILY". Three weeks later, Chase responded with another sign: "With such a family you really need a FRIEND".

7.1 Introduction

Vehicles, especially buses, are often shifted from one route to another (interlining) within a large-scale transit system, which frequently has them perform deadheading trips in order to operate a given timetable with the minimum vehicles required. As was noted in explaining Figure 1.2 in Chapter 1, it is desirable to analyse simultaneously the procedures for constructing timetables and for scheduling vehicles. However, because of the complexity of this analysis, the two procedures are treated separately. This chapter has two major foci: (1) to describe the task of vehicle scheduling and possible OR solutions for both a single transit route and a network of routes; (2) to proffer a graphical technique that is easy to interact with and that responds to practical concerns.

The motivation for the second foci arises from a problematic component of Egged, Israel's national bus carrier cooperative, which is presently engaged in scheduling about 4,000 buses over some 2,000 routes. Egged operates over an extensive geographical network composed of urban, suburban, regional and intercity routes, and performs an average of 50,000 daily trips – one of the world's largest schedules. Egged's crucial component of scheduling buses to trips was performed manually by about 60 schedulers using Gantt charts. A scheduler's duty was to list all daily chains (some deadheading) for a bus, ensuring the fulfilment of the timetable and the operator's requirements (refuelling, maintenance, etc.). However, because of frequent changes in the schedule and frequent additional imposed trips, the skilled schedulers were not capable of handling the bus-scheduling tasks efficiently. Consequently, the Egged management decided to test a fully computerized system. Their experience with this system is shown in Section 7.3. The experience led to the development of an informative graphical method, to be explicated in Sections 7.5 and 7.6.

In Adelaide, Australia, the newspaper headline on 14 April 2000, shown in Figure 7.1, gets our attention. We might think that a reverse headline, "Fewer Buses, No Waiting", would be the prudent one deserving a scoop. This chapter seeks ways of minimizing the number of vehicles and, at the same time, maintaining the best level of service set forth by the decided timetable.

The Advertiser

Adelaide, Friday, April 14, 2000 Metropolitan Edition www.news.com.au Phone (08) 8206 2000 80 cents

Official AFL collector's album and team photos
Bonus coupon inside today
 Album FREE or team photo \$1 with coupon on Page 32

MORE BUSES, NO WAITING

Go Zones, extra services for commuters

By State Political Reporter GREG KELTON

AN extra 40,000 services a year will be introduced on Adelaide's bus routes from Easter Sunday.

The new services will be spread across the metropolitan bus network and, on nine major routes, passengers will not have to wait more than 15 minutes for a bus during the day or longer than 30 minutes at night.

Those routes – to be known as Go Zones – are the O-Bahn, Main North Rd, Torrens Rd, Port Rd, Henley Beach Rd, Goodwood Rd, Unley Rd, Norwood Park and Payneham Rd. The extra services – and greater safety measures – will result from the virtual takeover of bus routes by private operators from April 23.

Extra night, weekend and public holiday services will be part of the package, along with increased security on buses, trains and at interchanges.

Security guards will travel on every train after 7pm, supported by increased video surveillance.

• Continued Page 8

Food more palatable under GST

By Political Reporter ANNABEL CRABB

AN AVERAGE basket of supermarket food should be about 1 per cent cheaper under the goods and services tax, according to the first detailed estimate.

The Australian Food and Grocery Council yesterday released a report into food prices under the new tax system, which found most products would have only tiny fluctuations in price.

Some bakery products, puddings and live fish will be up to 9 per cent more expensive – but soft drinks and cordial should fall by nearly 10 per cent.

The report by Access Economics said the deal between the Federal Government and the Australian Democrats last year to exclude basic food from the GST had made a substantial difference to the impact on the supermarket shopping basket.

“The original proposal had food going up by about 4.4 per cent and beverages up by about 1.1 per cent,” Access Economics director Geoff Carmody said.

• Continued Page 4

TAKING STOCK: Cathy Friebe and son Adam with the sample grocery basket yesterday.

Classifieds Index Page 76 Call Rhonda on 131 841 Metro forecast: Cloudy, 21° Index: Page 2

Figure 7.1 Main headline of an Adelaide newspaper on 14 April 2000. Wouldn't a real scoop be the reverse: "Fewer Buses, No Waiting"?

Approaches to vehicle scheduling world-wide, range from primitive decision-making to computerized mathematical-programming techniques. The latter is reviewed below in Section 7.3. As for primitive approaches, an example was observed in Santo Domingo. The drivers for a 40-bus company would gather very early each morning and create vehicle schedules as if was a lottery. That is, all trips in the timetable were represented by small pieces of papers (each trip with a different number), and the drivers simply picked up their piece of paper containing the chain of trips to be performed that day as though they were drawing a lucky number from a hat. Some drivers, of course, went back to sleep.

A common practice in vehicle scheduling is to use time-space diagrams, similar to the one from Europe appearing in Figure 7.2. Each line in the diagram represents a trip moving over time (x-axis) at the same average commercial speed represented by the slope of the line. Although many schedulers became accustomed to this description, it is cumbersome, if not impossible, to use these diagrams to make changes and improvements in the scheduling. It is impossible to use them for interlining because the space (y-axis) refers only to a single route, as can be seen in the figure. It is also difficult to use different average speeds for different route segments, in which the lines in the time-space diagram can cross one another; this is not to mention the inconvenience of using these diagrams manually for inserting deadheading trips and/or shifting departure times. These limitations of the time-space diagram caused us to look more closely into a more appealing approach – the one presented in Sections 7.5 and 7.6.

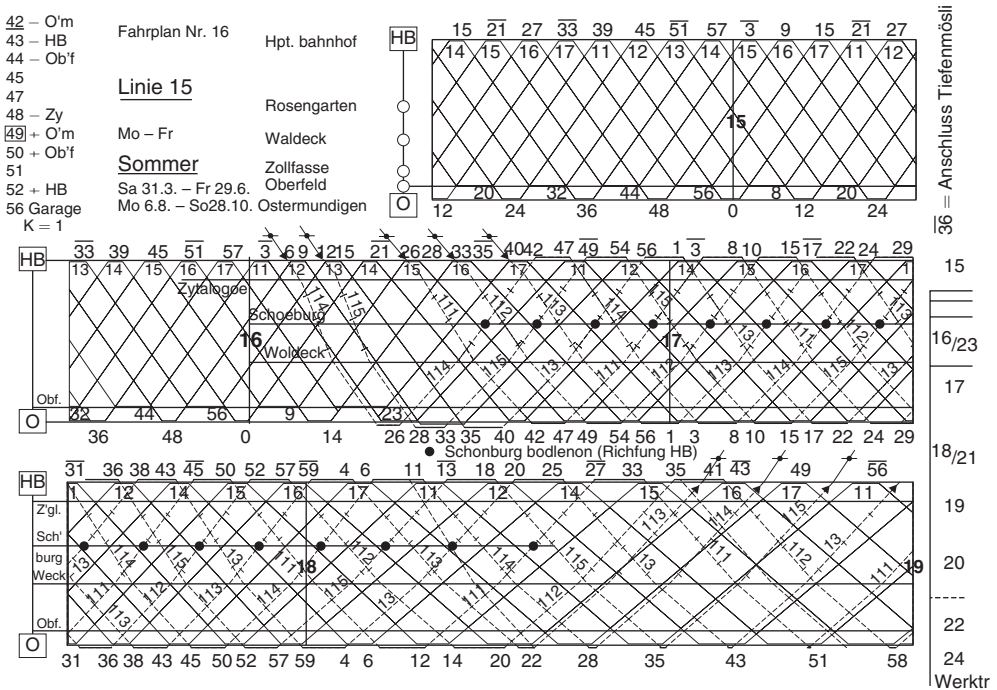


Figure 7.2 Typical time-space diagram of transit trips used to describe and change vehicle scheduling in many transit agencies

7.2 Fleet size required for a single route

This section considers the case in which interlinings and deadheading (DH) trips are not allowed and each route operates separately. Given the average round trip time and chosen layover time, the minimum fleet size for a radial route can be found according to the formula derived by Salzborn (1972). Specifically, let T_r be the average round-trip time, including layover and turn-around times, of a radial route r (departure and arrival points are the same). The minimum fleet size is equal to the largest number of vehicles that departs within T_r .

Although Salzborn's modelling provides the base for fleet-size calculation, it relies on three assumptions that do not hold up in practice: (i) vehicle departure rate is a continuous function of time, (ii) T_r is the same throughout the period under consideration, and (iii) route r is a radial route starting at a major point (e.g. Central Business District (CBD)). In practice, departure times are discrete (see Chapters 4–6), average trip time is usually dependent on time-of-day, and a single transit route usually has different timetables for each direction of travel. For that reason, we will now broaden Salzborn's model to account for practical operations planning.

Let route r have two end points: a and b . Let T_{ria} and T_{rjb} be the average trip time on route r for vehicles departing at t_{ia} and t_{jb} from a and b , respectively, including layover time at their respective arrival points. Let n_{ia} be the number of departures from a between t_{ia} , in which departure i_a is included, and $t_{i'a}$ in which departure $i'a$ is excluded. Thus, i_a arrives at terminal b , then continues with trip jb , the latter being the *first* feasible departure from b to a at a time greater than or equal to the time $t_{ia} + T_{ria}$; $t_{i'a}$ is the *first* feasible departure from a to b at a time greater than or equal to $t_{jb} + T_{rjb}$. Similarly n_{jb} may be defined for a trip j from b .

Lemma 7.1: In the case of no deadheading (DH) trips, n_{ia} departures must be performed by different vehicles at a , and n_{jb} must be performed by different vehicles at b , for all ia and jb in the timetables of r .

Proof: The proof is actually based on a contradiction. Let us assume that the same vehicle can perform two departures included in n_{ia} at a . However, in order to complete a full round trip, including layover times, this vehicle can only pick up the $i'a$ departure at a , which is not included (by definition) in n_{ia} ; hence, it is impossible for the same vehicle to execute two departures within n_{ia} .

Theorem 7.1: In the case with no interlining (between routes) and no DH trips, the minimum fleet size required for route r is

$$N_{\min}^r = \max\{\max_i n_{ia}, \max_j n_{jb}\} \quad (7.1)$$

Proof: Based on Lemma 7.1, $\max_i n_{ia}$ and $\max_j n_{jb}$ represent the maximum number of vehicles required to execute the timetables at a and b , respectively. Consequently the minimum fleet size for r is the greater of the two.

An example of deriving the required fleet size for a single transit route r is shown in Figure 7.3. In this figure, a single average travel time $T_{ria} = T_{rjb} = 15$ minutes is used throughout the timetable for both directions of r . The timetables contain 12 departures at b and 10 at a . The calculations for n_{ia} and n_{jb} are shown by arrows; starting with the departures at a for n_{ia} (using $T_{ria} = 15$), and starting at b for n_{jb} . The solid lines in Figure 7.3 represent

the direction from the starting time to the first feasible connection (after 15 minutes), and the dashed lines in the opposite direction link to the first feasible connection (also after 15 minutes) from the starting point. This leads to a determination of both n_{ia} and n_{jb} , and eventually the minimum fleet size, $N_{\min}^r = 5$, according to Equation (7.1). It should be mentioned that the same T_{ria} and T_{rjb} are used throughout the example only for the sake of simplicity. Varied T_{ria} and T_{rjb} can be utilized in the same manner for each departure.

Vehicle chains (blocks) can be constructed by using the FIFO (first-in, first-out) rule. That is, a block will start at a depot for the first assigned scheduled trip, and then will make the first feasible connection with a departure (based on the route's timetable) at the other end point of the route, and so on. The block usually ends with a trip back to the depot. The trips to and from the depot are often deadheading trips. In Figure 7.3, the five blocks can be constructed starting with the first departure (5:00) at b and using the FIFO (first-feasible connection) rule, then deleting the departures selected and continuing with another block until all departures are used. At the start of each step (at b), a check is made to see whether the next (in time) departure can

Calculating n_{ia} $T_{rja} = 15$ minutes			Calculating n_{jb} $T_{rjb} = 15$ minutes		
n_{ia}	Timetable at a	Timetable at b	Timetable at a	Timetable at b	n_{jb}
3		5:00	6:00	5:00	3
2		5:30	6:15	5:30	2
3		6:00	6:30	6:00	1
3		6:30	6:45	6:30	2
4		6:50	7:00	6:50	5
4		7:05	7:10	7:05	5
3		7:10	7:20	7:10	4
2		7:15	7:25	7:15	4
2		7:20	7:40	7:20	3
—		7:30	8:00	7:30	3
		7:40		2	
		8:00		—	
$\max n_{ia}$	4				
$\max n_{jb}$			5		
$N'_{\min}$	$\max (4, 5) = 5$				

Figure 7.3 Example of derivation of single-route fleet size with no deadheading trips

be connected to an earlier unused departure at a and, if so, whether this connection can be allowed. The five blocks, therefore, are as follows: $[5:00(a) - 6:00(a) - 6:30(b) - 6:45(a) - 7:05(b) - 7:20(a) - 7:40(b) - 8:00(a)]$; $[5:30(b) - 6:15(a) - 6:50(b) - 7:10(a) - 7:30(b)]$; $[6:00(b) - 6:30(a) - 7:10(b) - 7:25(a) - 8:00(b)]$; $[7:00(a) - 7:15(b) - 7:40(a)]$; $[7:20(b)]$. An earlier connection, linking the 7:15 departure at b to the 7:00 departure at a , is possible only in the fourth block. The above FIFO process can certainly start at a , as well. Note that the last block has only one trip; the five blocks can undergo changes, including swapping trips, between blocks. Each block can start and end at a depot or can be used as a segment in a larger block.

Finally, when deadheading (DH) trips between the ends of two routes and/or slightly shifting departure times are allowed, it is more complex to use the formulation developed above. Instead, the solution can be found using the graphical method presented in Section 7.5 and in Chapter 8.

7.3 Example of an exact solution for multi-route vehicle scheduling

Practical vehicle scheduling usually involves a network of transit routes with interlining between routes. The switch from one route to another can be executed either by a service trip when the two routes share the same end point (terminal) or by a DH (empty) trip. The most problematic part of the vehicle-scheduling task for a network of routes is to minimize the number of vehicles required to carry out the timetables. It is basically a cost-minimization problem. This section provides an example of the implementation of one solution.

The problem of scheduling vehicles in a multi-terminal (multi-route) scenario is known as the Multi-Depot Vehicle Scheduling Problem (MDVSP). This problem is complex (NP-complete, which is explained in Chapter 5, Section 5.3), and considerable effort is devoted to solving it in an exact way. A review and description of some exact solutions can be found in Desrosiers *et al.* (1995), Daduna and Paixao (1995), and in the literature review at the end of this chapter.

An example of the formulation of the MDVSP is as follows:

$$\text{Objective function: } \min_y \sum_{i=1}^{n+1} \sum_{j=1}^{n+1} c_{ij} y_{ij} \quad (7.2)$$

Subject to $\sum_{i=1}^{n+1} y_{ij} = 1, j = 1, 2, \dots, n$, and $\sum_{j=1}^{n+1} y_{ij} = 1, i = 1, 2, \dots, n$; $\sum_{j=1}^{n+1} y_{n+1,j} = 1$, and $\sum_{i=1}^{n+1} y_{i,n+1} = 1$; finally there are also constraints ensuring that trip end i can link to trip start j , and that y_{ij} is integer and greater than or equal to zero, where

i = end of a trip at time t_i ,

j = start of a trip at time t_j ,

$$y_{ij} = \begin{cases} 1, & \text{ending is a connection to start} \\ 0, & \text{otherwise} \end{cases}$$

For $i = n + 1$, then $y_{n+1,j} = 1$ if a depot supplies a vehicle for the j -th trip. For $j = n + 1$, then $y_{i,n+1} = 1$ if, after the end of the i -th trip, the vehicle returns to a depot and

$y_{n+1,n+1}$ = number of vehicles remaining unused at a depot. The cost function c_{ij} takes the form:

$$c_{ij} = \begin{cases} K & ; i = n + 1; j = 1, 2, \dots, n \\ 0 & ; i = 1, 2, \dots, n; j = n + 1 \\ L_{ij} + E_{ij} & ; i, j = 1, 2, \dots, n \end{cases} \quad (7.3)$$

where

K = saving incurred by reducing the fleet size by one vehicle,

L_{ij} = deadheading (DH) cost from event i to j ,

E_{ij} = cost of a driver's idle time between i and j .

This formulation, appearing in a similar form in Gavish *et al.* (1978), covers the chaining of vehicles in a sequential order from the depot to the transit routes, alternating with idle time and deadheading trips, and back to the depot. In operations research (OR) terms (see Section 5.3 in Chapter 5), this is a zero-one integer-programming problem in that it can be converted to a large-scale assignment problem. The objectives are to minimize fleet requirements (minimize linkages of routes) and minimize deadheading and idle time (maximize vehicle and driver utilization). In addition, the assignment of vehicles from the depots to the vehicle schedule generated in the foregoing chaining process can be formulated as a 'transportation problem', as known in the OR literature. The objectives of the assignment stage are to minimize the driving cost from the depot to the first trip on the schedule, and from the last trip to a depot.

The formulation presented partially in Equations (7.2) and (7.3) was applied by Egged (the Israeli national bus carrier). As mentioned earlier, Egged has about 4,000 buses and tried to gain more efficient schedule handling based on Equations (7.2) and (7.3). Egged examined the algorithm based on several sets of bus-scheduling data. It is worth noting that the program is capable of solving bus-scheduling problems involving up to 2,500 bus trips. The full test results and discussion of this examination were reported by Ceder and Gonen (1980). For example, the results of five sets are shown in Figure 7.4. Only the DH and idle times of the first data set are not available (indicated by N/A). These examples were selected for peak-hour bus schedules, and at a first look the results seemed quite attractive. However, when there was an attempt to implement the results, it became evident that several significant constraints remained unfilled. Applying these constraints results in a solution that is considerably different from that obtained by the optimal method. In Egged, as in most transit agencies, each bus is used by the same drivers. During the manual scheduling of buses to trips, it was necessary for the scheduler to also consider some of the drivers' constraints, although that is of secondary importance at this stage.

In summary, the major limitations of the programmable algorithm were that it could not consider the following: (1) integration of more than 2,500 trips, (2) availability of adequate bus type for each trip, (3) need for bus refuelling, (4) need for driver's meals, and (5) location and availability of drivers. The Egged schedulers agreed that using software as a black box was working blind as opposed to understanding the software and interacting with

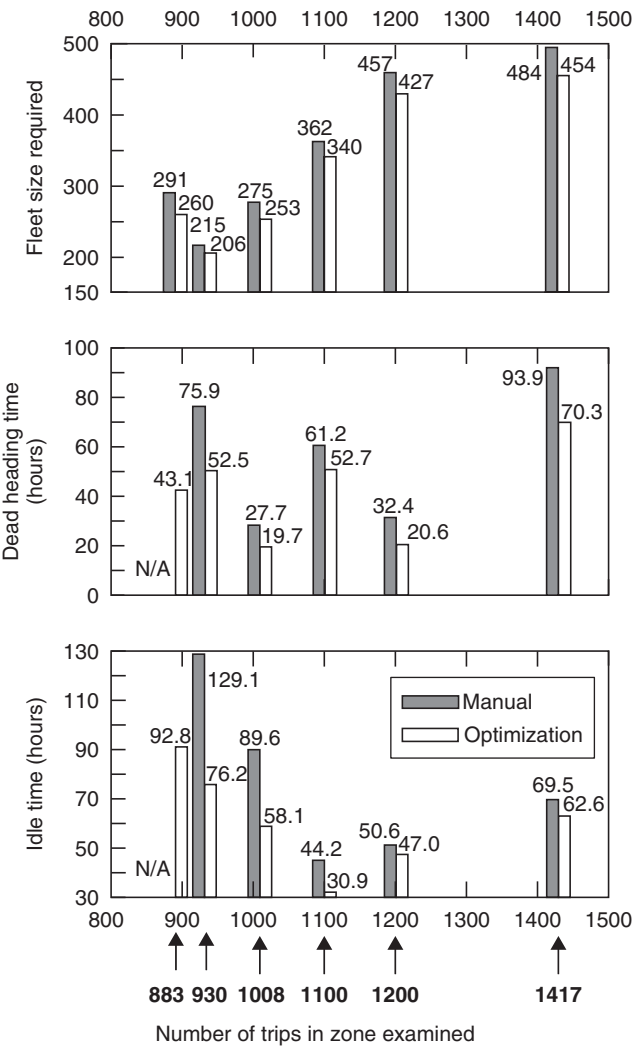


Figure 7.4 Comparison of manual and optimization (programmable) solutions for vehicle-scheduling planning at a large bus agency

it, which becomes a real eye-opener. In order to satisfy these limitations, it was decided to search for a procedure that would allow the inclusion of practical considerations that experienced schedulers might wish to introduce into the schedule. This procedure, based on the deficit-function graphical method, is described in Section 7.5.

7.4 Max-flow technique for fixed vehicle scheduling

This section shows that the procedure for solving the minimum fleet-size problem has its roots in the classic work of Ford and Fulkerson (1962). The approach described is based on

Ceder (1978) and Stern and Ceder (1983a) using network-flow techniques well known in the OR field.

7.4.1 Vehicle scheduling task

Consider a schedule (set of timetables) of required transit trips in which each trip is defined by a starting terminal, starting time, ending terminal and ending time. The problem is to find the minimum number of vehicles (the fleet size) that can carry out all the trips in the schedule. As this is most efficiently done if single vehicles are allowed to service several trips in succession, a second part of the problem is to find the set of trips assigned to each vehicle in the fleet. If trip j immediately follows i , then (a) the starting time of trip j must be greater than or equal to the ending time of trip i , and (b) the starting and ending terminal of j and i must be identical. If the second condition is not met, some systems allow the vehicle to run empty from i to j . If the travel time for this empty trip can be completed before the starting time of trip j , then both trips may appear in the same chain and be assigned the same vehicle. Bus operations offer an example of transportation systems that often carry out such empty or deadheading (DH) trips. On the other hand, airline companies rarely, if ever, 'deadhead' their aircraft because of the high cost of running empty.

7.4.2 Formulation and transformation of the max-flow problem

A trip-joining array for S may be constructed by associating the i -th row with the arrival event of the i -th trip, and the j -th column with the departure event of the j -th trip. Cell (i, j) will be admissible if i and j can feasibly be joined. Otherwise, (i, j) will be an inadmissible cell. Let x_{ij} be a 0-1 variable associated with cell (i, j) and I be the set of required trips; consider, then, the following problem:

Problem P1.

$$\text{Max } Z1 = \sum_{i \in I} \sum_{j \in I} X_{ij} \quad (7.4)$$

$$\text{Subject to: } \sum_{j \in I} x_{ij} \leq 1, \quad i \in I \quad (7.5)$$

$$\sum_{i \in I} x_{ij} \leq 1, \quad j \in I \quad (7.6)$$

$$\left. \begin{array}{ll} x_{ij} \in \{0, 1\}, & \text{all } (i, j) \text{ admissible} \\ x_{ij} = 0, & \text{all } (i, j) \text{ inadmissible} \end{array} \right\} \quad (7.7)$$

A solution with $x_{ij} = 1$ indicates that trips i and j are joined. The objective function maximizes the number of such joinings. Constraint (7.5) ensures that each trip may be joined with, at most, one successor trip. Similarly, constraint (7.6) indicates that each trip may be joined with, at most, one predecessor trip. The following theorem states that maximizing (7.4) in P1 is tantamount to minimizing the number of chains for a trip schedule of size n .

Table 7.1 Trip schedule S for the example problem

Trip number i	Departure terminal p^i	Departure time t_e^i	Arrival terminal q^i	Arrival time t_s^i
1	b	6:00	b	6:30
2	a	7:05	c	8:05
3	c	7:10	a	8:00
4	b	8:30	a	9:20
5	a	9:00	b	9:45

Table 7.2 Average DH travel-time (minutes) matrix for the example in Table 7.1

		Arrival terminal		
		a	b	c
Departure terminal	a	0	30	50
	b	35	0	45
	c	45	40	0

Theorem 7.2: Let N and n denote the number of chains and trips, respectively. Then, $\text{Min } N = n - \text{Max } Z1$.

Proof of Theorem 7.2 can be found in Ford and Fulkerson (1962).

The problem P1 is equivalent to a special arrangement of the maximum-flow (max-flow) problem. The max-flow algorithm that solves the vehicle-scheduling problem with DH trips is called an augmenting-path algorithm. It appears, in an extensive treatment, in the classic book by Ford and Fulkerson (1962). A complete description of the augmenting-path algorithm, which will be used for further applications in this volume, is shown in Appendix 7.A at the end of this chapter.

The vehicle scheduling problem can be transformed to a unit-capacity bipartite network, in which the solution time has the complexity of $O(n^{1/2}m)$ with n nodes (departure times) and m arcs. A bipartite network is a network in which the set of nodes is partitioned into two subsets of nodes such that directed arcs exist only between (not within) the two subsets. This solution time, first shown by Even and Tarjan (1975), was explicated by Ahuja *et al.* (1993). The complexity function is explained in Section 5.3 in Chapter 5.

We will use a simple numerical example as an explanatory tool to describe and interpret the construction and solution of the vehicle-scheduling problem transformed into a max-flow network problem. Consider the three-terminal problem defined by the data in Tables 7.1 and 7.2. The data in Table 7.1 are transformed into the upper part of the network-flow representation in Figure 7.5, which has two dummy nodes: a *source* node s and a *sink* node t .

The nodes, being connected from s , are the arrival times of the example, with an indication, in parenthesis, of the arrival terminal. The nodes connected to t are the departure times, with an indication of the departure terminal. Feasible connections between the arrival and departure times, utilizing the DH data in Table 7.2, establish the arcs between the left-side and right-side nodes, based on Equation 7.4. Each arc capacity represents the number of connections

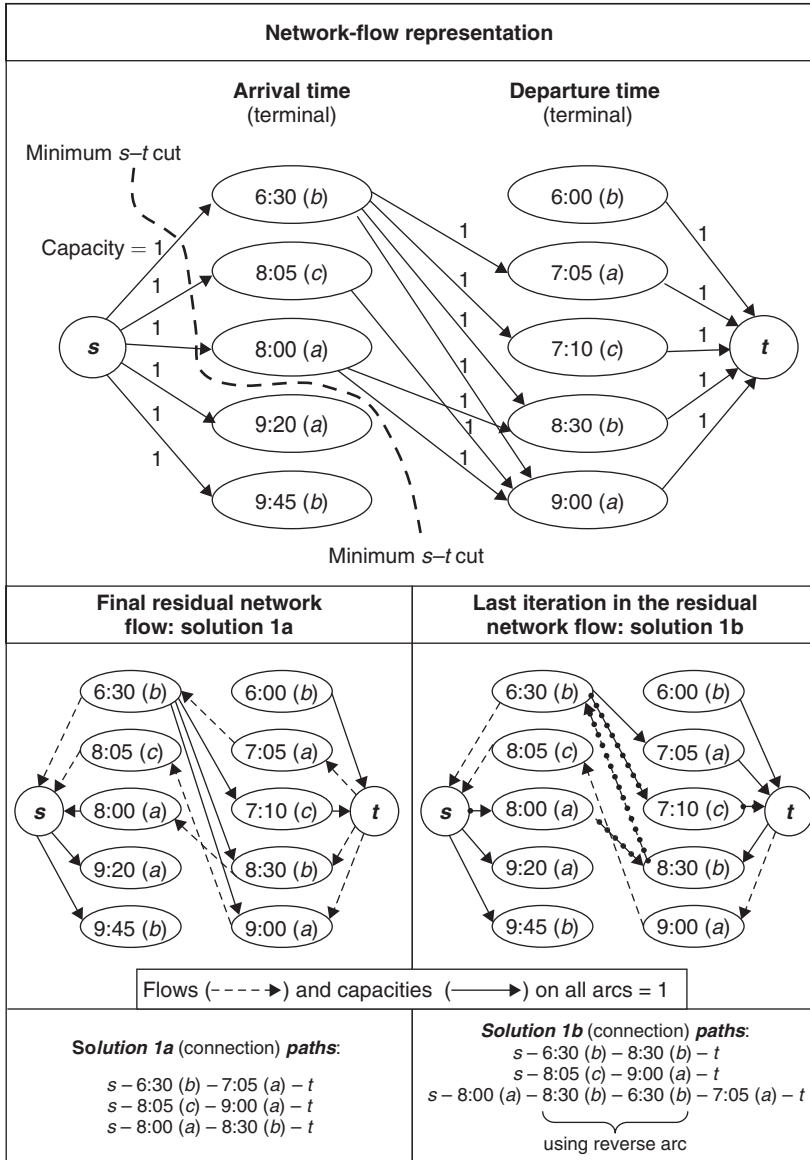


Figure 7.5 Solutions 1a and 1b of the vehicle-scheduling example using the augmenting-path algorithm appearing in Appendix 7.A

that can ‘flow’ through the arc. In our case, there is only a unit capacity assigned to each arc, because only one connection (if any) between a given arrival time and terminal and a given departure time and terminal is possible. The more flow created, the fewer chains will be required, as stated by Theorem 7.2. The objective function $Z1$ in $P1$ equals the flow to be created at s and absorbed at t . The maximum flow of four in Figure 7.5 means that all trips can be handled by a single chain, with four linkages between arrivals and departures.

The augmenting-path algorithm described in Appendix 7.A is applied for the example problem. Look for the minimum-arc path between s and t , assign a flow = 1, and create a residual network in which all arcs utilized have a reverse arc (dashed line in the two residual networks in Figure 7.5). All capacities and flows in the residual networks have a unit value. Then search for the next minimum-arc path between s and t , assigning to it flow = 1 on the residual network, and construct an updated residual network. The process continues until no s - t path is found. Two solutions, 1a and 1b, leading to the same optimal result (max-flow = 3) are shown in Figure 7.5. In solution 1b, a case in which a reverse arc is used in the augmenting-path algorithm is shown. Based on Theorem 7.A4 (in the Appendix), max-flow = minimum s - t cut in the original network-flow. This minimum s - t cut, explicated in Appendix 7.A, is shown in the upper part of Figure 7.5.

The result of the example is max-flow = max Z = 3, and, following Theorem 7.2, min N = $5 - 3 = 2$ chains. That is, the timetable in Table 7.1 can be carried out by a minimum of two vehicles having the connections shown in Figure 7.5. Explicitly, the two blocs, by their trip number in Table 7.1, are [1 – 2 – 5] and [3 – 4]. These two blocs have three DH trips for connecting arrival and departure terminals: b - a and c - a (in the first bloc) and a - b (in the second bloc), with the total of, respectively, $35 + 45 + 30 = 110$ minutes DH time.

7.5 Deficit-function model with deadheading trip insertion

The minimum fleet-size problem may be referred to with or without DH trips. When DH is allowed, we can reach the counter-intuitive result of decreasing the required resources (fleet size) by introducing more work into the system (adding deadheading trips). This approach assumes that the capital cost of saving a vehicle far outweighs the cost of any increased operational cost (driver and vehicle travel cost) imposed by the introduction of deadheading trips. This section offers a highly informative, graphical person–computer interactive technique based on a step function, called deficit function, that is simple to use and interact with. The lesson that is learned is that it is easier to deadhead vehicles or shift their departure time (see next chapter) when the problem is seen graphically. Perhaps the inspiration for this approach is the saying: “It is easier for an elephant to pass through the eye of a needle if it is lightly greased”. Explanatory background on a step-function approach follows, described by Ceder and Stern (1981) and Ceder (2003), for allocating the minimum number of vehicles to a given timetable.

7.5.1 Definitions and notations

Let $I = \{i: i = 1, \dots, n\}$ denote a set of required trips. The trips are conducted between a set of terminals $K = \{k: k = 1, \dots, q\}$, each trip to be serviced by a single vehicle and each vehicle able to service any trip. Each trip i can be represented as a 4-tuple (p^i, t_s^i, q^i, t_e^i) , in

which the ordered elements denote departure terminal, departure (start) time, arrival terminal, and arrival (end) time. It is assumed that each trip i lies within a schedule horizon $[T_1, T_2]$, i.e. $T_1 \leq t_s^i \leq t_e^i \leq T_2$. The set of all trips $S = \{(p^i, t_s^i, q^i, t_e^i): p^i, q^i \in K, i \in I\}$ constitutes the timetable. Two trips i, j may be serviced sequentially (feasibly joined) by the same vehicle if and only if (a) $t_e^i \leq t_s^j$ and (b) $q^i = p^j$.

If i is feasibly joined to j , then i is said to be the predecessor of j and j the successor of i . A sequence of trips $i_1, i_2, \dots, i_w$ ordered in such a way that each adjacent pair of trips satisfies (a) and (b) is called a chain or block. It follows that a chain is a set of trips that can be serviced by a single vehicle. A set of chains in which each trip i is included in I exactly once is said to constitute a vehicle schedule. The problem of finding the minimum number of chains for a fixed schedule S is defined as the minimum fleet-size problem.

Let us define a DH trip as an empty trip from some terminal p to some terminal q in time $\tau(p, q)$. If it is permissible to introduce DH trips into the schedule, then conditions (a) and (b) for the feasible joining of two trips, i, j , may be replaced by the following:

$$t_e^i + \tau(q^i, p^j) \leq t_s^j \quad (7.8)$$

A **deficit function (DF)** is a step function defined across the schedule horizon that increases by one at the time of each trip departure, and decreases by one at the time of each trip arrival. This step function is called a deficit function (DF) because it represents the deficit number of vehicles required at a particular terminal in a multi-terminal transit system. To construct a set of DFs, the only information needed is a timetable of required trips. The main advantage of the DF is its visual nature. Let $d(k, t, S)$ denote the DF for terminal k at time t for schedule S . The value of $d(k, t, S)$ represents the total number of departures minus the total number of trip arrivals at terminal k , up to and including time t . The maximum value of $d(k, t, S)$ over the schedule horizon $[T_1, T_2]$, designated $D(k, S)$, depicts the deficit number of vehicles required at k .

The DF notations are presented in Figure 7.6, in which $[T_1, T_2] = [5:00, 8:30]$. It is possible to partition the schedule horizon of $d(k, t)$ into a sequence of alternating hollow and maximum intervals $(H_0^k, M_1^k, H_1^k, \dots, H_j^k, M_{j+1}^k, \dots, M_{n(k)}^k, H_{n(k)}^k)$. Note that S will be deleted when it is clear which underlying schedule is being considered. Maximum intervals $M_j^k = [s_j^k, e_j^k]$, $j = 1, 2, \dots, n(k)$ define the intervals of time over which $d(k, t)$ takes on its maximum value. Index j represents the j -th maximum intervals from the left; $n(k)$ represents the total number of maximal intervals in $d(k, t)$, where s_j^k is the departure time for a trip leaving terminal k and e_j^k is the time of arrival at terminal k for this trip. The one exception occurs when the DF reaches its maximum value at $M_{n(k)}^k$ and is not followed by an arrival, in which case $e_{n(k)}^k = T_2$.

A hollow interval H_j^k , $j = 0, 1, 2, \dots, n(k)$ is defined as the interval between two maximum intervals: this includes the first hollow, from T_1 to the first maximum interval, $H_0^k = [T_1, s_1^k]$; and the last hollow, which is from the last interval to T_2 , $H_{n(k)}^k = [e_{n(k)}^k, T_2]$. Hollows may contain only one point; if this case is not on the schedule horizon boundaries (T_1 or T_2), the graphical representation of $d(k, t)$ is emphasized by a clear dot.

The sum of all DFs over k is defined as the overall DF, $g(t) = \sum_{k \in K} d(k, t)$. This function $g(t)$ represents the number of trips that are simultaneously in operation; i.e. a count, from a bird's-eye view at time t , of the number of transit vehicles in actual service over the entire transit network of routes. The maximum value of $g(t)$, $G(S)$, is exploited for a determination

of the lower bound on the fleet size (see Chapter 8). An example of a two-terminal operation, a fixed schedule of trips, and the corresponding set of DFs and notations is illustrated in Figure 7.6.

7.5.2 Fleet-size formula

Determining the minimum fleet size, $D(S)$, from the set of DFs is simple enough – one merely adds up the deficits of all the terminals. In the example in Figure 7.6, without DH

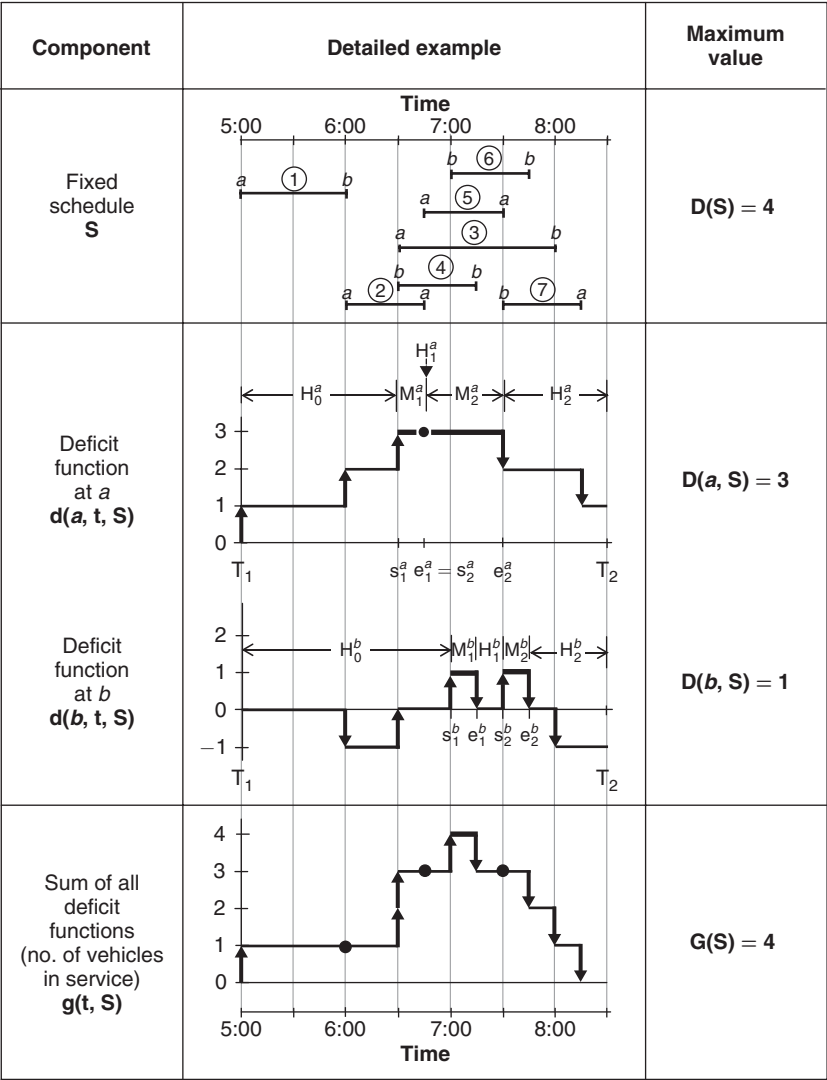


Figure 7.6 Illustration of two-terminal fixed schedule with associated deficit functions and their sum, including notations and definitions

trips, $D(S) = D(a) + D(b) = 4$. This result was apparently derived independently by Bartlett (1957), Salzborn (1972, 1974), and Gertsbach and Gurevich (1977). It is formally stated as Theorem 7.3.

Theorem 7.3: If, for a set of terminals K and a fixed set of required trips I , all trips start and end within the schedule horizon $[T_1, T_2]$ and no DH insertions are allowed, then the minimum number of vehicles required to service all trips in I is equal to the sum of all the deficits.

$$\text{Min } N = \sum_{k \in K} D(k) = \sum_{k \in K} \max_{t \in [T_1, T_2]} d(k, t) \quad (7.9)$$

Proof: Let F_k = the number of vehicles present in terminal k at the start of the schedule horizon T_1 ; let $s(k, t)$ and $e(k, t)$ be the cumulative number of trips starting and ending at k from T_1 up to and including time t . The number of vehicles remaining at k at time $t \geq T_1$ is $F_k - s(k, t) + e(k, t)$.

In order to service all trips leaving k , the above expression must be non-negative; i.e. $F_k \geq s(k, t) - e(k, t)$, $T_1 \leq t \leq T_2$. The minimum number of vehicles required at k is then equal to the maximum deficit at k . $\text{Min } F_k = \text{Max}_t[s(k, t) - e(k, t)] = \text{Max}_t d(k, t)$. Hence, the minimum number of vehicles required for all terminals in the system is equal to the total deficit $\text{Min } N = \sum_{k \in K} \text{Min } F_k = \sum_{k \in K} d(k) = D(S)$.

7.5.3 DH trip insertion: effect and initial fleet-size lower bound

This section follows Ceder and Stern (1981). A DH trip is an empty trip between two termini that is usually inserted into the schedule in order to: (i) ensure that the schedule is balanced at the start and end of the day, (ii) transfer a vehicle from one terminal where it is not needed to another where it is needed to service a required trip, and (iii) refuel or undergo maintenance. This section will consider case (ii), and Section 7.6 will comment on case (iii). We start by asking: Where and when are such trips needed? Usually, a trip schedule received from operating personnel includes such deadheading trips, and it is easy to apply the fleet-size formula to determine the minimum fleet size, followed by the first-in-first-out rule to construct each vehicle's schedule. The assumption is that the trip schedule S has been purged of all DH trips, leaving only required trips. From this point, the question of how to insert deadheading trips into the schedule in order to further reduce the fleet size will be examined. At first, it seems counter-intuitive that this can be achieved, since it implies that increased work (adding trips to the schedule) can be carried out with decreased resources (fewer vehicles). This section will show through an examination of the effect of such deadheading trip insertions on deficit functions that this is indeed possible.

Consider the example in Figure 7.7. In its present configuration, according to the fleet-size formula, four vehicles are required at terminal a , 0 at terminal b , and 1 at terminal c for a fleet size of five. The dashed arrows in the figure represent the insertion of DH_1 trip from b to a and DH_2 from c to b . After the introduction of these DH trips into the schedule, the DFs at all three terminals are shown updated by the dotted lines. The net effect is a reduction in fleet size by one unit at terminal a . It is interesting to examine the particular circumstances

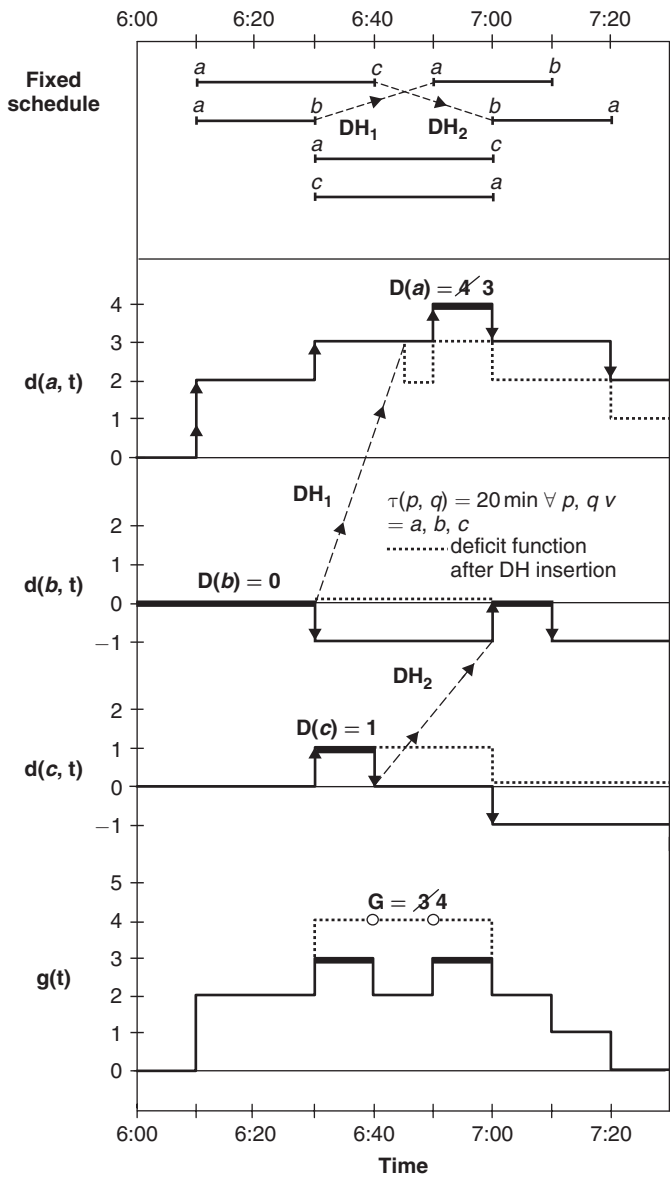


Figure 7.7 Description of six-trip, two-terminal example in which the fleet size is reduced by one using a chain of two DH trips (URDHC) and in which $g(t)$ is changed

under which this reduction was achieved. After adding an arrival point in the first hollow of terminal a before s_1^a , the maximal interval when using DH_1 is reduced by one unit, causing a unit decrease in the deficit at a . This arrival point becomes, therefore, e_1^a . Since the DH_1 departure point is added in the middle hollow of terminal b , at e_1^b , it is necessary to introduce a second DH trip, which will arrive at the start of the second maximum interval of b .

Fortunately, this DH_2 trip departs from the last hollow of c , where it could no longer affect the deficit at c . In general, it is possible to have a string of DH trips to reduce the fleet size by one unit: one ‘initiator trip’ and the others ‘compensating trips’.

All successful DH trip chains follow a common pattern. The initial DH trip is introduced to arrive in the first hollow of a terminal in which a reduction is desired. This DH trip must depart from some hollow of another terminal. Moving to the end of this hollow, another DH trip is inserted, such that its arrival epoch will compensate for the departure epoch added by the first DH trip. This is followed by additional compensating trips; however, in order to avoid looping, no more than one DH trip will be allowed to depart from the same hollow. Each time a DH trip is inserted (from p to q) to arrive at the end of a hollow H_i^q from the start of a hollow H_j^q , it must pass a feasibility test; i.e. $e_j^q + \tau(p, q) \leq s_{i+1}^q$. If the inequality is true with $<$, then there will be some slack time, during which the DH trip can be shifted. Let this slack time be defined as $\delta_{pq} = s_{i+k}^q - [e_j^q + \tau(p, q)]$. In practice, if the DH time plus the slack time are greater than, or equal to, the average service travel time, then a service trip may replace the DH trip. In this way, an additional service trip is introduced, thereby resulting in higher frequency (i.e. an improved level of service) at usually the same operational cost. Figure 7.8 illustrates a nine-trip schedule with three DH times required to reduce $D(b)$ by one. In this example, DH_2 can be converted to a service trip in which $\delta_{ac} = 30$ minutes.

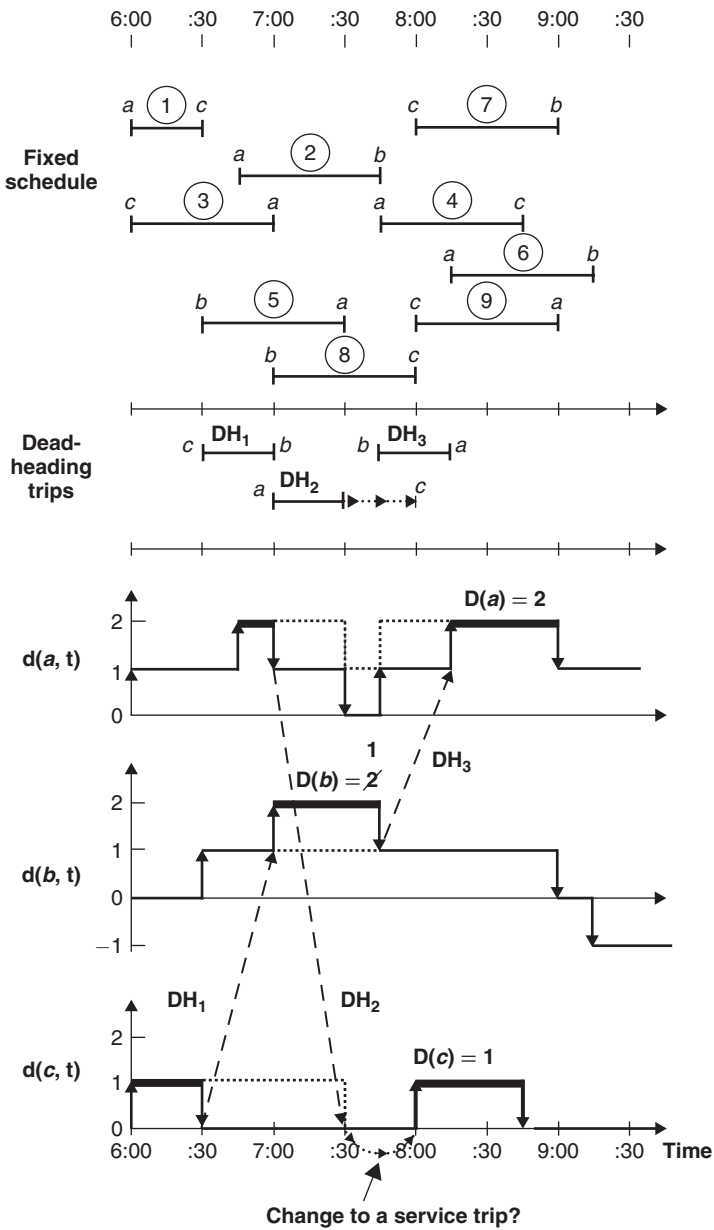
The process ends when a final hollow of some terminal q is reached (i.e. $H_1^q = H_{n(q)}^q$), after which no compensation is necessary. It is possible to arrive at a point where no feasible compensating DH trips can be inserted, in which case the procedure terminates or one may backtrack to the arrival point of the last DH trip added and try to replace it with another. This procedure results in a sequence of DH trips known as a unit reduction dead-heading chain (URDHC) if it ends successfully (i.e. if it reduces the fleet size by a unit amount). A URDHC is a sequence of DH trips of the form $(k_1, DH_1, k_2, DH_2, \dots, k_j, DH_j, k_{j+1}, DH_w, k_{w+1})$, where DH_j is a DH trip from terminal k_j to k_{j+1} . The DH trips are inserted into the deficit functions from hollow to hollow, with DH_1 arriving at the first hollow $H_0^{k_2}$ of $d(k_2, t)$ and DH_w departing from the last hollow of $d(k_w, t)$. The procedure inserts URDHCs until no more can be found or a lower bound on the minimum fleet is reached. The examples of Figures 7.7 and 7.8 include a single URDHC with two and three DH trips, respectively.

Obviously, the continued reward for such a search must stop, and the following ‘Initial Lower Bound Theorem’ provides a condition when it is futile to continue this search. The initial, and rather intuitive, lower bound on the fleet size is stated below. Further improvements in the fleet-size lower bound appear in Chapter 8.

Theorem 7.4: (a) For a set of terminals K and a trip schedule S , the maximum number of trips in simultaneous operation provides a lower bound on the minimum fleet size,

$$G = \max_{t \in [T_1, T_2]} g(t) \leq \text{Min } N \quad (7.10)$$

(b) If $G = \text{Min } N$, it is impossible to reduce the minimum fleet size through the introduction of DH trips.



- DH travel time between all terminals = 30 minutes (Service time = 60 minutes)
- Before (5 FIFO chains): [1–7], [2], [3–4], [5–6], [8–9]
- After (4 FIFO chains): [1–DH1–8–9], [2–DH3–6], [3–DH2–7], [5–6]

Figure 7.8 Example of a URDHC with three DH trips in which DH₂ can be converted to a service trip

Proof: (i) In general it is known that:

$$\max_{t \in [T_1, T_2]} \sum_{k \in K} d(k, t) \leq \sum_{k \in K} \max_{t \in [T_1, T_2]} d(k, t)$$

From the definition of $g(t)$ and (7.9), $\max_{t \in [T_1, T_2]} g(t) \leq \text{Min } N$

(ii) Let S represent the original schedule of given trips. Add some trip i starting at $t(s)$ and ending at time $t(e)$ to the schedule (this can be a DH trip). Call the new schedule (with i) S' . The symbol S or S' is added to the argument of the deficit and overall DFs to indicate the set of trips for which they are constructed.

Since

$$g(t, S') = \begin{cases} g(t, S), & \text{for } t \notin [t(s), t(e)] \\ g(t, S) + 1, & \text{for } t \in [t(s), t(e)] \end{cases}$$

$$\max_t g(t, S') \geq \max_t g(t, S)$$

$$G(S') \geq G(S)$$

From (i) $\text{Min } N(S') \geq G(S')$ and using the assumption in (b)

$$\text{Min } N(S) = G(S).$$

Therefore,

$$\text{Min } N(S') \geq \text{Min } N(S).$$

This theorem is quite useful, as it enables one to understand the situations, based on the deficits, in which we can be sure that no reduction in fleet size can be (further) achieved. Figures 7.6 and 7.7 show and indicate both $g(t)$ and G . Figure 7.7 also demonstrates the impact of DH insertion on the value of G . The result of Theorem 7.4 determines the extent to which we can expect (in a maximal successful case) to reduce a fleet size through the introduction of DH trips, compared to the minimum fleet size required when DH trip insertions are not allowed.

7.5.4 Heuristic algorithm for a DH trip insertion

The results of the prior section provide a number of clues for the insertion of DH trips into the schedule that can form the basis of a heuristic algorithm to reduce fleet size. As much freedom of choice exists in the selection of DH trips at any point in the chain-construction procedure, the algorithm can be programmed so that it may be employed in a conversational (manual mode), person-machine mode. See Appendix 8.A in Chapter 8 for such an algorithm, appearing currently on the website: www.altdoit.com. This allows the practical considerations of experienced schedulers to be brought into the process. In addition, the algorithm may operate in a purely automatic manner; i.e. selections of DH trips being made on the basis of several criteria, to be discussed below. The primary inputs required for the heuristic procedure are: (1) the initial fixed set of trips, S° , defined over a set of terminals K , and (2) the travel-time matrix for potential DH trips between each pair of terminals in the schedule. The output of the algorithm

is a new trip schedule that includes the set of DH trips inserted and the number of vehicles required at each terminal. This constitutes the fleet size. Final DF information is also available. In order to construct the trip schedule of each vehicle in the fleet, a second phase is required. This can be done by applying the chain-construction rules described in the next section.

A general framework of the heuristic algorithm is shown in a flow-diagram form, following Ceder and Stern (1981), in Figures 7.9(a) and 7.9(b). The parentheses { } denote a set of elements. The result of the lower-bound fleet-reduction theorem suggests that $G(S^0)$ should be examined before trying to insert DH trips into the schedule. This is called a lower-bound termination test in Figure 7.9(a). Note that after DH trip insertions are made, $G(S^0)$ may be increased as shown in the example of Figure 7.6. It is thus important to evoke the lower-bound termination test after each DH chain insertion. Nonetheless, in Chapter 8, $G(S)$ is improved (can be increased) to $G''(S'')$; hence, the lower-bound termination test will have $G''(S'')$, a non-updated value, instead of $G(S)$ to compare with $D(S)$. In the latter case, the updated $G(S)$ is kept in order to know the number of vehicles in service at any time.

The next step in Figure 7.9(a) is an attempt to reduce the fleet size at some terminal in the schedule. The selection of the terminal is made by using rule NT (next terminal). This may be done manually on the basis of depot capacity violations; i.e. the present fleet size required at a terminal exceeds the capacity of the terminal. A default rule selects the terminal whose first hollow is the longest. The rationale here is to try to offer the largest opportunity for the insertion of the first DH trip. This selection should, of course, be made from terminals that have not been previously examined; they should not be made from terminals previously examined for URDHCs with unsuccessful results. In cases of multiple longest-first hollows, the selection is based on a terminal whose overall maximum region (from the start of the first maximum interval to the end of the last one) is the shortest.

Once a terminal, u , is selected, the algorithm searches for a URDHC from that terminal. This is done in the sub-routine URDHC, which appears with a dashed line in Figure 7.9(b). If a URDHC is not found from u , this terminal is added to the set W of terminals examined. If all terminals have been examined ($W = K$), then the algorithm stops. Otherwise, a new terminal is selected and the search for a URDHC continues. If a URDHC has been found from u , in practice another test takes place. This concerns the DH cost involved as compared with the cost saving of a single vehicle; the comparison is usually computed by accountants. If the DH cost is higher than the saving cost, the URDHC is cancelled. Otherwise, the set of DH trips {DH} is added to the previous schedule; all DF information is updated for this new schedule, and a new iteration initiated. The algorithm must terminate after a finite number of steps because the fleet size in each iteration is reduced by one unit until either it reaches its lower bound or a URDHC cannot be found. Since the search for a URDHC is conducted many times in the algorithm, it has been placed in a sub-routine, called URDHC, which will now be described.

The sub-routine URDHC in Figure 7.9(b) starts by setting a dummy computer variable q to u . The number taken on by q represents the terminal for which the next DH trip is to be inserted. This trip is to arrive before time S_{j+1}^q ; i.e. before the end of the hollow H_j^q (in the case of $q = u$, $H_j^q - H_0^u$). The computer then determines a set of feasible candidate departure points $\{e^v\}$ from the set of terminals $V = \{u\}$, $q \notin V$. These points are identified by the start of hollows that lie to the right of the last hollow examined. That is, the selection of one of the candidate trips is made using rule R. When $R = 0$, the computer allows this selection to be inserted manually into a conversational mode. Otherwise, an automatic selection is made by the computer using one of the following criteria.

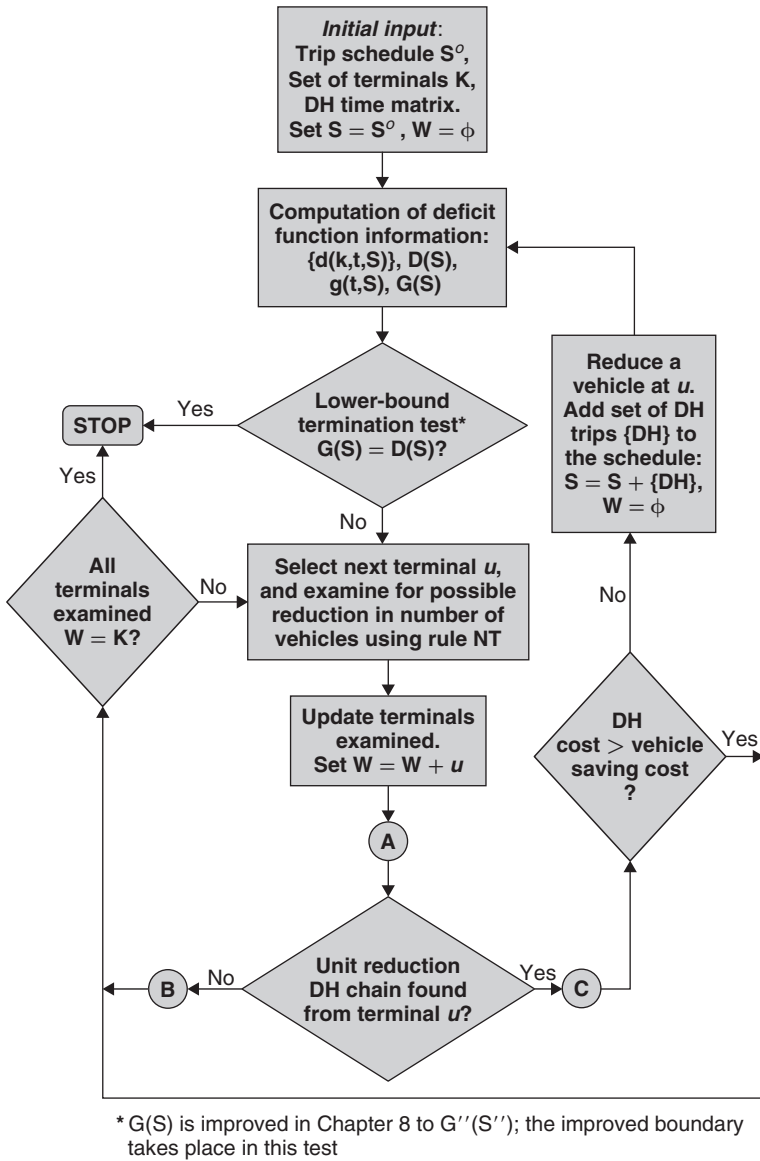


Figure 7.9(a) Flow diagram of the DH insertion heuristic algorithm

R = 1: the minimum DH trip time (MTT). Insert the candidate DH trip with the least travel time.

R = 2: the furthest start of a hollow (FSH). Insert a candidate DH trip whose hollow starts furthest to the right.

R = 3: the farthest end of a hollow (FEH). Insert a candidate DH trip whose hollow ends furthest to the right.

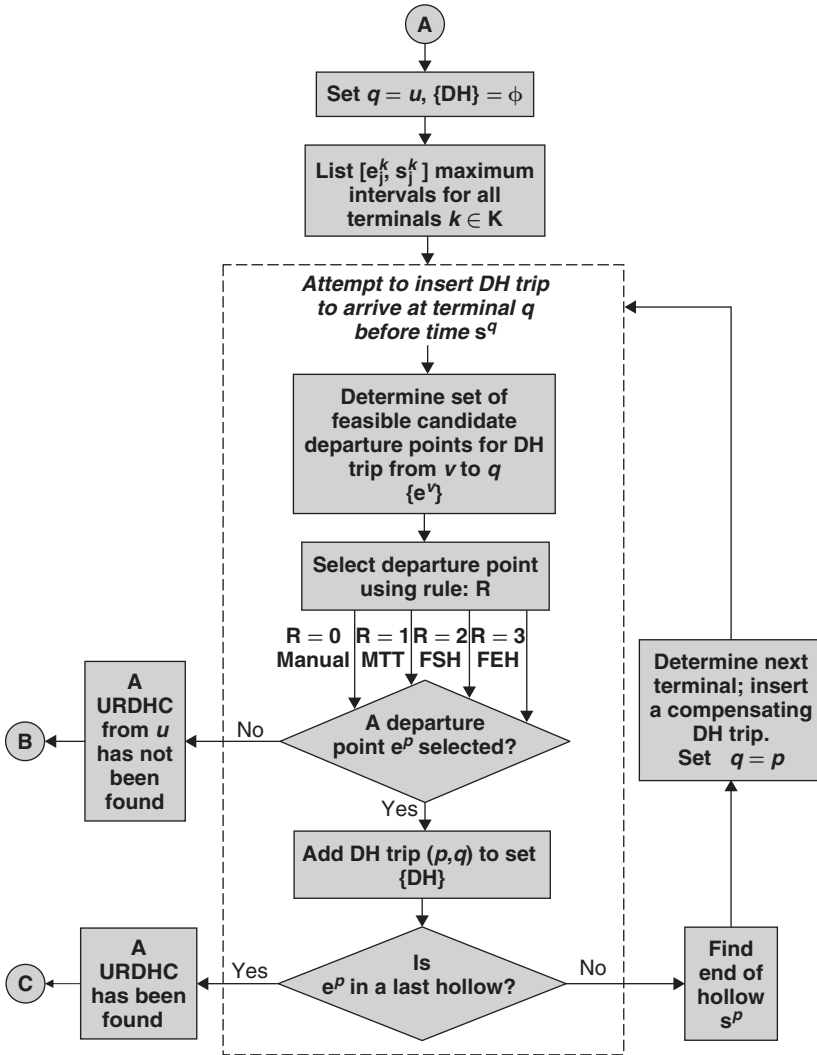


Figure 7.9(b) Part of the flow diagram in Figure 7.9(a), in which the URDHC sub-routine is illustrated

Figure 7.10 shows the effect of each criterion on three examples. For each example, the value of the total travel time of the DH trips in the URDHC is found, and the minimum value across all criteria circled. These examples show that the minimum DH travel time associated with the URDHC can be found by any of the criterion; no one criterion dominates the others for this performance measure.

If a DH trip cannot be selected and the completion of a URDHC is blocked, the algorithm backs up to the last candidate list with the last selection deleted; it then proceeds to select another DH trip from the candidate list. In the manual mode, the scheduler may choose a new terminal and initiate a new search for a URDHC instead of backing up. The option is also

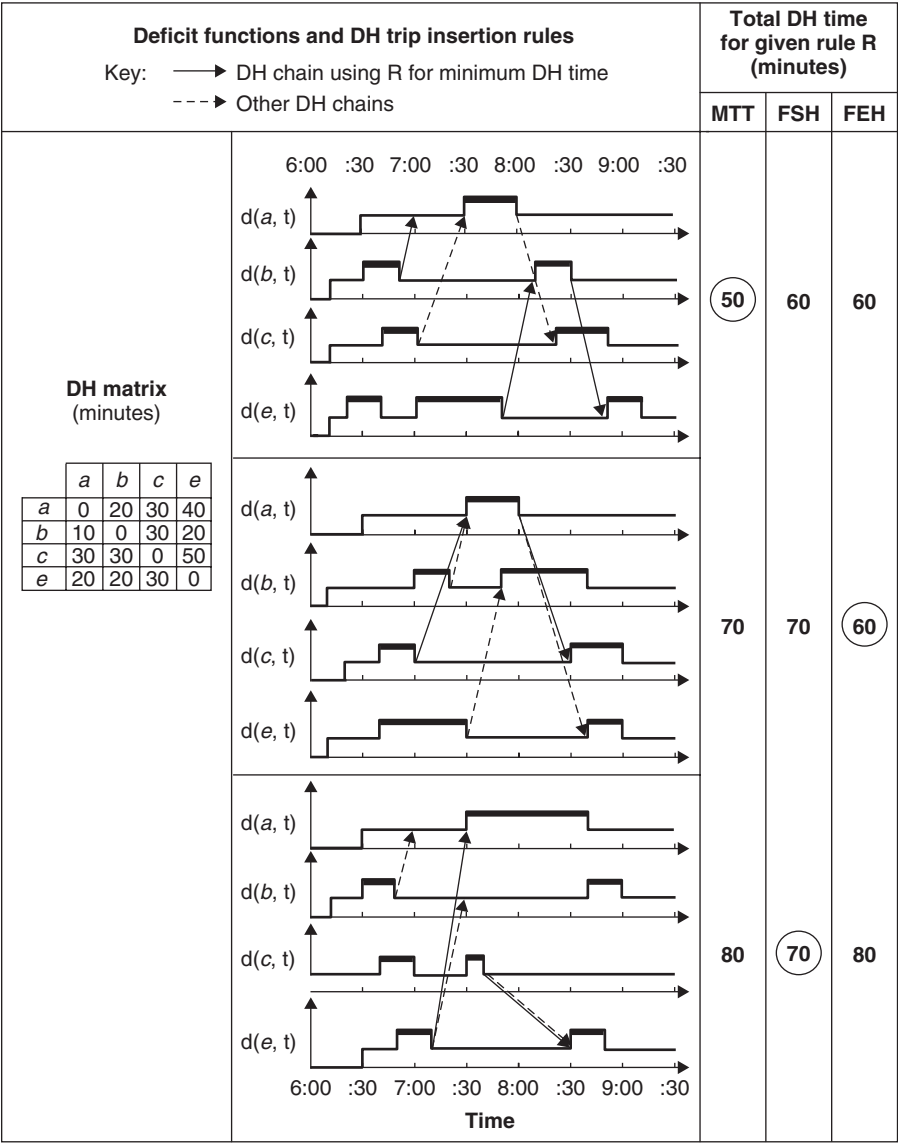


Figure 7.10 The effect on total deadheading time of three alternative heuristic rules for selecting candidate DH trip-departure terminals

available in the automatic mode to forego the backup feature when blocks occur in order to reduce excessive computation time. If the backup feature is allowed to continue, one may return to the initial hollow of terminal *u*, after which a new terminal may be selected to reinitiate the search for a URDHC. When the full backup feature is used, the procedure will find the optimal solution, which is equivalent to Ford and Fulkerson's (1962) max-flow procedure, which is explained in Section 7.4. If a final hollow is reached, a URDHC has been found, and

it is returned to the main program for introduction into the schedule. Otherwise, the search for a compensating trip is repeated until one trip is found that departs from a last hollow.

7.5.5 Constructing vehicle schedules (chains/blocks)

At the end of the heuristic algorithm, all trips, including the DH trips, are chained for constructing the vehicle schedule (blocks). Two rules can be applied for creating the chains: first-in-first-out (FIFO) and a chain-extraction procedure described by Gertsbach and Gurevich (1977). The FIFO rule simply links the arrival time of a trip to the nearest departure time of another trip (at the same terminal); it continues to create a schedule until no connection can be made. The trips considered are deleted and the process continues.

The chain-extraction procedure allows an arrival–departure connection for any pair within a given hollow (on each DF). The pairs considered are deleted, and the procedure continues. Figure 7.11 illustrates one DF at k . This $d(k, t)$ has four hollows, H_j^k , $j = 0, 1, 2, 3$, with H_1^k having arrivals of Trips 1, 2 and 3 and departures of Trips 4, 5 and 6. Below Figure 7.11 are the FIFO connections (within this hollow) as well as other alternatives; in all, the minimum fleet size at k , $D(k)$, is maintained. In addition two full FIFO chains are shown in Figure 7.6 before and after DH insertion procedure.

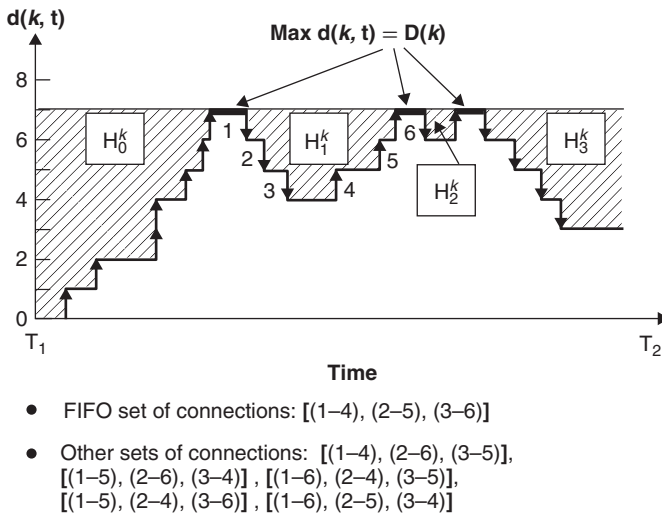


Figure 7.11 Example of creating trips connections within one hollow, H_1^k , using the FIFO rule and all other possibilities while maintaining the minimum fleet size attained

7.6 Depot-constrained vehicle scheduling

This section combines the description of previous sections with the methodology advanced by Stern and Ceder (1983a). A vehicle schedule (chain/block) is said to be balanced if for every terminal, k , the number of vehicles starting their schedules from k equals the number of vehicles ending their schedules at k (not necessarily the same vehicles). Otherwise, the

vehicle schedule is unbalanced. In practice, transit agencies seek this balanced condition at terminals with overnight parking (usually called depots).

7.6.1 Deficit function formulation

An example of a balanced schedule appears in Figure 7.12 for a solution without DH trips, using the data of Table 7.1. The solution indicates that exactly one vehicle chain starts and ends at each of the three terminals (albeit not the same chain for each terminal). However, a solution with DH trips, which saved a vehicle at terminal *a*, resulted in an unbalanced schedule (see lower part of Figure 7.12). A DH trip from *a* to *c* after 9:20 will balance the

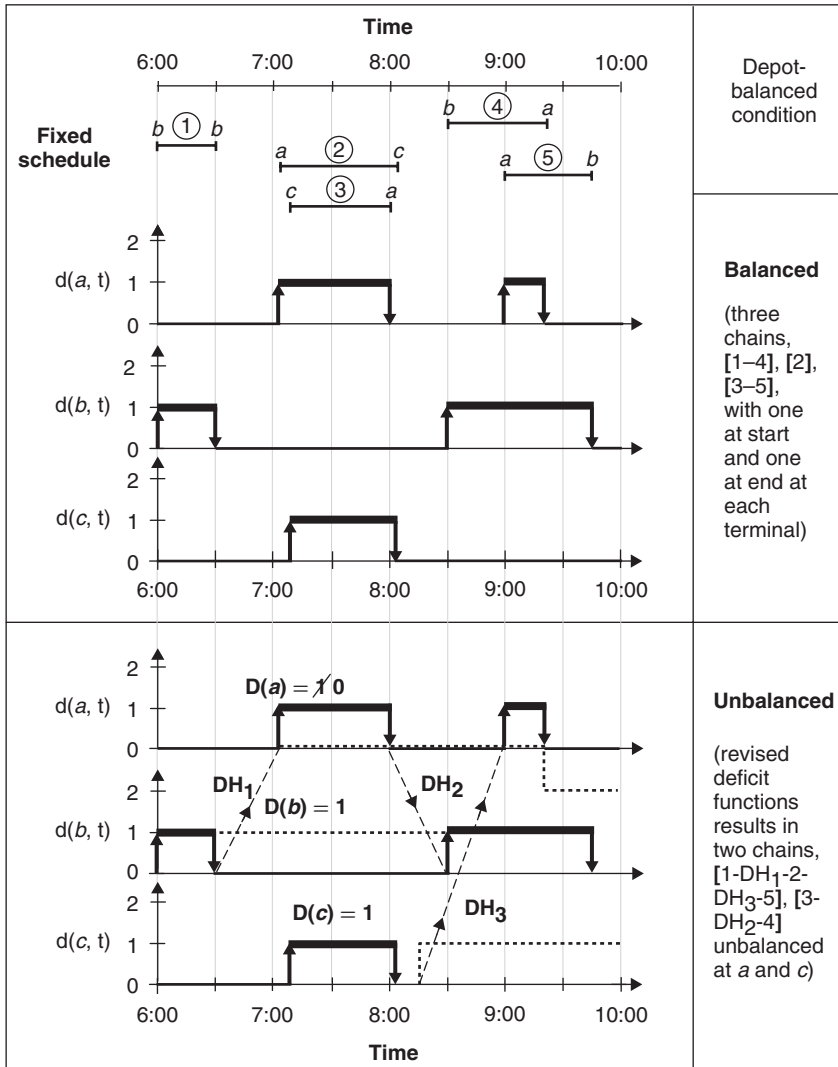


Figure 7.12 Deficit-function solution, with depot-balanced evidence, for the example in Tables 7.1 and 7.2

schedule. To determine whether a schedule is balanced, it is not necessary to carry out the actual chain constructions. The assessment may be made through an examination of the deficit functions of the schedules. If $vs(k)$ and $ve(k)$ are defined as the number of vehicles required at the start and remaining at the end of the schedule horizon, respectively, for terminal k . Then the schedule is balanced if $vs(k) = ve(k)$ for all k .

Theorem 7.5: A vehicle schedule is balanced if and only if the value of each deficit function at the end of the schedule horizon is zero; i.e.

$$d(k, T_2) = 0, k \in K$$

Proof: $D(k) - d(k, T_2)$ is the net number of arrivals (arrivals minus departures) from the end of the last maximum interval. As these arrivals are not followed by departures in any chain construction, they represent the number of chain ends at k ; i.e. $ve(k) = D(k) - d(k, T_2)$. Furthermore, it follows from Theorem 7.5 that $vs(k) = D(k)$. The requirement from a balanced schedule, $vs(k) = ve(k)$, implies $d(k, T_2) = 0$.

The conclusions drawn from Theorem 7.5 are as follow: (i) if $d(k, T_2) < 0$, then a surplus of $d(k, T_2)$ vehicles is present at k at the end of the day; (ii) if $d(k, T_2) > 0$, then a shortage of $d(k, T_2)$ vehicles occurs at k at the end of the day; and (iii) the following identity is true under a system that conserves vehicles:

$$\sum_{k \in K_1} d(k, T_2) = - \sum_{k \in K_2} d(k, T_2) \quad (7.11)$$

in which K_1 and K_2 represent the sets of shortage and surplus terminals, respectively.

One may notice, however, that in order to balance the schedule, it may not be enough to satisfy practical situations in which limitations exist on the number of overnight depots or parking facilities. Let $Q(k)$ represent such a limitation at terminal k . Then a feasible schedule must also satisfy the depot constraints:

$$vs(k) \leq Q(k), ve(k) \leq Q(k) \text{ for all } k$$

If the schedule is balanced, this reduces to $D(k) = Q(k)$, and the depot-constrained balance-schedule problem may be stated in deficit function form as:

Problem P2

Insert as many deadheading trips as necessary into the schedule of required trips in order to:

$$\text{Min } Z2 = \sum_{k \in K} D(k) \quad (7.12)$$

$$\text{Subject to: } d(k, T_2) = 0, \quad k \in K \quad (7.13)$$

$$D(k) \leq Q(k), \quad k \in K \quad (7.14)$$

Alternatively, using the identities of the proof of Theorem 7.5, the problem may be restated as

$$\text{Min } Z2 = \sum_{k \in K} vs(k) \quad (7.15)$$

$$\text{Subject to:} \quad \text{vs}(k) - \text{ve}(k) = 0, \quad k \in K \quad (7.16)$$

$$0 \leq \text{vs}(k) \leq Q(k), \quad k \in K \quad (7.17)$$

$$0 \leq \text{ve}(k) \leq Q(k), \quad k \in K \quad (7.18)$$

In this case, either (7.17) or (7.18) is redundant, but is still included in order to facilitate the development of the next section.

7.6.2 Mathematical programming formulation

In order to extend problem P1 (Section 7.4.2) to include depot and vehicle-balancing constraints, a number of dummy overnight trips may be defined that travel backwards in time. For each terminal k , let $n + k$ represent an undisclosed number of trips of the form $(n + k, t_s^{n+k}, n + k, t_e^{n+k})$ such that $t_e^{n+k} \leq T_1 < T_2 \leq t_s^{n+k}$. The times t_e^{n+k} and t_s^{n+k} represent the latest time at which a vehicle can arrive at depot k after servicing its last required trip and the earliest time a vehicle can leave depot k to service its first required trip, respectively. Such times may be stipulated in a driver-union contract. The following additional variables are now introduced:

$Y_{n+k,j} = 1$ if a vehicle departs for depot k (after completing dummy trip $n + k$) to service j , its first required trip; otherwise, equals 0.

$Y_{i,n+k} = 1$ if a vehicle services its last required trip as trip i before arriving at depot k to park for the night (start dummy trip $n + k$); otherwise, equals 0.

z_{kk} = an integer variable whose value represents the number of unused spaces at depot k ($0 \leq z_{kk} \leq Q(k)$, $k \in K$).

Let the travel time of the morning and evening DH trips from and to depot k be defined as $(n + k, p^j)$ and $(q^i, n + k)$, respectively. The joining $(n + k, j)$ will be considered admissible if $t_e^{n+k} + \tau(n + k, p^j) \leq t_s^j$ for $j \in I$, $k \in K$. Otherwise, $(n + k, j)$ will be inadmissible, and $Y_{n+k,j} = 0$. The joining $(i, n + k)$ is admissible if $t_e^i + \tau(q^i, n + k) \leq t_s^{n+k}$ for $i \in I$, $k \in K$. Otherwise, $(i, n + k)$ is inadmissible, and $Y_{i,n+k} = 0$. Note that $(n + k, j)$ and $(i, n + k)$ are always admissible if $p^j = k$ and $q^i = k$, respectively. Let A represent the set of admissible joinings of the above type in addition to those between required trips (i, j) as described earlier. The mathematical programming version of this problem with DH trips may now be stated.

Problem P3

$$\text{Max } Z3 = \sum_{i \in I} \sum_{j \in I} x_{ij} \quad (7.19)$$

Subject to:

$$\sum_{j \in I} x_{ij} + \sum_{k \in K} y_{i,n+k} = 1, \quad i \in I \quad (7.20)$$

$$\sum_{i \in I} x_{ij} + \sum_{k \in K} y_{n+k,j} = 1, \quad j \in I \quad (7.21)$$

$$\sum_{j \in I} y_{n+k,j} + z_{kk} = Q(k), \quad k \in K \quad (7.22)$$

$$\sum_{i \in I} y_{i,n+k} + z_{kk} = Q(k) \quad , k \in K \quad (7.23)$$

$$z_{kk} \in \{0, 1, \dots, Q(k)\} \quad , k \in K \quad (7.24)$$

$$\left. \begin{array}{l} x_{ij} \in \{0, 1\} \quad , (i, j) \in A \\ x_{ij} = 0 \quad , (i, j) \notin A \end{array} \right\} \quad (7.25)$$

$$\left. \begin{array}{l} y_{i,n+k} \in \{0, 1\} \quad , (i, n+k) \in A \\ y_{i,n+k} = 0 \quad , (i, n+k) \notin A \\ y_{n+k,j} \in \{0, 1\} \quad , (n+k, j) \in A \\ y_{n+k,j} = 0 \quad , (n+k, j) \notin A \end{array} \right\} \quad (7.26)$$

The set A represents the set of admissible trip joinings. The variables x_{ij} are the same as defined previously in problem P1, in which a solution with $x_{ij} = 1$ indicates that trips i and j are joined. Constraint (7.20) ensures that each required trip i is joined to exactly one successor trip. The successor trip may be either another required trip j or a dummy trip $n+k$. The latter implies that trip i is the last trip serviced by a vehicle before returning to depot k . Constraint (7.21) ensures that each required trip j must be joined to exactly one predecessor trip. The predecessor trip may be either a required trip i or a dummy trip $n+k$. If joined to the latter trip, j is the first trip serviced by a vehicle departing from depot k . The objective function (7.19) is identical to that of P1. From the definitions of $y_{n+k,j}$ and $y_{i,n+k}$, the first terms of (7.22) and (7.23) represent the following:

$\sum_{j \in I} y_{n+k,j}$ = the number of vehicles that start their daily schedule from depot k .

$\sum_{i \in I} y_{i,n+k}$ = the number of vehicles that end their daily schedule at depot k .

Both terms are also equal to the number of occupied overnight parking spaces (used capacity) at depot k .

Theorem 7.6: Problems P2 and P3 are equivalent. The P3 solution minimizes fleet size, subject to depot-capacity and depot-balance constraints.

Proof: Make the following substitutions in P3:

$$\left. \begin{array}{l} \sum_{j \in I} y_{n+k,j} = vs(k) \quad , k \in K \\ \sum_{i \in I} y_{i,n+k} = ve(k) \quad , k \in K \end{array} \right\}$$

Any solution satisfying (7.22) and (7.23) satisfies the balance constraint, since subtracting (7.23) from (7.22) for each k yields Equation (7.16):

$$vs(k) - ve(k) = 0, \quad k \in K$$

Constraints (7.22) and (7.23) are also equivalent to the depot constraints (7.17) and (7.18) when the non-negative slack variables z_{kk} are dropped. Summing up Equation (7.21) for all j shows that the objective functions for both problems are equivalent; i.e.

$$\max \sum_{j \in I} \sum_{i \in I} x_{ij} = n + \max \left(- \sum_{j \in I} \sum_{k \in K} y_{n+k,j} \right) = n - \min \sum_{k \in K} v_s(k)$$

It should be noted that P3 is a capacitated transportation problem, known in the OR field. As such, it has a bipartite graph representation similar to the example in Figure 7.5, but without s,t. The set of the first-column (supply) nodes are in one-to-one correspondence with the arrival epochs of trips $i = 1, 2, \dots, n$ and $n+1, n+2, \dots, n+k, \dots, n+q$. Similarly, the set of second-column (demand) nodes are in one-to-one correspondence with the departure epochs of each of the required and dummy trips. Since the supplies and demands are integer, the solution will be integer upon relaxing the integer requirements on the variables. Any of the standard transportation algorithms (see, for example, Ahuja *et al.* 1993) may be used to solve P3. However, it is possible that P3 is not feasible, since not every possible arc is admissible from the supply node set to the demand node set. Feasibility conditions are stated in the multi-terminal, supply-demand theorem found in Ford and Fulkerson (1962).

7.7 Literature review and further reading

Vehicle scheduling refers to the problem of determining the optimal allocation of vehicles to carry out all the trips in a given transit timetable. A chain of trips is assigned to each vehicle, although some of them may be deadheading (DH) or empty trips in order to reach optimality. The number of feasible solutions to this problem is extremely high, especially in the case in which the vehicles are based in multiple depots. Much of the focus of the literature on scheduling procedures is, therefore, on computational issues.

Dell Amico *et al.* (1993) developed several heuristic formulations, based on a shortest-path problem, that seek to minimize the number of required vehicles in a multiple-depot schedule. The algorithm presented is performed in stages, in each of which the duty of a new vehicle is determined. In each such stage, a set of forbidden arcs is defined, and then a feasible circuit through the network is sought that does not use any of the forbidden arcs. Computational efficiency is obtained by searching for the shortest path across a subset of all arcs in the network, rather than searching the entire network. Several modifications to the basic algorithms are offered that save computer time by substituting parts of the full problem with problems of a reduced size. These modifications include, for instance, solving the re-assignment of trips as a single-depot problem; an attempt to swap parts of duty segments; and an internal re-assignment of trips within each pair of vehicles associated with different depots.

Löbel (1998, 1999) discussed the multiple-depot vehicle scheduling problem and its relaxation into a linear programming formulation that can be tackled using the branch-and-cut method. A special multi-commodity flow formulation is presented, which, unlike most other such formulations, is not arc-oriented. A column-generation solution technique is developed, called Lagrangean pricing; it is based on two different Lagrangean relaxations. Heuristics are used within the procedure to determine the upper and lower bounds of the solution, but the final solution is proved to be the real optimum.

Kwan and Rahin (1999) described an object-oriented approach for bus scheduling, based on the VAMPIRES algorithm for iterative improvement of the solution presented by Smith and Wren in 1981. A key feature of VAMPIRES is the attempt to swap links at each stage of the

solution in order to improve the current schedule; Kwan and Rahin improved this feature by refining the swapping criteria. In addition, the methodology presented introduces a hierarchical classification of auxiliary activities: trip, layover, relocation, invalid layover, invalid relocation, depot return, depot start, depot end. This classification scheme for vehicle activities enables planners to improve the current solution more efficiently.

Mesquita and Paixao (1999) used a tree-search procedure, based on a multi-commodity network-flow formulation, to obtain an exact solution for the multi-depot vehicle scheduling problem. The methodology employs two different types of decision variables. The first type describes connections between trips in order to obtain the vehicle blocks, and the other relates to the assignment of trips to depots. The procedure includes creating a more compact, multi-commodity network-flow formulation that contains just one type of variables and a smaller amount of constraints, which are then solved using a branch-and-bound algorithm.

Banihashemi and Haghani (2000) and Haghani and Banihashemi (2002) focused on the solvability of real-world, large-scale, multiple-depot vehicle scheduling problems. The case presented includes additional constraints on route time in order to account for realistic operational restrictions such as fuel consumption. The authors proposed a formulation of the problem and the constraints, as well as an exact solution algorithm. In addition, they described several heuristic solution procedures. Among the differences between the exact approach and the heuristics is the replacement of each incorrect block of trips with a legal block in each iteration of the heuristics. Applications of the procedures in large cities are shown to require a reduction in the number of variables and constraints. Techniques for reducing the size of the problem are introduced, using such modifications as converting the problem into a series of single-depot problems.

Freling *et al.* (2001a) discussed the case of single-depot with identical vehicles, concentrating on quasi-assignment formulations and auction algorithms. A quasi-assignment is a reduced-size, linear problem in which some of the nodes and their corresponding arcs are not considered. An auction algorithm is an iterative procedure in which neither the primal nor the dual costs are obliged to show an improvement after each iteration. The authors proposed four different algorithms and compared their performance: an existing auction algorithm for the asymmetric assignment problem; a new auction algorithm for the quasi-assignment problem; an alternative, two-phase, asymmetric assignment formulation (valid in a special case), in which vehicle blocks are determined first and combined afterwards; and a core-oriented approach for reducing the problem size.

Freling *et al.* (2001b) and Huisman *et al.* (2005) presented an integrated approach for vehicle and crew scheduling for a single bus route. The two problems are first defined separately; the vehicle scheduling problem is formulated as a network-flow problem, in which each path represents a feasible vehicle schedule, and each node a trip. In the combined version, the network problem is incorporated into the same program with a set partitioning formulation of the crew scheduling problem (see Chapter 10).

Haase *et al.* (2001) formulated another problem that incorporated both crew and vehicle scheduling. For vehicle scheduling, the case of a single depot with a homogeneous fleet is considered. The crew scheduling problem (see Chapter 10) is a set partitioning formulation that includes side constraints for the bus itineraries; these constraints guarantee that an optimal vehicle assignment can be derived afterwards.

Haghani *et al.* (2003) compared three vehicle scheduling models: one multiple-depot (presented by Banihashemi and Haghani, 2000) and two single-depot formulations which

are special cases of the multiple-depot problem. The analysis showed that a single-depot vehicle scheduling model performed better under certain conditions. A sensitivity analysis with respect to some important parameters is also performed; the results indicated that the travel speed in the DH trip was a very influential parameter.

Huisman *et al.* (2004) proposed a dynamic formulation of the multi-depot vehicle scheduling problem. The traditional, static vehicle scheduling problem assumes that travel times are a fixed input that enters the solution procedure only once; the dynamic formulation relaxes this assumption by solving a sequence of optimization problems for shorter periods. The dynamic approach enables an analysis based on other objectives except for the traditional minimization of the number of vehicles; that is, by minimizing the number of trips starting late and minimizing the overall cost of delays. The authors showed that a solution that required only a slight increase in the number of vehicles could also satisfy the minimum late starts and minimum delay-cost objectives. To solve the dynamic problem, a ‘cluster re-schedule’ heuristic was used; it started with a static problem in which trips were assigned to depots, and then it solved many dynamic single-depot problems. The optimization itself was formulated through standard mathematical programming in a way that could use standard software.

Main features of the works reviewed are illustrated in Table 7.3.

Table 7.3 Characteristics of work covered in Section 7.8 concerning literature review and further reading

Source	Depots	Exact or heuristic	Special features
Dell Amico <i>et al.</i> (1993)	Multiple	Heuristic	
Löbel (1998, 1999)	Multiple	Both	
Kwan and Rahin (1999)	Single	Heuristic	Object-oriented program
Mesquita and Paixao (1999)	Multiple	Exact	
Banihashemi and Haghani (2000); Haghani and Banihashemi (2002)	Multiple	Both	
Freling <i>et al.</i> (2001a)	Single	Heuristic	
Freling <i>et al.</i> (2001b);			
Huisman <i>et al.</i> (2005)	Only one route	Exact	Vehicle and crew scheduling
Haase <i>et al.</i> (2001)	Single	Exact	Vehicle and crew scheduling
Haghani <i>et al.</i> (2003)	Multiple and single	Heuristic	Comparison of three formulations
Huisman <i>et al.</i> (2004)	Multiple	Heuristic	Dynamic scheduling

Exercises

7.1 Given a bus system consisting of only two cyclical routes (performing round trips) departing from and arriving at the same terminal. The mean trip time for **both** routes is given as **three hours**. The headways for each route are determined based on passenger demand. These headways are given by hour of day in the following table.

Route 1		Route 2	
Hour of day	Headway (minutes)	Hour of day	Headway (minutes)
6 a.m.–9 a.m.	7.5	6 a.m.–8 a.m.	6
9 a.m.–12 noon	12	8 a.m.–12 noon	7.5
12 noon–2 p.m.	10	12 noon–3 p.m.	10
2 p.m.–4 p.m.	15	3 p.m.–6 p.m.	7.5
4 p.m.–6 p.m.	6	6 p.m.–11 p.m.	10
6 p.m.–11 p.m.	10	–	–

What is the **minimum fleet size** required for this bus system? **Do not** use Deficit Function. Comment on the generalization of your approach.

Note: Buses may be switched from one route to another if necessary.

7.2 Given the following 8-trip, 4-terminal timetable and deadheading (DH) matrix. Find the minimum number of vehicles required to carry out the timetable, including their chains (blocks), by using the max-flow augmenting-path algorithm.

Trip number <i>i</i>	Departure terminal	Departure time	Arrival terminal	Arrival time
1	<i>a</i>	7:00	<i>b</i>	7:25
2	<i>d</i>	7:00	<i>a</i>	7:35
3	<i>b</i>	7:15	<i>c</i>	8:15
4	<i>c</i>	7:30	<i>d</i>	8:35
5	<i>a</i>	7:35	<i>d</i>	8:00
6	<i>d</i>	8:00	<i>a</i>	8:30
7	<i>a</i>	8:35	<i>d</i>	9:05
8	<i>c</i>	9:05	<i>d</i>	10:00

Average DH travel time (minutes) matrix

		Arrival terminal			
		<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>
Departure terminal	<i>a</i>	0	15	30	25
	<i>b</i>	20	0	40	30
	<i>c</i>	25	35	0	45
	<i>d</i>	20	25	35	0

- 7.3 The following problem uses the deficit function (DF) solution approach. Given a schedule S between 5:00 (T_1) and 9:30 (T_2), with two terminals (a, b) and one depot. An 8-trip schedule is shown in the table below. Note that trips 6 and 7 are conducted at the same time and in the same direction because of specific passenger demand. The deadheading (DH) average travel time for all trips between a and b (both directions of travel) is 30 minutes, and *there are no* DH trips between either a or b and the depot (both directions). The DH trips are to be assigned to the latest time possible.
- Construct the DFs for terminals a and b only; find the minimum number of vehicles required at terminals a and b using DH trip-insertion procedures between the two terminals.
 - Construct the sum function $g(t)$ for the entire schedule S ; Is $G(S)$ the minimum fleet size required for S ? Explain.
 - Find the minimum fleet size required for S and construct vehicle schedules (chains, blocs) using the FIFO rule.

Trip number	Departure terminal	Departure time	Arrival terminal	Arrival time
1	depot	5:00	<i>b</i>	6:00
2	depot	5:30	<i>a</i>	6:30
3	<i>a</i>	6:00	<i>b</i>	6:30
4	depot	6:40	<i>a</i>	8:00
5	<i>a</i>	7:00	depot	9:00
6	<i>a</i>	7:30	depot	9:30
7	<i>a</i>	7:30	depot	9:30
8	<i>b</i>	8:30	<i>a</i>	9:30

Note: Further exercises involving DFs, shifting departure times and DH-insertion procedures appear in the Exercises sections of Chapters 8, 9 and 15.

References

- Ahuja, R. K., Magnanti, T. L. and Orlin, J. B. (1993). *Network Flows*. Prentice Hall.
- Banihashemi, M. and Haghani, A. (2000). Optimization model for large-scale bus transit scheduling problems. *Transportation Research Record*, Issue Number **1733**, pp. 23–30.
- Bartlett, T. E. (1957). An algorithm for the minimum number of transport units to maintain a fixed schedule. *Naval Res. Logist. Quart.*, **4**, 139–149.
- Ceder, A. (1978). *Network Theory and Selected Topics in Dynamic Programming*. Dekel Academic Press (in Hebrew).
- Ceder, A. (2003). Public transport timetabling and vehicle scheduling. Chapter 2 in *Advanced Modeling for Transit Operations and Service Planning* (W. Lam and M. Bell, eds), pp. 31–57, Elsevier Ltd.
- Ceder, A. and Gonen, D. (1980). The operational planning process of a bus company. *UITP Review*, **29**, 199–218.
- Ceder, A. and Stern, H. I. (1981). Deficit function bus scheduling with deadheading trip insertion for fleet size reduction. *Transportation Science*, **15**(4), 338–363.
- Ceder, A. and Stern, H. I. (1985). The variable trip procedure used in the AUTOBUS vehicle scheduler. In *Computer Scheduling of Public Transport 2* (J. M. Rousseau, ed.), pp. 371–390, North-Holland Publishing Co.
- Daduna, J. R. and Paixao, J. M. P. (1995). Vehicle scheduling for public mass transit – an overview. In *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **430** (J. R. Daduna, I. Branco and J. M. P. Paixao, eds), pp. 76–90, Springer-Verlag.
- Dell Amico, M., Fischett, M. and Toth, P. (1993). Heuristic algorithms for the multiple depot vehicle scheduling problem, *Management Science*, **39**(1), 115–123.
- Desrosiers, J., Dumas, Y., Solomon, M. M. and Soumis, F. (1995). Time constrained routing and scheduling. In *Network Routing*. Volume 8 of *Handbooks in Operations Research and Management Science* (M. O. Ball, T. L. Magnanati, C. L. Monma and G. L. Nemhauser, eds), pp. 35–39, Elsevier Science B.V.
- Even, S. and Tarjan, R. E. (1975). Network flow and testing graph connectivity, *SIAM Journal on Computing*, **4**, 507–518.
- Ford, L. R. Jr and Fulkerson, D. R. (1962). *Flows in Networks*. Princeton University Press.
- Freling, R., Wagelmans, A. P. M. and Paixao, J. M. P. (2001a). Models and algorithms for single-depot vehicle scheduling, *Transportation Science*, **35**(2), 165–180.
- Freling, R., Huisman, D. and Wagelmans, A. P. M. (2001b). Applying an integrated approach to vehicle and crew scheduling in practice. In *Computer-aided Scheduling of Public Transport*. Lecture Notes in Economics and Mathematical Systems, **505** (S. Voss, S. and J. R. Daduna, eds), pp. 73–90, Springer-Verlag.
- Gavish, B., Schweitzer, P. and Shlifer, E. (1978). Assigning buses to schedules in a metropolitan area. *Computers and Operations Research*, **5**, 129–138.
- Gertsbach, I. and Gurevich, Y. (1977). Constructing an optimal fleet for transportation schedule. *Transportation Science*, **11**, 20–36.
- Gertsbach, I. and Stern, H.I. (1978). Minimal resources for fixed and variable job schedules. *Operations Research*, **26**, 68–85.
- Haase, K., Desaulniers, G. and Desrosiers, J. (2001). Simultaneous vehicle and crew scheduling in an urban mass transit system. *Transportation Science*, **35**(3), 286–303.

- Haghani, A. and Banihashemi, M. (2002). Heuristic approaches for solving large-scale bus transit vehicle scheduling problem with route time constraints. *Transportation Research*, **36A**, pp. 309–333.
- Haghani, A., Banihashemi, M. and Chiang, K. H. (2003). A comparative analysis of bus transit vehicle scheduling models, *Transportation Research*, **37B**, 301–322.
- Huisman, D., Freling, R. and Wagelmans, A. O. M. (2004). A robust solution approach to the dynamic vehicle scheduling problem. *Transportation Science*, **38**(4), 447–458.
- Huisman, D., Freling, R. and Wagelmans, A. O. M. (2005). Models and algorithms for integration of vehicle and crew scheduling. *Transportation Science*, **39**, 491–502.
- Kwan, R. S. K. and Rahin, M. A. (1999). Object oriented bus vehicle scheduling – the BOOST system. In *Computer-Aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **471** (N. H. M. Wilson, ed.), pp. 177–191, Springer-Verlag.
- Löbel, A. (1998). Vehicle scheduling in public transit and lagrangean pricing. *Management Science*, **44**(12), 1637–1649.
- Löbel, A. (1999). Solving large-scale multiple-depot vehicle scheduling problems. In *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **471** (N. H. M. Wilson, ed.), pp. 193–220, Springer-Verlag.
- Mesquita, M. and Paixao, J. M. P. (1999). Exact algorithms for the multi-depot vehicle scheduling problem based on multicommodity network flow type formulations. In *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **471** (N. H. M. Wilson, ed.), pp. 221–243, Springer-Verlag.
- Salzborn, F. J. M. (1972). Optimum bus scheduling. *Transportation Science*, **6**, 137–148.
- Salzborn, F. J. M. (1974). Minimum fleet size models for transportation systems. In *Transportation and Traffic Theory* (D. J. Buckley, ed.), pp. 607–624, Reed, Sydney.
- Smith, B.M. and Wrew, A. (1981) VAMPIRES and TASC: Two successfully applied bus scheduling programs. In *Computer scheduling for public transport* (A. Wrew, ed.), pp. 97–124. North-Holland Publishing.
- Stern, H. I. and Ceder, A. (1983a). The garage constrained-balance vehicle schedule minimum fleet size problem. In the *Proceeding of the 9th ISTTT – International Symposium on Transportation & Traffic Theory* (V. F. Hurdle, E. Hauer and G. N. Stewart, eds), pp. 527–556, University of Toronto Press.
- Stern, H. I. and Ceder, A. (1983b). An improved lower bound to the minimum fleet size problem. *Transportation Science*, **17**(4), 471–477.

Appendix 7.A: The maximum-flow (max-flow) problem

The following description is based on Ceder (1978). Further reading can be found in the fine book by Ahuja *et al.* (1993). The general augmenting-path algorithm described can solve the vehicle-scheduling problem. This Appendix supplements material for Section 7.4 and Chapter 12.

7.A1 Definitions

In the max-flow problem, one considers a directed graph (network) $G = \{N, A\}$, with a single node s as *source* and another node t as a *sink*. Each arc (i, j) has a capacity $c(i, j)$ = the maximum flow that can be traversed from i to j . This can be associated with a flow of people, vehicles, trains, bits of information, dollars, water, etc. An s - t flow is the amount of the flow that leaves s and arrives at t , provided that:

- (a) $0 \leq f(i, j) \leq c(i, j) \forall (i, j) \in A$,
- (b) and to maintain flow conservation, $\sum_{j|(i,j) \in A} f(i, j) = \sum_{k|(k,i) \in A} f(k, i) \forall i \in N, i \neq s, t$,
where $f(i, j)$ is the arc (i, j) 's flow.

Lemma 7.A1: Let f be an s - t flow on $G = \{N, A\}$. Then: $f(s, N) - f(N, s) = f(N, t) - f(t, N)$, where $f(U, V)$ is the flow between two subsets, U and V , such that $(U, V) = \{(i, j) \in A \mid i \in U, j \in V\}$ and $f(U, V) = \sum_{(i,j) \in (U,V)} f(i, j)$

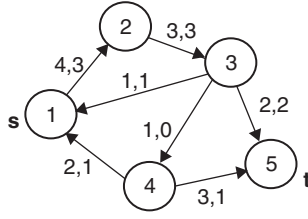
Proof: $f(N, N) = f(s, N) + f(t, N) + \sum_{\substack{i \in N \\ i \neq s, t}} f(i, N) = f(N, s) + f(N, t) + \sum_{\substack{i \in N \\ i \neq s, t}} f(i, N)$

Since the flow is conserved at each node, $f(i, N) = f(N, i) \forall i \in N, i \neq s, t$. Therefore, we obtain $V(f) = f(s, N) - f(N, s) = f(N, t) - f(t, N)$, where $V(f)$ is called the *value* of the flow.

Example:

The first number on each arc (see the network on page 204) represents its capacity and the second the flow. In this example, the s - t flow = 3 and can be measured at node 1 or 5.

s - t cut: If in a directed graph $G = \{N, A\}$ the set of nodes N is partitioned into two sets, X and $\bar{X}$ (i.e. $X \cap \bar{X} = \emptyset$ and $X \cup \bar{X} = N$), such that $s \in X$ and $t \in \bar{X}$, then $(X, \bar{X})$ = the set of all arcs in A connecting nodes in X to nodes in $\bar{X}$, is called the s - t cut. The capacity of the cut is defined as $c(X, \bar{X})$.

**Example:**

In the example network above:

$$\begin{aligned} X &= \{1, 2, 3\} \quad (X, \bar{X}) = \{(2, 4), (3, 5)\}, \quad c(X, \bar{X}) = 6 \\ X &= \{4, 5\} \quad \Rightarrow (\bar{X}, X) = \{(4, 1), (4, 3)\}, \quad c(\bar{X}, X) = 2 \end{aligned}$$

Lemma 7.A2: $V(f) = f(X, \bar{X}) - f(\bar{X}, X)$

Proof:

$$\begin{aligned} V(f) &= f(s, N) - (N, s) + \sum_{\substack{i \in \bar{X} \\ i \neq s}} [f(i, N) - f(N, i)] = f(X, N) - f(N, X) \\ &= [f(X, \bar{X}) + f(X, X)] - [f(\bar{X}, X) + f(X, X)] = f(X, \bar{X}) - f(\bar{X}, X) \end{aligned}$$

Theorem 7.A1: For any s-t flow f and any s-t cut $(X, \bar{X})$:

$$V(f) \leq c(X, \bar{X})$$

Proof: For every f and (x, x) : $f(X, \bar{X}) \leq c(X, \bar{X})$ and $f(X, \bar{X}) = 0$

Therefore, by Lemma 7.A2 $\rightarrow V(f) = f(X, \bar{X}) - f(\bar{X}, X) \leq c(X, \bar{X})$

Theorem 7.A1 leads to the fundamental result in network-flow theory: a flow exists that is equal to the minimum capacity cut; hence, it is the maximum flow. The latter claim will be shown following the description of the maximum-flow (max-flow) algorithm in the next section.

7.A2 The augmenting-path algorithm

For a given $G = \{N, A\}$ with $c(i, j) \forall (i, j) \in A$, and some s-t flow f (it is possible to start with $f(i, j) = 0 \forall (i, j) \in A$):

Step 1: Construct a *residual network* $G(f) = \{N, A(f)\}$ in which each arc is labelled with number $\alpha(i, j)$ as follows:

- (a) for $f(i, j) < c(i, j)$, then $(i, j) \in A(f)$ and $\alpha(i, j) = c(i, j) - f(i, j)$; (i, j) is called a *forward arc*;
- (b) for $f(i, j) > 0$ then $(j, i) \in A(f)$ and $\alpha(j, i) = f(i, j)$; (j, i) is called a *reverse arc*.

Step 2: If there is no path from s to t in $G(f)$, then terminate; f is the maximum s-t flow (see Theorem 7.A3 below).

Step 3: Let P be the shortest path (least number of arcs), and let $\alpha = \min_{(i,j) \in P} \alpha(i, j)$.

Define a new flow f' on G (see Theorem 7.A2 below):

- (a) if (i, j) is a forward arc on P , let $f'(i, j) = f(i, j) + \alpha$
- (b) if (i, j) is a reverse arc on P , let $f'(i, j) = f(i, j) - \alpha$
- (c) if (i, j) does not belong to P , let $f'(i, j) = f(i, j)$

Step 4: Let $f := f'$; go to step 1.

7.A3 Labelling procedure to find minimum-arc paths

In **Step 3** of the augmentation algorithm, we seek to find the shortest path from s to t . (An arbitrary path can also fit **Step 3**, but usually with more computations than the shortest path. The latter has a finite number, and normally fewer computations; hence, it is an improved version.) The labelling procedure is simple: at each stage, every node $i \in N$ is in one of three states: (i) unlabelled (indicated by a blank), (ii) labelled but not scanned (indicated by a single label, $l(i)$), and (iii) labelled and scanned (indicated by $l(i)$ and marked by $\checkmark$).

The procedure:

Step L_1 : Label the node s as $l(s) = s$

Step L_2 : If node t is labelled, stop; $[l(t), t]$ is the last arc on the shortest path from s to t . $[1(1, t), 1(t)]$ is the next-to-last arc, and so on until reaching node s .

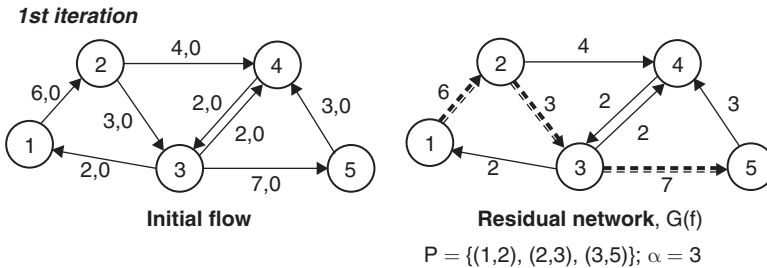
Step L_3 : If all labelled nodes are scanned, stop; no path exists from s to t .

Step L_4 : If all labelled, but non-scanned nodes i , use the 'first labelled, first scanned' policy. Label each unlabelled node j , provided $(j, i) \in A$, with $l(j) = i$. Mark i as a scanned node with a $\checkmark$.

Step L_5 : Go to **Step L_2** .

7.A4 Example of the augmenting-path algorithm

Given a network $G = \{N, A\}$ in which the first number on each arc represents the capacity, and the second the flow.

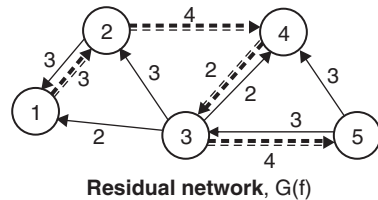
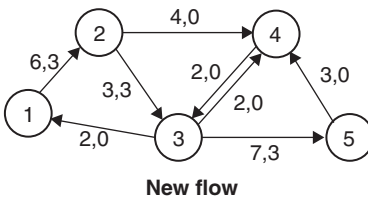


In order to demonstrate the labelling procedure, we will show how one can find P in the 1st iteration (it is used similarly in all other iterations).

Node	Iteration			
	1st	2nd	3rd	4th
$s = 1$	$1 = 1(1)$	$1 \checkmark$	$1 \checkmark$	$1 \checkmark$
2		$1 = 1(2)$	$1 \checkmark$	$1 \checkmark$
3			$2 = 1(3)$	$2 \checkmark$
4			$2 = 1(4)$	2
$t = 5$				$3 = 1(5)$

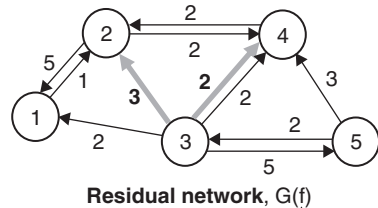
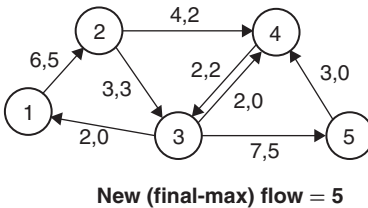
The shortest path is then reconstructed backward: $5 \rightarrow 3 \rightarrow 2 \rightarrow 1$ (see P above).

2nd iteration



$$P = \{(1,2), (2,4), (4,3), (3,5)\}; \alpha = 2$$

3rd iteration



because there is no more augmented s - t path

The final and optimum result is $\text{max-flow} = 5$. The bottleneck (minimum cut; see Theorems 7.A3 and 7.A4 below) consists of this max-flow, and is emphasized by a heavy grey line in the last residual network.

Note: When an augmenting path that includes a reverse arc (j^*, i^*) is detected, we can use it for increasing the s - t flow. However, according to **Step 3** (b) of the algorithm, we have to subtract the value of α from the previous flow on (i^*, j^*) in constructing the new flow.

7.A5 Additional theorems

The following theorems assure that the augmenting-path algorithm indeed solves the max-flow problem.

Theorem 7.A2: (a) the function f' (see **Step 3**) is an s - t (i.e. $0 \leq f'(i, j) \leq c(i, j) \forall (i, j) \in A$ and $f'(i, N) = f'(N, i) \forall i \in N, i \neq s, t$)
 (b) $V(f') > V(f)$

Proof: (a) The values of $\alpha(i, j)$ and $\alpha(j, i)$ are the maximum amount of flow with which we can increase or decrease $f(i, j)$, $(i, j) \in A$, respectively. The function f' is different from f only for $(i, j) \in P =$ augmenting path. Since $\alpha = \min_{(i,j) \in P} \alpha(i, j) = \alpha(q, k)$ then $f'(q, k) \leq c(q, k)$ and the capacity constraints hold for all other arcs in the augmenting path. Hence, $0 \leq f'(i, j) \leq c(i, j)$. The conservation flow required retains because for every node i along P (except s, t), the number of arcs entering i = number of arcs leaving i ; for (i, j) as a forward arc $f''(i, j) = f(i, j) + \alpha$; for (i, j) as a reverse arc $f''(i, j) = f(i, j) - \alpha$. (b) Similar to the arguments in (a), it is easy to show that: $f''(s, N) - f''(N, s) = f(s, N) - f(N, s) + \alpha$ and since $\alpha > 0 \rightarrow V(f') > V(f)$.

Theorem 7.A2 demonstrates the feasibility criterion in part (a) and the improvement criterion in part (b). The optimality criterion is shown in Theorem 7.A3.

Theorem 7.A3: If the augmenting-path algorithm is terminated with a flow f^* (see **Step 2**), then $V(f^*) = \min_{s-t \text{ cuts}(X, \bar{X})} c(X, \bar{X})$ and f^* has a maximum value.

Proof: Let Y be all the nodes belonging to the paths from s to i in the last augmented network (certainly $i \neq t$). Also, $\bar{Y} = N - Y$ and, hence, $t \in \bar{Y}$. Therefore, for all $(i, j) \in A$ such that $i \in Y$ and $j \in \bar{Y}$, $f(i, j) = c(i, j)$ (otherwise, we would have an augmented path between Y and $\bar{Y}$); and for all $(i, j) \in A$ such that $j \in \bar{Y}$, $j \in A$, $f(i, j) = 0$. Hence, $f(Y, \bar{Y}) = c(Y, \bar{Y})$ and $f(\bar{Y}, Y) = 0$.

For clarity, in the example above we obtain:

$$Y = \{1, 2, 4\}, \bar{Y} = \{3, 5\}$$

$$f(Y, \bar{Y}) = c(Y, \bar{Y}) = f(2, 3) + f(4, 3) = 5, f(\bar{Y}, Y) = f(3, 1) + f(3, 4) + f(5, 4) = 0.$$

Because $(Y, \bar{Y})$ is an s - t cut, then according to Lemma 7.A2:

$V(f^*) = f(Y, \bar{Y}) - f(\bar{Y}, Y) = c(Y, \bar{Y}) \geq \min_{s-t \text{ cuts}(X, \bar{X})} c(X, \bar{X})$. Thus, using Theorem 7.A1 ($V(f^*) \leq c(X, \bar{X}) \forall (X, \bar{X})$), we can show that $(Y, \bar{Y})$ is a minimum s - t cut and f has a maximum value.

Theorem 7.A4: $\min_{s-t \text{ flows } f} V(f) = \min_{s-t \text{ cuts}(X, \bar{X})} c(X, \bar{X})$, which is the known maximum-flow minimum-cut theorem.

Proof: By applying Theorems 7.A1 and 7.A3, the proof is straightforward.

7.A6 Linear programming (LP) formulation

For the sake of simplicity, let us add to $G = \{N, A\}$ an arc (t, s) with $c(t, s) = \infty$. This forms $G' = \{N, A'\}$ in which $A' = A \cup (t, s)$. If an arc (t, s) already exists, it can be removed without changing the problem. The LP formulation is as follows:

$$\max\{Z = f(t, s)\}$$

s.t.

$$\begin{aligned} f(i, N) - f(N, I) &= 0, \quad \text{for all } i \in N \\ f(i, j) - c(i, j) &\leq 0, \quad \text{for all } (i, j) \in A \\ f(i, j) &\geq 0 \quad \text{for all } (i, j) \in A' \end{aligned}$$

This LP has $|N|$ conservation constraints (for all $i \in N$), $|A|$ capacity constraints (one for each arc in G), and $|A'| = |A| + 1$ non-negativity constraints.

The LP formulation enables us to solve the max-flow problem by the simplex method (or one of its variants – see OR literature). However, we note that the coefficient matrix of the conservation constraints has a very special structure (this matrix is called the incidence matrix of G'). The rows and columns in this matrix correspond to the nodes and arcs in G' , respectively. Every column of this matrix contains exactly two non-zero elements: 1 and -1 . Because of this special structure of the matrix, it is possible to reduce significantly the computational effort required by the simplex method.

7.A7 Useful extensions

Undirected arcs

An undirected arc (i, j) with capacity $c(i, j)$ can be replaced by a pair of directed arcs (i, j) and (j, i) , each with a capacity of $c(i, j)$; that is,

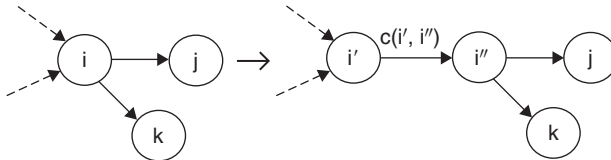


Unlimited capacities

Some arcs may have unlimited capacities, and the augmenting-path algorithm might end with an infinite flow. Therefore, in this case we calculate primarily $\max_{s-t \text{ paths}} \{\min \alpha_i\}$

Nodes with capacities

It is possible to establish an upper bound on the amount of flow traversing given nodes. In this case, the node i with capacity $c(i)$ is replaced by two new nodes i' and i'' with an arc having the capacity $c(i) = c(i', i'')$ in between; that is,

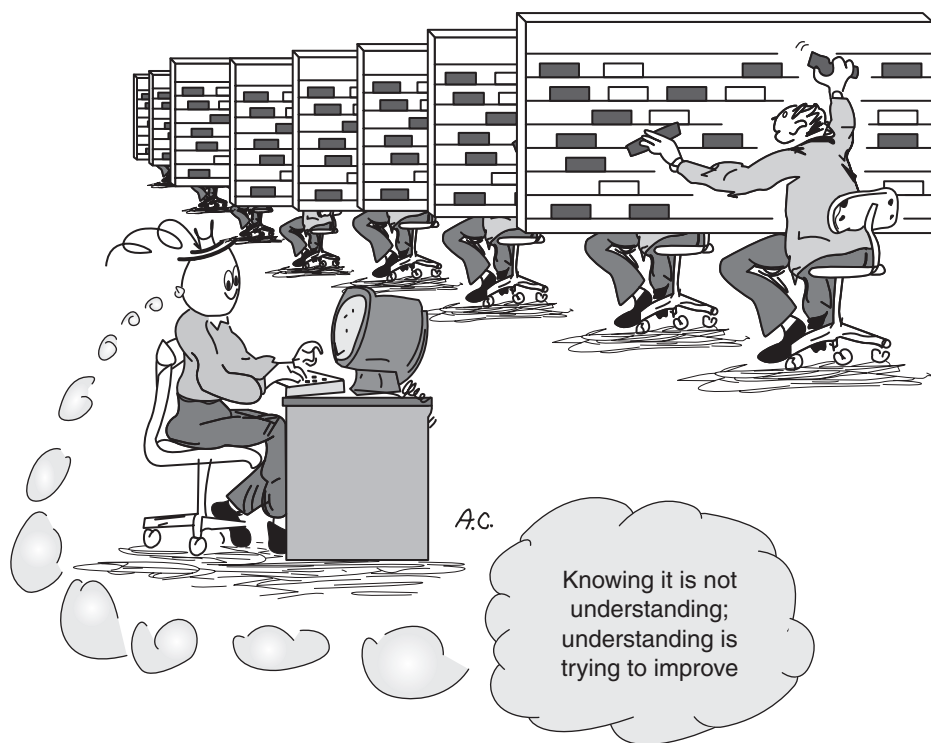


Lower bounds on the flow

For cases with a lower bound on the flow (e.g. assuring minimum flow to justify the permission to use the arc), the flow constraints are changed to $l(i, j) \leq f(i, j) \leq c(i, j)$, where $l(i, j)$ is the lower bound on the flow on arc (i, j) . This problem can be solved by converting the original network into an artificial network (satisfying both upper and lower bounds) called a transshipment scheme. Its solution appears in OR (Operations Research) literature, such as in Ahuja, *et al.* (1993).

This page intentionally left blank

8 Vehicle Scheduling II: Variable Schedules



Chapter 8 Vehicle Scheduling II: Variable Schedules

Chapter outline

- 8.1 Introduction
 - 8.2 Fleet-size lower bound for fixed schedules
 - 8.3 Variable trip-departure times
 - 8.4 Fleet-size lower bound for variable schedules
 - 8.5 Fleet-reduction procedures
 - 8.6 Experiences with bus schedules
 - 8.7 Examination and consideration of even-load timetables
 - Exercises
 - References
 - Appendix 8.A: Deficit-function software
-

Practitioner's Corner

This chapter extends the deficit-function approach to include possible modifications in the creation and editing of trip timetables and vehicle schedules (blocks). In addition to deadheading trip insertions, schedulers may consider a variable, instead of a fixed schedule; that is, shifting departure times based on some acceptable tolerances in the process of minimizing fleet size. Admittedly, this shifting effort is a very time-consuming task and usually not done in a systematic manner. This chapter will show how to use the graphical person-machine interactive approach to assist vehicle schedulers in selecting the most efficient shifts (in trip-departure times), with and without the inclusion of deadheading trip insertions. To that end, we can say that this will not be hard work or involve complex software that will wear out the scheduler. It's the knowledge that a small change can do the job.

The chapter contains six main parts, following an introductory section. Section 8.2 presents deficit-function procedures for fixed schedules in order to find rapidly the lower bound on the fleet size without creating vehicle schedules. This lower-bound calculation can help decision-makers in evaluating the number of vehicles required for any given change. Section 8.3 introduces the concept of maximum advance (early) and delay (late) from the scheduled departure time, given shifting tolerances (early, late) for each trip. Left (early) and right (late) shift limits are established, including the possibility of simultaneous left and right shifts in opposite directions. Section 8.4 provides an analysis and method of deriving the fleet-size lower bound when shifting departure times are allowed within their tolerances. Section 8.5 exhibits procedures in which a single or chain (multiple) of shifting is required to reduce the fleet size, with and without its integration with deadheading trip insertions. Section 8.6 illustrates the process of implementation through an example in which variable schedules are used to construct vehicle blocks in a large bus agency. Section 8.7 looks into the effect of shifting scheduled departure times on even-load timetables. A criterion for

maximum allowable change in the even-load is introduced. The chapter ends with exercises.

Although this chapter contains maths notations and proofs of derived conclusions, all sections are appropriate for practitioners. We especially recommend looking at the examples, illustrated in figures, to capture the interactive person-machine process. We end this corner, as we do almost customarily, with a story. A scheduler who couldn't insert deadheading trips because of only a few minutes difference and who made an effort not to increase the number of buses went to his supervisor. "Guess what solution I found (shifting)?" he asked. The supervisor told him that he reminded him of a new father who once told his friend: "I have a new born baby, guess what?" – "A boy", replied the friend, – "Guess again", said the father – "A girl". The new father exclaimed: "Who told you?"

8.1 Introduction

Chapter 7 dealt with fixed schedules, in which departure times could not be changed. In practical transit scheduling, however, schedulers should attempt to allocate vehicles in the most efficient manner possible, including the employment of small shifts in departure times. For example, in Egged, Israel's national bus carrier, the schedulers consider a variable instead of a fixed schedule in addition to deadheading (DH) trip insertions. That is, they consider shifting trip-departure times, based on some acceptable tolerances. Although confident of their work efficiency, the Egged schedulers agree that variable scheduling with DH trip insertions is a very complex process, not systematic, and a very time-consuming task; it can turn a good schedule into a soured one. This situation is similar to the following truth. If you put a spoonful of wine (an efficient chain of trips) into a barrel of sewage (problematic schedule), you get sewage. If you put a spoonful of sewage (a chain of trips impossible to execute) in a barrel full of wine (implemental schedules), you still get sewage (problematic schedule).

This chapter presents an extension of the deficit-function (DF) concept in an effort to provide an interactive computer technique to assist the scheduler in the planning task. Usually, only the timetable-construction and vehicle-scheduling (generation of chains or blocks) portions of the transit planning process are demonstrable. Given a multi-terminal trip schedule, the object is to minimize the required fleet size, subject to a number of constraints, including trip-departure time tolerances and insertable deadheading trips. Once the fleet size is determined, a number of alternative block constructions may be generated.

Another important feature, and one that will open this chapter, is the derivation of a fleet-size lower bound for both fixed and variable trip schedules. The DF lower-bound approach introduced can be used across almost all operational planning components. The importance of knowing how many vehicles are actually required applies to many transit-design and planning problems. Five such problems are as follows: (i) vehicle scheduling with different vehicle types (see Chapter 9); (ii) crew scheduling (see Chapter 10); (iii) design of operational transit-parking spaces (see Chapter 12); (iv) design of a new transit network or redesign of an existing one (see Chapters 12–14); and (v) design of efficient short-turn trips (see Chapter 15).

8.2 Fleet-size lower bound for fixed schedules

8.2.1 Overview and example

The initial lower bound on the fleet size with DH trip insertions was found by Ceder and Stern (1981) to be the sum of all DFs, $G(S)$, as shown in Figures 7.6 and 7.7 in Chapter 7. An improved lower bound was established later by Stern and Ceder (1983), based on extending each trip's arrival time to the time of the first feasible departure of a trip to which it may be linked or to the end of the finite time horizon. Further lower-bound improvements for fixed and variable schedules appear in Ceder (2002) and in Ceder (2005). The direct calculation of the fleet-size lower bound enables schedulers and transit decision-makers to ascertain more promptly how much the fleet size can be reduced by deadheading trip insertions and allowing shifts in departure times.

Figure 8.1 presents a nine-trip example with four terminals (a , b , c , and d). Table 8.1 shows the data required for the simple example used in this section and Sections 8.3 and 8.4 for demonstrating the lower-bound methods. Four DFs are constructed along with the over-all DF. According to the NT procedure (see Chapter 7), terminal d (whose first hollow is the

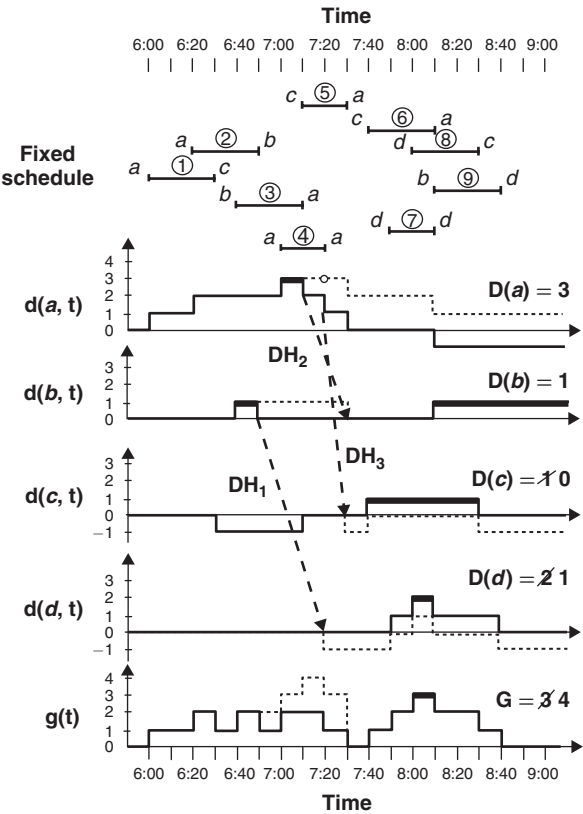


Figure 8.1 Nine-trip example with DH trip insertion for reducing fleet size

Table 8.1 Input data for the problem illustrated in Figure 8.1

Trip no.	Departure terminal	Departure time	Arrival terminal	Arrival time	Deadheading (DH) trips	
					Between terminals	DH time (same for both directions)
1	<i>a</i>	6:00	<i>c</i>	6:30	$a - b$	20 min
2	<i>a</i>	6:20	<i>b</i>	6:50		
3	<i>b</i>	6:40	<i>a</i>	7:10	$a - c$	10 min
4	<i>a</i>	7:00	<i>a</i>	7:20		
5	<i>c</i>	7:10	<i>a</i>	7:30	$a - d$	60 min
6	<i>c</i>	7:40	<i>a</i>	8:10		
7	<i>d</i>	7:50	<i>d</i>	8:10	$b - c$	30 min
8	<i>d</i>	8:00	<i>c</i>	8:30	$b - d$	30 min
9	<i>b</i>	8:30	<i>d</i>	9:00	$c - d$	20 min

longest) is selected for a possible reduction in $D(d)$. The DH-insertion process depicted in Figures 7.9(a) and 7.9(b) in Chapter 7 continues using the criterion $R = 2$. The first URDHC is $DH_1 + DH_2$, and the second DH_3 . The result is that $D(c)$ and $D(d)$ are reduced from 1 to 0 and from 2 to 1, respectively; hence, $N = D(S) = 5$, and G is increased from 3 to 4 using three inserted DH trips. The five FIFO-based blocks are as follows: [1-5- DH_2 -9], [2- DH_1 -7], [3- DH_3 -6], [4], [8].

8.2.2 Stronger fleet-size lower bound

While Stern and Ceder (1983) extended each unlinked trip's departure time, (i.e. one that cannot be linked to any trip's arrival time) to both T_1 and T_2 , Ceder (2002, 2005) proved that an extension only to T_2 was sufficient. The extension to the time of the first feasible departure time of a trip with which it may be linked, or to T_2 , results in an S' schedule and an overall DE , $g'(t, S')$, with its maximum value $G'(S')$.

While S' is being created, it is possible that several trip-arrival points are extended forward to the same departure point that is their first feasible connection. However, in the final solution of the minimum fleet-size problem, only one of these extensions will be linked to the single departure point. This observation provides an opportunity to look into further artificial extensions of certain trip-arrival points without violating the generalization of requiring all possible combinations for maintaining the fleet size at its lower bound.

Figure 8.2 illustrates three cases of multiple extensions to the same departure point. Case (i) shows two extensions, Trips 1 and 2, both with the same arrival point b , which is their first feasible connection at point a of Trip 3. Because only one of the two trips will be connected to Trip 3, the question is, Which one can be extended further? It is clear that Trip 1 has better DH

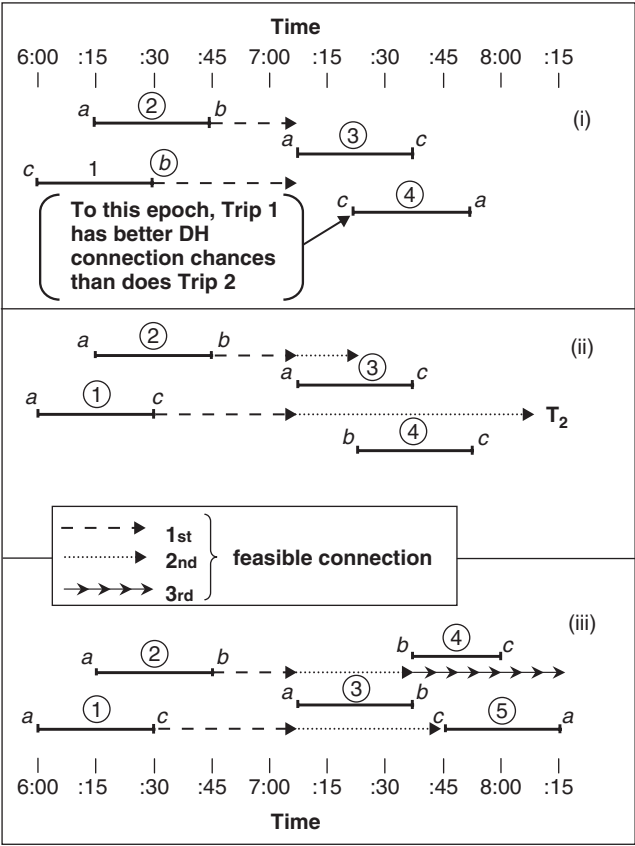


Figure 8.2 (i) shows why one should select the Trip 2 extension; (ii) shows that the argument in (i) cannot be used in case of multiple connections from different terminals; (iii) shows another case in which multiple connections cannot be applied for constructing the lower bound

chances to be connected to Trip 4 than to Trip 2 because of its longer DH time. Hence, Trip 1 can be further extended (2nd extension) to the start of Trip 4 if it is feasible. Case (ii), Figure 8.2, shows that Trips 1 and 2 do not end at the same point and that Trip 4 has different points than in Case (i). The argument of Case (i) cannot hold here, since the DH time differs between each two different points. In this case, the second feasible connection for Trip 1 is T_2 . By using the Case (i) argument, one can then create three possible chains [1], [2–3], [4], instead of two chains: [1–3], [2–4]. Case (iii) shows an opposite situation to that of Case (ii), with multiple extensions from different arrival points. If we link, in Case (iii), Trips 1 (longest DH time to the common departure point) and 3 and extend Trip 2 to Trip 4, we have another multiple extension case like Case (i), this one concerning the start of Trip 4 (linked to Trips 2 and 3). Following the Case (ii) argument, Trip 3 will be linked to Trip 4, and Trip 2 will have its third extension. This results in three possible chains: [1–3–4], [2], [5], instead of two: [1–5] and [2–3–4]. Cases (ii) and (iii) show why it is impossible to apply any general rule to a multiple extension of different arrival epochs. Consequently, further improvement of $G'(S')$ can be made only for Case (i) situations.

Following is the procedure for finding a stronger fleet-size lower bound.

1. Establish S' .
2. Select a case in which more than one extension is linked to the same departure time t_{sk}^j of trip j at terminal k . If no more such cases – STOP. Otherwise, select a group (two or more) of extensions with the same scheduled arrival terminal, u , and apply the following steps:
 - 2a. find a trip that fulfils: $\min_{i \in E_u} (t_{sk}^j - t_{eu}^i)$, where E_u = set of all trips arriving at u and extended to t_{sk}^j , and t_{eu}^i is the arrival epoch of trip i at terminal u ;
 - 2b. perform the second feasible extension for all trips $i \in E_u$, except the one selected in step 2a. Go to step 2.

Using this procedure, define the overall DF of the extended S' schedule by $g''(t, S'')$ with the maximum value $G''(S'')$. The following theorem and its proof establish that $G''(S'')$ is a stronger lower bound than $G'(S')$.

Theorem 8.1: Let $N_0(S)$ be the minimum fleet size for S with DH insertions. Let $G'(S')$ and $G''(S'')$ be the maximum value of the overall DF for S' and S'' , respectively. Then:

$$(i) \ G''(S'') \geq G'(S')$$

and

$$(ii) \ G''(S'') \leq N_0(S)$$

Proof: (i) The new overall DF, $g''(t, S'')$, has more extensions than $g'(t, S')$; i.e. $g''(t, S'') \geq g'(t, S')$. Therefore, $G''(S'') \geq G'(S')$.

(ii) According to the definition of S'' , at any time t in which $g''(t, S'') = G''(S'')$, there exist $G''(S'') - g'(t, S')$ trip extensions over S' . The additional extensions in S'' represent multiple extensions (2nd, 3rd, . . .), given that each extended trip is associated with another trip having the same arrival epoch and terminal, and has only one extension. In the optimal chain solution, a departure time t_s^* may or may not be linked to its nearest feasible arrival epoch (t_e^*) across all other points representing the same arrival terminal. Linkage to t_e^* complies with the procedure to construct S'' . Otherwise, t_e^* in S'' is further extended either to another trip or to T_2 while t_s^* is linked to $t_e^{**} < t_e^*$. We should note that t_e^{**} is linked to t_s^{**} when using the procedure described. Because t_e^* to t_s^* is the shortest link, the additional extension of t_e^* cannot be linked to a trip that starts before t_s^{**} (otherwise, t_e^{**} too will be linked to it, and not to t_s^*). Therefore, the additional extension of t_e^* in the optimal chain solution, $N_0(S)$, results in a greater overlap between trips (when constructing $g''(t, S'')$). Hence, $G''(S'') \leq N_0(S)$.

Figure 8.3 presents the schedule of Figure 8.1, with S' in its upper part, S'' in its middle part, and three overall DFs – $g(t, S)$, $g'(t, S')$, and $g''(t, S'')$ – in the lower part. For S' , it may be observed that Trips 3, 4 and 5 are extended to the same departure point as Trip 6 from the same arrival terminal a . According to the procedure for constructing S'' , the extension of Trip 5 is selected, and Trips 3 and 4 are further extended to the departure time of Trip 9. These additional extensions create another multiple connection associated with Trips 3 and 4, in which Trip 4 is the selected extension and Trip 3 is further (3rd time) extended. The initial lower bound is $G = 3$, the first improved lower bound is $G' = 4$, and the proposed improved lower bound is $G'' = 5$, which happens to be the optimal solution (see Figure 8.1).

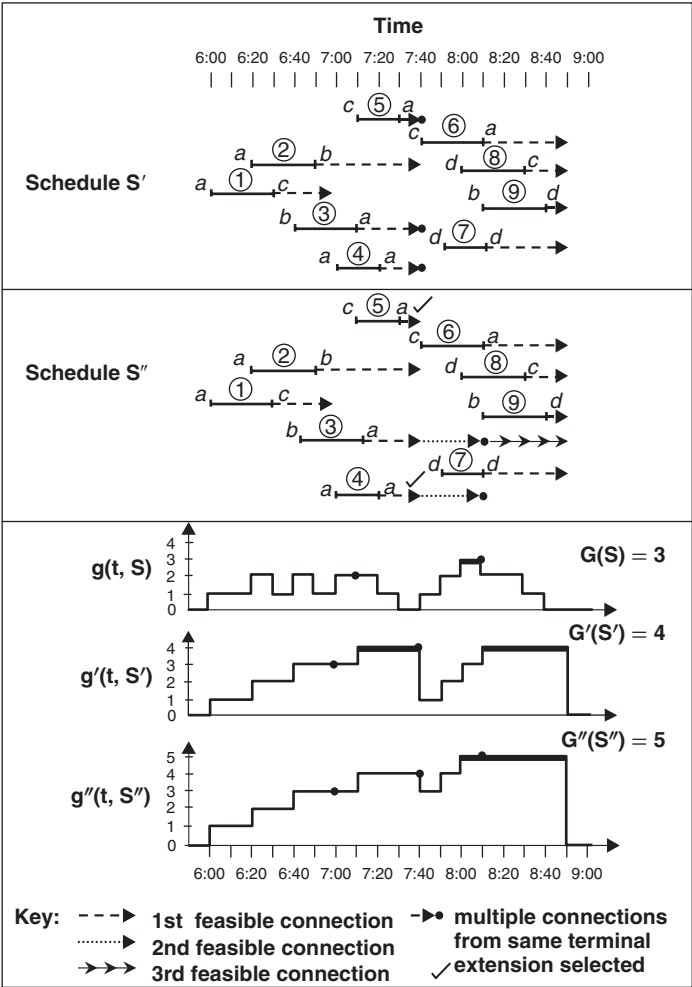


Figure 8.3 Lower-bound determination using the example shown in Figure 8.1, with the first and second improvement procedures

8.3 Variable trip-departure times

A small amount of shifting in scheduled departure times becomes almost common in practice when attempting to minimize fleet size or the number of vehicles required. This section presents methods, mostly according to the DF, to realize a variable trip schedule in an efficient manner. It may be recalled that Chapters 3 and 4 provided methods to improve the correspondence of vehicle departure times with passenger demand. The construction of timetables in Chapters 4 and 5 was based on either even headways or even average loads, entailing situations in which even headways result in uneven passenger loads. Shifting departure times, therefore, may unbalance these desirable features in the timetable while

favouring resource (vehicle) saving. Nonetheless, the last section of this chapter introduces a method to eliminate, at least to some extent, the possibility of too drastic changes in the even-load timetable requirement. Similar to the subject order of Chapter 7, we will start this section with reference to a single transit route and then continue with a minimum-fleet-size analysis for a network of routes.

8.3.1 Single-route minimum fleet size

Section 7.2 in Chapter 7 describes a method for ascertaining the minimum fleet size required for a given single route without interlining. We will now extend this method to account for possible shifting in departure times for given backwards and forwards shifting tolerances (in minutes) for each trip. This additional flexibility, which is employed in practice, can reduce the fleet size required as the primary objective. A secondary objective is to minimize the length of the shifting within their given tolerances.

In practice, departure times are shifted without any systematic method. Shifting tolerances are usually determined by rule of thumb although it makes sense to correlate them with the headways between departures. A proposed method appears in Table 8.2, in which the length of the shifting tolerance is headway dependent. That is to say, the longer the headway, the shorter is the tolerance. If the shifting is backward, the preceding headway is considered as H ; if it is forward, the next headway is considered.

Table 8.2 *Shifting tolerances as headway dependent*

Headway (H , in minutes)	Percentage of H for tolerance determination (%)	Tolerance length as H -dependent (minutes)
<10	50	0.5 H
10–20	40	0.4 H
21–40	30	0.3 H
>40	20	0.2 H

A new process needs to be designed for applying the shifting capability of departure times on single routes. This process simply attempts, through possible shifting of the relevant departure times, to reduce the minimum fleet size required. We will use the same notation for route r as in Section 7.2, but the symbol r is deleted, because it is clear which underlying route is being referred to. Thus, a and b are the end points; T_{ia} and T_{jb} are the average trip times on the route for vehicles departing at t_{ia} and t_{jb} from a and b , respectively, including layover time at their respective arrival points; n_{ia} is the number of departures from a between t_{ia} , in which departure ia is included, and $t_{i'a}$, in which departure $i'a$ is excluded. Trip ia arrives at terminal b , then continues with trip jb , the latter being the *first* feasible departure from b to a at a time greater than or equal to the time $t_{ia} + T_{ia}$; and $t_{i'a}$ is the *first* feasible departure from a to b at a time greater than or equal to $t_{jb} + T_{jb}$. Similar notations are defined for a trip starting from b .

Let $[t_{ia} - \Delta^{ia(-)}, t_{ia} + \Delta^{ia(+)}]$ be the tolerance time interval of the departure time of trip ia , in which: $\Delta^{ia(-)}$ = maximum advance of the trip's scheduled departure time (the case of

an early departure), and $\Delta^{ia(+)}$ = maximum delay from the scheduled departure time (the case of a late departure). Note that $t_{ik} + \Delta^{ik(+)} < t_{(i+1)k}$ and $t_{ik} - \Delta^{ik(-)} > t_{(i-1)k}$, for all $k \in K$. The minimum fleet size, $N_{\min}$, is then attained by construction, using the procedure illustrated in a flow diagram in Figures 8.4(a) and 8.4(b). The procedure described fits the case of Equation (7.1), in which $N_{\min} = \max_i n_{ia}$. For the case in which $N_{\min} = \max_j n_{jb}$ (determined by a trip starting from b), the same procedure is applied, but with b replacing a and j replacing i . The procedure first identifies the departure ia (or one of a few) referring

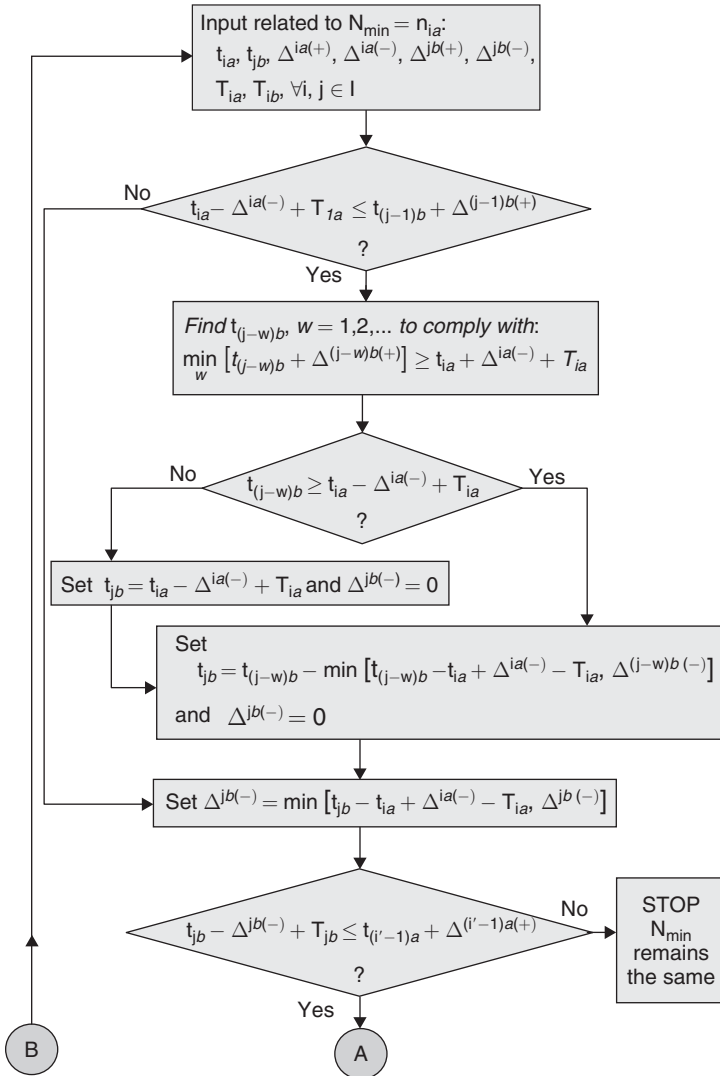


Figure 8.4(a) Flow diagram of the shifting departure-times process for reducing the minimum fleet size $N_{\min}$ determined at terminal a (for terminal b , the same process is used with a change of symbols)

to $N_{\min} = n_{ia}$; then it attempts through shifting t_{ia} to arrive at b before or at $t_{(j-1)b}$, and most important – to arrive before or at $t_{(i'-1)a}$. If the process manages to reduce n_{ia} by one or more units, it looks for the next $n_{ia} = N_{\min}$ or $n_{jb} = N_{\min}$ to continue. A successful process is that in which $N_{\min}$ is reduced. In addition, the procedure depicted in Figures 8.4(a) and 8.4(b) minimizes the length of shifting departure times, except for the shifting of the first departure, t_{ia} .

The interpretation of the shifting procedure may be assisted by the example in Figure 7.3 in Chapter 7. Here, $N_{\min} = 5$, resulting from the fifth and sixth departures from b . When b replaces a and j replaces i in Figures 8.4(a) and 8.4(b), we can then use the procedure described and start with the 6:50 b departure. Given $\Delta^{6:50b(-)} = 5$ minutes, then

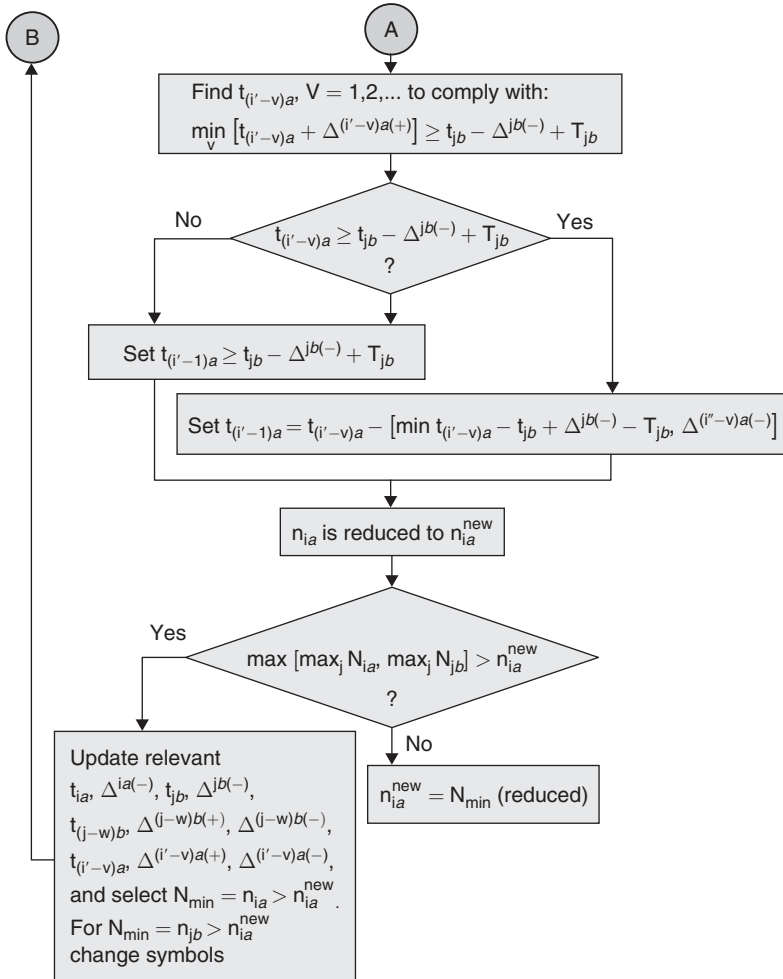


Figure 8.4(b) Flow diagram continued from Figure 8.4(a)

$\Delta^{7:00a(+)} = \Delta^{7:20b(+)} = \Delta^{7:15b(+)} = \Delta^{7:10b(+)} = 3$ minutes. The first check in the Figure 7.3 example, results in shifting 6:50 to 6:45 from b having $\Delta^{6:45b(-)} = 0$ minutes. Then from a , the first feasible connection is at 7:03 (including a forward tolerance). The second check, $7:00 \geq 6:45 + 15$, results in setting the departure time from a at 7:00 with $\Delta^{7:00a(-)} = 0$. The third check, $7:00 + 15 \leq 7:20 + 3$, leads to finding the first feasible connection to 7:00 to be from a . That is, $\min[7:20 + 3, 7:15 + 3, 7:10 + 3] \geq 7:15$ is 7:18. In the fourth check, $t_{(j'-1)b} = 7:15$. Hence, $n_{6:45b} = 3$, instead of the previously $n_{6:50b} = 5$.

We now move to the 7:05 departure from b , in which $n_{7:05b} = 5$. Given are $\Delta^{7:05b(-)} = 0$ minutes and $\Delta^{7:10a(+)} = \Delta^{7:20b(-)} = \Delta^{7:30b(+)} = 5$ minutes. In the first check, an early departure from a is impossible. The second check, $7:20 - 0 + 15 \leq 7:30 + 5$, results in setting $t_{(j'-1)b} = 7:35$, and $n_{7:05b} = 4$. The result is a multi-case of $N_{\min} = n_{ia} = n_{jb} = 4$, but without any further possibility of improvement, because $n_{7:05b}$ cannot be further reduced; this case is based on the procedure constructed in Figures 8.4(a) and 8.4(b).

8.3.2 Variable scheduling using deficit-functions

The transit scheduler who employs shifting in trip-departure times is not always aware of the consequences that could arise from these shifts. This section develops a formal algorithm to handle the complexities of shifting departure times. The algorithm is intended for both automatic and man-computer conversational modes.

According to the definitions in Section 7.5.1 in Chapter 7, s_j^k and e_j^k are the start and end of the j -th maximal interval, M_j^k , $j = 1, 2, \dots, n(k)$, at terminal k , $k \in K$, and N is the number of vehicles (chains, blocks) determined. If the length of M_j^k is denoted as $\overline{M}_j^k = e_j^k - s_j^k$, then s_j^k and e_j^k are associated with t_s^i and $t_e^{i'}$, respectively. That is, s_j^k refers to the departure time of a trip designated by I , and e_j^k to the arrival time of a trip designated by i' (where i, i' can be selected from several trips that depart at time s_j^k and arrive at e_j^k , respectively). The shifting terms defined in the foregoing section will also be used.

The nine-trip example illustrated in Figure 8.1 is used for possible shifting departure times in Figure 8.5, which employs the DF display. The tolerances of this example are $\Delta^{i(+)} = \Delta^{i(-)} = 5$ minutes for all trips in the schedule. Starting with shifting Trip 3 backwards and Trip 4 forwards by 5 minutes results in reducing $D(a)$ from 3 to 2. This may be continued with shifting Trips 7 and 8 to reduce $D(d)$ from 2 to 1.

Because no further shifting in departure times is feasible for the given tolerances, the process becomes one of searching for URDHC using DH trip insertion. This yields three DH trips resulting in $\text{Min } N = G(S^{\text{new}}) = 3$, in which S^{new} is the new schedule. The three blocks are determined by FIFO: [1-5-DH₂-9], [2-DH₁-7-8], and [3-4-DH₃-6]. In case a DH trip insertion is not allowed, the shifting process will end with $\text{Min } N = 5$ and the FIFO-based blocks: [1-5], [2-9], [3-4], [7-8], [9].

Another seven-trip example is depicted in Figure 8.6. It demonstrates the possible chaining effect of shifting departure times. In this example, $\Delta^{i(+)} = \Delta^{i(-)} = 10$ minutes. Hence, it is possible at the outset to reduce $D(b)$ by one unit through the shifts of t_s^3 to the right and t_s^4 to the left. However, these shifts increase $D(a)$, and the net saving is zero. Consequently, another iteration is needed in which t_s^7 is shifted to the right. Only then can a total saving be obtained of one vehicle. Given the desire to reduce a maximal interval M_j^k by shifting a maximum of two trips, let us consider the following three cases: (A) shift only trip i to the right, (B) shift only trip i' to the left, and (C) shift both trips i and i' in opposite directions (see Figure 7.6 for definitions used).

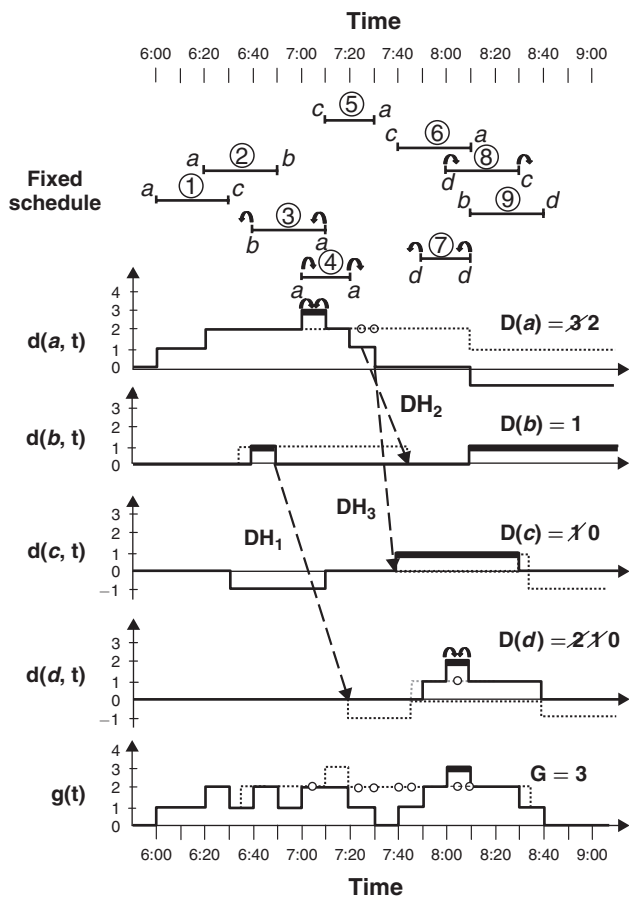


Figure 8.5 The nine-trip example (of Figure 8.1), first with shifting and second with DH trip insertion, for reducing fleet size

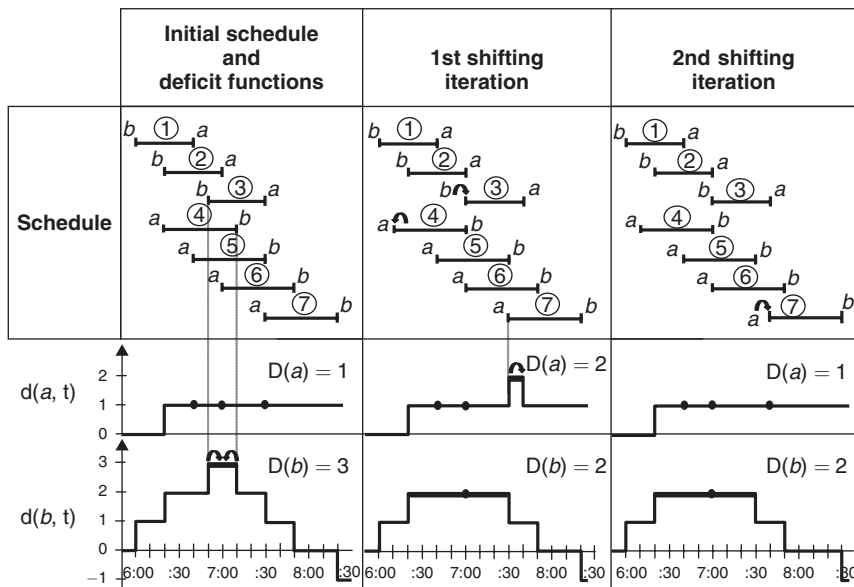


Figure 8.6 Example of two shifting iterations to reduce the required fleet size from 4 to 3

Case A: Right-shift limit

Let trip i (between terminals k and m), which starts at $t_s^i = s_j^k$, arrive at hollow H_{q-1}^m (preceding M_q^m) at time t_e^i . Shift trip i to the right as close to e_j^k as possible without increasing the maximal interval M_q^m or without exceeding $\Delta^{i(+)}$. Let this right-shift limit be defined as

$$\delta(+) = \min \left\{ s_q^m - t_e^i, \Delta^{i(+)}, \overline{M}_j^k \right\} \quad (8.1)$$

If $\delta = s_q^m - t_e^i$, then the shift has reached M_q^m and any further right shift will increase $D(m)$. If $\delta(+) = \Delta^{i(+)}$, then the shift is stopped by the tolerance limit of trip i . If $\delta(+) = \overline{M}_j^k$, then a successful shift has occurred and $D(k)$ is reduced by one.

Case B: Left-shift limit

Let trip i' (between terminals m and k) depart from hollow H_{q-1}^m at time $t_s^{i'}$. Shift trip i' to the left as close to s_j^k as possible, without increasing the maximal interval M_{q-1}^m or without exceeding $\Delta^{i'(-)}$. Let this left-shift limit be defined as

$$\delta(-) = \min \left\{ t_s^{i'} - e_{q-1}^m, \Delta^{i'(-)}, \overline{M}_j^k \right\} \quad (8.2)$$

If $\delta(-) = t_s^{i'} - e_{q-1}^m$, then the shift has reached M_{q-1}^m , and any further shift left will increase $D(m)$. If $\delta(-) = \Delta^{i'(-)}$, then the shift is stopped by the tolerance limit of trip i' . If $\delta(-) = \overline{M}_j^k$, then $D(k)$ is reduced by one.

Case C: Shift both trips

Without loss of generality, $D(k)$ can be reduced by shifting both trips i and i' in opposite directions. Assume that the procedure starts with an attempt to shift trip i to the right and is unsuccessful, $\delta(+) < \overline{M}_j^k$. Now perform Case B, with the length of $\overline{M}_j^k = \overline{M}_j^k - \delta(+)$, reduced from $\overline{M}_j^k$. Similarly the procedure can start with Case B and continue with Case A.

These three cases can be incorporated into a formal SDT (shifting departure time) algorithm. Two points should be noted: (i) if a unit reduction shifting chain (URSC) is allowed, as in Figure 8.6, the first criterion in Equations (8.1) and (8.2) is dropped (i.e. $s_q^m - t_e^i$ and $t_s^{i'} - e_{q-1}^m$); the process for URSC is finite, as we always move towards T_1 and/or T_2 and stop there (for an unsuccessful URSC). (ii) A criterion for minimum shifting can be established (i.e. a least square of the shifting length) in order to assure minimum deviation from scheduled departure times (Case C is affected by such a criterion). The next section analyses fleet-size lower bound with shifting departure times; this will be followed by a section on SDT algorithms.

8.4 Fleet-size lower bound for variable schedules

In regard to a lower bound on the fleet size, with allowed shifts in trip departure time, it is certain that where S can be shifted in $g(t, S)$, the lower bound will be equal to or less than $G(S)$. The main part of this section covers an algorithm that was devised for giving the opportunity to reach the lower bound through the combination of shifting and DH trips.

This procedure follows Ceder (2002, 2005), but with further adjustments and revisions. In addition, the end of the section presents a special lower-bound case in which only shifting, but not DH trips, is permitted.

8.4.1 Lower bound using shifting and DH trips

The algorithm for the lower bound consists of three phases. Phase A is an LB-SHIFT procedure for reducing $G(S)$. Phase B is an LB-DH&SHIFT procedure for constructing a new $g'(t, S')$ with shifting considerations; and Phase C is the construction of a new $g''(t, S'')$. Phase A involves the construction of a new temporary schedule, with shifts to be termed S_{sf} , that will end with $G(S_{sf}) \leq G(S)$. Phase B constructs a new temporary schedule with shifts combined with DH trips, to be termed S'_{sf} , that will end with $G'(S'_{sf}) \leq G'(S_{sf}) \leq G'(S')$. Note that $G'(S_{sf})$ is an extended schedule that follows the construction of $g'(t, S')$ in Section 8.2, but with an S_{sf} , in which no shifts are made in constructing $g'(t, S_{sf})$. Finally, Phase C applies the procedure for constructing $g''(t, S'')$ in Section 8.2 to S'_{sf} without any additional shifts. All notations used here are applicable to the procedures that follow.

Phase A: LB-SHIFT procedure

- Step 0 (initialization):** Define $M_j = [s_j, e_j]$, $j = 1, 2, \dots, n(g)$, and $\overline{M}_j = e_j - s_j$ as the j -th maximum interval and its length, respectively, of $g(t, S_{sf})$. Let $s_j = t_s^i$, $e_j = t_e^{i'}$, and $S = S_{sf}$, in which s_j refers to the departure time of a trip designated by i and e_j to the arrival time of a trip designated by i' . Set $j = 0$.
- Step 1 (selecting the next maximum interval):** Let $j = j + 1$; if $j > n(g)$ stop, otherwise continue.
- Step 2 (feasibility check):** If $\overline{M}_j \leq \Delta^{i'(+)} + \Delta^{i'(-)}$ continue, otherwise go to Step 1.
- Step 3 (right shift):** Compute $\delta(+) = \min \{\Delta^{i'(+)} + \overline{M}_j\}$ for $g(t, S_{sf})$; if $\delta(+) = \overline{M}_j$, go to Step 5, otherwise continue.
- Step 4 (left shift):** Compute $\delta(-) = \min \{\Delta^{i'(-)} + \overline{M}_j\}$ for $g(t, S_{sf})$; if $\delta(-) = \overline{M}_j$, go to Step 5; otherwise if $\Delta^{i'(-)} \geq \overline{M}_j - \Delta^{i(+)}$ (Case C for $g(t, S_{sf})$ in Section 8.3), set $\delta(-) = \overline{M}_j - \Delta^{i(+)}$, $\delta(+) = \overline{M}_j - \Delta^{i(+)}$ and go to Step 5. Otherwise, go to Step 1.
- Step 5 (reduce the lower bound):** Update S_{sf} and $g(t, S_{sf})$ with the confirmed shifts; go to Step 1.

Phase B: LB-DH&SHIFT procedure

- Step 0 (initialization):** Construct $g'(t, S_{sf})$ from $g(t, S_{sf})$. Define $M'_j = [s'_j, e'_j]$, $j = 1, 2, \dots, n(g')$, and $\overline{M}'_j = e'_j - s'_j$ as the j -th maximum interval and its length, respectively, of $g'(t, S_{sf})$. Let $s'_j = t_s^i$, $e'_j = t_e^{i'}$, and $S_{sf} = S'_{sf}$, in which s'_j refers to the departure time of a trip designated by i and e'_j to the arrival time of a trip designated by i' . Set $j = 0$.
- Step 1 (selecting the next maximum interval):** Let $j = j + 1$; if $j > n(g')$ stop, otherwise continue.
- Step 2 (right shift check):** If trip i of $s'_j = t_s^i$ can be shifted by $\Delta^{i(+)}$ (was not shifted in Phase A) and results in an increase in $G'(S_{sf})$, go to Step 4; otherwise continue.
- Step 3 (artificial shift):** Let t_s^i be shifted to the right by $\Delta^{i(+)}$ and included in S'_{sf} ; set $t_s^i = t_s^i + \Delta^{i(+)}$.

Step 4 (trip extensions with shifts): Let E be the set of all trips in which their artificial extension overlaps with M'_j , $i \notin E$; examine each trip u , $u \in E$, to determine whether its first feasible linkage point (departure time of a trip with which it may be linked) is t_s^i by performing a shift to the left ($\leq \Delta^{u(-)}$); if no linkage exists, stop; no improvement can be made. Otherwise continue.

Step 5 (new shifts and extensions): Perform for all $u \in E$ that can be shifted to the left (and linked to t_s^i) the new shifts and extensions to be included in S'_{sf} ; go to **Step 1**.

If Phase B indeed improves the lower bound, $G'(S'_{sf}) < G'(S_{sf})$, the foregoing S'_{sf} is then subject to Phase C. The latter has further extensions of arrival epochs if more than a single extension exists from the same terminal; this results in construction S''_{sf} from S'_{sf} according to the analysis of the construction of S'' in Section 8.2.

To better comprehend Phases A and B, the example problem of Figures 8.1 and 8.3 is presented. This problem undergoes these phases in Figure 8.7, with $\Delta^{i(+)} = \Delta^{i(-)} = 5$ minutes for all $i \in S$. Figure 8.7 shows the schedule and shifting required to construct $g(t, S_{sf})$ and $g'(t, S'_{sf})$. The process of constructing $g(t, S_{sf})$ exhibits Phase A,

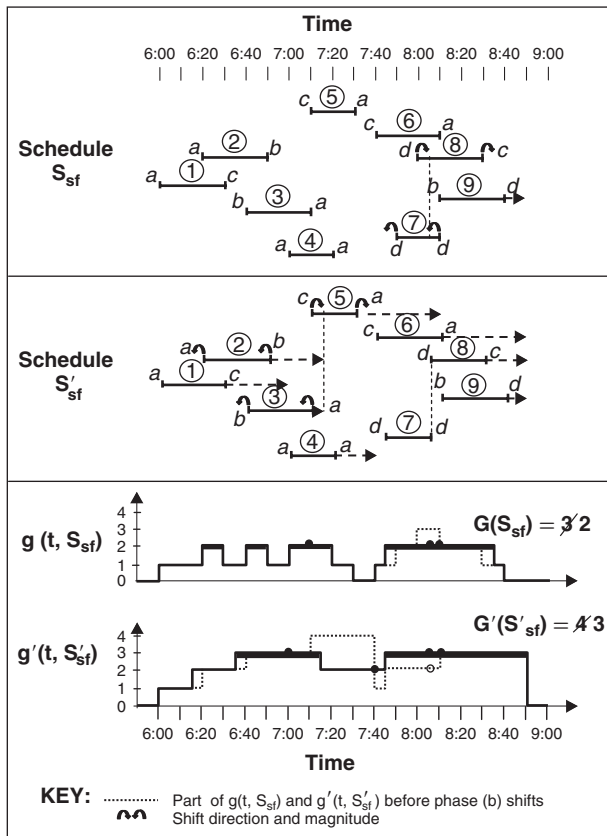


Figure 8.7 Lower-bound determination using the example in Figure 8.1, with LB-SHIFT and LB-DH&SHIFT procedures

in which two trips are shifted in opposite directions, Trips 7 and 8, to allow for $G(S_{sf}) = 2 < G(S) = 3$; this change is represented by a dotted line. The function $g'(t, S'_{sf})$ exhibits Phase B; the dotted line in $g'(t, S'_{sf})$ shows that the departure time of Trip 5 is the start of the only maximum interval that exists in $g'(t, S_{sf})$ following the two shifts in Phase A. Shifting Trip 5 to the right by 5 minutes will open up the possibility for Trips 2 and 3 to be extended to Trip 5 while being shifted to the left by 5 minutes, instead of extended as it was previously to Trip 6. This is not the case with Trip 4, whose original extension overlaps, too, with the maximum interval. Finally, it should be observed that by shifting Trip 5 to the right, it can no longer be extended to Trip 6, and its new extension is to Trip 9. However, this new extension does not create a new maximum interval (see Step 2 in the LB-DH&SHIFT procedure).

The final Phase C applies the procedure to attain $G''(S''_{sf})$ of the function $g''(t, S''_{sf})$. That is, to start with schedule S'_{sf} and allow further extensions of arrival points if more than a single extension exists from the same terminal. In the example shown in Figure 8.7, Phase C cannot be implemented. Assuming, however, that Trip 3 starts and ends at the same terminal b , then Trip 2 could be further extended to Trip 6 (the shortest extension of Trip 3 is the one selected to remain). Nonetheless, the latter extension of Trip 2, based on the above assumption, will not increase $G'(S'_{sf}) = 3$. This final lower bound of three is the optimal number of chains (blocks) that can be derived in the example problem. This compares with five chains with only a DH trip insertion. The shifting of departure times in the LB-SHIFT and LB-DH&SHIFT procedures can be used in the example to create the optimal three trip chains : [1-4-6], [3(shifted)-5(shifted)-9], [2-7(shifted)-8(shifted)], in which the shift of Trip 2 in Figure 8.7 is not required, because it links to Trip 7, not Trip 5.

A special scheduling case occurs when only shifting, but not a DH trip insertion, is permitted. In practice, this situation arises in transit systems without interlining and in which passenger demand does not differ greatly in either direction of the route (e.g. a rail operation). The lower-bound algorithm for this case follows the LB-DH&SHIFT procedure, but without a DH extension consideration. That is, each trip's arrival time is extended to the time of the first feasible departure time of a trip to which it may be linked, by having the same arrival and departure terminal as the extension, or to the end of the finite time horizon.

Let $\bar{S}'_{sf}$ and $\bar{S}''_{sf}$ be the extended schedules that are similar to S'_{sf} and S''_{sf} , but with a possible extension to link only the same terminals or to the end of the time horizon. We then obtain $G'(\bar{S}'_{sf}) \geq G'(S'_{sf})$ and $G''(\bar{S}'_{sf}) \geq G''(S'_{sf})$, and of course $G''(\bar{S}'_{sf}) \geq G'(\bar{S}'_{sf})$. Figure 8.8 uses the example in Figure 8.1 for constructing $\bar{S}'_{sf}$, which has the same results as $\bar{S}''_{sf}$. In this example, therefore, $G''(\bar{S}'_{sf}) \geq G'(\bar{S}'_{sf}) = 6$.

8.5 Fleet-reduction procedures

This section will describe some of the considerations in incorporating the SDT (shifting departure time) algorithm into the fleet-reduction procedures. The primary inputs required for the heuristic procedures are these:

1. an initial set of fixed trips, S^0 , defined over a set of terminals K ;
2. a tolerance matrix of trip-departure times;
3. a travel time matrix of potential DH trips (if applicable).

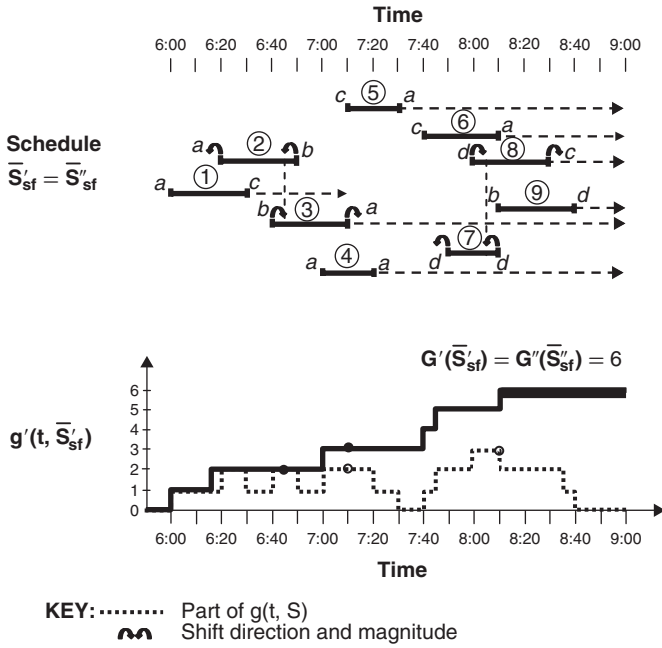


Figure 8.8 Lower-bound determination using the example in Figure 8.1 with the LB-EXT SHIFT procedures

The output of the algorithm will be a new vehicle schedule, including the shifts of trip-departure times, the set of DH trips inserted (if allowed), and the required number of vehicles at each terminal. The trips assigned to each vehicle can be constructed in a second phase using the FIFO rule or the chain-extraction procedure explained in Section 7.5.5 in Chapter 7.

8.5.1 SDT algorithm

- Step 0 (initialization):** Let E represent the set of unexamined terminals; set $E = K$.
- Step 1 (select the next terminal):** If $E = \Phi$ stop. Otherwise, select a terminal k from E and remove it; update $E = E - \{k\}$; $j = 1$.
- Step 2 (select the next maximum interval):** Let $j = j + 1$; if $j > n(k)$, go to **Step 6**.
- Step 3 (feasibility check):** If $\bar{M}_j^k \leq \Delta^{i(+)} + \Delta^{i'(-)}$ continue, otherwise go to **Step 1**.
- Step 4 (right shift):** Compute $\delta(+)$; if $\delta(+)=\bar{M}_j^k$, go to **Step 6**, otherwise continue.
- Step 5 (left shift):** Compute $\delta(-)$, and if $\delta(-)=\bar{M}_j^k-\delta(+)$, go to **Step 6**; otherwise, set $\bar{M}_j^k-\bar{M}_j^k-\delta(+)$ and go to **Step 4** (see Case C in Section 8.3).
- Step 6 (reduce the fleet):** Perform all shifts and update the DF; if $G(S_{sf})=D(S)$, or $G''(S_{sf}'')=D(S)$ in the case with DH trips, stop. Otherwise, go to **Step 1**.

The foregoing traces the basic structure of the SDT algorithm. In the operational algorithm, the order in which terminals and maximum intervals are selected for examination is

determined by various heuristic priority rules. As noted at the end of Section 8.3, a criterion for minimum shifting can be established and will affect Steps 3–5 in the SDT algorithm. Chaining shifts, as shown by the example of Figure 8.6, constitutes a second-level phase examination of all the terminals. This is not described here in order to make understanding of the underlying basic principles of the algorithm easier.

Another possibility is to consider variable trip-departure times along with the DH trip-insertion procedure. In that case, the feasibility requirement shown in Equation (7.8) in Chapter 7 for a feasible joining of two trips, i and j , is changed to

$$t_e^i - \Delta^{i(-)} + \tau(q^i, p^i) = t_s^j + \Delta^{j(+)} \quad (8.3)$$

in which $\tau(q^i, p^i)$ is the DH time of trip i from terminal q to terminal p .

8.5.2 The order of sub-routines

In the variable-schedule problem, two major sub-routines are identified in order to reduce the fleet size: (a) shifting trip-departure times (SDT) within their tolerances; and (b) searching for a unit reduction DH trip chain (URDHC). In some practical cases, DH trips are not allowed and only an SDT algorithm is utilized. The URDHC procedure could be made both with and without consideration of the SDT sub-routine. Following is a discussion of these sub-routines.

The selection of the SDT sub-routine needs to be examined primarily in light of the maximum possible saving of vehicles. Figure 8.9 illustrates an example of three terminals using the DF representation. The first column, ‘URDHC start’, begins with the URDHC procedure and ends with a modified URDHC (mixed with the SDT) procedure. The second column, ‘SDT start’, uses first the SDT procedure and second the modified URDHC procedure. In this example, $\Delta^{i(+)} = \Delta^{i(-)} = 5$ minutes for all trips i in the schedule; the DH trip time for all possible DH trips is 10 minutes. It is also assumed that the departure times for trips shifted to the left and the arrival times for trips shifted to the right are outside the time scale for the SDT procedure.

The first-column solution succeeds in reducing $D(a)$ from 4 to 2 in Step 1 of Figure 8.9. In Step 2, $D(b)$ is reduced by one through four shifts of trip-departure times and two DH trips. The final schedule can be constructed from the DFs in the last step, in which $D(S)$, the fleet size, has been reduced from 11 to 8 vehicles. In the second-column solution (SDT start), $D(b)$ is reduced from 4 to 3 by four trip-departure time shifts. Then in Step 2, $D(a)$ is reduced from 4 required vehicles to 2 through two shifts of the trip-departure time and two DH trips. The final schedule has $D(S) = 8$.

Note that in the final DF configuration in both columns of Figure 8.9, no further reductions in fleet size can be achieved based on the formal SDT algorithm and on the feasibility requirement for DH trip insertion indicated in Equation (8.3). In consideration of the secondary objectives, it appears that the second column has priority over the first column (4 shifts and 2 DH trips as against 4 shifts and 6 DH trips, respectively). Intuitively, it seems worthwhile to start with the SDT procedure. However, the next example shows that this is not always the case.

Figure 8.10 presents two examples with shift tolerances and a DH trip time as in Figure 8.9. In Example 1, $d(a, t)$ and $d(b, t)$ are the same as in Figure 8.9, but a small change is

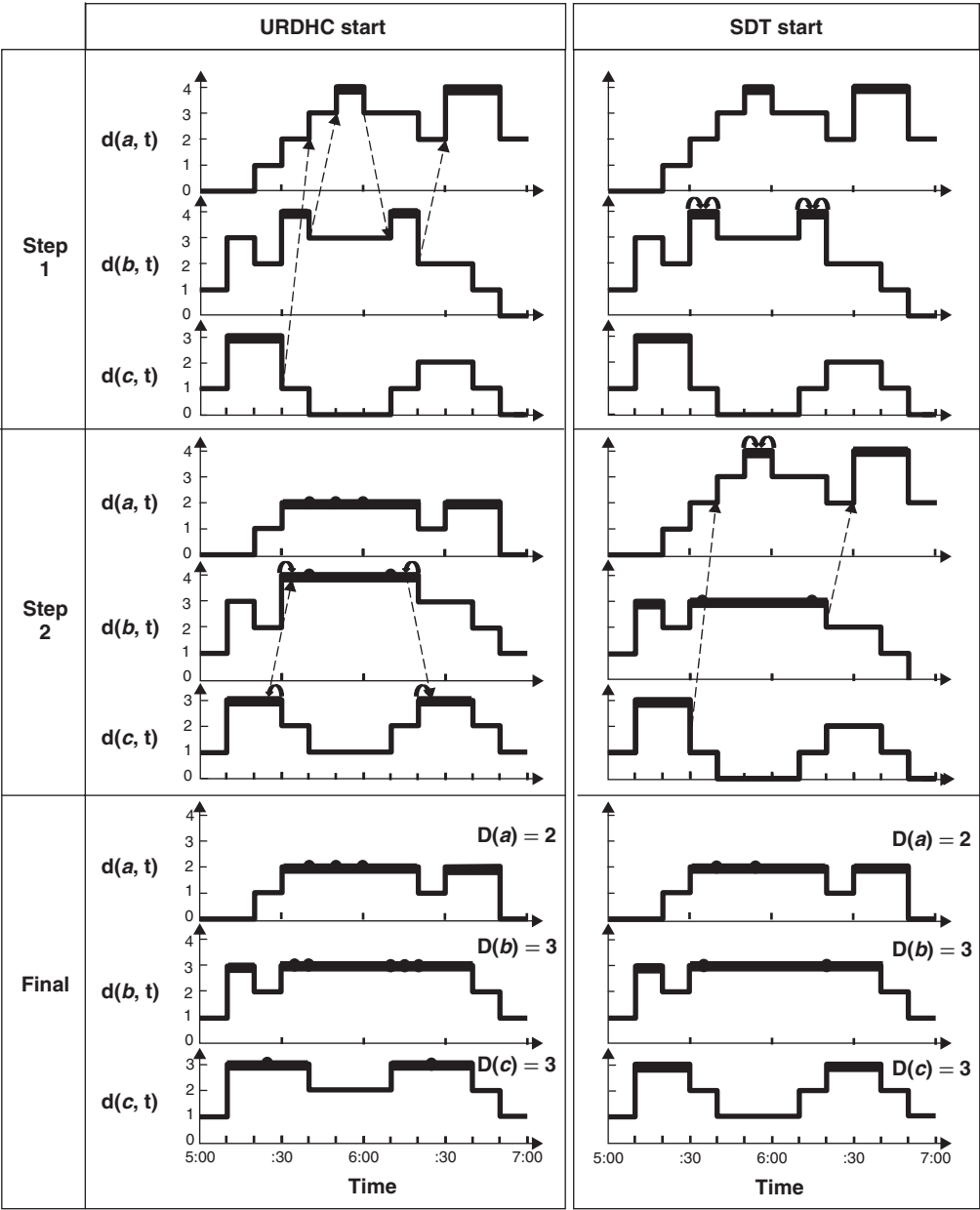


Figure 8.9 Two solutions for reducing fleet size; the first starts with the URDHC sub-routine, and the second with the SDT sub-routine; both continue with the modified URDHC procedure

introduced in $d(c, t)$. The trip, which departs at 6:20, moves to depart at 6:10 from terminal c . In Example 1, only two vehicles are saved (see Figure 8.10) by starting with the URDHC procedure. This is in comparison with a saving of three vehicles if we begin with the SDT procedure (see the better solution indicated under Example 1 in Figure 8.10). In Example 2,

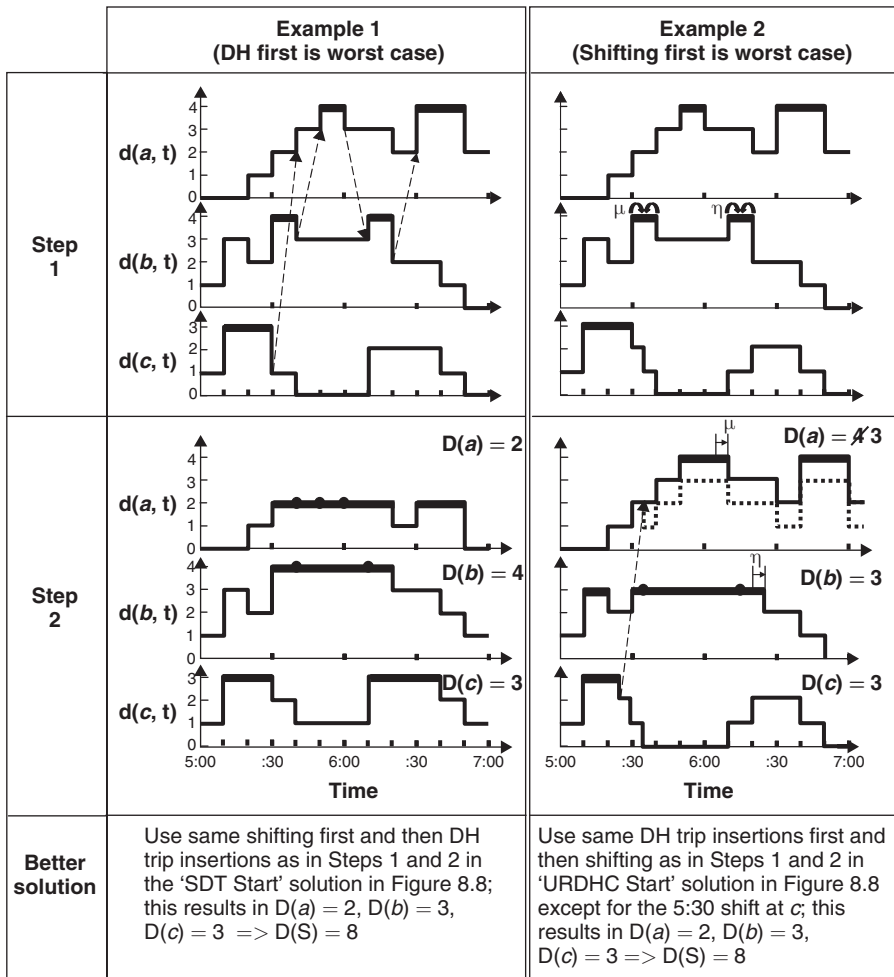


Figure 8.10 Two examples demonstrating no explicit priority for starting with either the URDHC or the SDT procedure

$d(a, t)$ and $d(b, t)$ are the same as in Figure 8.9; and in $d(c, t)$, one trip arrives at 5:35 instead of 5:30. In this example, we relax the assumption that the SDT procedure influences departure and arrival times outside the time scale (though this assumption holds in Example 1).

In Step 1 of Example 2, in Figure 8.10, four shifts are illustrated. The shifts that are denoted by μ and η (forward shifts) affect the DFs at terminals a and b , respectively. Shift μ refers to a trip between terminals b and a with a trip time of 30 minutes, and shift η refers to a round trip from terminal b with a length of 10 minutes. Thus a DH trip between terminals c and a reduces $D(a)$ by one. Consequently, we can see in Step 2 (Example 2) in Figure 8.10 that by starting with the SDT procedure, it is possible to save only two vehicles at terminals b and a . This compares to a saving of three vehicles when applying the better solution indicated in Example 2 in Figure 8.10.

A DH departure has the effect of increasing the departure terminals' DF and also of reducing the deficit of some maximum intervals at their destination terminals (either by reducing the fleet requirement or by compensating for an earlier DH trip departure). In this way, maximum intervals are levelled and expanded at both the departure and the arrival terminals. This reduces the opportunities for a further reduction of the fleet size through the SDT procedure. On the other hand, similar consequences are observed when starting with the shifting of trip-departure times. That is, the SDT procedure generally reduces the opportunities for a further reduction of the fleet size through DH trips (although some shifts might open up new opportunities because of a larger time interval for a DH trip insertion).

8.5.3 Description of the procedure

For the fleet-reduction procedure, when DH trip insertion is allowed, we use a heuristic rule and start with the SDT sub-routine, followed by the modified URDHC (mixed with SDT) sub-routine. This rule can represent to some extent the viewpoint of the transit agency that wishes to minimize operational costs.

The basic fleet-reduction procedure that contains only the SDT algorithm is presented in the flow diagram in Figure 8.11 and is designed for a person-computer interactive system. It does not interact with the DH trip-insertion process. The selection of a terminal u can be made by the scheduler by inspecting the DFs on a graphical display. The search for a reduction in the fleet requirement at terminal u (see Figure 8.10) can be performed manually by the scheduler or by procedures based on the formal SDT algorithm. If a unit reduction shifting chain (URSC) is found, all DF information is updated for this new schedule, and a new iteration initiated. The SDT sub-routine continues until either the lower-bound test is successful or all terminals are examined.

The modified URDHC (mixed with SDT) sub-routine, shown schematically in a flow diagram in Figure 8.12, is based on the URDHC procedure described in Section 7.5.4 in Chapter 7. The upper part of the flow diagram is the SDT sub-routine described in Figure 8.11. In the lower part of Figure 8.12, the lower-bound determination suggests that $G''(S''_{sf})$ should be examined before trying to insert DH trips into the schedule. In this modified procedure, the feasibility requirement for DH trip insertion is based on Equation (8.3). Another point that should be mentioned is that DH trips added to schedule S should include a shifting tolerance for the next iterations of the SDT and modified URDHC procedures. Finally, if a URDHC has been found, the DH chain cost involved is compared with the saving cost of a single vehicle. If the DH cost is higher than the saving cost, the URDHC is cancelled. Otherwise, the set of DH trips, designated $\{DH\}$, is added to the previous schedule, and all DF information, including shifting, is updated for a new schedule. The modified URDHC sub-routine continues with updated S until $G''(S''_{sf}) = D(S)$ or until all terminals are examined.

The foregoing sub-routines are implemented in a practical algorithm appearing on a website: www.altdoit.com and explicated in Appendix 8.A. The interaction with this algorithm ostensibly can give rise to the graphical person-machine dialogue in either a manual or an automatic mode. This computer program also contains the search for a URSC, in which a shift in departure time at a given terminal increases the maximum interval for another terminal.

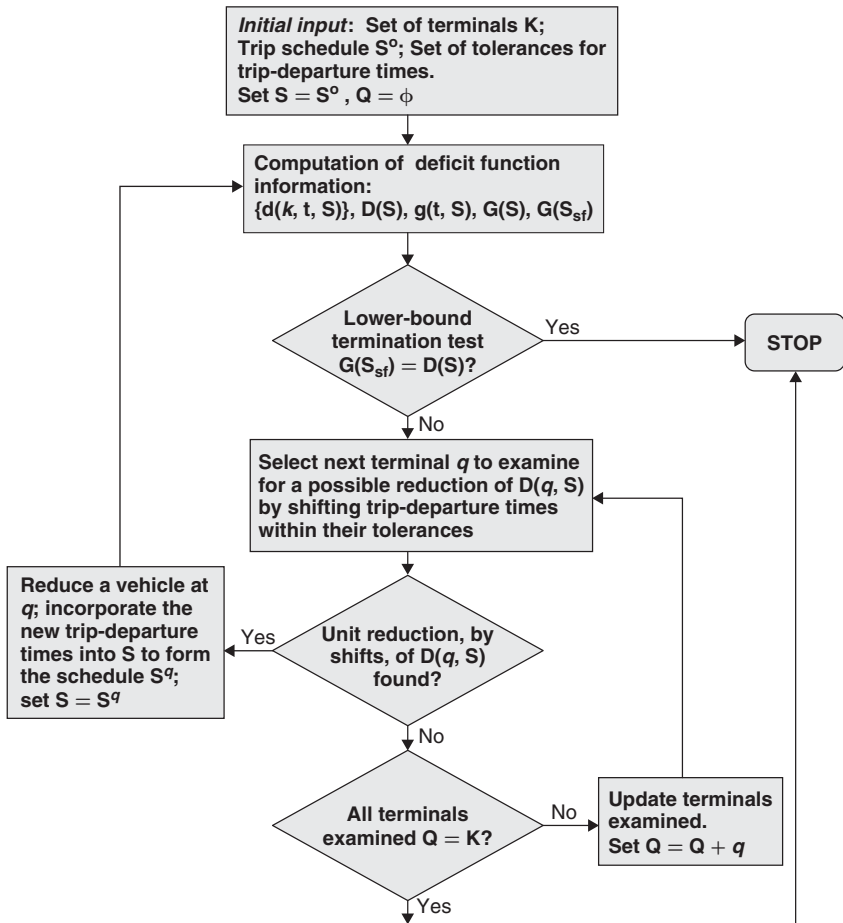


Figure 8.11 Flow diagram of a fleet-reduction procedure involving only shifting of departure times (SDT algorithm)

8.6 Experiences with bus schedules

As mentioned in the Introduction, fundamental problems in designing variable schedules arise from practice. An example of a transit agency that faced these problems is Egged, Israel's national bus carrier, which was discussed in Section 7.3 in Chapter 7. The Egged fleet of about 4,000 buses operates over a spread-out national network with some 2,000 routes; it performs an average of 50,000 daily trips, among them 12,000 DH trips. Bus schedules are produced by about 60 schedulers using Gantt charts through a trial-and-error approach. The need for a quicker response to timetable changes has led the Egged management to investigate the use of a fully computerized system. As is explained in Section 7.3, attempts to implement the computer-generated schedule have failed because of an inability to meet a number of necessary practical constraints.

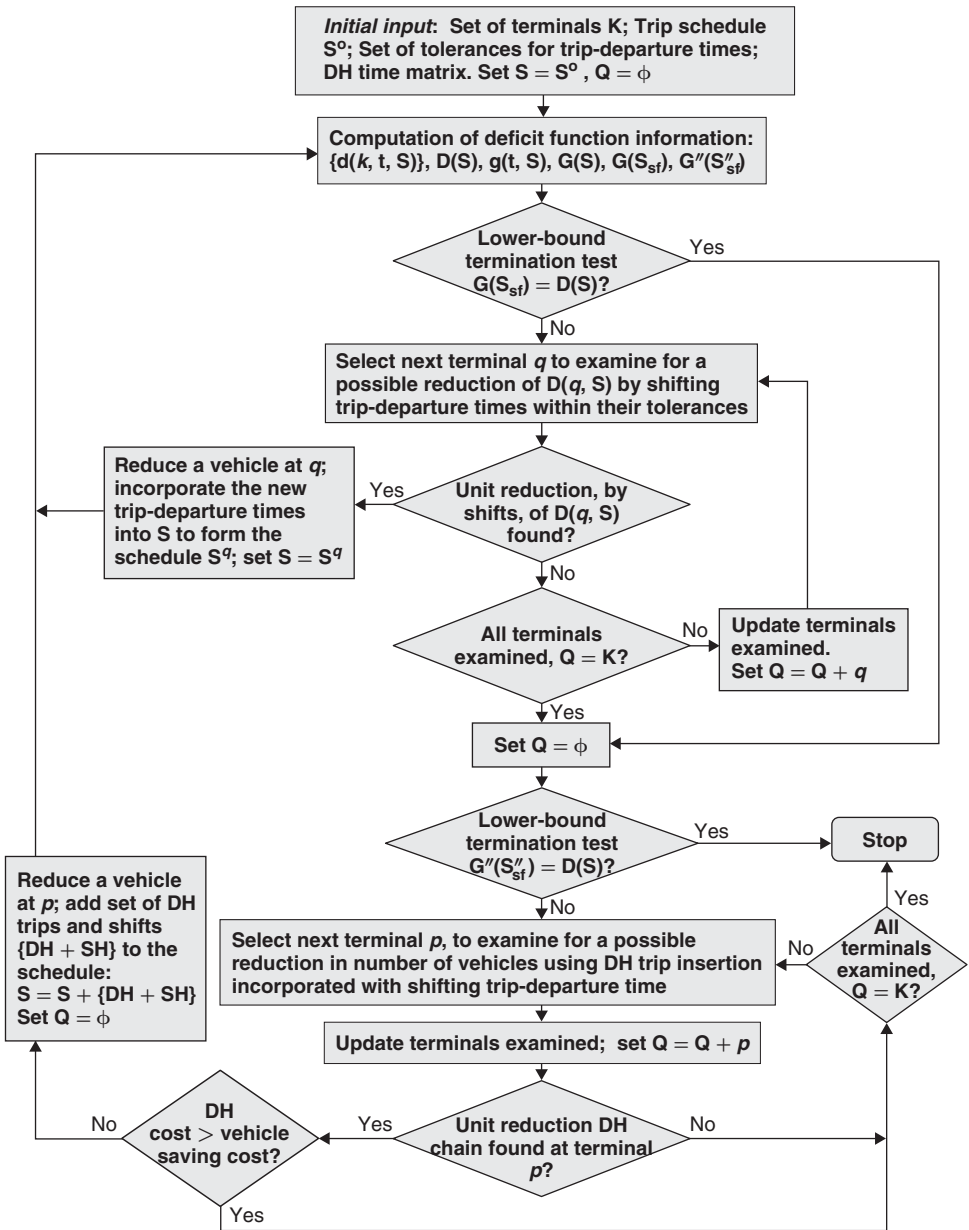


Figure 8.12 Flow diagram of fleet-reduction procedure involving shifting of departure times and DH insertions (modified URDHC sub-routine)

It was therefore decided to continue the search for an approach that would combine the advantages of modern electronic computers while, at the same time, allow the scheduler to make his or her own contribution to the scheduling task. Because of its visual nature, a DF approach was selected for use with a person-machine interactive system. The implementation

of the DF approach was introduced gradually so that the schedulers could gain confidence in this approach and reach the conclusion that this method was very useful for increasing the speed and accuracy of the scheduling tasks.

Two simple, real-life examples are described below to demonstrate the implementation stage at Egged. In the first example, illustrated in Figure 8.13, the schedulers claimed at first that it was impossible to further reduce fleet size from their Gantt chart scheduling results. Figure 8.13 shows only a small part of the Gantt chart and the corresponding DFs, those that are undergoing change. The schedulers allow for an acceptable shift in trip-departure time for all trips by $\Delta^{i(+)} = \Delta^{i(-)} = 3$ minutes.

By illustrating $d(k, t)$, however, we saved one bus by shifting six trips, each by 3 minutes. Before the changes in Figure 8.13, trips are identified in the Gantt chart by letters. This is for the sake of clarity when referring to shifts and for reconstructing the Gantt chart chains after

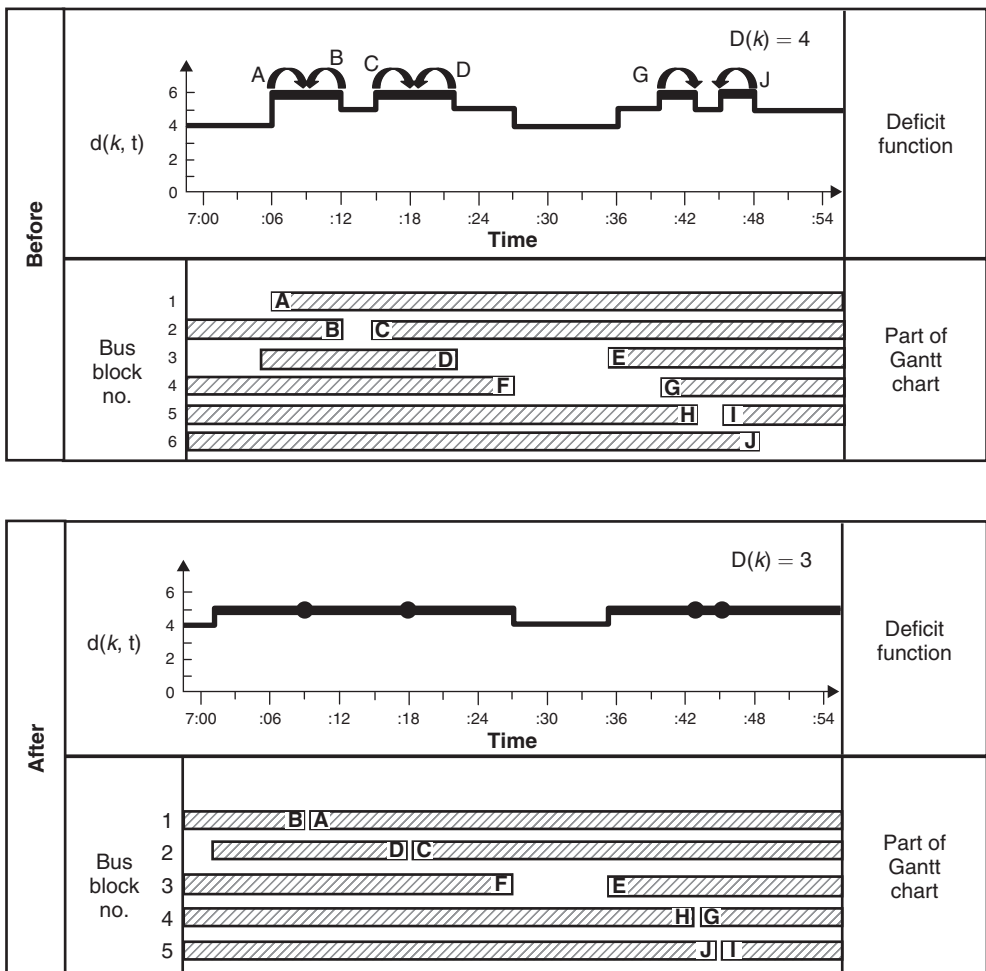


Figure 8.13 Demonstration of superiority of deficit-function representation over the Gantt-chart approach.

the changes. From Figure 8.13, the problem appears easy to handle. However, before the changes in part of Figure 8.13, only 6 of 52 portions of the bus blocks were shown, and those six rows were spread among the 46 other rows in the Gantt chart.

Following this demonstration, the schedulers were still not wholly convinced. They argued that with a little more effort on their part, they too could have saved the bus, as in Figure 8.13. Therefore, a more complex example was decided on, as shown in Figure 8.14. This second example refers to an afternoon schedule for two Egged branches: Ramle, terminal k , and Lod, terminal m . On the left-hand side of Figure 8.14, only trips that involved changes are exhibited in the before-and-after Gantt chart representation; trips are designated by letters. On the right-hand side, the DFs of the complete schedule are illustrated, including trips not shown in the left-hand side. The schedulers again claimed that no further reductions could be achieved from the $D(k) + D(m) = 57 + 19 = 76$ fleet-size requirement. The given information was that $\Delta^{i(+)} = \Delta^{i(-)} = 2$ minutes, and that the DH trip time between the k and m terminals (both directions) was seven minutes.

As seen in Figure 8.14, six shifts in trip-departure times and a single DH trip were required in order to save two buses and to reduce the total fleet requirement to 74 buses. It was only after this second demonstration that the schedulers began to take a serious interest in the DF model. This was particularly due to its simplicity and visual nature. The schedulers expressed a positive feeling about the valuable aid of this gradual approach.

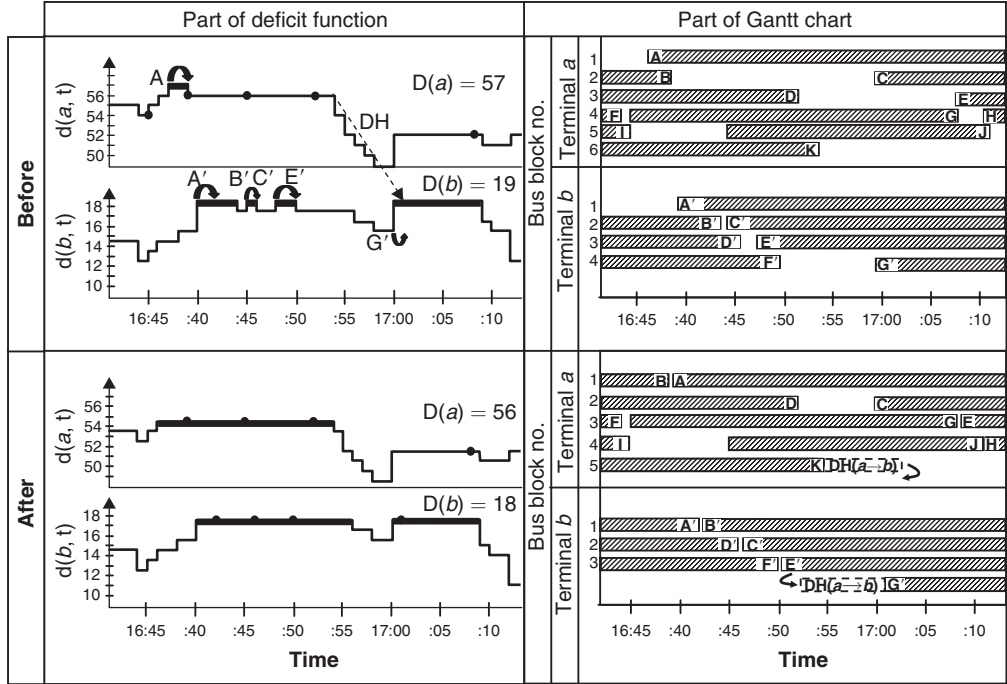


Figure 8.14 Two-terminal case, in which two buses are saved through shifting departure times and modifying URDHC procedures

8.7 Examination and consideration of even-load timetables

We noted in Section 8.3 that the construction of timetables in Chapters 4 and 5 was based on either even headways or even average loads, entailing situations in which even headways resulted in uneven passenger loads. Shifting departure times, therefore, may unbalance these desirable features in the timetable while favouring cost (vehicle) saving. This section introduces the possibility of establishing a load-tolerance criterion for even-load timetables.

Following the notation in Section 5.2.2 (Chapter 5), let $SL_i(t)$ = the slope of $L_i(t)$ at t , in which $L_i(t)$ is the cumulative load curve at stop i , $i = 1, 2, \dots, n$; Δd_τ = a given positive tolerance (in passengers) of the desired occupancy d_τ at time interval τ , $\tau = 1, 2, \dots, v$; $t_{1j}^* = t_{qj} - T_{q\tau}$ is a departure time determined by Principle 3 in Section 5.2.2 (even-load on individual vehicles) at stop q , in which t_{qj} is the j -th candidate departure time from stop q $j = 1, 2, \dots, m$; and $T_{q\tau}$ is the average service travel time between the departure terminal $q = 1$ and stop q when t_{qj} is at interval τ (note that by definition $T_{1\tau} = 0$), and $\Delta^{j\tau(-)}$ and $\Delta^{j\tau(+)}$ are given positive tolerances (in minutes) for maximum shifting t_{1j}^* to the left (early departure), and to the right (late departure), respectively, for each interval τ , $j = 1, 2, \dots, m$; $\tau = 1, 2, \dots, v$.

When d_τ is added along $L_i(t)$, the minimum time that intersects one of the cumulative load curves (see Principle 3 in Section 5.2.2) is determined at stop q . Hence,

$$d_\tau = L_q(t_{1j}^*) - L_q(t_{1,j-1}^*) \quad (8.4)$$

The shifts in departure times made by the DF model are defined as follows:

$$\begin{aligned} t_{1j}^- &= t_{1j}^* - \Delta^{j\tau(-)} \\ t_{1j}^+ &= t_{1j}^* + \Delta^{j\tau(+)} \end{aligned} \quad (8.5)$$

for all $j = 1, 2, \dots, m$, and relevant τ for t_{1j}^- and t_{1j}^+ .

In order to avoid excess average loads beyond $d_\tau + \Delta d_\tau$ at each trip's critical point, two criteria may be established for early and late departures. What follows is a formal derivation of the early-departure criterion, while a similar derivation for the late-departure criterion is presented as an exercise at the end of the chapter.

Early departure criterion

According to the foregoing definitions:

$$L_i(t_{1j}^-) = L_i(t_{1j}^*) - \Delta^{j\tau(-)}SL_i(t_{1j}^*), \quad (8.6)$$

in which $SL_i(t_{1j}^*) = SL_i(t_{1j}^-)$, $i = 1, 2, \dots, n$, and t_{1j}^- belongs to τ .

In case the slope is changed within a $\Delta^{j\tau(-)}$ shift for any stop i , Equation 8.6 should consider two (or more) decreased portions, each related to a different slope and its associated part of $\Delta^{j\tau(-)}$.

The loads at t_{1j}^- can be expressed as:

$$L_i(t_{1j}^-) - L_i(t_{1,j-1}^*) < d_\tau, i = 1, 2, \dots, n, t_{1j}^- \text{ belongs to } \tau. \quad (8.7)$$

These loads for the new t_{1j}^- departure across all stops are based on Principle 3 (Section 5.2.2) in which the desired occupancy d_τ is attained for t_{1j}^* only at q and the load is less than d_τ at all other stops. Using a $\Delta^{j\tau(-)}$ shift will further reduce these loads. However, the loads at each stop i for the departure times adjacent to t_{1j}^- at $t_{1,j+1}^*$ will increase the loads if the shifting takes place. This increase is $\Delta^{j\tau(-)}SL_i(t_{1j}^*)$ for all i and relevant τ , or it is the sum of portions of the slope that are changed within $\Delta^{j\tau(-)}$.

The increased new loads at $t_{1,j+1}^*$ need to be checked against $d_\tau + \Delta d_\tau$ across all stops. This check is applied for the maximum load increase, and hence the early departure criterion for accepting $\Delta^{j\tau(-)}$ is this:

$$\max_{i=1,2,\dots,n} [L_i(t_{1,j+1}^*) - L_i(t_{1j}^-)] \leq d_\tau + \Delta d_\tau \quad (8.8)$$

in which $t_{1,j+1}^*$ belongs to interval τ , and $L_i(t_{1j}^-)$ is obtained by Equation 8.6.

Figure 8.15 illustrates a simple example of Principle 3 in Section 5.2.2. Given a transit line $A \rightarrow B \rightarrow C$, with average travel times of 15 minutes between A and B, three departures, at 6:15, 6:45 and 7:10, and a desired occupancy of 50 passengers. The average observed on-board loads on the 6:15 vehicle are 30 passengers at stop A and 65 at stop B; on the 6:45 vehicle, 80 and 35 passengers at A and B, respectively; and on the 7:10 vehicle, 25 and 80 passengers at A and B, respectively. Figure 8.15 shows $L_A(t)$ and $L_B(t - 15)$ as the cumulative load curves of the three vehicles, in which the curve at B is shifted by 15 minutes to allow for an equal time basis (at the route's departure point) in the analysis. The procedure explicated in Principle 3 in Section 5.2.2 results in individual even-load departures at 6:11, 6:31 and 6:56. This procedure is indicated by the lines with arrows in Figure 8.15.

Figure 8.16 presents the example of Figure 8.15 with a $\Delta^{j\tau(-)}$ shift in part (a) of the figure, and a $\Delta^{j\tau(+)}$ shift in part (b), utilizing $\Delta d_\tau = 10$ passengers for both parts. In part (a), the shifting is $\Delta^{j\tau(-)} = 3$ minutes. The solid lines (with arrows) show graphically how to determine the new loads on the (new) 6:28 and 6:56 departures at both stops A and B. The slope at A between 6:28 and 6:31 is $8/3$ and times 3 minutes; this results in an average load of 8 more passengers on the 6:56 departure, and 8 fewer on the 6:28 departure. At stop B, the slope is $7/6$, and the change in load is 3.5 passengers. The criterion in Equation 8.8 considers the load at A with $120 - 72 + 8 = 56$, and at B with $134 - 84 + 3.5 = 53.5$ passengers for the 6:56 departure. The maximum of the two is 56; because $\Delta d_\tau = 10$ (the average load can reach 60 passengers), the value of $\Delta^{j\tau(-)} = 3$ minutes is accepted.

Part (b) of Figure 8.16 presents an example of $\Delta^{j\tau(+)} = 3$ minutes for the 6:31 departure. The solid lines (with arrows) show graphically how to determine the new loads on the (new) 6:34 and 6:56 departures at both A and B. The slope at A is the same as for part (a) of Figure 8.16, and results in a difference of 8 passengers, while the outcome of the slope at B is 3.5 passengers. The increased load at A is $50 + 8 = 58$, and at B is $84 + 3.5 - 50 = 37.5$ passengers for the new 6:34 departure. Consequently, and similar to the early-departures case, the shift of $\Delta^{j\tau(-)} = 3$ minutes is accepted.

This section provided a procedure to integrate two operational planning components (timetabling and vehicle scheduling), based on the deficit-function model and given tolerances. The outcome may be a set of efficient schedules from both the passenger and operator perspectives.

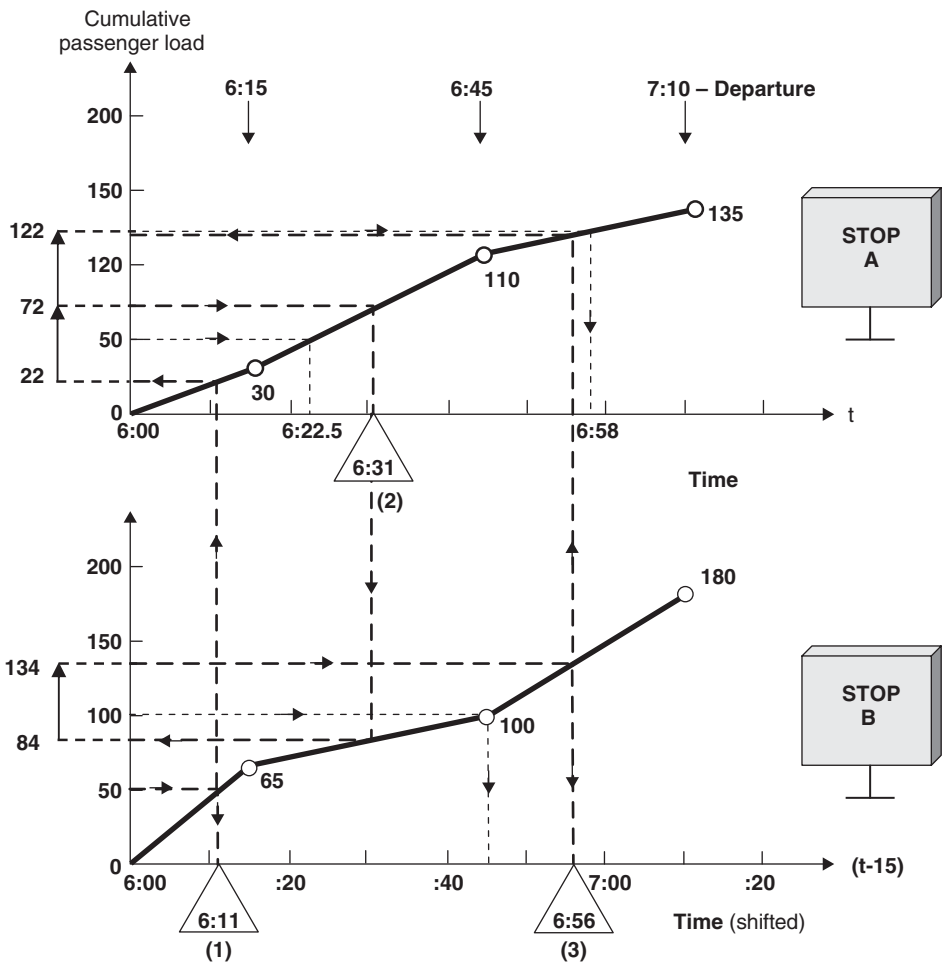
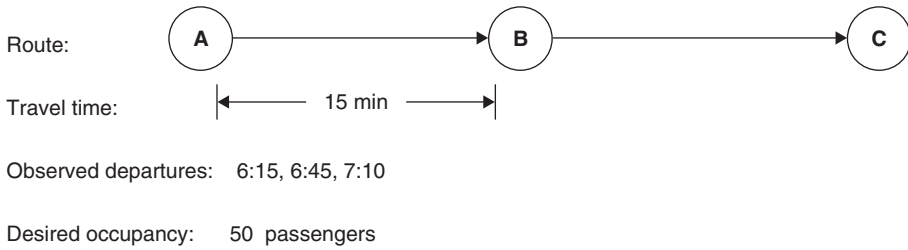


Figure 8.15 Example using the even max load procedure on individual vehicles (see Section 5.2)

Exercises

8.1 Given:

- A network with 5 terminals – a, b, c, d, e – shown in the figure below.
- A table with trip number, departure terminal, arrival terminal, average travel times (and a total of 47 trips).
- A table with average travel times (including layover times) between terminals (same in both directions); the numbers in parentheses are the DH times between each terminal.
- Tolerances for possible departure-time shifting are ± 5 minutes for all trips.

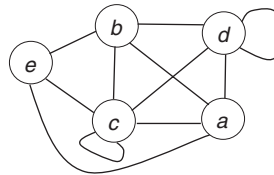
Perform the vehicle-scheduling task with this data with the aim of minimizing fleet size, using the following three approaches:

- only shifting trip-departure times,
- only DH trip insertion,
- both shifting trip times and DH trip insertion.

For each of the three approaches, the final outcome should embody

- a final DF for each terminal, marked by changes made;
- chains of trips (blocks, vehicle schedule) using
 - FIFO approach;
 - ‘Within a Hollow’ approach (example of other possible chains).

Basic route network:



Average travel times (DH times in parentheses), in minutes, for both directions of travel

	a	b	c	d	e
a		35 (20)	60 (40)	70 (45)	50 (40)
b			55 (40)	40 (30)	25 (20)
c			25	35 (25)	30 (25)
d				20	
e					

Trip schedules

Trip no.	Departure terminal	Arrival terminal	Departure time		Trip no.	Departure terminal	Arrival terminal	Departure time
1	d	d	6:30		25	b	c	7:30
2	b	a	6:30		26	c	c	7:30
3	a	d	6:30		27	d	a	7:35
4	e	a	6:30		28	e	b	7:35
5	c	c	6:30		29	b	d	7:45
6	b	a	6:45		30	b	a	7:45
7	d	c	6:45		31	a	c	7:45
8	e	a	6:50		32	a	e	7:50
9	b	a	6:55		33	d	a	7:55
10	d	d	7:00		34	d	b	8:00
11	a	d	7:00		35	a	b	8:00
12	b	a	7:00		36	c	a	8:00
13	b	d	7:00		37	c	d	8:00
14	a	b	7:00		38	a	d	8:00
15	c	a	7:00		39	c	b	8:05
16	c	b	7:05		40	e	a	8:10
17	a	c	7:10		41	b	a	8:15
18	c	e	7:25		42	b	a	8:30
19	e	c	7:25		43	b	e	8:30
20	d	d	7:30		44	c	e	8:30
21	b	e	7:30		45	e	c	8:30
22	b	a	7:30		46	e	b	8:30
23	a	b	7:30		47	a	b	8:35
24	e	a	7:30					

- 8.2 Given two shuttle (radial) routes, each trip starting and ending at a major train station; the following is their trip schedule, coordinated with the train schedule:

Shuttle route 1		
Trip no.	Departure time	Arrival time
1	6:30	7:05
2	6:45	7:25
3	7:00	7:28
4	7:20	7:50
5	7:40	8:05
6	7:55	8:15
7	6:45	7:35
8	7:00	7:50
9	7:20	8:00
10	7:50	8:15

- (1) Find the minimum fleet size required to execute the entire trip schedule, using:
 - (i) the modelling of single routes; (ii) the DF model.
- (2) Repeat (1) with shifting possibilities in trip-departure times, in which for all trips the forward shifting tolerance ($\Delta^{i(+)}$) is three minutes and the backward shifting tolerance ($\Delta^{i(-)}$) is six minutes.
- (3) Construct the DF after the shifting; can the shifted schedule have an adverse effect on the Shuttle 2 route operation?
- (4) What is the improved lower bound of the trip schedule? Is it needed?
- (5) Apply the FIFO rule to create vehicle chains (blocks), using the shifted schedule determined in (2).

8.3 Find the three levels of lower bound – $G(S)$, $G'(S')$ and $G''(S'')$ – for the 10-trip schedule, given as follows.

Trip no.	Departure terminal	Departure time	Arrival terminal	Arrival time	Deadheading (DH) trips	
					Between terminals	DH time (same for both directions)
1	b	6:00	c	6:30	$a - b$	60 min
2	a	6:10	a	6:50	$a - c$	30 min
3	d	6:10	b	7:10		
4	b	7:00	b	7:30	$a - d$	30 min
5	c	7:10	b	7:30		
6	c	7:40	b	8:10	$b - c$	10 min
7	d	7:50	a	8:30		
8	a	8:00	a	8:30	$b - d$	50 min
9	a	8:30	d	9:10		
10	b	9:00	c	9:20	$c - d$	20 min

- 8.4 For the 10-trip schedule in problem 8.3, consider shifting departure times by five minutes for both forward and backward shifts ($\Delta^{i(+)} = \Delta^{i(-)} = 5$) for all trips.
- (i) Find $G'(S'_{sf})$ and $G''(\bar{S}''_{sf})$ by constructing $g(t, \bar{S}'_{sf})$ and $g''(t, \bar{S}''_{sf})$.
 - (ii) Find $G(S_{sf})$, $G'(S'_{sf})$, $G''(S''_{sf})$ using the LB-SHIFT and LB- DH&SHIFT procedures.
- 8.5 Derive a load-tolerance criterion for even-load timetables for the case of a late departure. End the derivation with a formulation similar to Equation 8.8.

References

Ceder, A. (2002). A step function for improving transit operations planning using fixed and variable scheduling. In *Transportation and Traffic Theory* (M. A. P. Taylor, ed.), pp. 1–21, Elsevier Ltd.

Ceder A. (2005). Estimation of fleet size for variable bus schedules. *Transportation Research Record*, **1903**, 3–10.

Ceder, A. and Stern, H. I. (1981). Deficit function bus scheduling with deadheading trip insertion for fleet size reduction. *Transportation Science*, **15**, 338–363.

Stern, H. I. and Ceder, A. (1983). An improved lower bound to the minimum fleet size problem. *Transportation Science*, **17**, 471–477.

Appendix 8.A

Deficit-function software

Based on the substance and articles mentioned in Chapters 7 and 8, Altdoit Software Solutions Ltd created deficit-function (DF) software, which is partially implemented in the following website: www.altdoit.com. Its software project, called PT-Manager,[®] was designated as the *Optimal Determination of Vehicle Schedules*. The main screen of the software (as of this writing) is shown in Figure 8A.1. This website can be used for undertaking both exercises and small-size practical schedules; most probably, the website will be continuously updated.

PT-Manager contains three stages: (1) *Preprocessing*, in which the necessary data are prepared and the required setting and constraints are configured; (2) *Processing*, in which both custom- and auto-planning functions are available; and (3) *Post processing*, in which the optimized schedule can be reported in detail. The four sections that follow cover the principal features of this software.

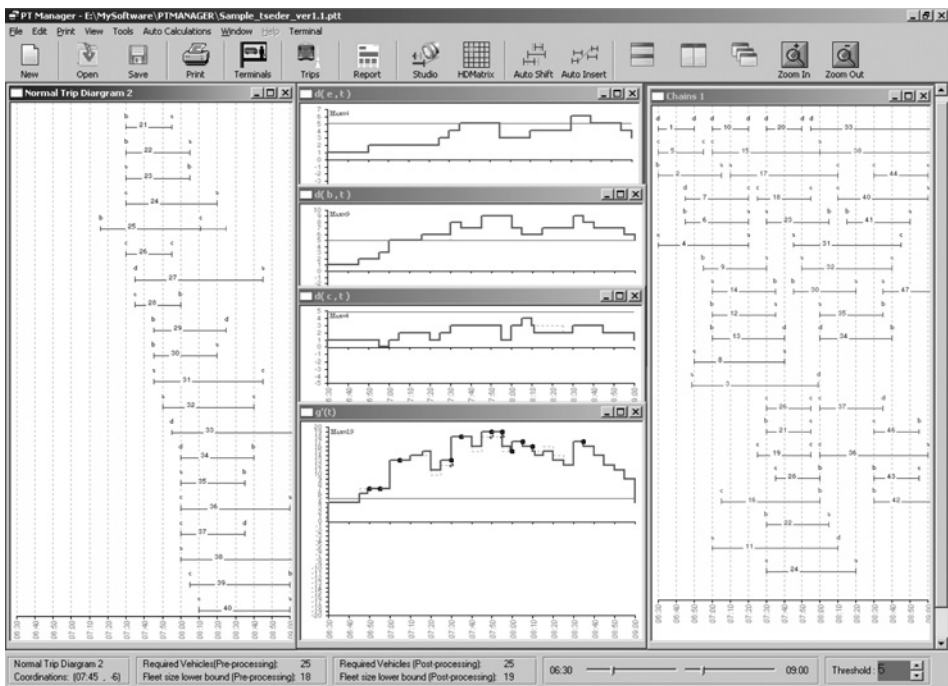


Figure 8A.1 Main screen of PT-Manager

8A.1 Key features

The key tasks of the PT-Manager software are as follow:

- Represent schedules by DFs
- Monitor large multi-terminal schedules
- Detect challenged points
- Consider a wide range of practical constraints
- Work with flexible workspace for manual optimization
- Run heuristic algorithms for automatic optimization
- Construct vehicle chains/blocks
- **Generate detailed reports.**

Data from SQL servers, TCP streaming (real time), and other external data sources can be processed. The data processing and configuration include the following elements:

Trips, deadheading (DH) trips and terminals information

- Handling trips, import, add, edit
- Handling terminals, import, add, edit
- Handling DH trips.

Traffic information

- Handling traffic congestion templates, import, add, edit
- Average DH time matrix between each trip's end and start locations (by time-of-day).

General constraints and configurations

- Maximum block length
- Maximum DH operational cost allowed per block
- Trip recovery-time tolerances (maximum and minimum waiting-time allowed for next-trip preparation)
- Trip departure-time tolerances (maximum-time deviation allowed for departure shifting)
- Handling pullouts (average DH trip time from a depot/garage to the first trip)
- Handling pull-ins (average DH trip time from the end of last trip until the arrival at a depot/garage).

8A.2 Main Dialogs

Trips: Service and DH trips are defined by: (i) their departure and arrival times and departure and arrival terminals; (ii) assigned vehicle types; and (iii) constraint data (e.g. departure shifting tolerances; see Figure 8A.2). Figure 8A.3 presents a group of trips sorted by time of day.

Terminals: Each terminal is defined by name and its constraint data (e.g. available parking spaces; see Figure 8A.4).

Figure 8A.5 shows DF in a single terminal.

Figure 8A.6 shows the sum of all DFs (function g), which is the number of vehicles simultaneously in service (or in DH) at any point in time.

Trip Dialog

Trip Properties

Basic Information:

Trip ID: 41

Trip Name: t41

Vehicle Type: Bus

Trip Type: Normal Trip

Shift Tolerance:

Backward Shift Tolerance: 0 [min]

Forward Shift Tolerance: 0 [min]

Shift Status:

Shift Value: 0 [min]

Departure / Arrival:

Terminal Departure: b (Terminal)

Time Departure: 08:15 AM

Terminal Arrival: a (Terminal)

Time Arrival: 08:50 AM

Extra Information:

Trip Description:

TAG:

Update Cancel

Figure 8A.2 Trip properties dialog

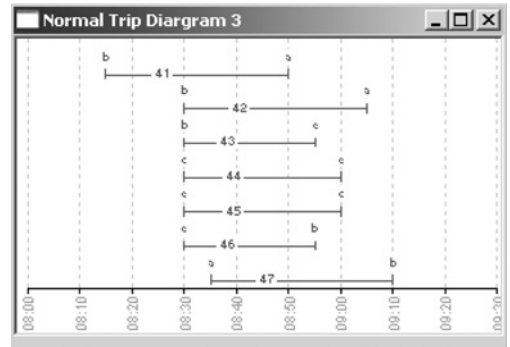


Figure 8A.3 Trips in graphical form

Terminal Dialog

Terminal Properties

Terminal ID: 1

Terminal Name: a

Shift Tolerance: 0 [min]

Vehicle Capacity: 2

Terminal Description: Terminal

Update Cancel

Figure 8A.4 Terminal properties dialog

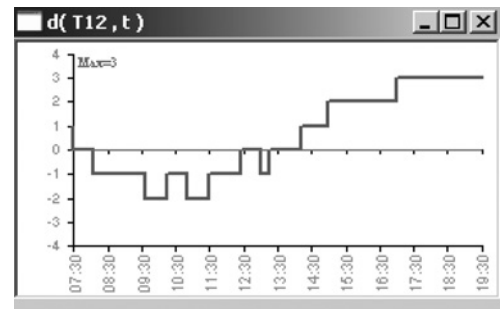


Figure 8A.5 DF at one terminal

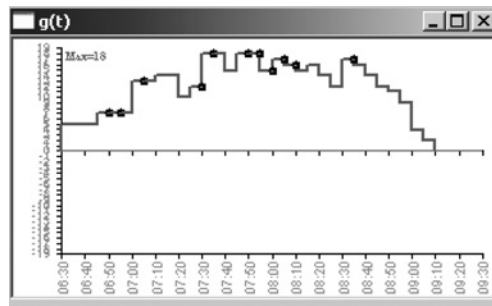


Figure 8A.6 Sum of all DFs

8A.3 Main cases utilized

Shifting trip departure time

The DF in Figure 8A.7 has a level line set to four; that is, the user is not interested in dealing with terminals whose maximum DF value (deficit-max) is less than four. This feature is helpful for monitoring a large number of terminals. Moving the cursor to the deficit-max area highlights the black line of the trips involved.

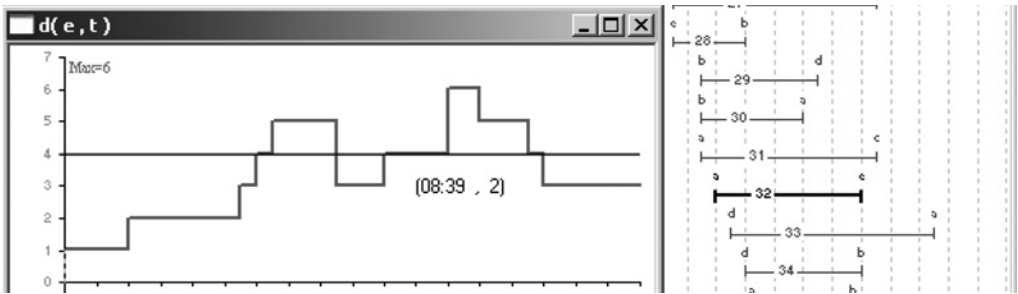


Figure 8A.7 DF with $D(e) = 6$

The selected trip in Figure 8A.7 can be shifted (within the tolerance range). Figure 8A.8 shows graphically the result of shifting the departure time. The original display will remain on the screen, but with a level-line indicating Trip 32.

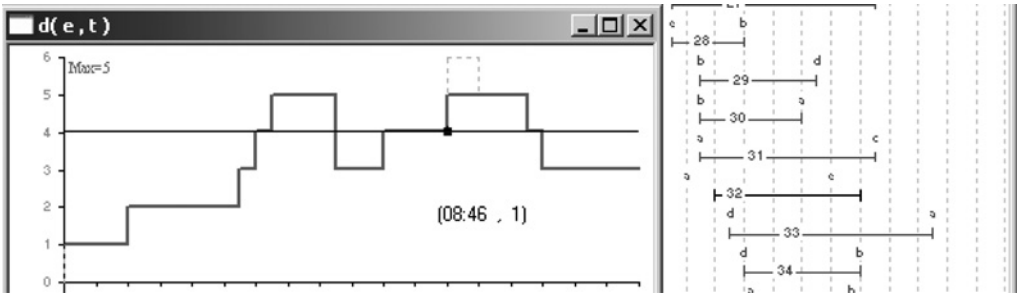


Figure 8A.8 DF and trip marks change following a shift

Inserting a DH trip

Inserting a DH trip between two terminals will affect the DFs of both terminals and the sum of all DFs; thus, it is important to evaluate this DH insertion. The DH dialog is designed to show three rows: (1) DF of the first terminal; (2) DF of the second terminal; and (3) the sum of the two DFs. Figure 8A.9 shows that the DH insertion dialog will immediately respond to any data insertions prior to accepting them; this will allow the user to amend or cancel such changes.

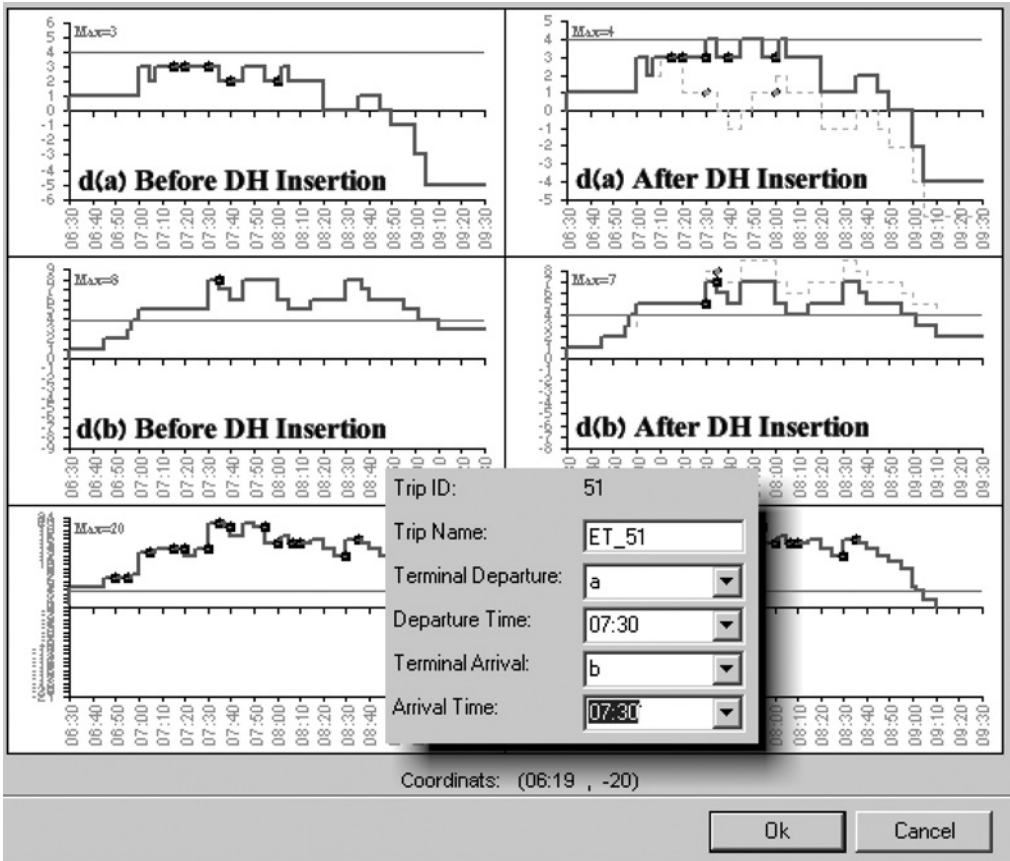


Figure 8A.9 DH Studio

Running the automatic features

Automatic processing is based on the selection of appropriate items from the main menu. However, the user can make manual changes at any time during this process. The features that can run automatically as follow:

Auto Shift

- Calculate shifting tolerance; e.g. headway-dependent (see Table 8.1)
- Run the SDT algorithm (see Figure 8.11).

Auto DH Insertion

- Run the auto DH trip-insertion algorithm (see Figure 7.9 (a, b) in Chapter 7)
- Run the SDT and DH insertion-trip algorithm (see Figure 8.12).

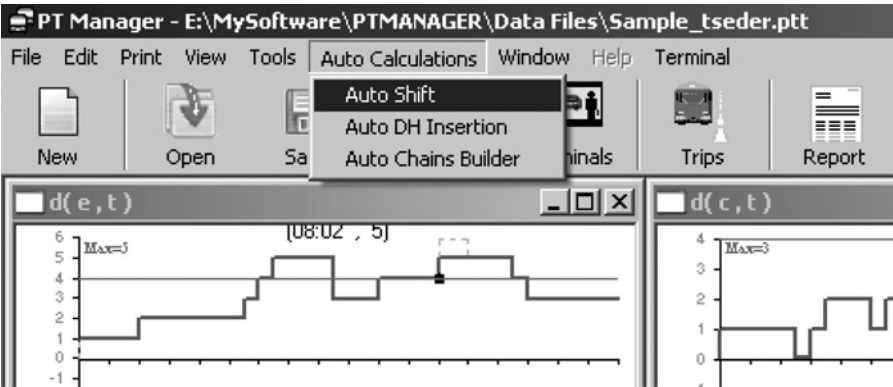


Figure 8A.10 Toolbar indicating automatic tasks

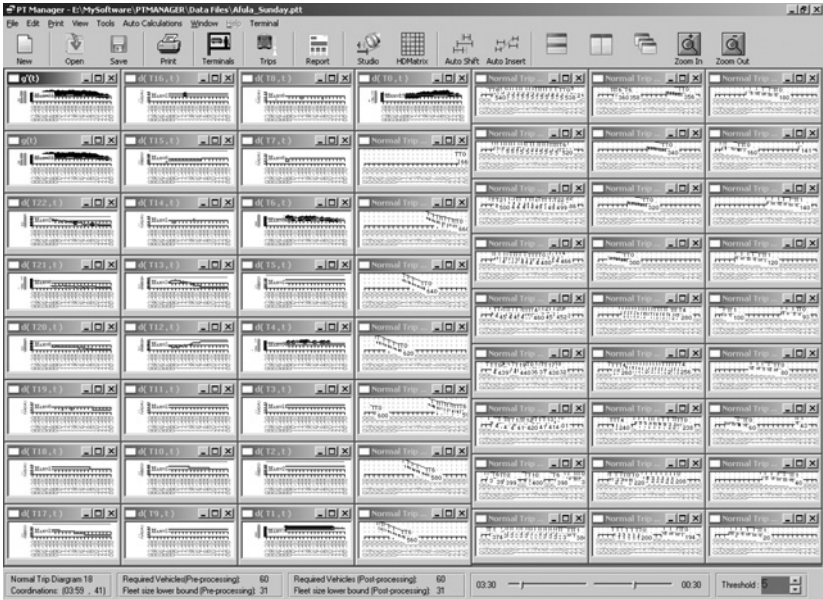


Figure 8A.11 Example of large multi-terminal displayed by the PT-Manager

Auto chain builder

- Construct vehicle schedules (chains/blocks) automatically utilizing the FIFO and ‘within-hollow’ methods (see Section 7.5.5, Chapter 7).

These three automatic features are selected on the toolbar shown in Figure 8A.10.

8A.4 Advanced graphical display

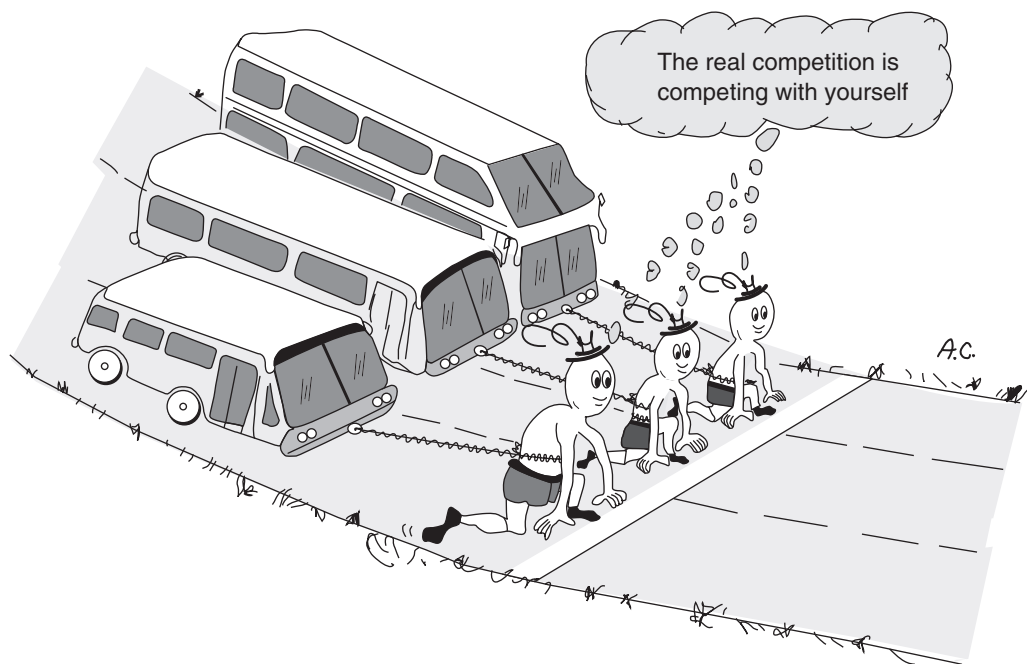
The PT-Manager GUI (graphical user interface) concept is designed to display a large amount of graphical objects, such as DFs, trips and blocks. The user can customize the workspace layout by choosing from the following display options:

- Critical parameters (e.g. fleet-size lower bound before and after processing) are always shown on the status bar.
- Resize the windows as to focus on certain processing segments.
- Store workspace layout with workspace files.
- All DFs and/or trips are shown on a multi-terminal document.
- Set focus/zoom by time-of-day for all windows.
- Auto-arrange windows by maximum DF values.
- All original data continue to be shown using dotted lines to allow for a visual comparison between the changes and the original schedule.

Utilizing the interactive DF window enables observation of the length and location of a $D(k)$ for all $k \in K$; then a shifting, DH insertion, or both can be implemented. The following screen, in Figure 8A.11, manifests the difficulty in observing simultaneously a large multi-terminal case. The idea behind GUI is to enable the user to avoid such a display and to focus on specific DFs, from which one may be more likely to succeed with the vehicle reduction process. Further information is given on the website mentioned.

This page intentionally left blank

9 Vehicle-type and Size Considerations in Vehicle Scheduling



Chapter 9 Vehicle-type and Size Considerations in Vehicle Scheduling

Chapter outline

- 9.1 Introduction
 - 9.2 Optimization framework
 - 9.3 Procedure for vehicle scheduling by vehicle type
 - 9.4 Examples
 - 9.5 Vehicle-size determination
 - 9.6 Optimal transit-vehicle size: literature review
 - Exercises
 - References
-

Practitioner's Corner

This chapter contains two main, almost independent parts. First, it addresses the problem of how to allocate vehicles efficiently for carrying out all the trips in a given transit timetable, while taking into account the association between the characteristics of each trip (urban, peripheral, intercity, etc.) and its required vehicle type. Second, it provides an overview and tools for the determination of, and a practical decision concerning, transit-vehicle size. Both parts are directly related to the prominent act of purchasing a vehicle or a fleet of vehicles.

Ostensibly when dealing with vehicle type, there is need for more operational details. Paying attention to details is similar to answering the following riddle. (Do so without using a calculator.) You are driving a bus from terminal *a* to terminal *b*. At the first stop, 10 people get on the bus and 2 get off. At the second stop, 12 get on and 7 get off. At the third stop, 5 people get on and 10 get off. At the fourth stop, 11 people get off and 1 gets on. You then arrive at terminal *b*. What is the name of the driver? (See end of this corner.)

The chapter consists of five sections, following the introduction. Section 9.2 describes an optimization framework for the vehicle-scheduling problem, in which categories of vehicle types are arranged in decreasing order of vehicle cost and comfort level. The assignment of vehicles to trips using DH (deadheading) and shifting procedures within the optimization framework is discussed in Section 9.3; the objective is to minimize the total purchasing cost. Section 9.4 presents two detailed examples (one of which is real life) of the optimization procedures developed. Section 9.5 provides basic tools for a determination of vehicle size, including a known square-root formula. Section 9.6 continues, more in depth, with the overview of the tools and formulation for vehicle-size determination. The chapter ends with exercises.

There is no doubt that the successful deployment of optimal or adequate vehicle types requires a considerable amount of analysis. After all, success comes before work only in the dictionary. Finally, as to the riddle, it said: “**You** are driving a bus . . .”.

9.1 Introduction

The vehicle-scheduling task described in Chapters 7 and 8 considers only one type of transit vehicle. In practice, however, more than one type is used; e.g. a bus operation may employ minibuses, articulated and double-decker buses, and standard buses with varying degrees of comfort and different numbers of seats. Commonly, the consideration of vehicle type in transit-operations planning involves two considerations: first, determining the suitable or optimal vehicle size; second, choosing vehicles with different comfort levels, depending on trip characteristics. Certainly a multi-criteria effort may treat both considerations simultaneously, but this is seldom done in practice. The issue of what vehicle type to consider arises when purchasing a vehicle or a fleet of vehicles, an undertaking that is not performed frequently. There is a saying that good judgement comes from experience, and experience comes from bad judgement. Another is that experience is not what is happening to us, but what we do with what is happening to us. In our case, the experience accumulated focuses on the need for an analysis framework in order to decide how many and which type of vehicle to purchase. This chapter attempts to introduce (a) a cost-effective framework for choosing a vehicle's comfort level, and (b) the basic tools for analysing optimal vehicle size.

The purpose of the first part of the chapter is to address the vehicle-scheduling problem, while taking into account the association between the characteristics of each trip (urban, peripheral, intercity, etc.) and the vehicle type required for the particular trip. This means complying with a certain level of service for that trip: degree of comfort, seat availability and other operational features. The purpose of the second part of the chapter (Sections 9.5 and 9.6) is to review a formulation for trade-offs between vehicle size and operational variables that can be used as a tool for the determination of optimal vehicle size for a given frequency of service.

9.2 Optimization framework

This and the two following sections use the notation and procedures of DF (deficit function) described in Chapters 7 and 8. The problem addressed concerns the assignment of vehicles to trips using DH and shifting procedures, in which categories of types are arranged in decreasing order of vehicle cost and comfort. The process described follows Ceder (1995a, 1995b).

The problem, entitled the vehicle-type scheduling problem (VTSP), is based on a given set S of trips (schedule) and set M of vehicle types. The set M is arranged in decreasing order of vehicle cost so that if $u \in M$ is listed above $v \in M$, it means that $c_u > c_v$, where c_u , c_v are the costs involved in employing vehicle types u and v , respectively. Each trip $i \in S$ can be carried out by vehicle type $u \in M$ or by other types listed prior to u in the above-mentioned order of M .

The problem can be formulized as a cost-flow network problem, in which each trip is a node and an arc connects two trips if, and only if, it is possible to link them in a time sequence with and without DH connections. On each arc (i,j) , there is a capacity of one unit and an assigned cost C_{ij} . If the cost of the lower-level vehicle type associated with trip i is higher than the cost of the vehicle type (even if of a lower level) required for trip j , then $C_{ij} = c_i$. That is, $C_{ij} = \max(c_i, c_j)$. The use of such a formulation was implemented by Costa

et al. (1995), who employed three categories of solutions: (a) a multi-commodity network flow; (b) a single-depot vehicle-scheduling problem; and (c) a set-partitioning problem with side constraints. The mixed-integer programming of these problems is known to be NP-complete as may be seen, for example, in Bertossi *et al.* (1987). The maths-formulation concepts for the third category are further explained in Chapter 10.

Because of the complexity involved in reaching an optimal solution for a large number of trips in S , a heuristic method is considered a more practical approach. The heuristic procedure developed is called the VTSP algorithm. It begins by establishing lower and upper bounds on fleet size. The upper bound is attained by creating different DFs, each associated with a certain vehicle type $u \in M$, which includes only the trips whose lower-level required vehicle type is u . Certainly, this scheduling solution reflects a high cost, caused by the large number of vehicles demanded. The lower bound on the fleet size is attained by using only one vehicle type: the most luxurious one with the highest cost that can clearly carry out any trip in the timetable. Between these bounds on fleet size, the procedure searches for the best solution, based on the properties and characteristics of DF theory.

This optimization framework is illustrated in Figure 9.1, with (C_1, N_1) and (C_2, N_2) representing the lower- and upper-boundary solutions, respectively. The following are added notations to DF theory (described in Chapters 7 and 8):

- M = set of all vehicle types u , $u = 1, 2, \dots, m$, arranged in decreasing order of vehicle cost;
- S_u = set of required trips (schedule) for vehicle types u , $u = 1, 2, \dots, m$;
- $S = \cup_{u=1}^m S_u$, which is the combined schedule of all trips;
- t_{su}^i = start time of trip $i \in S_u$ of type $u \in M$;
- t_{eu}^i = end time of trip $i \in S_u$ of type $u \in M$;
- $\Delta_u^{i(+)}$ = maximum delay tolerance from the scheduled departure time of trip $i \in S$ of type $u \in M$;
- $\Delta_u^{i(-)}$ = maximum advance tolerance of the trip scheduled departure time of trip $i \in S$ of type $u \in M$;

$d_u(k, t, S_u)$, $D(k, S_u)$, $D(S_u)$, $[s_{iu}^k, e_{iu}^k]$, $g(t, S_u)$, $G(S_u)$ = refer to the DF definitions (see Chapters 7, and 8) and but only with respect to trips of type $u \in M$:

- c_u = cost involved in employing a vehicle of type $u \in M$;
- N_1 = minimum number of vehicles required to service all trips in S , using a single vehicle of type $u = 1$;
- n_{2u} = minimum number of vehicles required to service all trips in S_u , $u \in M$;
- $N_2 = \sum_{u=1}^m n_{2u}$ = sum of all minimum numbers of vehicles required when treating each type separately;
- $C_1 = c_1 N_1$ = total cost involved in employing N_1 vehicles of type 1 (most luxurious);
- $C_2 = \sum_{u=1}^m c_u n_{2u}$ = total cost involved in employing N_2 vehicles (for each type separately);

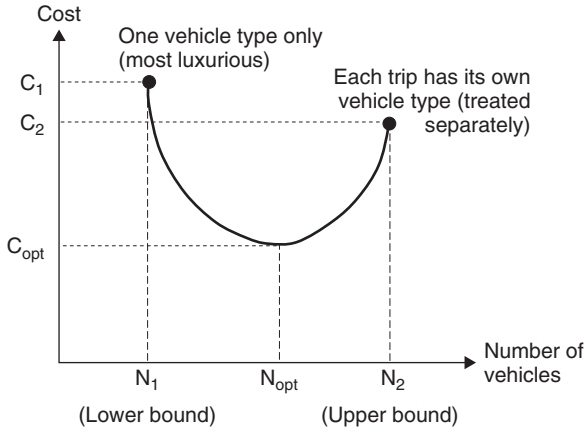


Figure 9.1 Trade-off between cost of purchasing vehicles and number of vehicles

$C = \sum_{u=1}^m c_u n_u$ = objective function (total cost) following algorithm VTSP, with n_u vehicles required of type u for all $u \in M$;

$N = \sum_{u=1}^m n_u$ = total number of vehicles incurring cost C .

9.3 Procedure for vehicle scheduling by vehicle type

The algorithm VTSP developed is heuristic in nature while incorporating all DF components. Because of the graphical features associated with DF theory, the algorithm can be applied in an interactive manner or in an automatic mode, along with the possibility of examining its intermediate steps. The following is a general description of algorithm VTSP in a stepwise manner:

9.3.1 Algorithm VTSP

- Step 0:** Arrange the set of vehicle types M in decreasing order of vehicle cost (so that if $u \in M$ is listed above $v \in M$, it means that $c_u \geq c_v$).
- Step 1:** Solve the problem as a single-vehicle-type problem using DF theory, including the DH and shifting procedures (see Figures 7.9a, 7.9b and 8.12), to obtain N_1 vehicles, considered as type 1, with a total cost of C_1 .
- Step 2:** Divide the trips by their associated type and apply the DF methodology with the DH and shifting procedures for each type separately. Add up the number of vehicles derived to obtain the total of N_2 vehicles with a total cost of C_2 .
- Step 3:** If $N_1 = N_2$, stop. Use the solution in Step 2.
- Step 4:** Consider $d_u(k, t)$ as in Step 2 for all $k \in K$ and $u \in M$.

- Step 5:** Perform the URSC (shifting only) procedure for shifting departure times within their tolerances (see Figure 8.11).
- Step 6:** Find a URDHC (see Figures 7.9(a), 7.9(b), 8.12), such that each DH trip (with possible shifting) in this chain fulfils condition (a) and/or (b):
- (a) The DH trip is from DF of vehicle type u to DF of type v , such that $u \leq v$, meaning that $c_u \geq c_v$ (see Proposition 2);
 - (b) The URDHC aims at saving a vehicle of type w and $-c_w + \sum_{q,r \in E} (c_q - c_r) \leq 0$, in which the set E is composed of all DH trips included in the URDHC; each DH trip is from a DF of vehicle type r to a DF of vehicle type q , in which $q < r$ (see Proposition 3). If no URDHC can be found, stop.
- Step 7:** Examine whether the total cost of the URDHC (DH cost) is less than the cost of saving one vehicle (of the type considered). If it is not, delete this possibility and go to Step 6. Otherwise, update $d_u(k, t)$ for all $k \in K$ and $u \in M$.
- Step 8:** Apply the improved lower-bound check. If $D(S) = G''(S''_{sf})$, stop; otherwise, go to Step 5.

Among the eight steps of algorithm VTSP, the conditions specified in Step 3 and particularly Step 6 deserve further attention. The following four propositions clarify and interpret these conditions:

Proposition 1 (for Step 3): If $N_1 = N_2$, then $C_2 \leq C_1$.

Proof: Given $M \geq 2$ and $c_1 \geq c_2 \dots \geq c_m$, the proof is straightforward because

$$C_1 = N_1 c_1, \quad C_2 = \sum_{j=1}^m N_{2j} c_j, \quad N_2 = \sum_{j=1}^m N_{2j}, \quad \text{and} \quad N_1 = N_2$$

Proposition 2 (for Step 6(a)): Any DH trip connection from a DF of vehicle type u to a DF of type v , such that $u \leq v$ within any URDHC, does not increase C .

Proof: Any DH trip from u to v will link, in the final vehicle chain (block), an arrival epoch at u and a departure epoch at v , including possible idle times at u and v . Because $c_u \geq c_v$, this DH trip connection cannot lead to an upgrade of the vehicle type; therefore, it cannot increase the objective function C .

Proposition 3 (for Step 6(b)): Any URDHC that aims at saving a vehicle of type w does not increase C if $-c_w + \sum_{\text{for all } q,r \in E} (c_q - c_r) \leq 0$, where the set E is composed of all the DH trips included in the URDHC; each DH trip is from a DF of vehicle type r to a DF of type q in which $q < r$.

Proof: In the URDHC, there may be DH trips that comply with Propositions 1 and 2 and, therefore, do not increase C . Each of the other DH trips is from a DF of vehicle type r to a DF of type q , in which $q < r$, i.e. $c_q \geq c_r$. This DH trip connection may upgrade the vehicle type (carrying out this DH trip) from r to q . That is, the result may be a saving of a vehicle of type w , along with several vehicle-type upgrades. In order to ensure that C does not increase, the

condition set forth is that the cost saving of vehicle type w is greater than, or equal to, the additional cost required by the possible upgrades.

Algorithm VTSP is shown in Figure 9.2 in two parts. It starts in Figure 9.2(a) by calculating N_1 and C_1 , and then N_2 and C_2 . If $N_1 = N_2$, we stop with $N = N_2$ and $C = C_2$, indicating that each trip is carried out by its own designated type. Otherwise, the algorithm continues by using Figure 9.2(b), in which Step 6, with its two conditions, is utilized. The algorithm ends once all terminals are examined.

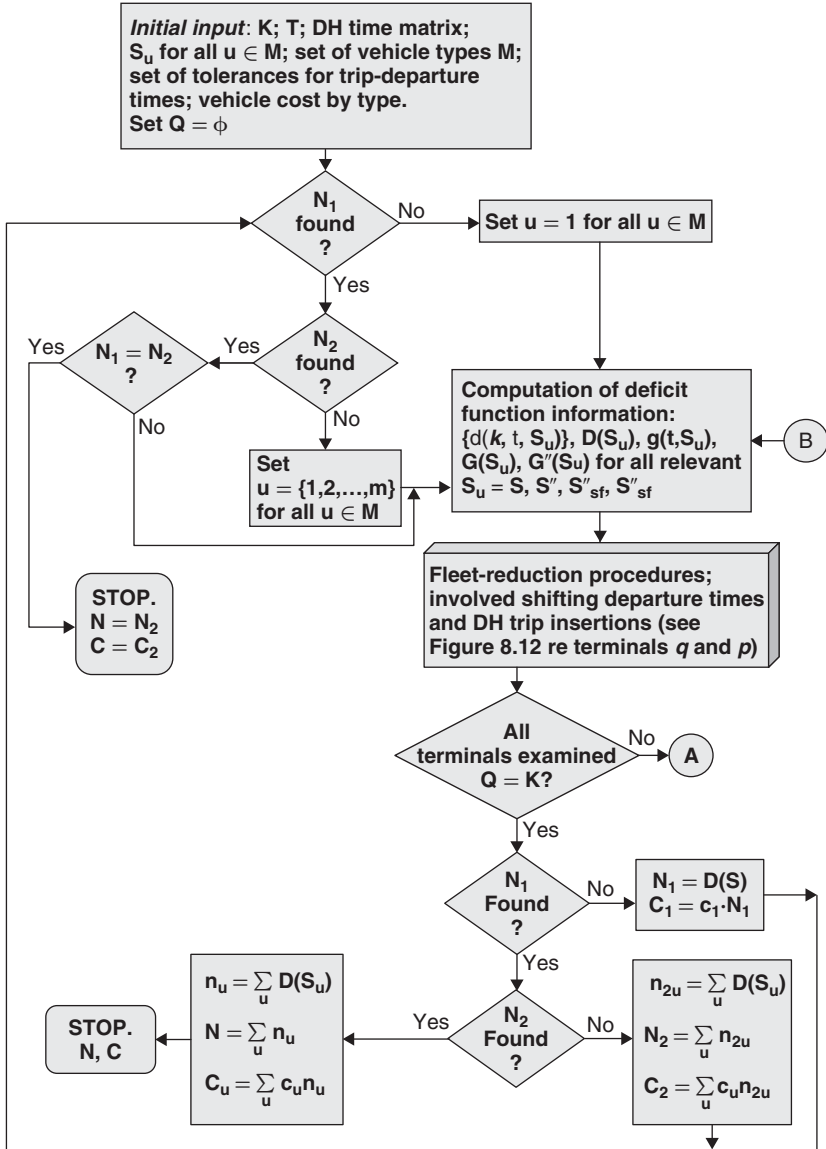


Figure 9.2(a) Flow diagram of the VTSP algorithm

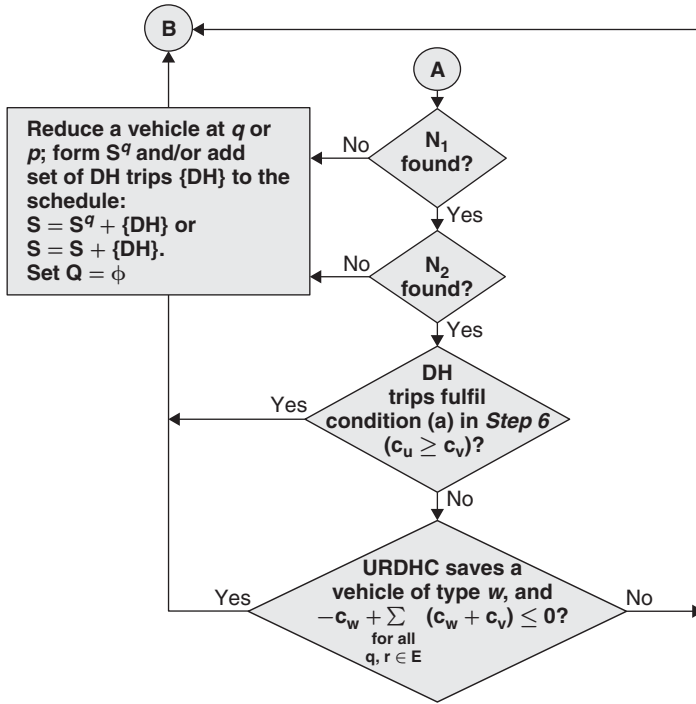


Figure 9.2(b) Continuation of the flow diagram of the VTSP algorithm

9.3.2 Sensitivity analysis

The VTSP presents a vehicle-scheduling problem in which the objective is to minimize the cost involved; primarily it aims at vehicle-purchase cost although indirectly it also affects operational cost. Often we want to know the result of the solution when changes occur in vehicle cost. Alternatively we may want to know the extent to which c_u for all $u \in M$ can change and still have the same scheduling solution. Certainly changes in vehicle cost by type can shift the location of $u \in M$ in decreasing order of vehicle cost. The examination of the solution for these cost-change possibilities is usually called sensitivity analysis.

The first observation in algorithm VTSP is that as long as the arrangement in M of the decreasing order of vehicle cost is preserved, the scheduling solution will not change. That is, any change of c_u – it can be of more than one vehicle type – that will not change the order of vehicle cost will result (except in special cases) in a change of C , but not in the blocks attained. The proof of this observation is straightforward because of unchanged situations applied to conditions (a) and (b) in Step 6 in algorithm VTSP.

The second observation concerns changes in which the order of vehicle costs is also changed. Following the cost change(s), three cases will be analysed: (1) a DH trip that fulfils condition (a) in Step 6 of algorithm VTSP, changing from a DF of vehicle type u to a DF of type v in which $c_u < c_v$; (2) a DH trip that fulfils condition (b), thus becoming fulfilled condition (a) in Step 6 of algorithm VTSP; and (3) a chain of DH trips that fulfils condition

(b) in *Step 6* of algorithm VTSP. Case (1) will not change the scheduling solution – it will only change the value of C – because the former fulfilled condition (a) now becomes fulfilled condition (b). That is, saving a vehicle of type v and converting vehicle type u to v results in $-(c_v + \Delta_v) + (c_v + \Delta_v) - c_u < 0$, in which Δ_v is the increase in c_v that yields $c_u \Delta_v c_v$. Case (2) will also not change the scheduling solution because of the existence of condition (a). In Case (3), condition (b) in *Step 6* of algorithm VTSP needs to be checked. If fulfilled, the scheduling solution remains the same; otherwise, algorithm VTSP needs to be repeated.

9.4 Examples

Three examples are presented in this section for comprehending the underlying principles of the VTSP algorithm. The first two are detailed examples, with and without shifting departure times; the third conveys the results of a real-life example.

9.4.1 Detailed example 1

Example 1, which is illustrated in Figure 9.3, consists of 8 trips, three terminals (a, b, c), and three types of vehicles, with the cost of 12, 5 and 3 cost-units, respectively. Figure 9.3(a) presents the simple network of the routes, in which the DH travel time between each two terminals is 20 minutes, and shifts in departure times are not allowed. The timetable and trip travel times are shown in Figure 9.3(b) according to vehicle type. The DFs of *Step 1* of algorithm VTSP for Example 1 are depicted in Figure 9.3(c); all trips, it should be recalled, are served by the same vehicle type (type 1). For inserting a DH trip, the NT rule (the first hollow is the longest) is applied; this results in the selection of terminal b (for the DH insertion procedure, see Section 7.5.4 in Chapter 7). The URDHC procedure with $R = 2$ (furthest start of a hollow) then results in three DH trips, in which DH_2 is used for maintaining the level of $D(a)$. Figure 9.4 constructs $g'(t, S') = g''(t, S'')$, which determines $G'(S') = G''(S'') = 2$. Thus, *Step 1* stops when $D(S) = N_1 = 2$. Two vehicle chains are then created, using the FIFO [1-3-DH₁-5-7] and [2-4-DH₂-6-DH₃-8], and the total cost is $C_1 = 24$.

Algorithm VTSP continues in *Step 2*, in which vehicle types are treated separately. Figure 9.5, which illustrates this step, is marked by (d) to show that this step follows Figure 9.3(c). The maximum DF of types 1 and 2 are reduced by one, using DH_1 and DH_2 , respectively; the number of type 3 vehicles remains the same. Thus, $N_2 = 1 + 1 + 2 = 4$, and the four following chains are derived by using the FIFO rule: [1-DH₁-4] (vehicle type 1), [2-5-DH₂-8] (vehicle type 2), [3-7], and [6] (vehicles of type 3); this results in a total cost of $C_2 = 23$.

The next step in algorithm VTSP compares $N_1 = 2$ with $N_2 = 4$, and then moves to the relevant *Step 6* because of not allowing shifting departure times. Figure 9.6 illustrates the process of *Step 6*, using its two conditions. Note that S_u will be deleted (as in Figure 9.6) when it is clear which underlying vehicle type is being considered. This step again applies the NT rule and the URDHC procedure with $R = 2$ (furthest start of a hollow), but this time (especially condition (a)) with the possibility of inserting any DH trip from a DF with a more expensive vehicle type to a DF with a less expensive type. Following the flow diagram in Figure 9.2, the first terminal selected is b , based on $d_2(b, t)$, from which DH_1 is determined

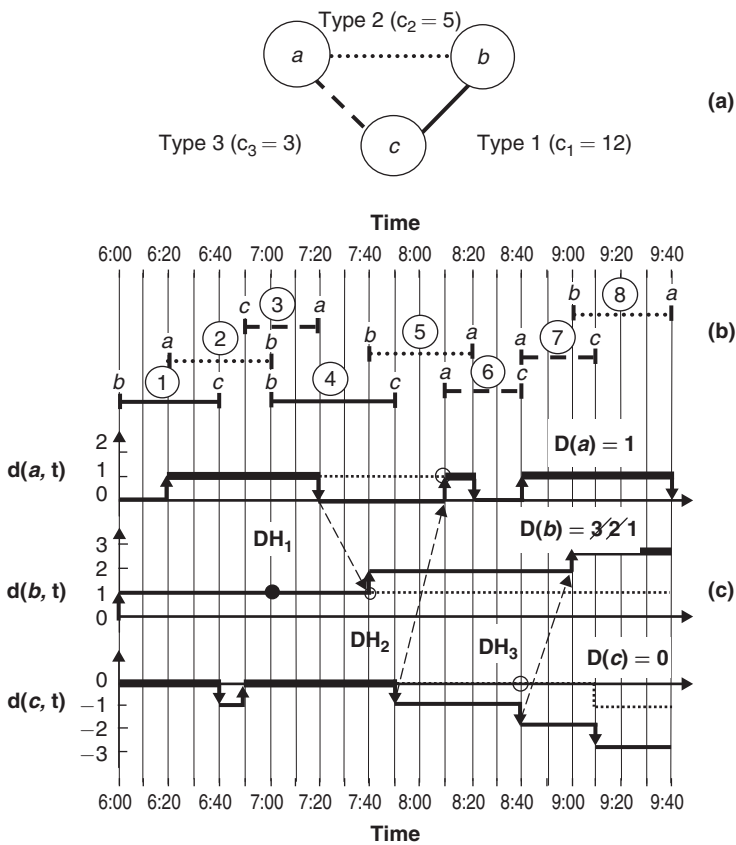


Figure 9.3 Example 1: (a) network of routes, vehicle types, and cost; (b) 8-trip schedule; (c) DFs with DH insertions for a single vehicle

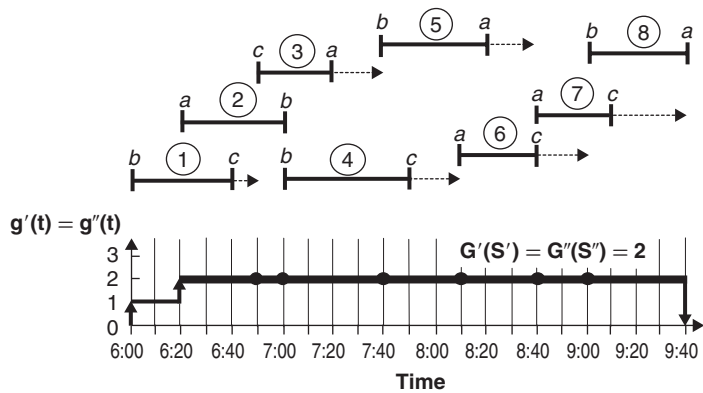


Figure 9.4 Determination of the lower bound corresponding to Step 1 of algorithm VTSP for Example 1

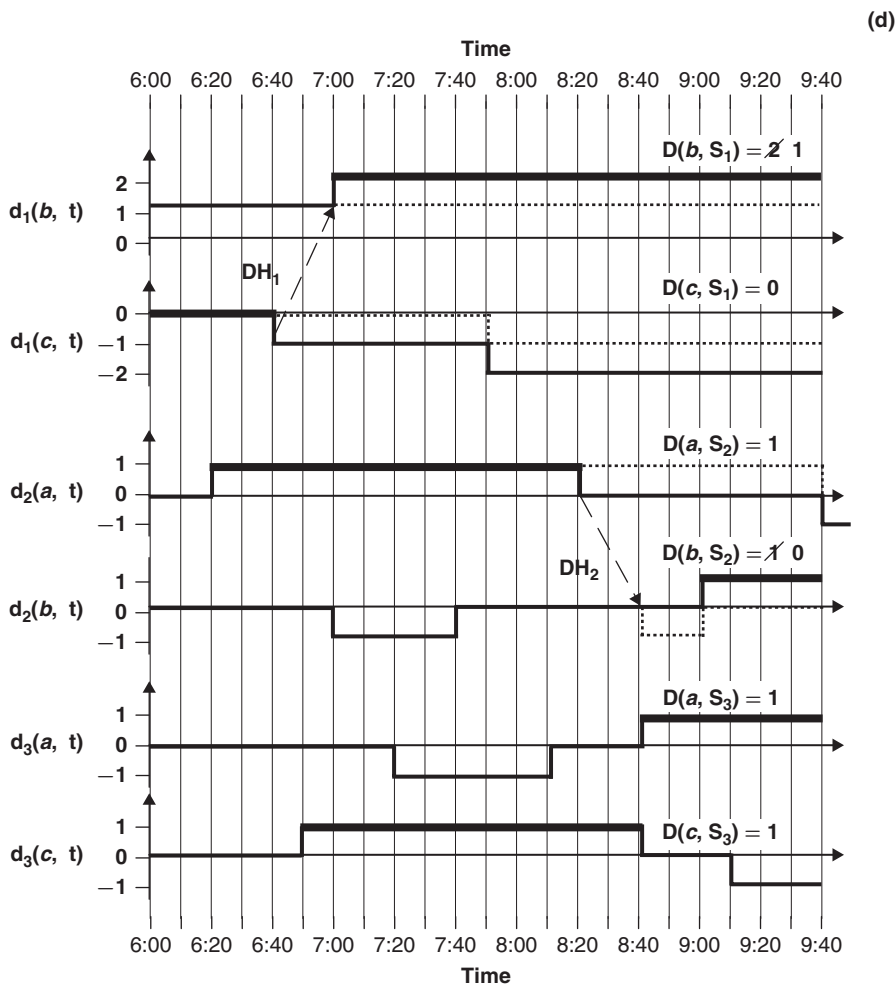


Figure 9.5 Example 1 (cont.): (d) DFs and DH insertions for each vehicle type

from terminal a . The DFs are then updated and the next terminal is again b , but related to $d_1(b, t)$; DH_2 is inserted from terminal c . We continue with the next selected terminal c , based on $d_3(c, t)$; however, no DH trip can be inserted into its maximum-interval starting point, including the check for both conditions (a) and (b) of Step 6. Thus, a is selected next, based on $d_3(a, t)$, and DH_3 is inserted to arrive from c , based on the updated $d_1(c, t)$. This terminates Step 6 and results in the three following (FIFO) chains: [1-DH₂-4-DH₃-6] (vehicle type 1), [2-5-DH₁-8] (vehicle type 2), and [3-7] (vehicle type 3), with a total cost of $12 + 5 + 3 = 20$.

For the sake of simplicity, we assume that a check of Step 7 of algorithm VTSP allows for the latter solution. Moving to Step 8, one may then see that $D(S) > G''(S'')$, but the process

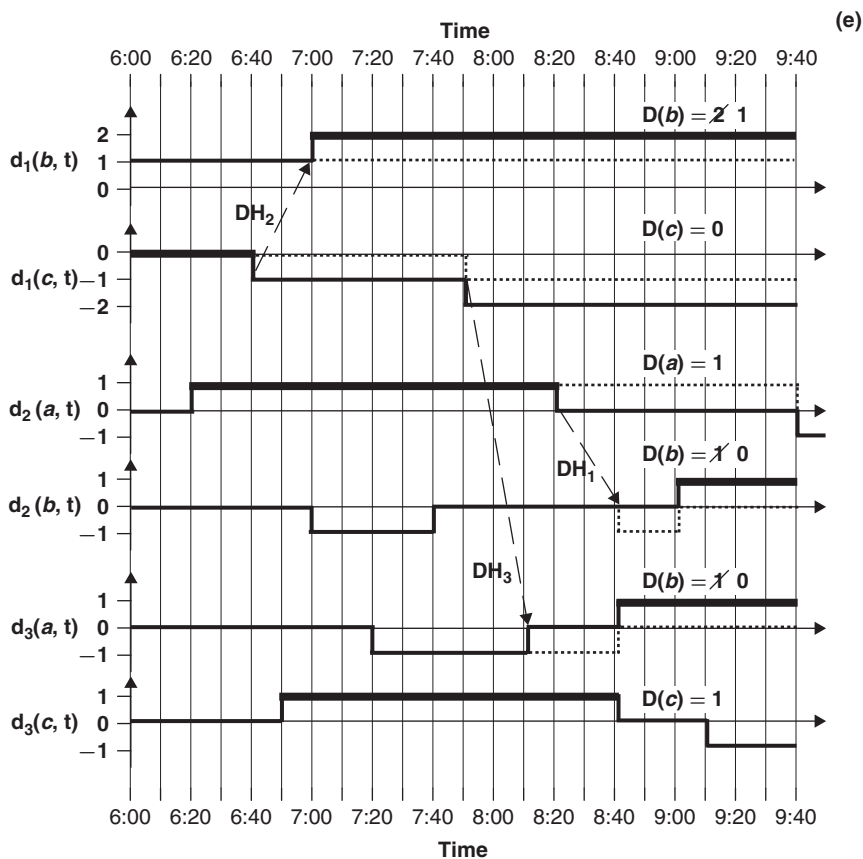


Figure 9.6 Example 1 (cont.): (e) DFs and DH insertions for optimal solution

stops by reiterating Step 6, because no URDHC can be found. At this stage, we may conclude that the 3-chain (blocks) solution with $C = 20$ is the best one attained.

9.4.2 Detailed example 2

The example demonstrated in this section exploits shifts in departure times and condition (b) in Step 6 of algorithm VTSP. Example 2, illustrated in parts (a) and (b) of Figure 9.7, consists of 14 trips, 3 terminals and 3 types of vehicles, with costs of 10, 3 and 2, respectively. Part (c) of Figure 9.7 shows the fleet-reduction procedure, using both DH trip insertion and shifting departure times. For simplicity, all DH travel times between the three terminals are 15 minutes, $\Delta_u^{i(+)} = \Delta_u^{i(-)} = 15$ minutes, for all $u \in M$, and priority in shifting departure times is given to late departures.

The DF analysis with only type 1, in Figure 9.7(c), results in one shift, enabling a reduction of one vehicle; the lower-bound determination, shown in Figure 9.8, includes possible

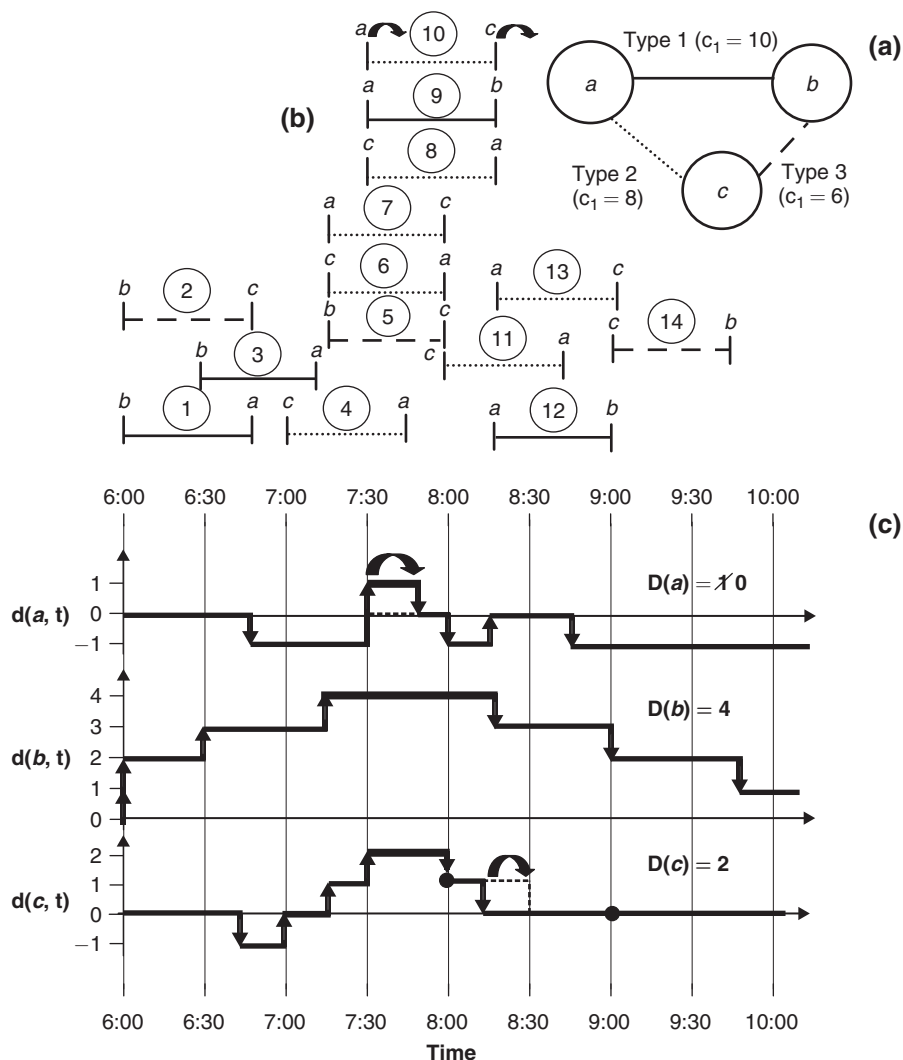


Figure 9.7 Example 2: (a) basic input of a network of routes, vehicle types, and cost; (b) complete 14-trip schedule; (c) DFs with Step 1 of algorithm VTSP

shifts of trip times combined with feasible extensions. The first analysis terminates with $N_1 = 6$ (lower bound) and $C_1 = 60$. The analysis proceeds with finding $N_2 = 7$ and $C_2 = 2 \cdot 10 + 4 \cdot 8 + 1 \cdot 6 = 58$, in which $n_{21} = 2$, $n_{22} = 4$, and $n_{23} = 1$, using Step 2 of algorithm VTSP. This is illustrated with one shift in departure time and one DH trip in Figure 9.9. Following the flow diagram in Figure 9.2(a), we can now check whether $N_1 = N_2$ to obtain that $N_1 < N_2$; then continuing with Steps 4 and 5 of the algorithm, resulting in a single shift forward (late departure) of trip 10 at 7:30 as is shown in Figure 9.10.

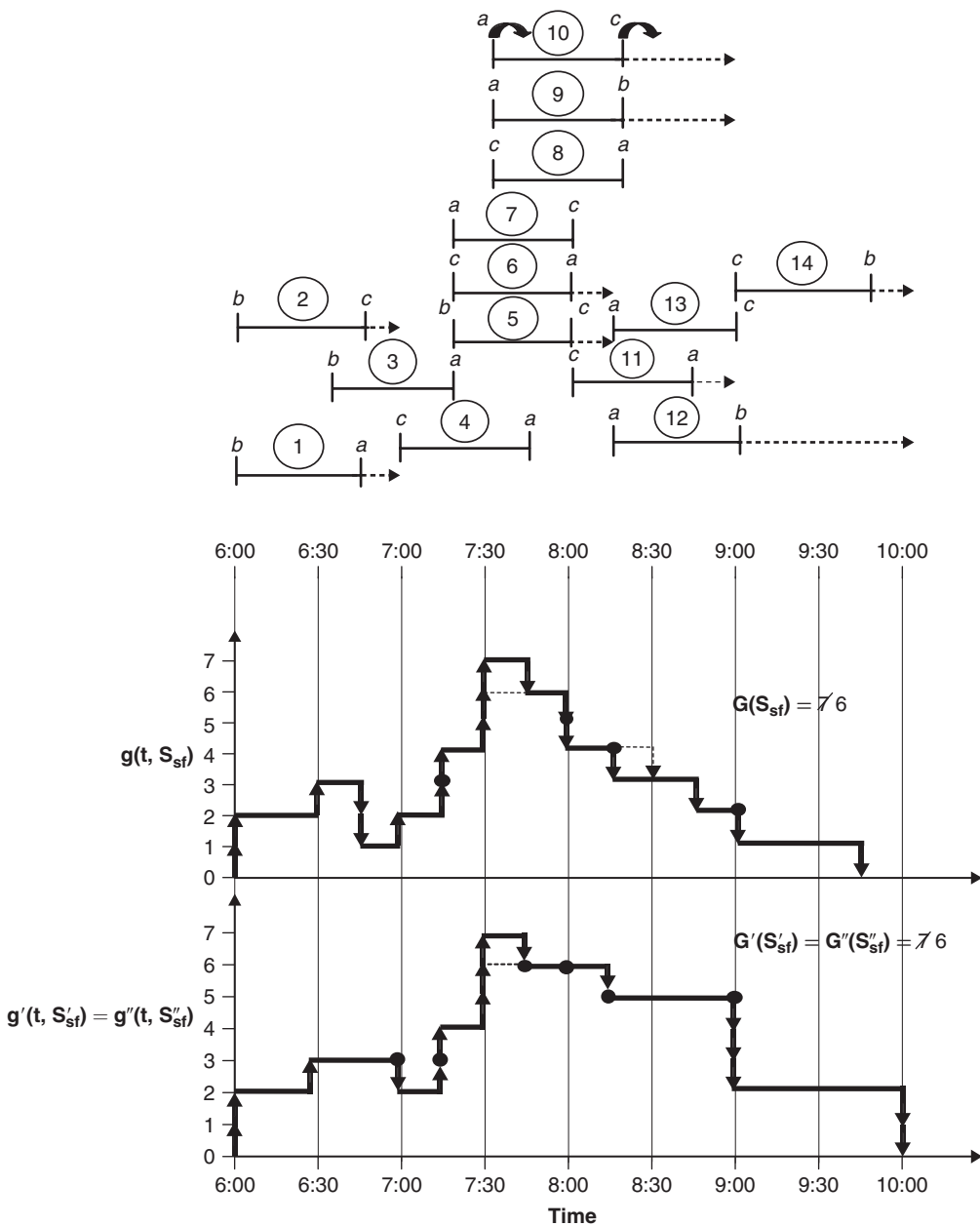


Figure 9.8 Determination of the lower bound corresponding to Step 1 of algorithm VTSP for Example 2

Because $D(S_{sf}) = 8 \neq G''(S''_{sf}) = 6$, the algorithm advances to a search of URDHC, including possible shifts in departure times. The NT rule determines terminal c of type 2 as a candidate location for reducing a vehicle; this results in DH_1 following condition (a) of Step 6 of the algorithm. However, in order to avoid the increase of $D(a)$ in $d_1(a, t)$, another DH trip is

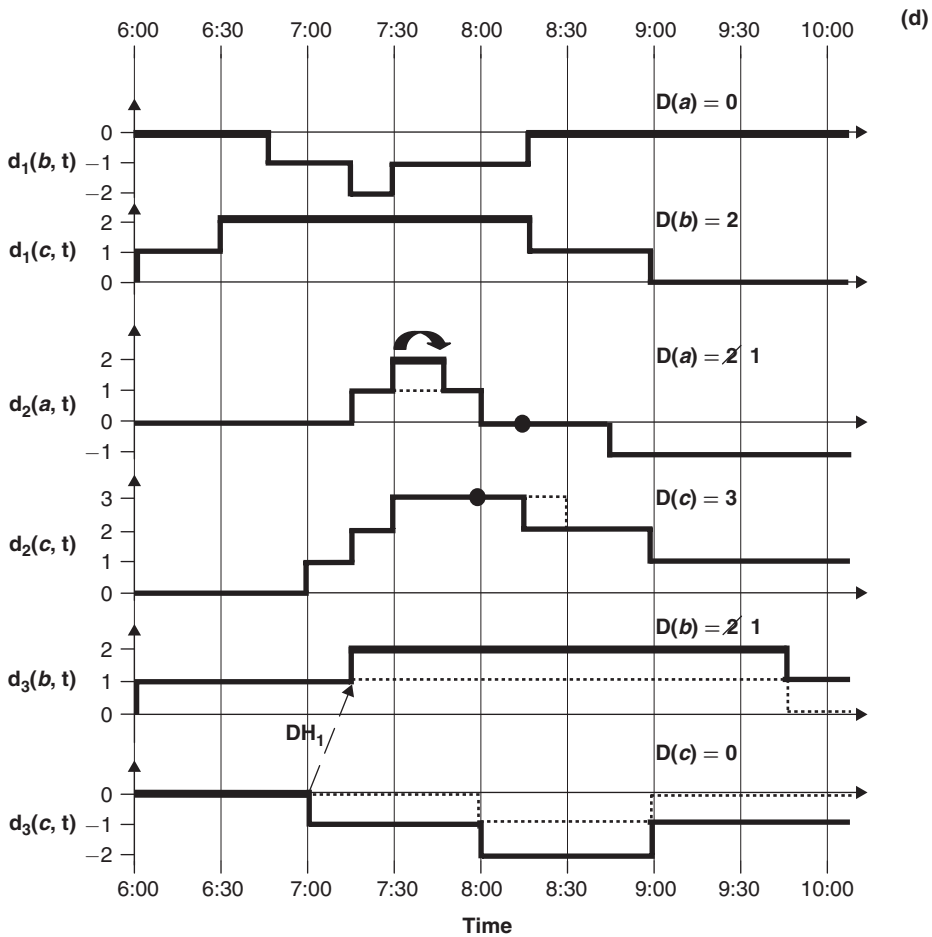


Figure 9.9 Example 2 (cont.): (d) DFs describing Step 2 of algorithm VTSP

found, DH_2 , but this time using condition (b) of Step 6 of the algorithm. The search for another URDHC allows for DH_3 and DH_4 , using conditions (b) and (a), respectively, of Step 6 of the algorithm. The procedure then stops because the sum of all $D(k) = 2 + 2 + 2 = 6$, $k = a, b, c$, equals the lower bound attained.

Unlike Example 1, the types of vehicles corresponding to the resultant six blocks cannot be derived by the sum $D(k)$, $k = a, b, c$, associated with each type. First, the blocks need to be constructed using the inserted DH trips; second, the type of each block is determined by its highest (most expensive) type trip. Following the FIFO rule, for instance, the first three blocks are of type 1: [1-9], [3- DH_1 -8] and [5- DH_2 -12] because there is at least one trip of this highest type in each block (Trips 8 and 5 are of types 2 and 3, respectively). The second three blocks are of type 2: [2- DH_3 -7-11- DH_4 -14], [6-13], [4-10] (Trips 2 and 14 are of type 3).

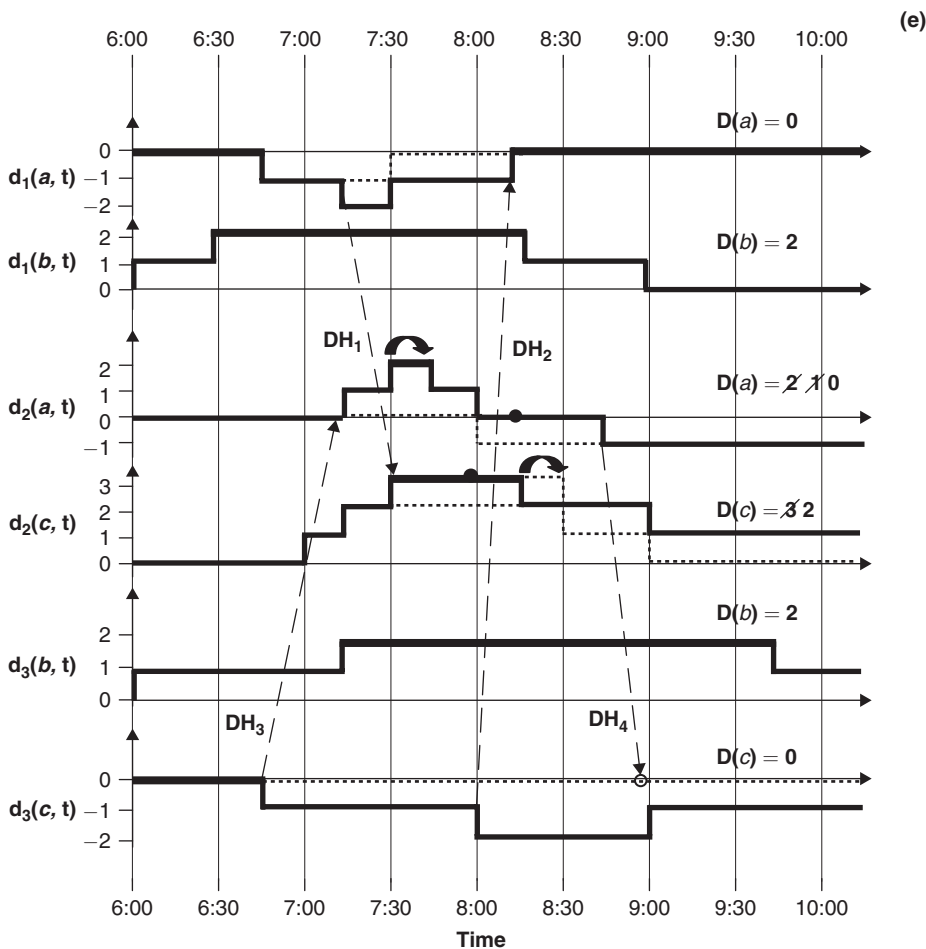
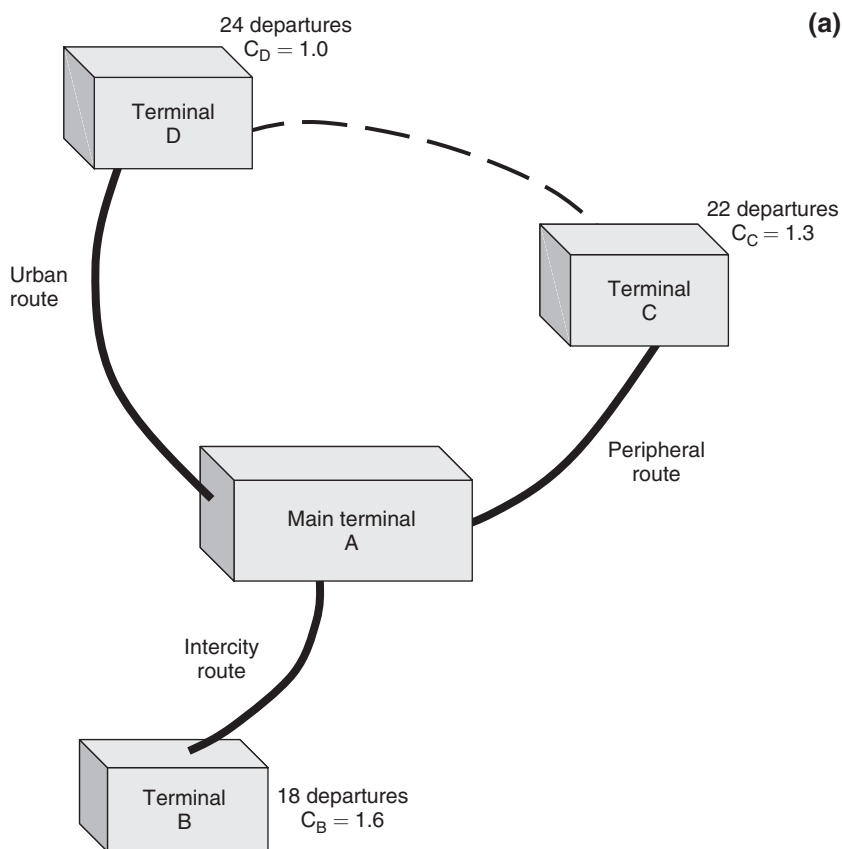


Figure 9.10 Example 2 (cont.): (e) DFs describing Step 5 and 6 of algorithm VTSP

Thus, in the best solution found, $n_1 = 3$, $n_2 = 3$, $n_3 = 0$, $N = 6$, and $C = 54$, compared with $C_1 = 60$ and $C_2 = 58$.

9.4.3 A real-life example

Algorithm VTSP was used to examine a real-life scheduling problem. The problem selected pertains to the Egged bus company, which has three different bus routes departing from a main terminal in Haifa. The three routes are shown schematically in part (a) of Figure 9.11: intercity, peripheral and urban. The daily timetable of the intercity route is characterized by 18 departures, 120 minutes being the average travel time and 105 minutes the average DH



(b)

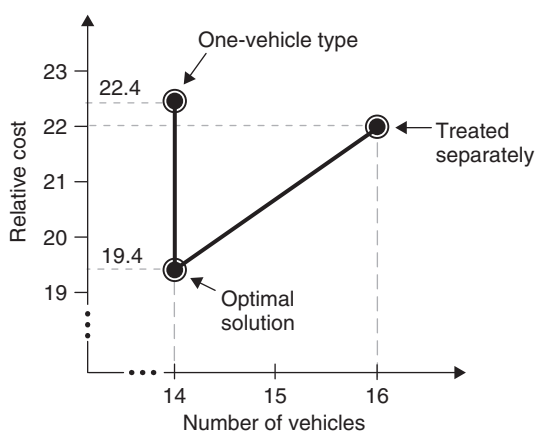


Figure 9.11 (a) Schematic description of a real-life, four-terminal, three-vehicles type example; (b) trade-off situation between total number of vehicles and relative cost associated with five disparate solutions

time between terminals A and B in each direction. The peripheral route has 22 daily departures, with 45 minutes the average travel time between terminals A and C, 24 minutes the average DH time between A and C, and 36 minutes between C and D. The urban route has 24 daily departures, with 30 minutes the average travel time and 15 minutes the average DH time between A and D. The relative costs of the intercity, peripheral and urban vehicles are 1.6, 1.3 and 1.0, respectively. The allowed shifts are $\Delta_u^{i(+)} = \Delta_u^{i(-)} = 3$ minutes for all trips i and vehicle type u .

The VTSP algorithm results in the optimal solution shown in part (b) of Figure 9.11, with $C = 19.4$ and 14 vehicles required: 7 intercity, 4 peripheral, and 3 urban vehicles, respectively. Steps 1 and 2 of algorithm VTSP result in $C_1 = 22.4$ (14 intercity vehicles) and $C_2 = 22$ (7 intercity, 6 peripheral and 3 urban vehicles), respectively. This outcome of the algorithm is circled in part (b) of Figure 9.11. The results of the heuristic method suggest that algorithm VTSP can be used, too, for large transit agencies, ensuring an efficient allocation of the different vehicles to trips while reducing the cost involved to a minimum level. The use of DF may be subject to the interference of the schedulers in the process whenever they think it justifiable.

9.5 Vehicle-size determination

There is doubt that there is a connection between different vehicle types and the size of vehicles used in public transit services. In the foregoing section, it was assumed that all vehicle types possessed a similar size so that the frequencies and timetables determined remained unchanged. However, practice often entails making a decision about vehicle size and the amount available for purchasing. That is, a bus fleet can consist of vehicles of the same size, thereby simplifying operations planning, or have a mix of sizes; i.e. minibuses, standard buses, articulated/double-decker buses, etc. This section provides some preliminary analytical tools for evaluating trade-offs between different vehicle sizes and for reviewing the models developed.

9.5.1 Example: standard bus versus two minibuses

Given a bus route operated by standard 50-seat buses, can the standard bus be exchanged more cost effectively and viably for two minibuses with 25 seats each? The first line of thought is directed to the question: What operational strategy should be conducted for the two minibuses? Basically two operational alternatives can be drawn: (1) Let the two minibuses depart at the same time, with each minibus picking up and dropping off passengers at each other stop. This will result in reducing travel time while retaining the timetable (frequency). (2) Let the minibus headway be half that of the standard bus and let each minibus pick up and drop off at each stop. This alternative will result in reducing passenger waiting time while maintaining average travel time.

The following basic analysis assumes that for alternative (1) we have a long bus route (end effects ignored) that is a typical commuter collector – e.g. to the CBD – and the ridership is unchanged. The assumptions for alternative (2) are that the average waiting time is half the headway (deterministic case; on headway distribution see Chapter 17), that the

probability of no one being at the stop is the same for both the minibus and the standard bus, and that the ridership is fixed.

For an example of alternative (1), the following data are used (given in US\$): standard bus (SB) operating cost = \$15/hr, minibus (MB) operating cost = \$12/hr, SB travel speed denoted by μ miles/hr = $3/4$ MB travel speed, P = average number of passengers on board SB, and the value of a passenger's riding time is $\$(1/3)/\text{hr}$. Hence, SB operating cost per μ miles = $15 \times 1 = \$15$, and 2 MB operating cost per μ miles = $2 (12 \times 3/4) = \$18$. The problem arises: Who is going to pay the additional cost for the 2 MB (\$3 per μ miles)? A possible answer is the passenger, because their riding time is reduced. A saving in travel time by MB per μ miles is $1 - 3/4 = 1/4$ hr. The value of the total saving is $1/4 \cdot P \cdot 1/3 = P/12$. Requiring that the saving be greater than the additional cost results in $P/12 > \$3$ or $P > 36$ passengers. In other words, a two-minibus operation is preferable in alternative (1) for average loads greater than 36 on the standard bus.

For an example of alternative (2), the following data are used: SB and MB operating cost and P are the same as in the foregoing example, SB headway = 10 minutes, MB headway = 5 minutes, and the value of waiting time per passenger = \$1/hr. The number of hourly on-board passengers is $6P$, and the saving in average waiting time when using MB is $10/2 - 5/2 = 2.5$ minutes per passenger. Total hourly saving, $6P \cdot 2.5/60 \cdot 1 = \$(1/4)P$, to be viable must be greater than the additional MB hourly operating cost. That is, $(1/4)P > 2 \cdot 12$, which results, as in the example for alternative (1), in $P > 36$ passengers.

9.5.2 Vehicle-size square root formula

A simple derivation of transit vehicle size accentuates a trade-off between vehicle capacity cost and passenger waiting-time cost. We assume that operating cost per vehicle-hour is the same (independent of vehicle size and load carried), average waiting time is half the headway, riding time is independent of vehicle size, and travel time and stopping time are independent.

The following notation is used:

Z = desired average occupancy in the max load segment, which corresponds to *vehicle size*

C_b = operating cost per vehicle-hour

P = number of average hourly passengers carried in the max load segment

C_w = value of hourly waiting time.

The hourly frequency, utilizing Method 2 in Chapter 3, is P/Z , and the headway (in hours) is Z/P . Hence, the total operating cost is $C_b \cdot P/Z$, and the passenger waiting-time cost is $\frac{1}{2} Z/P \cdot P \cdot C_w = \frac{1}{2} Z C_w$. The overall cost, therefore, is $C_b \cdot P/Z = \frac{1}{2} Z C_w$.

The optimal vehicle size, Z_0 , is attained by minimizing the overall cost using the first derivative of the overall cost with respect to $Z = 0$. This results in the following formula:

$$Z_0 = \sqrt{\frac{2C_b P}{C_w}} \quad (9.1)$$

Equation (9.1), called the *square root formula*, signifies, generally speaking, that the optimal vehicle size varies according to the square root of the number of passengers carried. In addition, this optimal size is sensitive to changes in C_b and C_w . The next section reviews more advanced vehicle-size modelling.

9.6 Optimal transit-vehicle size: literature review

The following notation is used throughout this section (unless mentioned otherwise):

- Z = optimal vehicle size
- a = constant of operational cost function (vehicle operational cost as a function of size, $a + bZ$)
- b = slope of operational cost function
- w = value of passenger waiting time
- x = value of access/walking time
- r = value of riding/in-vehicle time
- L = route length
- P = total passenger flow
- Q = peak passenger flow
- V = vehicle speed
- φ = maximum allowed load factor.

In some cases, parameters are given in a different notation in order to enable a convenient comparison.

An early reference to bus-size optimization was made by Jansson (1980), who paid significant attention to the cost of operating similar services at peak and off-peak during daylight. The following formula for optimal bus size was developed:

$$Z = \frac{1}{\varphi} \sqrt{\frac{ahJQ}{\frac{\beta Ew}{2} + tQ \left(\beta Er + \frac{b}{\varphi} \right)}} \quad (9.2)$$

where:

- α = ratio of off-peak to peak passenger flow
- h = running and transitional time per km
- J = average trip length
- β = ratio of mean flow rate for the whole day to the peak-hour flow
- E = extent of peak + off-peak periods per day
- t = boarding and alighting time per passenger.

Calculation of the optimal bus size, given various levels of passenger flows, leads to the conclusion that most buses in actual operation are too big.

Gwilliam, *et al.* (1985) describe a simple formula for an optimal bus size:

$$Z = 2\sqrt{aP} \quad (9.3)$$

Based on assigning common values of a and P to their formula, the authors claim that operating much smaller buses than those commonly used is not justified.

A more sophisticated model that aims at determining optimal bus size is provided by Oldfield and Bly (1988). Unlike most other models, this one assumes elastic demand; that is, demand that changes with passenger-trip cost. The model also takes into account the influence of changes in bus demand on road congestion; it further assumes that in-vehicle

trip time increases with average bus load because of boarding times. Significant attention is given to the effect of capacity constraints on average passenger waiting time in cases in which passengers are unable to board the first bus that arrives because it is full. Complex expressions are introduced explaining the dependence of headway on bus size. The expression developed for optimal bus size is the following:

$$Z^2 = AY_i \left\{ \frac{aY_i^{\frac{1}{2}}W_i}{(1+\beta)Z} + \frac{nP}{K} + U + R \right\}^{-g} \quad (9.4)$$

where:

- K = number of bus-km provided
- U = average cost of walking time to/from bus stop
- R = cost of time spent in the bus
- n = time-value constant that takes into account the extra time that a big bus spends at stops owing to a large number of boardings
- g = elasticity with respect to generalized cost (positive)
- A, β , Y_i , Z_i = constants (expressions for calculating them are developed in the paper).

Assigning common values of urban bus services in the UK, assuming no subsidies, the authors find that the optimal bus size lies between 55 and 65 seats. Under various other conditions, the model suggests that the optimum size at typical urban levels of demand will not be fewer than about 40 seats. Although this model is very sensitive to various phenomena that influence passenger demand and operator costs, it seems that the preliminary modelling and calibrating efforts required for using it are too intensive for practical use.

Jansson (1993) presents a model that simultaneously optimizes vehicle size, frequency, and journey price. All passengers having the same origin and destination along the route are referred to as a group, and each group may have its own value of time. Optimal vehicle size is computed as follows:

$$Z = \frac{\sum_i X_i \sum_{mi} \sum_j X_{j/mi} \frac{\partial r}{\partial R_m} \frac{h_m}{FNZ}}{F \left(\frac{\partial c_c}{\partial Z} h + L \frac{\partial c_\gamma}{\partial Z} \right)} \quad (9.5)$$

where:

- X_i = number of passengers belonging to group i
- $X_{j/mi}$ = number of passengers in group j who travel on link m , where passengers in group i also travel
- R_m = number of passengers per seat on link m
- F = frequency of service
- N = number of cars in the train if the vehicle discussed is a train, otherwise 1
- h = round-trip time
- h_m = time on link m

c_c = operating cost, which increases with vehicle size

c_γ = operating cost, which increases with distance travelled.

Implementation of this model may be somewhat more difficult compared with other models, since it requires detailed data about the origin and destination of all passengers.

Shih and Mahmassani (1994) developed a vehicle-sizing model which assumes that the total demand matrix of the whole route system is given, not the demand for each line. The load profile of each line is determined in a transit assignment. The optimal bus size is computed as follows:

$$Z = \frac{Q_k}{\varphi} \sqrt{\frac{2aL_k}{wP_k}} \quad (9.6)$$

where:

k = route index.

An iterative process is suggested that includes re-assigning the total demand matrix on the route system after the determination of optimal vehicle size and frequency for each route. Each iteration of the transit assignment yields corrected values of $(Q_k)_{\max}$ and P_k , so that new optimal vehicle sizes can be calculated. The iterative process does not seek to minimize system-wide cost, but to determine the optimal bus size for each separate line. Implementation of the procedure is illustrated using data from a transit network in Austin, Texas. Of 40 bus routes, results suggest that 37 have an optimal bus size of below 25 seats.

A model developed by Lee *et al.* (1995) attempts to find the optimal bus size not only for each route but also for each period of day, so that more than one bus size can be used on one route. The model also tries to determine the conditions under which it is better to use one bus size or, alternatively, a mixed-size fleet. The bus size that gives the minimum total operator and user cost on one line is determined as follows:

$$Z = \sqrt{\frac{2aLQ^2}{wVP}} \quad (9.7)$$

If only one bus size is operated on all routes during all periods, the system-wide optimal size is found as follows:

$$Z = \sqrt{\frac{\sum_{r=1}^n \sum_{t=1}^m 2aL_r Q_{rt} / V_{rt}}{\sum_{r=1}^n \sum_{t=1}^m wP_{rt} / Q_{rt}}} \quad (9.8)$$

where:

r = route index

n = number of routes

t = time-period index

m = number of time periods.

If two bus sizes (Z_1 and Z_2) are used, the following test can determine which is better for each route:

$$\frac{Q^2L}{PV} = \frac{wZ_1Z_2}{2a} \quad (9.9)$$

If the left-hand side is greater than the right side, bigger buses should be used. If the right-hand side is greater, the operation of smaller buses is justified. For cases in which Z_1 and Z_2 are not given, the paper describes an algorithm for determining the two bus sizes that give minimum cost. The algorithm is illustrated in a simple 4-route network with a 2-period demand. The optimal fleet is concluded to consist of 15 buses with 33 seats and 29 buses with 20 seats. The total fleet size required for a network operation with this mix is smaller than that needed for operating the same route system with buses all the same size. It is also shown that under the conditions of the given example, a mixed-fleet operation is preferable if the ratio of peak demand to off-peak demand is more than 1.92. If there is no significant demand variation between periods of the day, then the operation of a uniform bus size is preferable.

Gronau (2000) presents this model for determining the optimal bus size for a specific route:

$$Z = \sqrt{\frac{\alpha_0 + \beta_0 t_0}{\frac{\lambda r}{P} + t_1(\beta_1 + r)}} \quad (9.10)$$

where:

$\alpha_0 + \alpha_1 Z$ = distance-related operating cost

$\beta_0 + \beta_1 Z$ = time-related operating cost

$t_0 + t_1 Z$ = bus travel time

λ = ratio of the value of waiting time to the value of vehicle time, divided by the ratio of the headway to the waiting time (expected value: 1.5–2).

This model, introduced as a basis for a series of mathematical developments, aims at examining the option of using two bus sizes on one route that serves passengers with different values of time. The usefulness of this option is investigated in detail.

Tisato (2000) analyses the variation of public transit subsidy levels among several constraint cases with respect to bus size and load factor. Four cases are analysed: fixed-load factor with variable bus size; fixed bus size with variable load factor; both fixed; both variable. Conditions for maximum economic surplus are determined for each of the cases. In the case in which bus size is the only variable, the expression derived for optimal bus size is this:

$$Z = \sqrt{\frac{aLPA^2(1.25 - 1.65\varphi)}{15w\varphi^2}} \quad (9.11)$$

where:

A = average passenger-trip length, divided by bus-trip length.

In addition, it is shown how optimal bus size changes with a varying target-load factor.

A sub-group of vehicle-size optimization tools consists of the formulation of bus-sizing models as part of a comparison between fixed-route and flexible-route services. Chang and Schonfeld (1991), who make such a comparison, argue that a rectangular service area should be connected to an adjacent transportation terminal. If a fixed-route service is provided, the optimal bus size is the following:

$$Z = \left(\frac{8a^2gSDL^2}{xwV^2} \right)^{\frac{1}{3}} \quad (9.12)$$

where:

- g = access speed
- S = length of service area
- D = passenger-demand density in service area.

If the bus route is flexible, the optimal size is this:

$$Z = \left(\frac{ua^3L_T^3D}{wk^2V(b + r/2)^2} \right)^{\frac{1}{5}} \quad (9.13)$$

where:

- u = average number of passengers per pickup point
- L_T = equivalent-line haul distance
- k = constant (estimated value described in the paper).

Using these formulas, the authors show that flexible routes require smaller buses than do conventional services.

Another model that compares conventional and flexible routes is presented by Chien *et al.* (2001). They assume a given probabilistic demand function and a non-additive value of time; that is, the cost of one ten-minute wait is higher than the total combined cost of ten 1-minute waits. The model is not solved analytically, so that there is no explicit formula for optimal bus size. The optimum size in a fixed-route system is calculated by minimizing the following expression, c , with decision variables h and Z :

$$\min c = \frac{L(a + bZ)}{VZ} + r \left(\frac{M}{V} \right)^2 + w \left(\frac{Z}{2hSq(Q', \sigma)} \right)^2 + x \left(\frac{r + d}{4g} \right)^2 \quad (9.14)$$

where:

- h = route spacing
- M = average passenger-trip distance
- S = length of service area
- $q(Q', \sigma)$ = probabilistic demand-density function
- d = stop spacing
- g = average walking speed.

If a flexible-route system is operated, the optimal bus size is calculated by solving the following problem, with Z and A as decision variables:

$$\begin{aligned} \min c = & \frac{L_T(a + bZ)}{VS} + \frac{k\sqrt{A/(uZ)}(a + bZ)}{V} + v \left(\frac{L_T}{2V} \right)^2 \\ & + v \frac{AZ}{u} \left(\frac{k}{2V} \right)^2 + w \left(\frac{Z}{2Aq(Q', \sigma)} \right)^2 \end{aligned}$$

where:

A = service zone area

L_T = equivalent bus-trip distance

k = constant (estimated value described in the paper)

u = average number of passengers per pickup point.

The models reviewed are summarized in Table 9.1 according to their characteristics.

9.6.1 Literature-review conclusions

The following may be concluded from the review of the literature on optimal transit-vehicle size:

- According to most of the models, an increase in the value of passenger time decreases the optimal vehicle size. Vehicle size in some of the models is proportionally opposite to the square root of the time value.
- There is no consistency among models concerning which trip-time elements influence the optimal vehicle size. Each element (access, waiting in-vehicle) appears as a variable in some models, but no one element appears in all of them.
- In the majority of the models reviewed, the constant operating cost (parameter a of operating-cost function $C = a + bZ$) influences optimal vehicle size. In most of these cases, optimal vehicle size is proportional to the square root of this constant. Parameter b (slope of operating-cost function) appears in some of the models but not in a consistent manner.
- Dependency of optimal vehicle size on route length or time shows, in general, that the longer the transit route, the bigger are the vehicles required.
- Different models show opposite viewpoints on the relationship between demand and vehicle size: in some, a high passenger flow will increase the vehicles needed; in others, it will justify using smaller vehicles.
- Calibration of an operating-cost function is required for each of the models discussed. In some of the models (such as the one presented by Gronau), different functions are needed for distance-related and time-related operating costs. Calibrated values of passenger time are usually compulsory. In most of the models reviewed, no additional special effort is needed to calculate the optimal vehicle size other than using data that are readily available to most transit agencies. The model presented by Oldfield and Bly (1988) is different in that it requires much more sophisticated data. Similarly, the model developed by Jansson (1993) requires detailed information about passenger demand.

Table 9.1 Summary of the characteristics of the models reviewed

Source Comparison subject	Jansson (1980)	Gwilliam <i>et al.</i> (1985)	Oldfield and Bly (1988)	Jansson (1993)	Shih and Mahmassani (1994)	Lee <i>et al.</i> (1995)	Gronau (2000)	Tisato (2000)	Chang and Schonfeld (1991)	Chien <i>et al.</i> (2001)
Principles and assumptions Elastic route demand, with/without a fixed, system-wide, demand constraint	No	No	Yes, without	No	Yes, with	No	No	No	No	The model enables the use of any demand function
Demand influences road congestion	No	No	Yes	No	No	No	No	No	No	No
In-vehicle time increases when in-vehicle load increases	No	No	Yes	No	No	No	Yes	No	No	No
Waiting time is influenced by the chance of a full-vehicle arrival	No	No	Yes	No	No	No	No	No	No	No
Possibility of more than one vehicle size on one route	No	No	No	No	No	Yes	Yes (but not in the formula quoted here)	No	No	No
Differences between periods of the day	Yes	No	No	No	No	Yes	No	No	No	No

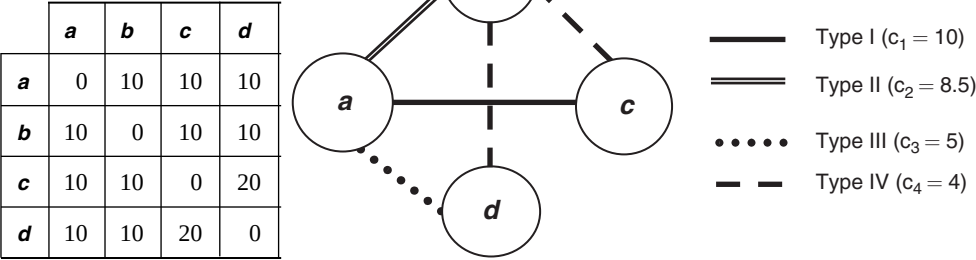
Total seats per hour (product of vehicle size and frequency)	Fixed	Fixed	Both fixed and not-fixed cases are examined	Fixed	Fixed	Fixed	Fixed	Both fixed and not-fixed cases are examined	Fixed	Fixed
Required effort in preparing input data	Reasonable	Easy	Intensive	Intensive. Detailed demand data are needed.	Reasonable. Transit-assignment model is needed.	Reasonable	Reasonable	Reasonable	Reasonable	Reasonable
Optimal bus size depends on: subsidy level	No	No	Yes	No	No	No	No	No	No	No
Fare	No	No	Yes	No	No	No	No	No	No	No
Cost of waiting	Yes	No	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Cost of access to/from bus stop	No	No	Yes	No	Yes	No	No	No	Fixed-route: Yes Flexible: No	Fixed-route: Yes Flexible: No
Cost of in-vehicle time	Yes	No	Yes	Yes	Yes	Yes	Yes	No	Fixed: No Flexible: Yes	Yes
Maximum allowed bus occupancy	Yes	No	No	No	Yes	No	No	Yes	No	No
Constant (a) of operating-cost function	No	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Yes
Slope (b) of operating-cost function	Yes	No	No	Yes	No	No	Yes	No	Fixed: no Flexible: yes	Yes
Route length or time	Yes	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes

Exercises

- 9.1 Given: a set of terminals, a , b , c and d ; four types of vehicles with their unit cost c_i ; $i = 1, 2, 3, 4$ (in monetary units); DH travel-time matrix; and a fixed schedule of trips.
- (a) Find the number of vehicles for each type required to obtain the minimum total cost; use algorithm VTSP.
 - (b) Find the solution to (a) after changing the costs of type II and type III from 8.5 and 5 to 10 and 4 cost units, respectively; use algorithm VTSP, but think about steps that do not need to be repeated.

Trip number	Type of vehicle	Departure time	Departure terminal	Arrival time	Arrival terminal
1	I	6:00	a	6:30	c
2	II	6:20	a	6:50	b
3	IV	6:20	b	6:50	c
4	III	6:40	d	7:10	a
5	I	7:10	a	7:40	a
6	II	7:10	b	7:40	a
7	III	7:20	d	7:50	a
8	I	7:40	c	8:10	a
9	IV	7:50	d	8:10	b
10	IV	8:00	b	8:30	c
11	IV	8:10	b	8:30	d
12	III	8:20	a	8:50	d

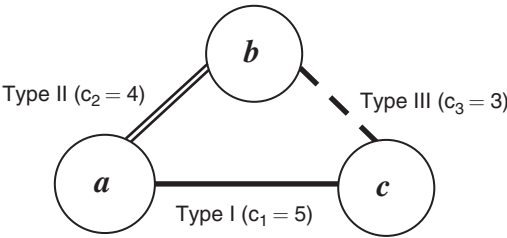
DH travel time (minutes)



- 9.2 Solve Example 1 in Section 9.4.1 using the following three new sets of input cost values; use algorithm VTSP, but think about steps that do not need to be repeated:
- (a) $c_1 = 10, c_2 = 5, c_3 = 3$
 - (b) $c_1 = 10, c_2 = 11, c_3 = 6$
 - (c) $c_1 = 12, c_2 = 8, c_3 = 9$
- 9.3 Given: a set of terminals, a, b, c ; three types of vehicles with unit costs c_i ; $i = 1, 2, 3$ (in monetary units); DH travel-time matrix; shifting tolerance of -15 minutes $\leq \Delta \leq 15$ minutes; and a fixed schedule of trips. Find the number of vehicles for each type required to obtain the minimum total cost; use algorithm VTSP.

DH travel time (minutes)

	<i>a</i>	<i>b</i>	<i>c</i>
<i>a</i>	0	30	30
<i>b</i>	30	0	30
<i>c</i>	30	30	0



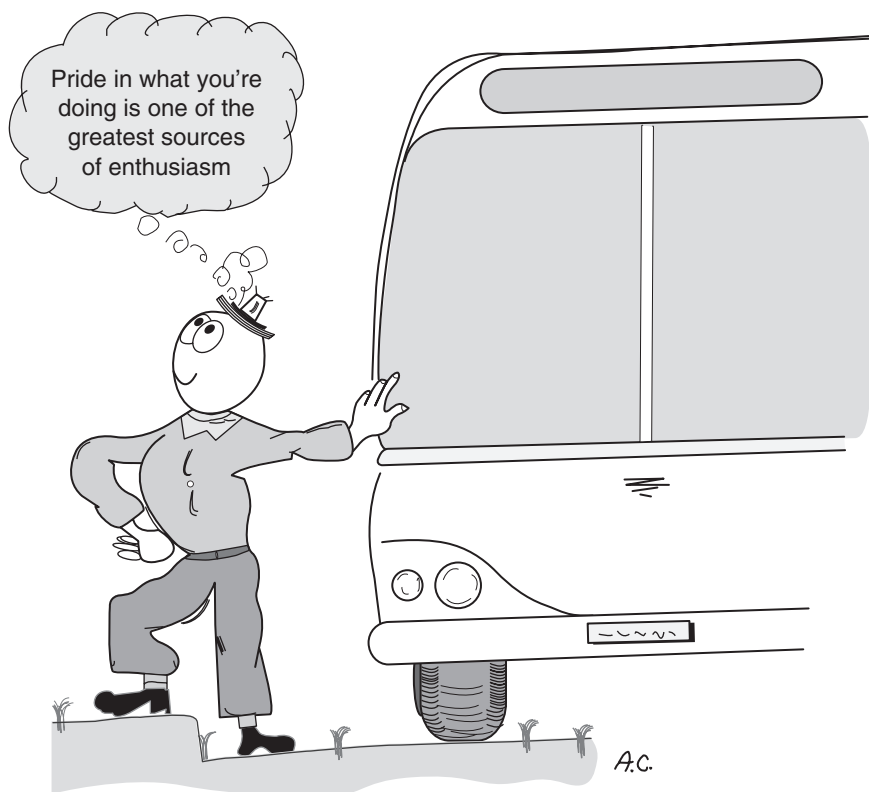
Trip number	Type of vehicle	Departure time	Departure terminal	Arrival time	Arrival terminal
1	I	6:00	<i>a</i>	7:00	<i>c</i>
2	II	6:00	<i>a</i>	7:00	<i>b</i>
3	III	6:00	<i>b</i>	7:00	<i>c</i>
4	III	6:00	<i>b</i>	7:00	<i>c</i>
5	II	6:30	<i>a</i>	7:30	<i>b</i>
6	II	6:30	<i>b</i>	7:30	<i>a</i>
7	III	7:00	<i>c</i>	8:00	<i>b</i>
8	II	7:00	<i>b</i>	8:00	<i>a</i>
9	I	8:00	<i>a</i>	9:00	<i>a</i>
10	III	8:00	<i>c</i>	9:00	<i>b</i>
11	I	8:00	<i>c</i>	9:00	<i>a</i>
12	I	8:30	<i>c</i>	9:30	<i>a</i>
13	II	8:30	<i>a</i>	9:30	<i>b</i>
14	II	9:00	<i>a</i>	10:00	<i>b</i>

- 9.4 Examine the trade-off between operating an articulated bus (AB) with 75 seats and operating three minibuses (MB) with 25 seats each. The following data and information are given: AB operating cost = \$25/hr, MB operating cost = \$15/hr, value of passenger-riding time is \$ $\frac{1}{2}$ /hr, frequency of AB and MB services are the same, MB travel speed is below AB travel speed of some operational tactics for the MB service are utilized, and for $P > 30$, in which P is the average number of passengers on-board AB, use of the MB service is preferred.
- Find the values of P for preferring the MB service if the MB travel speed (described above) is reduced by 10 per cent.
 - What are the upper and lower bounds on the travel speed of the AB service for positive P values?

References

- Bertossi, A., Carraresi, P. and Gallo, G. (1987). On some matching problems arising in vehicle scheduling. *Networks*, **17**, 271–281.
- Ceder, A. (1995a). Transit vehicle-type scheduling problem. *Transportation Research Record*, **1503**, 34–38.
- Ceder, A. (1995b). Minimum cost vehicle scheduling with different types of transit vehicles. In *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **430** (J. R. Daduna, I. Branco and J. M. P. Paixao, eds), pp. 102–114, Springer-Verlag.
- Chang, S. K. and Schonfeld, P. M. (1991). Optimization models for comparing conventional and subscription bus feeder services. *Transportation Science*, **25**(4), 281–298.
- Chien, S., Spasovic, L. N., Elefsiniotis, S. S. and Chhonkar, R. S. (2001). Evaluation of feeder bus systems with probabilistic time-varying demands and non-additive value of time. *Transportation Research Record*, **1760**, 47–55.
- Costa, A., Branco, I. and Paixao, J. (1995). Vehicle scheduling problem with multiple types of vehicles and a single depot. In *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **430** (J. R. Daduna, I. Branco and J. M. P. Paixao, eds), pp. 115–129, Springer-Verlag.
- Gronau, R. (2000). Optimum diversity in the public transport market. *Journal of Transport Economics and Policy*, **34**, 21–42.
- Gwilliam, K. M., Nash, C. A. and Mackie, P. J. (1985). Deregulating the bus industry in Britain – (B): The case against. *Transport Reviews*, **5**(2), 105–132.
- Jansson, J. O. (1980). A simple bus line model for optimization of service frequency and bus size. *Journal of Transport Economics and Policy*, **14**(1), 53–80.
- Jansson, K. (1993). Optimal public transport prices, service frequency, and transport unit size. *Selected Proceedings of the Sixth World Conference on Transport Research*, 1591–1602.
- Lee, K. K. T., Kuo, S. H. F. and Schonfeld, P. M. (1995). Optimal mixed bus fleet for urban operations. *Transportation Research Record*, **1503**, 39–48.
- Oldfield, R. H. and Bly, P. H. (1988). An analytic investigation of optimal bus size. *Transportation Research*, **22B**, 319–337.
- Shih, M. C. and Mahmassani, H. S. (1994). Vehicle sizing model for bus transit networks. *Transportation Research Record*, **1452**, 35–41.
- Tisato, P. (2000). A comparison of optimisation formulations in public transport subsidy. *International Journal of Transport Economics*, **27**, 199–228.

10 Crew Scheduling



Chapter 10 Crew Scheduling

Chapter outline

- 10.1 Introduction
 - 10.2 Vehicle-chain construction using a minimum crew-cost approach
 - 10.3 Mathematical solutions
 - 10.4 A case study: NJ commuter rail
 - 10.5 Crew rostering
 - 10.6 Literature review and further reading
 - Exercises
 - References
 - Appendix 10.A: The shortest-path problem
-

Practitioner's Corner

One of the most time-consuming and cumbersome scheduling tasks is assigning the crew (drivers) to vehicle blocks. The task requires the service of imaginative, experienced schedulers, and usually it is performed automatically. Consequently, it is not surprising to learn that most of the commercially available transit scheduling software packages concentrate primarily on crew-scheduling activities. After all, from the transit agency's perspective, the largest single-cost item in the budget is the driver's wage and fringe benefits. Because of the important implications of crew scheduling for providing good transit service, practitioners ought to comprehend the root of the problem, and be equipped with basic tools to be able to arrive at a solution.

This chapter consists of four principal parts, following an introductory section. Section 10.2 uses the DF (deficit function) properties for constructing vehicle chains (blocks) that take into account maximum paid idle time (swing time which is an unpaid break in a split duty). The latter consideration is aimed at helping to construct crew schedules with minimum cost. Section 10.3 presents the basic mathematical formulation used in solving crew-scheduling and rostering problems. In addition, this section accentuates an approximate solution to produce low-cost duty pieces of a crew schedule. Section 10.4 describes a case study performed for the New Jersey Transit Corporation that needed a tool for analysing the marginal cost impacts of changes in train schedules and for producing near optimal crew assignments (duties), given a schedule and a set of work rules. Finally, Section 10.5 discusses and provides examples of crew rostering. A roster is a pattern of duties to be fulfilled for a certain number of consecutive days; commonly the pattern repeats itself in cycles (of a week, a month, or any other period). The chapter ends with a literature review and exercises.

Practitioners are encouraged to visit all the sections while skipping only the mathematical formulation. They should pay special attention to the examples and their figures.

A crew schedule is called a pick. Often the selection of picks involves conflicts. Some may say that a selection without a conflict is almost as inconceivable as a nation

without crises. Two short stories: first, about being told that the pick is excellent. A man bought a newspaper for 50 cents. He found, however, that the price marked on the paper was 35 cents. He returned to the kiosk and asked about it. The reply was: don't believe everything written in newspapers. Second: two rabbits arrive at the planning department of a transit agency, and one stopped to cry. "Why are you crying?" she was asked. And she answered, "Because here they cut the fifth leg of every rabbit". "But you have only four legs, so there's nothing to worry about", she was told. She replied: "Yeah, but here they cut first and then count".

10.1 Introduction

The functional diagram of a typical transit-operation planning process, Figure 1.2 in Chapter 1, ends its fourth and last planning activity with crew scheduling, the aim being to assign drivers according to the outcome of vehicle scheduling. This activity is often called driver run-cutting (splitting and recombining vehicle blocks into legal driver duties, shifts, runs, or assignments). Part of a vehicle block is called a duty piece or a task. This crew-assignment process must comply with some constraints, which are usually dependent on a labour contract. The purpose of the assignment function is to determine a feasible set of driver duties in an optimal manner. Usually, the objective is to minimize the cost of duties so that each duty piece is included in one of the selected duties. It should be noted that vehicle-scheduling activity for railways is not cumbersome; however, it is important to consider how to build the work schedule of the train crews (drivers and conductors together) efficiently.

The criteria for crew scheduling are based on an efficient use of manpower resources while maintaining the integrity of any work-rule agreements. The construction of the selected crew schedule is usually a result of the following sub-functions: (i) duty piece analysis; (ii) work-rules coordination; (iii) feasible duty construction; and (iv) duty selection. The duty-piece analysis partitions each vehicle block at selected relief points into a set of duty pieces. These duty pieces are assembled in a feasible duty-construction function. Other required information: travel times between relief points and a list of relief points designated as required duty stops and start locations.

Theoretically, each relief point may be used to split the vehicle block into new duty pieces. Usually, it is more efficient to use one or more of the following criteria to select the relief points: (a) minimum duty-piece length; (b) next relief point, selected as close as possible to the maximum duty-piece time (maximum time before having a break); (c) only a few (say, two) relief points in each piece; and (d) operator decisions. In order to utilize any crew-scheduling method, a list of work rules to be used in the construction of feasible driver duties is required. The work rules are the result of an agreement between the drivers (or their unions) and the public transit agency (and/or public authorities).

The determination of different feasible sets of duties may be selected on the basis of, for example, one or more of the following performance measures: (1) number of duties (drivers), (2) number of split duties, (3) total number of changes, (4) total duty hours, (5) average duty length, (6) total working hours, (7) average working time, (8) number of short duties, and (9) costs. Sections 10.2–10.4 provide tools for handling both the preparation for and analysis of the construction of efficient crew schedules (duties).

Once the set of duties are established, it is common to group them into rosters. A roster is defined as a duty assignment over a certain amount of consecutive days, guaranteeing that all the trips are covered for a certain (usually cyclic) period. Commonly, a roster contains a subset of duties covering six consecutive days (called weeks). The length of a roster is typically between 30 and 60 days (5 to 10 weeks). The usual rostering problem is to find a feasible set of rosters to cover all the duties, using one or more of these four objectives: (a) minimum number of crews required, (b) minimum sum of roster costs, (c) minimum of the maximum roster duration, and (d) balancing (the equity of) workload and days off. The rostering problem is presented in Section 10.5.

10.2 Vehicle-chain construction using a minimum crew-cost approach

There are two predominant characteristics in transit-operations planning: (a) different resource requirements between peak and off-peak periods, and (b) working during irregular hours. These characteristics result in split duties (shifts) with unpaid periods in-between. Often it is called swing time. The inconvenience accompanying split duties, led driver (crew) unions to negotiate for an extension of the maximum allowed driver's idle time for which the driver can still get paid. It is common, therefore, to have a constraint in a labour union agreement specifying this maximum paid idle time (swing time), to be termed $T_{\max}$.

The crew-scheduling problem from the agency's perspective is known to be the minimum crew-cost problem. With this minimum-cost orientation in mind, we can use the DF (deficit function) properties to construct vehicle chains (blocks) that take into account $T_{\max}$. In other words, to maximize idle times (swing times) that are larger than $T_{\max}$, and hence to reduce crew costs.

10.2.1 Arrival-departure joinings within hollows

The following description uses the notation and definitions associated with the DFs of Section 7.5.1 in Chapter 7. Each hollow of a DF, $d(k, t)$ at terminal k , contains the same number of departures and arrivals, except for the first and last hollow at the beginning and end of the schedule horizon. This is due to the fact that each arrival reduces $d(k, t)$ by one, and each departure increases it by one, so that the hollow starts and ends at $D(k)$.

For a given hollow, H_m^k , let I_m^k be the set of all arrival epochs t_e^i in H_m^k , and let J_m^k be the set of all departure epochs t_s^j in H_m^k . The difference in time between departure and arrival is defined as $\Delta_{ij} = t_s^j - t_e^i$ for $t_s^j > t_e^i$ in H_m^k . The joining (connection) between t_e^i and t_s^j in a vehicle block is effectively the idle time between trips; hence, Δ_{ij} may represent this idle time. We may also define local peak uv within hollow H_m^k as $d(k, t_{uv})$ between t_s^u and t_e^v , in which $e_m^k < t_s^u \leq t_{uv} \leq t_e^v < s_{m+1}^k$, where H_m^k starts and ends at e_m^k and s_{m+1}^k , respectively. Note that if the start and/or end of a local peak, uv , has more than one departure or arrival, then it suffices to refer to only one of them (as u or v). Let $d_{uv}^{k,m}$ be the number of departures in H_m^k before and including t_s^u , and $a_{uv}^{k,m}$ be the number of arrivals in H_m^k before t_e^v .

Lemma 10.1: The number of arrival-departure joinings in hollow H_m^k before t_s^u must be $d_{uv}^{k,m}$.

Proof: If some departure epochs before a local peak, uv , are left without a joining, it will be impossible to connect them with arrival epochs after t_e^v . That is, each departure epoch

before and including t_s^u must have a joining to an earlier arrival time within H_m^k . This can be seen in Figure 10.1(a).

Lemma 10.2: The number of arrival–departure joinings that can be constructed after t_e^v within H_m^k is $(a_{uv}^{k,m} - d_{uv}^{k,m})$.

Proof: Given hollow H_m^k and local peak uv , then based on *Lemma 10.1* and the characteristics of local peaks in hollows, $d_{uv}^{k,m}$, departure epochs must and can be joined to earlier arrival epochs in H_m^k ; hence, the number of arrival epochs left over without joinings is $(a_{uv}^{k,m} - d_{uv}^{k,m})$ for all local peaks. Figure 10.1(b) displays this explanation.

Lemma 10.3: The sum of all idle times within any hollow is a fixed number and independent of any procedure aimed at joining arrival and departure epochs; that is $\sum_{i,j} \Delta_{ij} = \text{constant}$.

Proof: Let H_m^k have n arrivals and n departures. We noted previously that the number of departures and arrivals are the same within each middle hollow (i.e. excluding the first and last hollows). Let two different n -joining arrangements with idle times Δ_{ij}^1 and Δ_{ij}^2 for all joinings between $i \in I_m^k$ and $j \in J_m^k$ be expressed as follows:

$$\begin{aligned} \sum_{i,j} \Delta_{ij}^1 &= \sum_{i,j} (t_{s1}^j - t_{e1}^i) = \sum_j t_{s1}^j - \sum_i t_{e1}^i; \text{ and similarly} \\ \sum_{i,j} \Delta_{ij}^2 &= \sum_j t_{s2}^j - \sum_i t_{e2}^i \end{aligned}$$

We know that the sum of all departure or arrival times in a hollow is a fixed number; hence $\sum_{i,j} \Delta_{ij}^1 = \sum_{i,j} \Delta_{ij}^2 = \text{constant}$. Figure 10.1(c) further clarifies this argument.

10.2.2 Objective function and formulation

In constructing the blocks at each DF, we wanted to maximize the number of times in which $\Delta_{ij} \geq T_{\max}$; in other words, to reduce crew cost. At the same time, however, for cases in which $\Delta_{ij} < T_{\max}$, we wanted, from a crew's fairness perspective, to attempt to have equitable paid idle times. This was for a simple reason, to eliminate a situation in which some drivers will have long and some short paid idle times. We will start with a formulation of the main objective and continue with a secondary objective.

For a given hollow H_m^k at terminal k , let x_{ij} be a 0–1 variable associated with a trip-joining between the arrival of the i -th trip to H_m^k and the departure from H_m^k of the j -th. The problem of finding the maximum number of idle times greater than or equal to $T_{\max}$ in hollow H_m^k is as follows:

Problem P4.

$$\text{Max } Z4 = \sum_{i \in I_m^k} \sum_{j \in J_m^k} x_{ij} \quad (10.1)$$

Subject to:

$$\sum_{j \in J_m^k} x_{ij} \leq 1, \quad i \in I_m^k \quad (10.2)$$

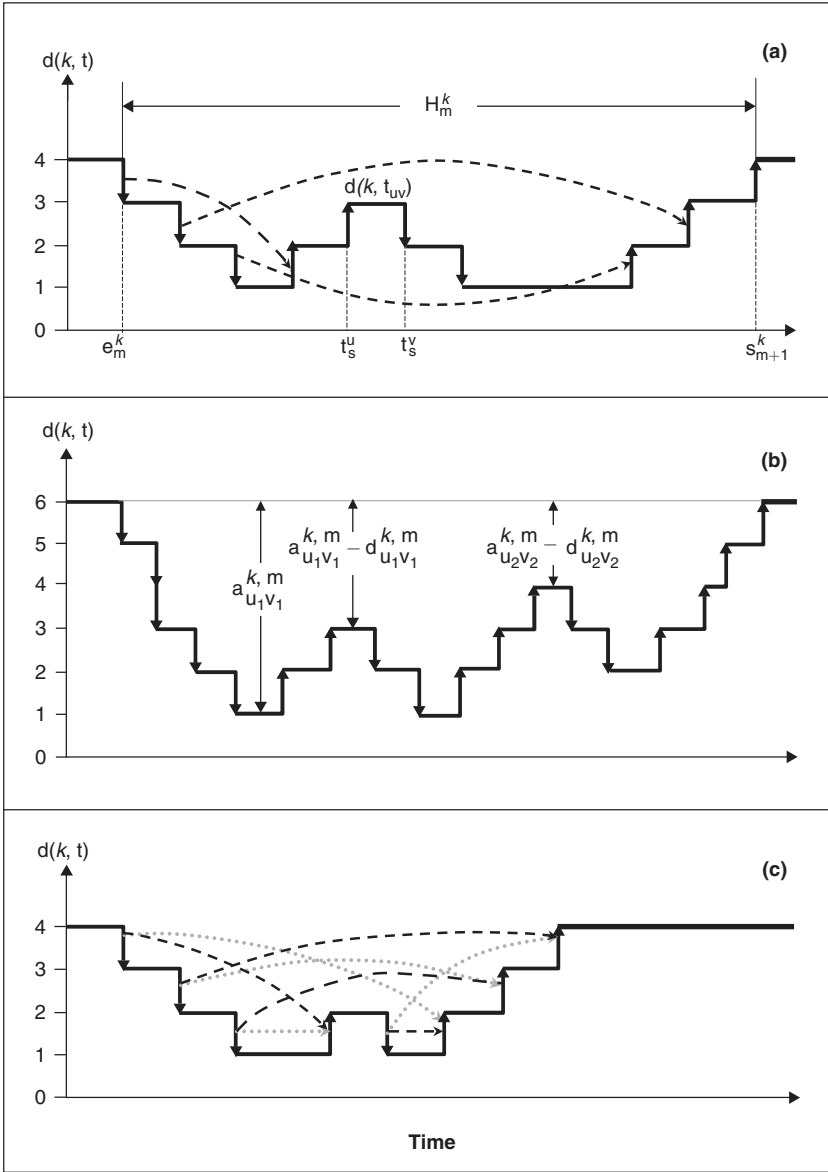


Figure 10.1 Part (a) describes examples of arrival–departure joining to support Lemma 10.1; part (b) interprets Lemma 10.2; part (c) shows two 4-joining examples of Lemma 10.3

$$\sum_{i \in I_m^k} x_{ij} \leq 1, \quad j \in J_m^k \quad (10.3)$$

$$x_{ij} = \{0, 1\}, \quad i \in I_m^k, \quad j \in J_m^k, \quad (10.4)$$

The binary decision variables are determined by:

$$x_{ij} = \begin{cases} 1, & t_s^j - t_e^i \geq T_{\max} \\ 0, & \text{otherwise} \end{cases}$$

A solution with $x_{ij} = 1$ indicates that joining trips i (arrival epoch) and j (departure epoch) results in an idle time larger than or equal to $T_{\max}$. Constraints (10.2) and (10.3) insure that each trip in H_m^k may be joined with, at most, one successor trip, and one predecessor trip, respectively.

Trips that were not joined in the solution of P4 are subject to a secondary objective: equitable paid idle times. It is shown below that joinings with this secondary objective are based on the FIFO rule. Balancing Δ_{ij} for $\Delta_{ij} < T_{\max}$ is the same as minimizing the difference between each Δ_{ij} and its average $\bar{\Delta}_{ij}$ either by absolute difference or by least-square difference. The FIFO rule used for this balancing is stated in the following theorem.

Theorem 10.1: Minimizing the least-square differences between $\bar{\Delta}_{ij}$ and each Δ_{ij} for all $i \in I_m^k$ and $j \in J_m^k$ in H_u^k is accomplished by constructing joinings using the FIFO rule.

Proof: It is sufficient to prove Theorem 10.1 on a simple, but generalized example, as illustrated in Figure 10.2, with a hollow containing two arrivals and two departures.

We may show that

$$(\Delta_1 - \bar{\Delta})^2 + (\Delta_2 - \bar{\Delta})^2 < (\Delta_3 - \bar{\Delta})^2 + (\Delta_4 - \bar{\Delta})^2 \quad (10.5)$$

where $\bar{\Delta}$ is the average arrival–departure joining length in the example. Using an algebraic expression, then (10.5) becomes

$$\Delta_1^2 + \Delta_2^2 - \Delta_3^2 - \Delta_4^2 < 2\bar{\Delta}(\Delta_1 + \Delta_2 - \Delta_3 - \Delta_4) \quad (10.6)$$

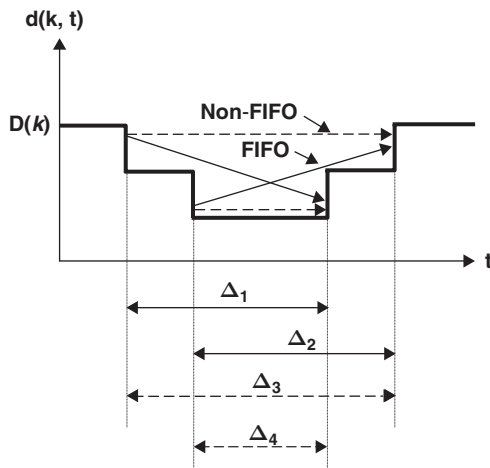


Figure 10.2 Example of a comparison of joinings based on FIFO and other rules

Lemma 10.3 states that $\Delta_1 + \Delta_2 = \Delta_3 + \Delta_4$, and hence the right-hand side of expression (10.6) is zero. From *Lemma 10.3* we can further obtain $(\Delta_1 + \Delta_2)^2 = (\Delta_3 + \Delta_4)^2$ or $\Delta_1^2 + \Delta_2^2 - \Delta_3^2 - \Delta_4^2 = 2\Delta_3\Delta_4 - 2\Delta_1\Delta_2$. The latter is inserted into (10.6) to yield

$$\Delta_3\Delta_4 < \Delta_1\Delta_2 \quad (10.7)$$

Based, again, on *Lemma 10.3*, let $\Delta_3 - \Delta_1 = \Delta_2 - \Delta_4 = B$ or $\Delta_1 = \Delta_3 - B$ and $\Delta_4 = \Delta_2 - B$; these last two equations are inserted into (10.7) to obtain $\Delta_3(\Delta_2 - B) < \Delta_2(\Delta_3 - B)$, which yields $\Delta_3 > \Delta_2$. The last result must be correct from Figure 10.2, and therefore it agrees with expression (10.5).

10.2.3 Procedure for determining maximum number of unpaid idle times

The mathematical programming formulation in Equations (10.1)–(10.4) is aimed at maximizing the number of idle times that are longer than or equal to $T_{\max}$. However, this formulation may involve a very large number of computations (NP-Complete, see Section 5.3 in Chapter 5), hence entailing the use of another (more simplified) procedure. Such a procedure is described in a flow diagram in Figure 10.3 and contains both $T_{\max}$ and FIFO rule considerations; the latter is for joining arrivals and departures with paid idle times. Let us call this procedure algorithm $T_{\max}F$.

The input for algorithm $T_{\max}F$ for each terminal k consists of two arrays, the arrival and departure arrays, and a given $T_{\max}$. This input enables constructing DF at k and obtaining $D(k)$ following the insertion of DH trips and the shifting of departure times for minimizing fleet size (see Chapters 7 and 8). Because $D(k)$ vehicles are required at k , we assume their arrivals there to be at (or before) T_1 (the start of the schedule horizon). Algorithm $T_{\max}F$ moves by steps on $d(k, t)$, in which each step refers to a change in $d(k, t)$ or the detection of a dot on $d(k, t)$; the dot means that arrival and departure epochs at k overlapped at t .

Algorithm $T_{\max}F$ continues with a check of the end of the schedule horizon and detects the nature of the change (or dot) in $d(k, t)$. For each departure epoch, Δ_{ij} is examined to determine whether it is greater than or equal to $T_{\max}$; if greater, then an unpaid joining array is added, otherwise a disjointed departure time array is added. Each arrival epoch (detected in a step move in Figure 10.3) is added as a disjointed arrival array. If a departure epoch is identified in a step move, the algorithm looks for a possible dot on $d(k, t)$, adding its arrival epoch to the list of disjointed arrival arrays. At the end of the process, the algorithm constructs joining arrays from the disjointed arrival and departure arrays, using the FIFO rule. The complete process is shown in Figure 10.3.

An example of constructing vehicle chains (blocks), including the employment of algorithm $T_{\max}F$, is shown in Figures 10.4–10.6. The example, consisting of three terminals and a 24-trip schedule, is exhibited in Figure 10.4, including DH travel time matrix, shifting tolerance, $T_{\max}$, and schedule horizon. It should be noted, though, that DH travel time between terminals b and c is considered in both directions although there is only a service route between c and b . The fleet-reduction procedure, involving the shifting of departure times and DH trip insertions, is shown in Figure 8.12 in Chapter 8; here it is applied to the example in Figure 10.5. Two DH trips and two shifts are introduced into the process to reduce $D(a)$

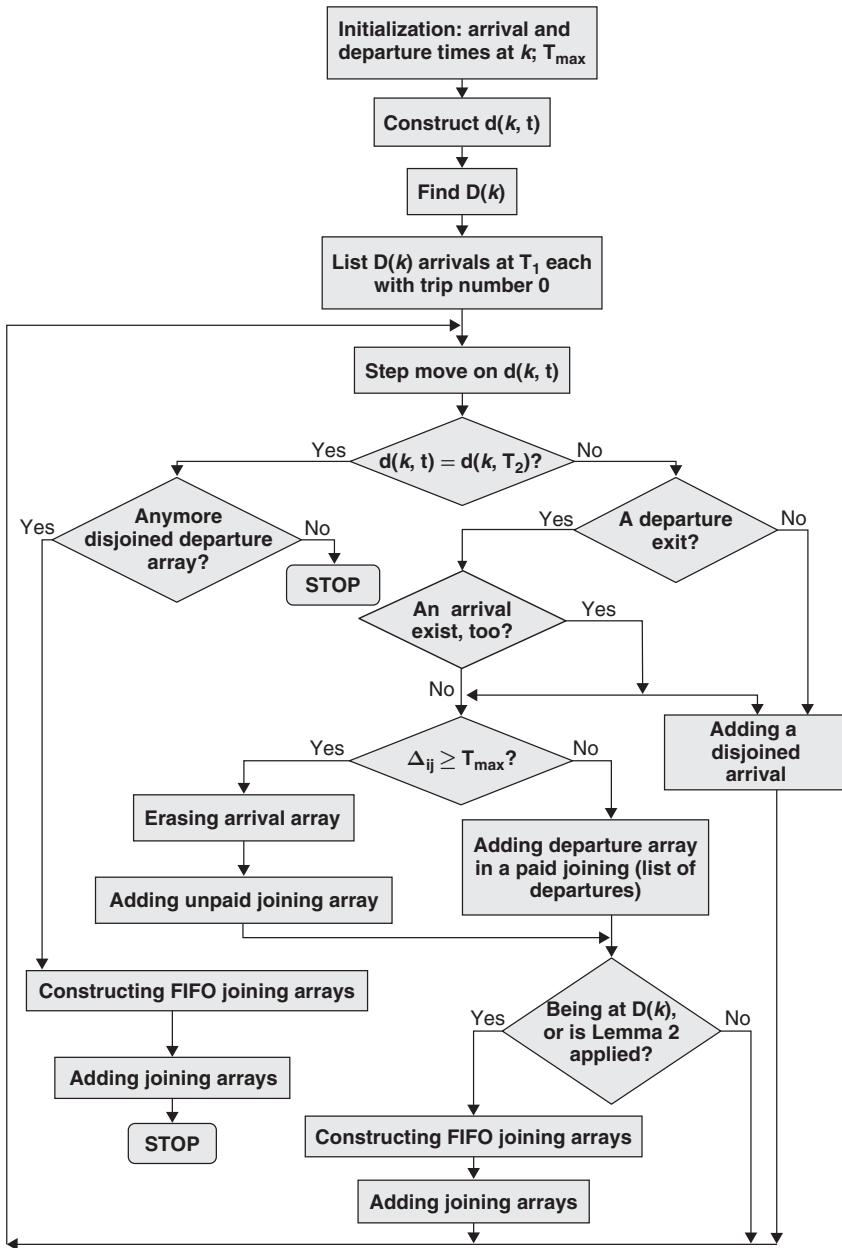


Figure 10.3 Flow diagram of algorithm T_mF

and $D(c)$ from four to three, resulting in a fleet size of 11 vehicles. The shifts are shown in Figure 10.5 by their shifting length and trip number. We can see from this figure that the only middle hollow containing more than a single departure is the second hollow of $d(b, t)$; hence, only this hollow is subject to the process of algorithm T_mF .

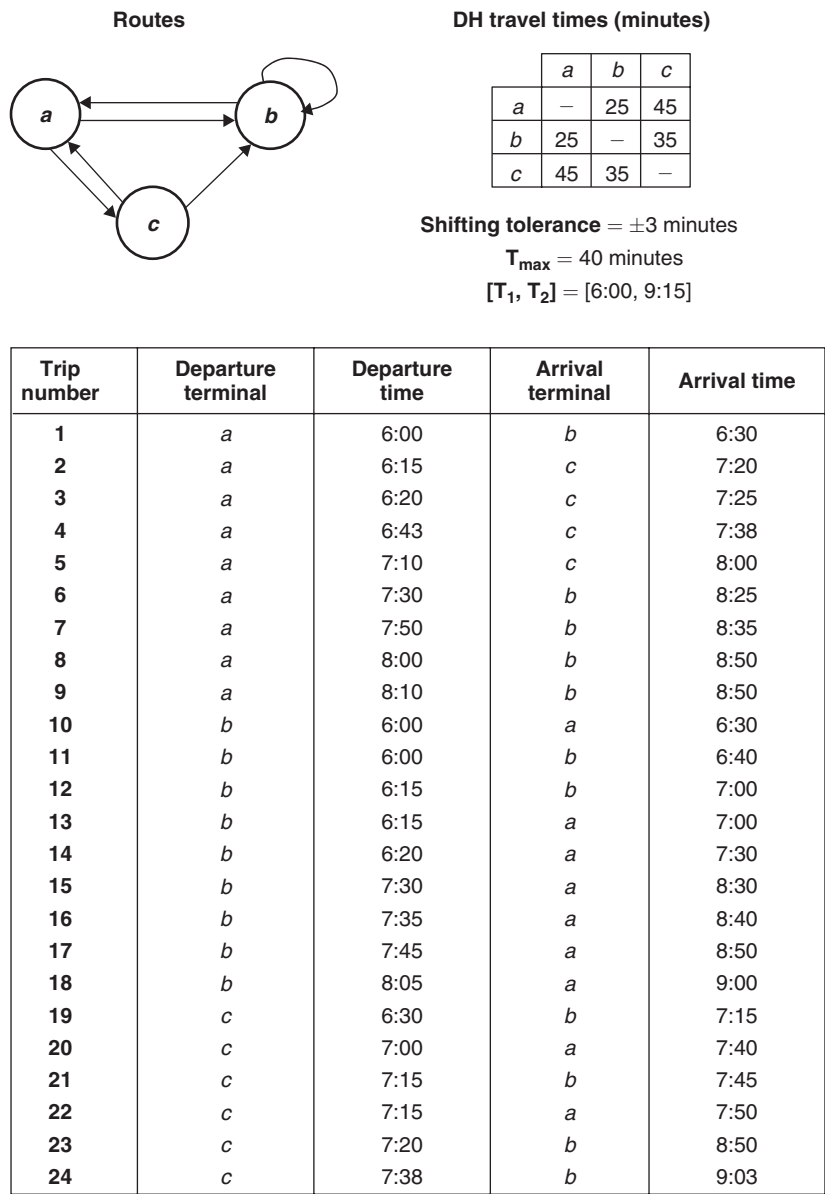


Figure 10.4 Example consisting of 24 trips and 3 terminals for constructing vehicle chains with the $T_{\max}$ constraint

Figure 10.6(a) describes the solution for algorithm T_mF in comparison with a solution based only on the FIFO rule in Figure 10.6(b). The trip numbers of the example, appearing in Figure 10.4 are added to Figure 10.6. Algorithm T_mF results in two unpaid joinings between the arrivals of Trips 11 and 12 and the departures of Trips 15 and 17, respectively. In both cases, $\Delta_{ij} > T_{\max} = 40$ minutes. The remaining joinings in Figure 10.6(a) are based on the

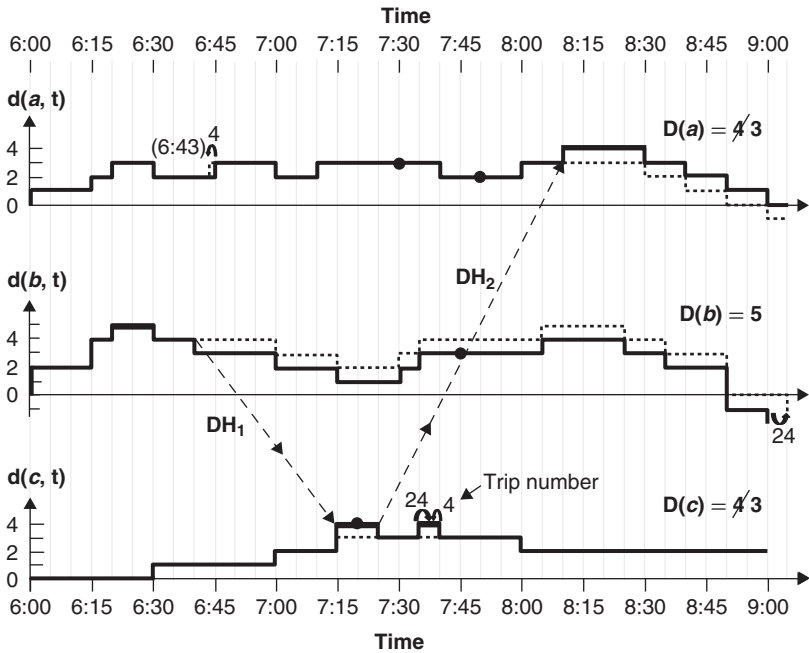


Figure 10.5 The 24-trip example (depicted by three DFs) undergoing a DH trip-insertion procedure, combined with shifting departure times.

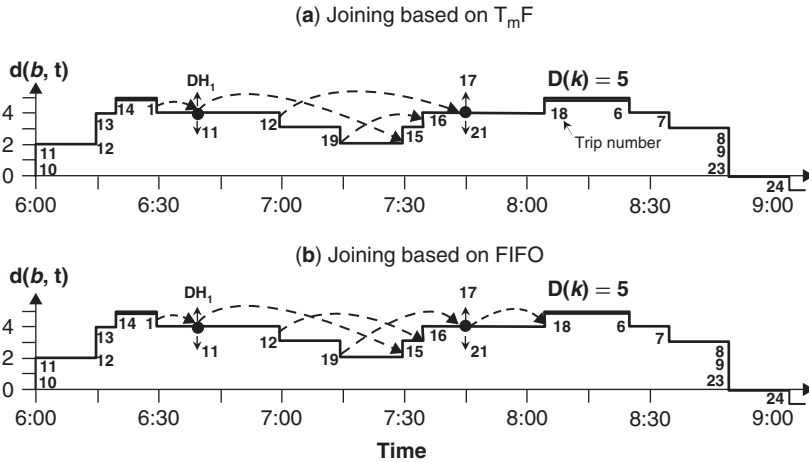


Figure 10.6 Arrival–departure joinings for constructing vehicle chains (blocks) in the middle hollow of terminal b , utilizing in (a) the T_{mF} algorithm, and in (b) the FIFO rule

FIFO rule. The use of only the FIFO rule for the entire process results in only one unpaid joining (that between Trips 11 and 15) as is shown in Figure 10.6(b).

The final phase of the arrival–departure joining process is to construct vehicle blocks. This will contain the joinings created and other FIFO-based joinings in order to make a complete set

of blocks. The 11 blocks of the 24-trip-schedule example, based on algorithm T_mF (at terminal b), are given by their numbers in the following list: [1-DH₁-22-7], [10-4-24], [11-15], [2-23], [12-17], [13-5], [3-DH₂-9], [14-6], [19-16], [20-8], [21-18]. The process based only on the FIFO rule results in the same blocks, except for the 5th and 9th blocks, which become [12-16] and [19-17], respectively.

10.3 Mathematical solutions

Crew scheduling and rostering are extensively treated mathematically in the books edited by Wren (1981), Rousseau (1985), Daduna and Wren (1988), Desrochers and Rousseau (1992), Daduna *et al.* (1995), Wilson (1999), and Voss and Daduna (2001), and in a forthcoming volume by Hickman *et al.* (2007). Specific detailed studies are reviewed below, in Section 10.6.

The basic OR formulation of the crew-scheduling problem (CSP) is a zero-one integer linear programming, called a set partitioning problem (SPP). SPP has as its objective the selecting of a minimum-cost set of feasible duties, such that each duty piece is included in exactly one of the duties. CSP is often illustrated by rows and columns, in which the rows are the duty pieces and the columns are the duties, each accompanied by a duty cost. In the latter, SPP is aimed at minimizing the cost of a set of columns, such that each row is included exactly once in one of the columns in the solution. The following is the SPP formulation.

Problem P5.

$$\text{Min } Z5 = \sum_{q \in Q} c_q x_q \quad (10.8)$$

$$\text{Subject to:} \quad \sum_{q \in Q(j)} x_q = 1, \quad j \in J \quad (10.9)$$

$$x_q = \{0,1\}, \quad q \in Q \quad (10.10)$$

where Q is the set of all feasible duties, c_q is the cost of duty $q \in Q$, and $Q(j) \in Q$ is the set of duties covering duty piece $j \in J$.

A binary zero-one variable x_q is used to indicate whether duty q is selected in the solution or not. Constraint (10.9) assures that each piece will be covered by exactly one duty. The running time of P5 belongs to the class of NP-Complete (see Section 5.3).

An easier way to solve the crew-scheduling problem is to relax constraint (10.9):

$$\sum_{q \in Q(j)} x_q \geq 1, \quad j \in J \quad (10.11)$$

Equation (10.11), together with Equations (10.8) and (10.10), represents a new problem, called a set covering problem (SCP). SCP is usually solved first (before the SPP), and the solution is changed to handle SPP by deleting overlapping trips. This experienced-based deletion process involves changes in the duties, in the SCP solution, and in a crew member who is so assigned and will make the trip as a passenger. Freling *et al.* (1999) explain that such a change affects neither the feasibility nor the cost of the duties considered.

One way to speed up the running time of the SPP of CSP is by the use of the *Lagrangian-relaxation* method. In this method, the complicated set of constraints shown by Equation (10.9) is removed and put into the objective function (10.8). Each constraint removed is weighted by a given Lagrange multiplier. Another useful and commonly used method for reducing the SPP complexity is the *column-generation* technique. The number of columns (representing duties) in the SPP is usually a very large number. The column-generation technique, instead of examining all the columns over and over again, selects subsets of columns and either checks each subset for optimality or solves several sub-problems. Some of the articles reviewed in Section 10.6 cover these methods.

A known heuristic approach to solve CSP within a framework of resolving both CSP and the vehicle-scheduling problem is discussed by Bodin *et al.* (1983) and presented by Ball *et al.* (1981, 1983). This approach is chosen to illustrate the inherent combinatorial difficulties that exist in CSP. The basic steps of the approach are as follow: (1) generate all feasible pieces of work (to be derived from the vehicle blocks); (2) establish an internal piece cost (based on piece characteristics and past experience); (3) create an acyclic network from each block, in which nodes are the relief points and arcs represent the cost of each feasible piece in the block; (4) solve the shortest-path problem in order to establish the best (minimum-cost) pieces; and (5) solve a matching problem while using a two or three legal-piece combination and, if the solution is not feasible, reiterate part of the process by updating the piece-combination cost and redo the shortest path (step 4) until the solution is feasible and satisfactory. Figure 10.7 presents the procedure schematically.

The example presented in Figure 10.4 is used for demonstrating the shortest-path and matching (SPM) approach shown in Figure 10.7. The 11-block results of the T_mF algorithm of

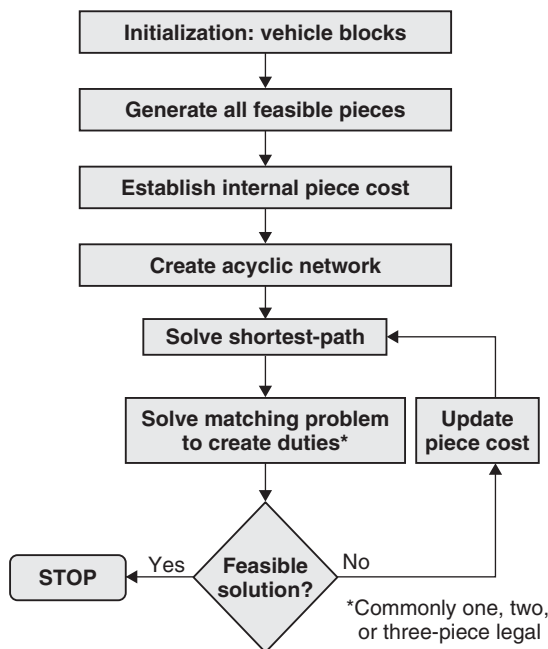


Figure 10.7 Flow diagram of the SPM heuristic approach to creating duties

this example appear in Table 10.1; they are extracted from the analysis shown in Figures 10.5 and 10.6(a). Given that all three terminals *a*, *b* and *c* are relief points, each block can be considered a driver schedule (duty) or be partitioned into alternative pieces covering all possible combinations. Figure 10.8 shows how to partition the blocks into combinations of pieces. In addition, each piece is assigned an internal piece cost, based on the piece’s characteristics (e.g. time of day, arrival and departure locations, type of vehicle required) and past experience (the cost of a similar piece in past crew schedules). Two sets of blocks have the same pieces and costs (numbered 3 and 5, and 9 and 11). It should be noted that the first set includes unpaid idle time as a result of algorithm T_mF .

Table 10.1 *Solution for vehicle blocks and routing of the example appearing in Figures 10.4, 10.5 and 10.6(a), using the algorithm T_mF*

Block number	Trips in block (see Figure 10.4)	Block routing (see Figure 10.4) (# represents DH trip)
1	1-DH ₁ -22-7	<i>a-b-#c-a-b</i>
2	10-4-24	<i>b-a-c-b</i>
3	11-25	<i>b-b-a</i>
4	2-23	<i>a-c-b</i>
5	12-17	<i>b-b-a</i>
6	13-5	<i>b-a-c</i>
7	3-DH ₂ -9	<i>a-c-#a-b</i>
8	14-6	<i>b-a-b</i>
9	19-16	<i>c-b-a</i>
10	20-8	<i>c-a-b</i>
11	21-18	<i>c-b-a</i>

The right-hand column of Figure 10.8 contains the acyclic network of each block, representing all the possible pieces and their internal costs. This acyclic network has undergone a shortest-path analysis, such as the known algorithm, by Dijkstra (1959). For the sake of clarity, the Dijkstra algorithm is described with an example in Appendix 10.A; it will also be used in Chapters 12–15. The results of the Dijkstra procedure are emphasized in Figure 10.8, along with the minimum piece cost needed to cover the whole block. These results are illustrated in Figure 10.9 for each block. Given that each piece is eligible to be covered by one driver, then Figure 10.9 shows that the 11 blocks require 17 drivers at a total cost of 82 (the cost units are meaningless for the example). However, there are more possibilities for matching some pieces if they are legal (from the labour-agreement perspective) and can reduce the total cost.

Taking into account the deadheading travel times in Figure 10.4 as a measure of moving between *a*, *b* and *c*, there are five possible matchings that need to be examined for the first piece of block 1, and one possible matching for the first piece of block 6. These possibilities are

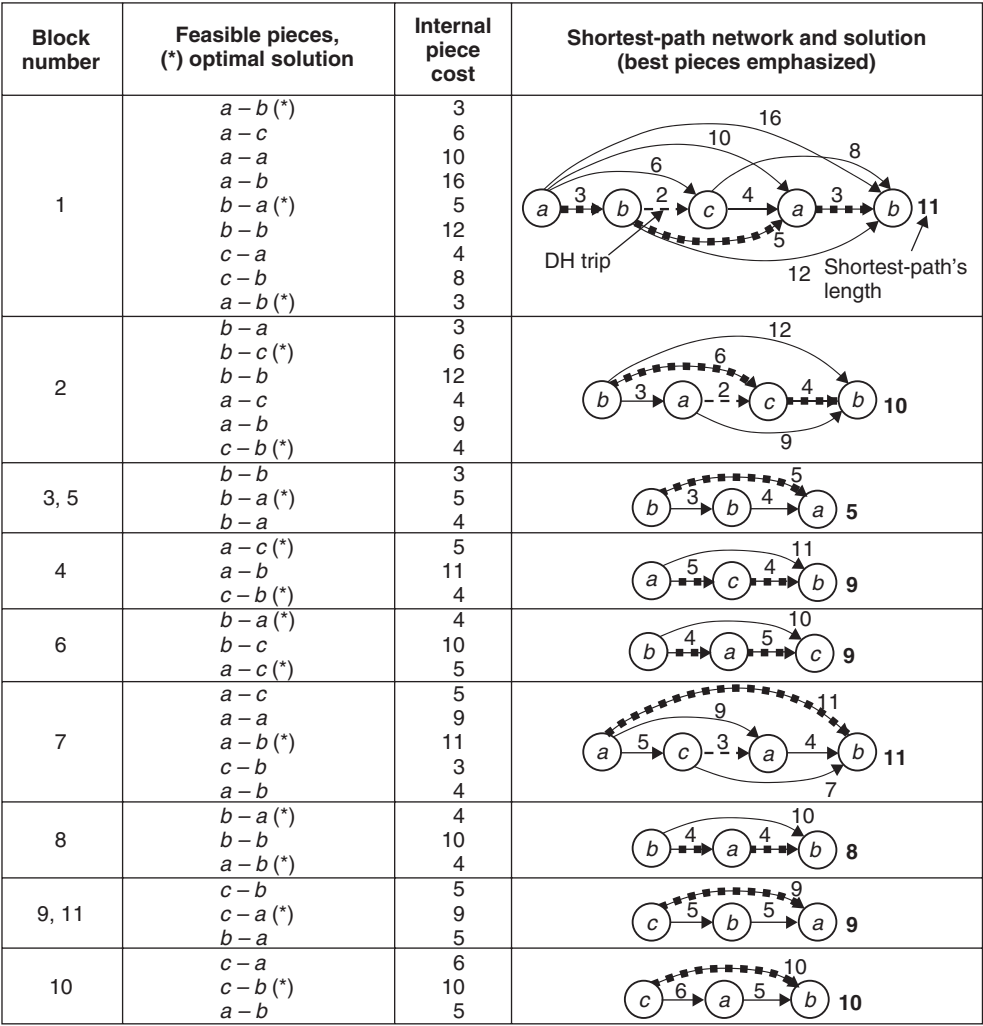


Figure 10.8 Partitioning of blocks into minimum-cost pieces

shown by the dashed lines in Figure 10.9. Usually, if the best (minimum-cost) feasible matching results in a cost reduction (i.e. less than 82), then this matching is selected and the number of drivers can be reduced by one or two (to 16 or 15). The more pieces examined for a single duty, the more running time the matching process requires (Ball *et al.*, 1983). Commonly the check is performed on two and three legal pieces. More details on possible optimal and heuristic matching procedures can be found, for example, in Ahuja *et al.* (1993).

The simplified example using an SPM approach demonstrates some of the complexities involved in handling CSP; hence, justifying the use of OR software. Moreover, in addition to the combinatorial problems inherent in the scheduling tasks, there are human dissatisfaction issues concerning the crews that deserve attention and make this undertaking even more cumbersome.

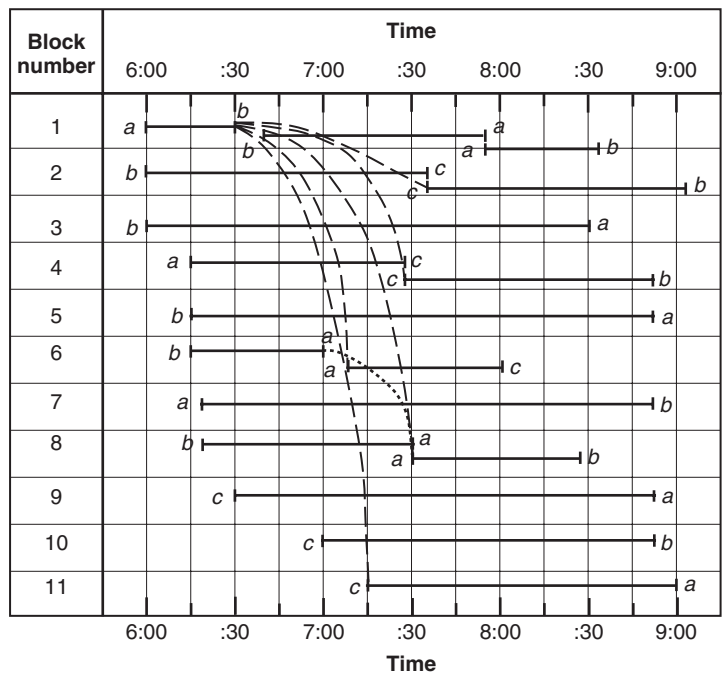


Figure 10.9 Result of partitioning the blocks into best pieces with feasible matching alternatives

10.4 A case study: NJ commuter rail

This section describes a case study that was performed for the New Jersey (NJ) Transit Corporation. In this study (Tykulsker *et al.*, 1985), we will experience that ... all computers wait at the same speed; what counts is the flexibility of the solution and the alternatives that can be produced.

10.4.1 Background

NJ Transit provides a large commuter rail operation, carrying over 70,000 passengers and running more than 400 trains daily. At the time of the study, NJ Transit had decided to take under its wing both the service and labour force then maintained by Conrail. The collective bargaining agreements then in effect between Conrail and the various bargaining units operating in the labour force would then no longer remain in effect. Thus, NJ Transit was placed in the position of negotiating new collective bargaining agreements.

With labour-crew costs representing a major component of commuter rail costs, it was critical that NJ Transit management understand the cost ramifications of different work rules. Given the variations in these rules that already existed for the different NJ commuter rail lines, and the multitude of permutations, it was important that NJ Transit have some mechanism for quickly analysing options. Thus, the objective of the project was to provide the NJ Transit management with a tool that would enable them to quickly analyse the implications of work-rule changes for labour costs. This tool would be used during the negotiation process, since it would

provide management with the ability to quantify the impacts of various proposals and counter-proposals. In addition, NJ Transit needed a tool that could be used on an ongoing basis in order to: (1) analyse the marginal cost impact of changes in train schedules; and (2) produce near optimal crew assignments (duties), given a schedule and a new set of work rules.

There were two primary bases of pay applicable to commuter rail. The first pay basis was similar to an industrial pay basis: pay is based on hours worked, with a guarantee of eight hours per day. Any time in excess of eight hours is paid at time and a half. The second pay basis was called the dual basis of pay. Employees were guaranteed pay for a 'standard day', defined as eight hours and a prescribed number of miles (100 for engineers and 150 for trainmen, in the case of NJ Transit); if either the time or the mileage limits were exceeded, additional payments were made. Therefore, it was possible to receive both overtime and overmile payments, a double penalty (for the company) for long assignments.

There were also two primary work rules, other than those incorporated in the basis of pay, that had a large impact on labour costs. The more important of the two relates to interlining restrictions. In many cases, crew members could only be assigned to a single line or group of lines within their seniority district. This limited the ability to assign crews in the most efficient manner available. The other work rules that affected labour costs involved manning requirements. For example, firemen may not be needed for the operation of trains, but may be required, in certain seniority districts, because of prior collective bargaining agreements. Also, the minimum number of trainmen (conductors, brakemen and ticket collectors) is usually specified by contract, but the maximum number is determined by a management decision relating to the need to protect revenue, based on passenger volume and trip duration.

10.4.2 Crew-assignment procedures

Overview of model

To address the issues described, a crew-assignment/work-rules model was developed. The approach consisted of three steps: (1) defining train schedules (arrival and departure times and locations, trip distances, and manpower requirements) and work rules (interlining restrictions, deadheading rules, minimum layovers, hours of service limits, and the locations of the beginning and ending of duties); (2) generating a range of alternative crew assignments; and, (3) selecting the set of 'best' assignments (duties).

Crew assignments were generated, starting with the earliest departing train, by finding the first, second, third, etc., feasible departing connections at each location. From each of these trains, additional connections were explored; thus, a tree of assignments was found for each train. The size of the tree can be limited by several user-supplied parameters. If the generation is limited, the entire procedure is repeated, starting with the latest arrivals and using feasible connections. This is to ensure that a good set of morning and evening assignments is generated. Assignments were then costed by applying the appropriate bases of pay. The costing procedure allowed for hourly or mileage-pay bases; daily, weekly, and monthly guarantees; both overtime and overmile payments; spread premiums; and split duties.

The last step was to determine the actual crew assignments. Several heuristic techniques were developed to reduce the magnitude of the computational problem. An optimization package was used separately for weekdays, Saturdays and Sundays. For each day, a relaxed (non-integer) SCP was formulated and solved, assignments in the solution were set with overlapping trains, additional assignments were made without the overlap, and this reduced-set covering

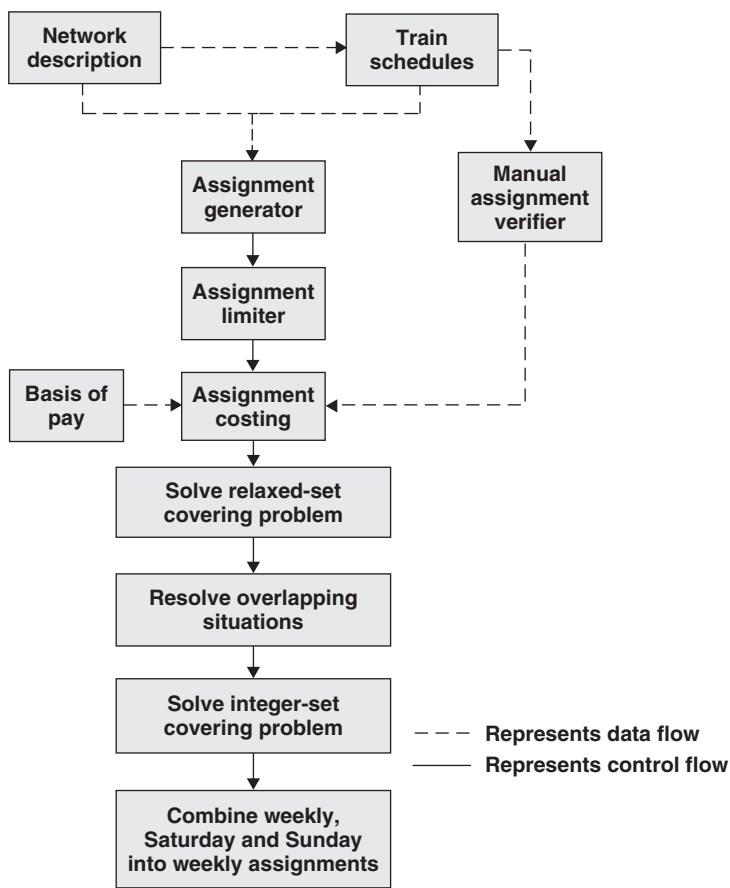


Figure 10.10 Overview of the CSP process in the NJ case study

problem was solved for an integer solution. Finally, the separate daily solutions were combined, producing assignments with the required relief days. Figure 10.10 presents an overview of the model’s components.

Model development

The first stage of the project was to identify an appropriate methodology. To accomplish the assignment-selection procedure, careful consideration was given to various mathematical programming and heuristic techniques. It is worth mentioning that in a comparison of bus and rail-crew scheduling, interlining and DH (deadheading) issues are more appropriate to bus-crew assignments. The combinatorial complexities of bus-crew scheduling, introduced by interlining and DH, may dictate sequential optimization techniques. For the case study, use of the first general technique of eliminating assignments prior to optimization was found to be sufficient.

We have already noticed that it is time consuming and expensive to solve large-size problems by exact, integer-programming procedures. Consequently, it is common either to relax the integer restriction (solving SCP) or to use heuristics. The SCP formulation was deemed more appropriate for the rail CSP of the case study, since crews can and do deadhead on trains. The SCP is formulated in P5 in Equations (10.8), (10.10) and (10.11); in the case study, the elements in P5 have the following interpretation: the set Q is all feasible duties (assignments), c_q is the crew cost of duty (assignment) $q \in Q$, $Q(j) \in Q$ is the set of duties (assignments) covering a set of trains (or trips) $j \in J$, and x_q is used for indicating whether duty q is to be worked (selected in the solution) or not.

In commuter-rail operations, crews occasionally perform DH trips while travelling on service trains. In bus operations, crews may deadhead via a route other than the service route, with or without their vehicle. Usually, bus-scheduling programs include a special DH travel-time matrix so that the DH crew/vehicle can reach the next departure point prior to the arrival of a service trip (with which they overlap in the SCP solution). Although the early availability of DH crews may create new opportunities for bus operations, this situation does not apply to commuter-rail operations; the problematic issue with the latter is the existence of DH (overlapping) loops in the solution.

A DH loop refers to a particular set of consecutive trains on a given assignment that fulfils the two following conditions: (a) the departure station of the first train on the loop coincides with the arrival station of the last train on the loop; (b) each train on the loop is covered by other assignments in the SCP solution. The loops are identified by first creating a train-coverage table, which shows how many times each train is covered. The procedure is then iterated over all the assignments in the solution set and over all trains in each assignment. When a train is covered more than once, all candidate loops are reiterated (initially for each assignment, there are no candidate loops). If the train completes a loop, the trains on the loop are eliminated and the loopless assignment is found. This train then starts a new, candidate loop. Eliminating DH loops will usually lower the value of the objective function. One should note, however, that it is possible to obtain a minimum-cost solution that contains DH loops if these assignments are already costed at the minimum daily pay (i.e. loop elimination cannot lower the cost). From an operations and safety perspective, though, all DH loops should be eliminated.

To overcome the undesirable situation of DH loops, the process chosen used three main steps, as shown in Figure 10.10: (1) solving a relaxed (linear) SCP formulation; (2) resolving DH loops by forming new, loopless assignments; and (3) solving an exact (integer) SPP formulation with the incumbent solution plus the new, loopless assignments.

Model implementation and results

The model was implemented as a set of ‘assignment filters’: each program was separate and acted to process or filter a file of assignments. This provided a structured, modular package. In addition, it was found necessary to build extensive checks for error and inconsistency into this procedure, since the optimization procedure assumes that all assignments presented to it are feasible. The existing assignments were then costed and used as ‘seeds’ for the optimization procedures. Figure 10.11 illustrates the computational process of the model in a flow diagram.

A major component of the model was the assignment generator. Assignments were generated without regard to crew cost. This is due to the fact that the model was explicitly designed to investigate alternative bases of pay. The generation was carried out in two passes, forwards

assignment. Several measures can be used: pay hours per total hour (effective hourly rate), pay dollars per train hour, pay dollars per train, total hours per train hour, and so forth. It was decided to use total hours per train hour because it was felt to be politically less sensitive to use a measure divorced from pay. The assignments were then sorted by productivity measures. The reduction was carried out by specifying a coverage limit – how many times each train must be covered before an assignment is dropped. The assignments were processed in order of productivity measure, updating a coverage count for each train on the assignment. If all the trains on an assignment are already covered to the limit, then that assignment is dropped. However, if at least one train is not at the limit, then the assignment is passed on. In this way it is possible for some trains to be covered more than the coverage parameter. Furthermore, a train could be covered less than the coverage limit if the assignment generator did not produce that many assignments for this train. In practice, a coverage limit of 20 was found to work well.

The assignment-costing procedure was designed to be as flexible as possible. It must be very simple in order to recost some or all of a set of assignments, using alternative bases of pay. More important, it must be easy to modify the procedure to incorporate completely new concepts in costing assignments. The mechanics of costing are as follow: for each assignment, select the appropriate pay basis (from the line-basis-of-pay definition), apply the appropriate rates to the miles and hours, and compute any guarantees. New pay bases are easily added by inserting the necessary code and referring to a previously unassigned parameter.

After costing, the assignments were ready to be optimized. As described previously, the process was to solve a relaxed SCP; resolve overlapping by forming new, loopless assignments; recost; and then solve an SPP using both the new assignments plus the relaxed SCP solution. The SPP was much smaller than the SCP; typically it limits the relaxed SCP to 20 assignments per train and solves the problem with an average of 4 trains per assignment. Thus, the number of assignments for the SPP could be reduced by two orders of magnitude.

Another technique used to reduce the problem dimension is to partition the weekly crew-assignment problem into three components: weekday, Saturday and Sunday. Three separate solutions were generated and solved, then combined with the three solutions to produce a full, weekly set of assignments.

The optimization steps used the software package MPSX/370 with mixed-integer programming. The number of assignments (columns) ranged between 3,379 and 7,591 covering 250 trains (rows). The CPU times were in the order of 100 seconds, during which it was found that the DH loop-elimination procedure was needed to produce good solutions: the objective-function reductions ranged from 0 to 20 per cent, but typically were around 2 to 5 per cent. The second linear optimization (for the relaxed solution plus the loopless assignments) usually terminated the integer, eliminating the need for branch and bound. For cases requiring the employment of branch and bound, the value of the objective function usually increases by less than 1 per cent.

The model generally produced very good results; typical savings for manually generated crew assignments ranged from 2 to 20 per cent. However, the DH loop-elimination procedure occasionally produced assignments with very low productivity. For example, an assignment with only two trains would uniquely cover only one train. The reason is that there are no penalties for these poor assignments: the choice among competing assignments during optimization is done solely on a cost basis, and not productivity. NJ Transit used the model extensively during the transition period and during recent labour negotiation to evaluate alternative bases of pay and work rules. The Operations Department has exhaustively reviewed the assignments produced, with positive reactions.

10.5 Crew rostering

The CSP establishes duties in an optimal manner with the aim of minimizing crew assignments for a given set of constraints. Usually transit agencies then arrange the duties in a set of patterns for a specified time horizon. Each pattern is defined as a *roster*, containing duties to be fulfilled over a certain number of consecutive days. Commonly, the pattern repeats itself in cycles (whether a week, a month, or any other period). The crew-rostering problem (CRP) is usually to find a feasible set of rosters to cover all duties, using one or more of these four objectives: (a) minimum number of crews required, (b) minimum sum of roster costs, (c) minimum of the maximum roster duration, and (d) balancing (equity of) workload and days off.

10.5.1 Literature review and problem definition

A broad literature survey on CRP by Ernst *et al.* (2004) reveals that the number of articles on rostering in public transit is relatively small. The survey refers to a total of 193 articles on staff scheduling and rostering, with a classification of applications, problems and models. Basically, the solution methods offered for CRP in this survey are similar to those of CSP. That is, maths-programming (e.g. SPP, SCP), heuristic and metaheuristic approaches. The last is based on methods drawn from artificial intelligence, neural engineering, biological evolution and more (see, for example, Lourenco *et al.*, 2001). Among the few studies on CRP, Carraresi and Gallo (1984) and Bianco *et al.* (1992) utilized network concepts for solving CRP. Catanas and Paixao (1995) used a SPP and SCP formulation combined with heuristic rules for CRP. Caprara *et al.* (1998) determined rosters complying with the minimum number of weeks in which each duty is carried out only once a day; mixed-integer programming was used as a base to develop an efficient heuristic algorithm. Sodhi and Norris (2004) divided the CRP at the London underground into two stages: the first established a roster pattern for each depot, followed by the insertion of duties into each of the patterns; the second was formulated as an assignment problem (a known problem in OR; a special case of the minimum-cost flow problem on networks) with side constraints.

The CRP involves compliance with work and legal rules, institutional requirements and individual preferences. The rules and requirements (some are safety oriented) commonly established in labour contracts necessitate constraints, as in CSP, that are difficult to treat analytically. For instance, the minimum number of hours required between two consecutive working days, exact (or between minimum and maximum) number of days off per week, minimum number of weekend days in days off per month, maximum working hours per day, equity of weekend working hours and more.

The rosters are commonly arranged in cyclic patterns; the crew can have the same cyclic pattern or rotate between patterns, depending on the arrangement between the transit agency and their employees. The upper part of Figure 10.12 illustrates seven weekly roster types covering all possible rosters complying with two consecutive weekly days off. Each duty is denoted by d^q_j , in which q is the day in the week and j is the index of a specific duty. Certainly some of the duties (among a total of 35) can be the same; e.g. if there are fewer than five duties for a given day. Moreover, more rosters of the same type can be used when more duties are involved. A single crew member may be tied to a single roster or switched between rosters every month (or every few weeks). In the latter case, the number of days off in the transition period can be fewer or greater than the required two days.

		Mon.	Tue.	Wed.	Thur.	Fri.	Sat.	Sun.
Roster type	R₁	d ₁ ^M	d ₂ ^{Tu}	d ₃ ^W	d ₄ Th	d ₅ ^F	off	off
	R₂	off	d ₆ ^{Tu}	d ₇ ^W	d ₈ Th	d ₉ ^F	d ₁₀ ^{Sa}	off
	R₃	off	off	d ₁₁ ^W	d ₁₂ Th	d ₁₃ ^F	d ₁₄ ^{Sa}	d ₁₅ ^{Su}
	R₄	d ₁₆ ^M	off	off	d ₁₇ Th	d ₁₈ ^F	d ₁₉ ^{Sa}	d ₂₀ ^{Su}
	R₅	d ₂₁ ^M	d ₂₂ ^{Tu}	off	off	d ₂₃ ^F	d ₂₄ ^{Sa}	d ₂₅ ^{Su}
	R₆	d ₂₆ ^M	d ₂₇ ^{Tu}	d ₂₈ ^W	off	off	d ₂₉ ^{Sa}	d ₃₀ ^{Su}
	R₇	d ₃₁ ^M	d ₃₂ ^{Tu}	d ₃₃ ^W	d ₃₄ Th	off	off	d ₃₅ ^{Su}
Complete cycle of duties assigned to a single crew member, by week	1	d ₁ ^M	d ₂ ^{Tu}	d ₃ ^W	d ₄ Th	d ₅ ^F	off	d ₆ ^{Su}
	2	d ₇ ^M	d ₈ ^{Tu}	d ₉ ^W	d ₁₀ Th	off	d ₁₁ ^{Sa}	d ₁₂ ^{Su}
	3	d ₁₃ ^M	d ₁₄ ^{Tu}	d ₁₅ ^W	off	d ₁₆ ^F	d ₁₇ ^{Sa}	d ₁₈ ^{Su}
	4	d ₁₉ ^M	d ₂₀ ^{Tu}	off	d ₂₁ Th	d ₂₂ ^F	d ₂₃ ^{Sa}	d ₂₄ ^{Su}
	5	d ₂₅ ^M	off	d ₂₆ ^W	d ₂₇ Th	d ₂₈ ^F	d ₂₉ ^{Sa}	d ₃₀ ^{Su}
	6	off	d ₃₁ ^{Tu}	d ₃₂ ^W	d ₃₃ Th	d ₃₄ ^F	d ₃₅ ^{Sa}	d ₃₆ ^{Su}

Figure 10.12 Roster types with two weekly, consecutive days off and possible duties assigned to a single crew member with one weekly day off for a sequence of quintuplets (five consecutive working days).

The lower part of Figure 10.12 presents a sequence of quintuplets (five consecutive working days) for a single crew member, which certainly results in one weekly day off. The complete cycle of these quintuplets covers 36 duties in six weeks, some of which can be the same. Observation of the upper part of Figure 10.12 will show quintuplets if the same crew member repeats the same roster, but with two weekly days off and not one. Rosters generally established in transit agencies may contain a combination of quintuplets, quadruplets, and triplets of working days.

10.5.2 A heuristic approach

This section provides a heuristic approach to a simplified CRP; this is done for the sake of illustrating the different planning elements involved and for illustrating the complexity of the analysis. The heuristic method is presented by an example in which the rosters are built on the basis of a four-day week rather than the usual week of seven days. That is, the planning horizon considered for the example is four days.

Table 10.2 Input data for the example problem.

		Day 1	Day 2	Day 3	Day 4	
Duties distributed		d_1^1	d_1^2	d_1^3	–	
		d_2^1	–	d_2^3	d_2^4	
		–	d_3^2	d_3^3	d_3^4	
		–	–	d_4^3	–	Maximum roster hours
Type of roster	R_1	V	V	V	–	25
	R_2	–	V	V	–	18
	R_3	–	–	V	V	18

Duty for day q, q = 1, 2, 3, 4	Start and end times	Duty length (hours), L_j
d_1^q	6:00–16:00	10
d_2^q	12:00–20:00	8
d_3^q	16:00–23:00	7
d_4^q	18:00–midnight	6

Table 10.2 provides the input data of the example. The objective is to find which rosters will determine the minimum number of crew members (drivers) required to cover all the duties, subject to given constraints. All together there are four different duties, d_j^q , $j = 1, 2, 3, 4$, in four days, $q = 1, 2, 3, 4$, to be assigned to three possible types of rosters. The lower part of Table 10.2 lists the start and end times, as well as the length of each duty. Two constraints exist: one on the minimum number of (rest) hours required between two consecutive working days, which is 10 hours; the second, on the maximum allowed roster hours for each roster type (see Table 10.2; each roster in this table contains one duty per day for days marked with a ‘V’).

The following heuristic procedure intends to find rosters with the maximum workload possible, thus fulfilling the objective of arranging the minimum number of crew members.

Procedure roster

Step 0: Initialization; determine index j of R_j by arranging all roster types in a list in decreasing order of their maximum possible lengths. For cases with the same length, assign a higher ranking in the list to roster types that contain more consecutive days (duties).

- Step 1:* Select the first non-treated roster type from the list in *Step 0*; construct a feasible network $G = \{N, A\}$ consisting of a set of nodes N and a set of arcs A per roster type, using a source node 's' and a target node 't'; let the arc's value be $v(i, j) = \begin{cases} L_j, & j \in N, j \neq t \\ 0, & \text{otherwise} \end{cases}$, where L_j is the duty length. Label the (unlabelled) nodes (say, by 'V') contained in List-1 and List-2 paths; if no more rosters – go to *Step 4*.
- Step 2:* Solve a longest-path problem by using algorithm RNN in Appendix 10.A (by multiplying all $v(i, j)$, $(i, j) \in A$ by (-1) and applying the shortest-path procedure); if no s-t path exists, go to *Step 1*.
- Step 3:* Delete and store in List-1 paths (arcs and nodes) belonging to the solution attained, except nodes s and t; if the current solution contains labelled node(s), store their path and the sum W in List-2, where $W = \sum_{j \in p^*, j \neq s, t} L_j$ and p^* is a longest-path determined, and delete only the incoming arcs to the labelled nodes (in p^*); go to *Step 2*.
- Step 4:* Determine the best solution using the paths (rosters) with the highest W values from List-1 and List-2, such that each duty is covered only once. List all nodes (duties) that remain outside any path solution for a manual rostering-assignment process.

The example problem in Table 10.2 undergoes Roster procedure. The index j assigned to the three types in Table 10.2 conforms to *Step 0* of the procedure. A preliminary stage to *Step 1* in Roster is a feasibility check. Table 10.3 lists by roster type all infeasible duty connections; these infeasibilities affect *Step 1*. The three feasible networks constructed for R_1 , R_2 , and R_3 appear in Figure 10.13. The longest-path solution is attained in *Step 2* by algorithm RNN (rearranging node number), presented in Appendix 10.A in Section 10.A2. The RNN procedure is designed for finding the shortest path in networks with any arc value (it can also be

Table 10.3 *Infeasible consecutive duty connections*

Roster type	Infeasible connection (q = 1, 2, 3, 4)	Dominant constraint
R₁	$d_1^q - d_1^q - d_1^q$	Maximum roster hours
	$d_1^q - d_1^q - d_2^q$	
	$d_1^q - d_1^q - d_3^q$	
	$d_1^q - d_1^q - d_4^q$	
	$d_2^q - d_1^q - d_1^q$	
R₂	$d_1^q - d_1^q - d_2^q$	Maximum roster hours
	$d_1^q - d_1^q$	
	$d_3^q - d_1^q$	Minimum hours between consecutive working days
R₃	none	–

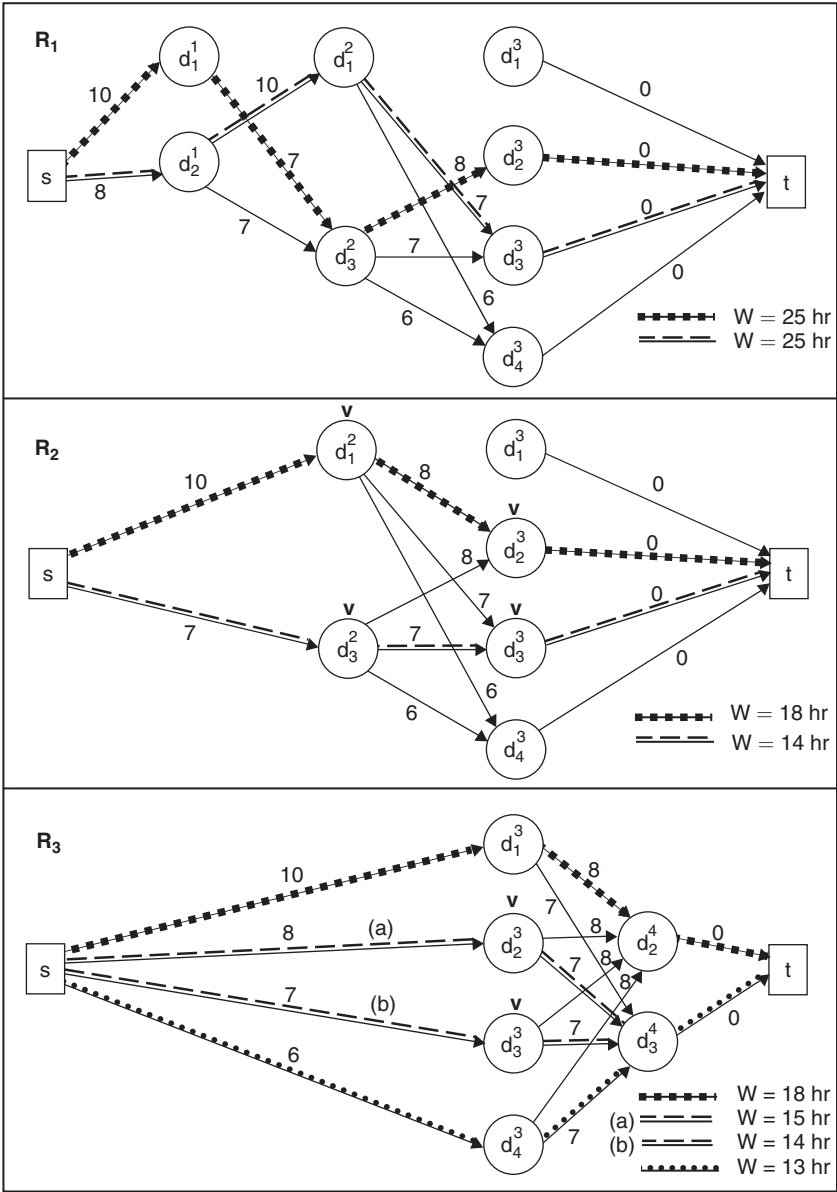


Figure 10.13 Heuristic rostering process for the example problem, using a longest-path approach for each roster type (R_1, R_2, R_3)

negative) and especially without cycles (in order to avoid the possibility of an infeasible solution derived from negative cycles). The latter characteristic is compatible with the networks constructed in *Step 1*. The longest path is determined by the shortest path using RNN, in which $v(i,j)$ is replaced by $-v(i,j)$.

The first feasible network for R_1 in Figure 10.13 results, with the use of Roster, in a two-path solution, both with $W = 25$ hours; these paths are inserted into List-1. The second network for R_2 also results in two paths of pairs of duties (second-type rosters), but all four nodes – besides s and t – in the solution are labelled; that is, they were included in a previous solution. An analysis of R_2 provides the longest path, $s-d_1^2-d_2^3-t$ with $W = 18$ hours. Then according to *Step 3* of Roster, arc $s-d_1^2$ is deleted and the second longest-path is found. Both paths of R_2 are inserted into List-2. The third and last roster-type network, R_3 , begins *Step 2* with a non-labelled path, $s-d_1^3-d_2^4-t$ with $W = 18$ hours. The second solution, marked (a) in Figure 10.13, contains a labelled node; thus only arc $s-d_2^3$ is deleted before continuing with the solution marked (b). In the latter case, arc $s-d_3^3$ is deleted, thereby allowing a fourth longest path with $W = 13$ hours.

The first three steps of Roster conclude with List-1 containing the solutions of R_1 and first and last solutions of R_3 , and List-2 containing the solutions of R_2 and the second and third solutions of R_3 . If we continue now to *Step 4* of Roster, List-1 (usually having paths with the highest W values) covers all ten duties and, therefore, presents the ultimate solution. Explicitly, four (as a minimum) crew members are required, two of roster type R_1 and two of R_3 . The rosters are as follow: $[d_1^1-d_2^2-d_3^3]$, $[d_1^1-d_2^1-d_3^3]$, $[d_1^3-d_2^4]$ and $[d_3^4-d_4^4]$ with $W = 25, 25, 18$ and 13 , respectively. The total hours of all duties are a given fixed number (81 hours in the example). It should be noted that: (1) if d_4^3 , for instance, is removed from the input data, then d_3^4 will be left for manual assignment, and (2) List-2, although it may seem unnecessary, could open up opportunities for practical solutions utilizing a partial manual-assignment process.

10.6 Literature review and further reading

The crew-scheduling problem (CSP) has been discussed abundantly both in and out of the transportation literature; relevant papers are found in journals related to mathematics, computing, operations research and specialized scheduling resources. We will focus here mainly on the latest developments in this field. The literature for the crew-rostering problem (CRP) is sparse (see previous section).

CSP is often formulated as a set covering problem (SCP), as is explicitly shown in Equations (10.8), (10.10) and (10.11). For this purpose, a large set of driver-workdays is defined, and a subset is then chosen that attempts to minimize costs, subject to constraints that make sure that all the necessary driving duties are performed. Most CSP formulations also verify that the labour-agreement rights of all drivers are maintained. A full, classic CSP normally includes the generation of the feasible workdays as a first, separate stage, and then the choice of a subset that best satisfies all needs as a second stage. In most practical transit networks, there can be thousands of driving duties to be carried out, and hence the need for millions of possible workdays. It is, therefore, common to construct in advance only a limited number of feasible workdays.

Most traditional CSP formulations concentrate on the work of bus drivers. Some of these formulations are transferable to rail use. This fact is emphasized throughout the review of papers that refer specifically to rail crews. Mitra and Darby-Dowman (1985) proposed a set covering solution for CSP, using integer linear programming (ILP). ILP was introduced to CSP-solution practice mainly during the 1980s as a result of improved computer power. A special

feature of the formulation presented is its willingness to allow some uncovered duties for cases in which an optimal covering solution is not found.

Smith and Wren (1988) introduce another CSP formulation based on SCP, solved by using ILP. The elements covered are duty pieces; the covering process uses slack and surplus variables. Using pieces rather than entire duties helps to reduce the computational effort involved. The approach presented by Desrochers and Soumis (1989) also uses SCP, but with a column-generation-solution procedure, rather than ILP. The first stage of the column-generation process seeks a subset that best covers all duties within the range of the feasible workdays, which are already known. A second stage follows, in which new feasible workdays that improve the existing solution are generated iteratively.

Paixao (1990) formulates another SCP, using dynamic programming. The search for solutions employs a state-space relaxation method. This approach is extended in Paias and Paixao (1993), who show that this method finds the lower-bound solution. Carraresi *et al.* (1995) propose another column-generation approach, one that starts with a feasible set of workdays and iteratively replaces some workdays to obtain a better solution. The pre-constructed workdays are built of duty pieces, and the solution uses a Lagrangean-relaxation method. Another, somewhat similar column-generation method is proposed by Fores *et al.* (1999).

Clement and Wren (1995) introduce a solution for the CSP using a genetic algorithm: a group of chromosomes, each of which represents a feasible crew schedule, is subject to repeated mutations, crossovers and other actions, based on the idea that the search for an optimal solution can follow rules similar to a genetic survival mechanism. Several 'greedy' algorithms are used for assigning duties to pieces of work. Another CSP solution procedure based on a genetic algorithm is presented by Kwan *et al.* (1999).

The approach demonstrated by Beasley and Cao (1996) does not follow an SCP concept; only a single type of workday is considered rather than attempting to search a broader range. A Lagrangean-relaxation method provides a lower-bound solution, which is later improved by using sub-gradient optimization. Next, a tree-search algorithm is used to obtain the final optimum. Beasley and Cao (1998) again use a similar approach but, instead of the Lagrangean-relaxation tool, seek the optimal lower bound by using a dynamic programming algorithm.

Another method that does not rely on SCP is suggested by Mingozzi *et al.* (1999). The authors describe two different duty-based heuristic solution procedures in which relaxed problems are formulated; their solutions also solve the original CSP. A third proposed solution procedure is based on SPP. The dual concept of linear relaxation programming is used to obtain a lower-bound solution. The number of variables in this problem is then reduced by using this lower bound; finally, the reduced-size problem is solved through a branch-and-bound technique. Haase *et al.* (2001) present an approach for the simultaneous solution of vehicle-scheduling and crew-scheduling problems in a single depot with a homogeneous vehicle fleet. The CSP is based on an SPP formulation that incorporates side constraints for the bus itineraries. Their proposed approach consists of a column-generation process for the crew schedules integrated into a branch-and-bound scheme.

Lourenco *et al.* (2001) bring a multi-objective CSP, led by the concept that in practice there is need to consider several conflicting objectives when determining the crew schedule. The multi-objective problem is tackled using metaheuristics, a Tabu-search technique, and genetic algorithms. Banihashemi and Haghani (2001) formulate a CSP as a duty-based

network-flow problem. First, minimum cost is sought in a binary programming problem that is a relaxed version of the original, omitting labour-rights constraints. Then, an iterative column-generation procedure using two sets of constraints is performed: ‘hard’ constraints restricting the building of specific workdays (e.g. too long), and ‘soft’ constraints penalizing non-efficient workdays, but not strictly forbidding their inclusion in the solution.

Freling *et al.* (2001a) and Huisman *et al.* (2005) present an integrated approach to solve a vehicle-scheduling problem (VSP) and a CSP on a single bus route (i.e. they assume that different routes do not share the same drivers or vehicles). This combined methodology enables an examination of such scenarios as when drivers are allowed to change vehicles between any two runs. First, CSP and VSP are defined separately. VSP is described as a network-flow problem, in which each path represents a single feasible vehicle schedule, and each trip a node. CSP is an SCP that is solved after the VSP to form a sequential solution. An integrated approach is then presented, in which the network formulation of the VSP is combined with an SPP formulation for the CSP. The integrated model is one programming problem instead of two, but it cannot use pre-generated workdays, since there are too many options; therefore, duty pieces are used.

Shen and Kwan (2001) introduce a process that involves partitioning a predetermined vehicle schedule into a set of driver duties. The focus is on refining an existing small set of workdays; hence, the methodology does not include the common stage of generating all feasible solutions. A Tabu search is used to improve the given crew schedule. Tabu search is a class of metaheuristic that tries to avoid being trapped in a local optimum solution by basing the solution choice in each iteration on a few-iterations-back analysis; sometimes, this means that a solution is chosen even if it leads to a poorer performance than the previous iteration. Fores *et al.* (2001) describe a traditional ILP formulation of the CSP, with some added flexibility. The formulation accepts different objective functions (minimize the number of duties, minimize costs, or a combination), different optimization techniques (primal column-generation or dual-steepest edge techniques), and different criteria for reducing the number of feasible workdays. The optimization technique chosen is used to solve a relaxed non-integer problem; a branch-and-bound process then finds an integer solution.

Freling *et al.* (2001b) discuss differences between bus and train CSPs and propose a methodology for the scheduling of train crews. The problem, formulated as an SCP with additional constraints, is solved using a heuristic branch-and-price algorithm. Branch-and-price is a special application of branch-and-bound, in which a column-generation technique is used to solve linear programming relaxations with a huge number of variables. Feasible workdays are generated in a network, in which each node corresponds to a duty piece, and each path through the network to a feasible duty; the optimal solution is sought through dynamic programming, that of a resource-constraint, shortest-path algorithm. Finally, Kroon and Fischetti (2001) present an SCP for railway crews that allows some flexibility in specifying penalties for undesirable types of workdays. A dynamic column-generation procedure is used; hence, duties are not generated *a priori* but in the course of the solution process. Re-generation and re-selection of workdays are carried out in each iteration. Generation is reformed in a network in which trips are represented by arcs. To solve the SCP, a Lagrangean-relaxation method and sub-gradient optimization are used instead of the common linear programming.

The main features of the methodologies reviewed are summarized in Table 10.4.

Table 10.4 Summary of the characteristics of the methodologies review

Source	Approach	Solution method	Comments
Mitra and Darby-Dowman (1985)	Set covering	ILP	Allows leaving some tasks uncovered
Smith and Wren (1988)	Set covering	ILP	
Desrochers and Soumis (1989)	Set covering	Column generation	
Paixao (1990); Paias and Paixao (1993)	Set covering (dynamic)	ILP; state-space relaxation	
Carraresi <i>et al.</i> (1995)	Set covering	Column generation; Lagrangean relaxation	
Clement and Wren (1995)	Genetic algorithm	Genetic algorithm	
Beasley and Cao (1996)	Single-type workdays	Lagrangean relaxation; sub-gradient optimization; tree search	
Beasley and Cao (1998)	Single-type workdays	Dynamic program	
Kwan <i>et al.</i> (1999)	Genetic algorithm	Genetic algorithm	
Mingozzi <i>et al.</i> (1999)	Set partitioning	Linear relaxation; branch-and-bound	
Hasse <i>et al.</i> (2001)	Set partitioning	Column generation; branch-and-bound	Simultaneous crew and vehicle scheduling
Lourenco <i>et al.</i> (2001)	Multi-objective program	Tabu search; genetic algorithm	
Banihashemi and Haghani (2001)	Network-flow problem	Column generation	
Freling <i>et al.</i> (2001a); Huisman <i>et al.</i> (2005)	Set covering; set partitioning	Column generation	Simultaneous crew and vehicle scheduling
Shen and Kwan (2001)	Set partitioning	Tabu search	Workday generation; method not included

Table 10.4 Summary of the characteristics of the methodologies review (Continued)

Source	Approach	Solution method	Comments
Fores <i>et al.</i> (2001)	Set covering	Primal column generation; dual steepest edge; branch-and-bound	Varying objective functions accepted
Freling <i>et al.</i> (2001b)	Set covering; network-flow problem	Branch-and-price	Rail crews
Kroon and Fischetti (2001)	Set covering (dynamic); network-flow problem	Dynamic column generation; Lagrangean relaxation and sub-gradient optimization	Rail crews

Exercises

- 10.1 Given 29 arrival epochs and 29 departure epochs (a total of 58 trips) at terminal k between 6:00 and 8:10 a.m. (see the following table), and a maximum paid idle time of $T_{\max} = 35$ minutes:
- (i) Apply algorithm T_mF and find $D(k)$ and the maximum number of joinings at k in which $\Delta_{ij} \geq T_{\max}$; list all joinings found.
 - (ii) Use the FIFO rule for constructing the joinings at k ; list all joinings and compare the results with those found in (i).

Departures		Arrivals	
Trip number	Departure time	Trip number	Arrival time
1	6:00	12	6:25
2	6:00	13	6:30
3	6:00	14	6:30
4	6:05	15	6:30
5	6:10	16	6:35
6	6:10	17	6:35
7	6:15	21	6:50
8	6:15	22	6:55
9	6:20	23	6:55

(Continued)

Departures		Arrivals	
Trip number	Departure time	Trip number	Arrival time
10	6:20	24	6:55
11	6:20	25	7:00
18	6:40	29	7:15
19	6:45	30	7:15
20	6:45	34	7:30
26	7:05	35	7:35
27	7:05	36	7:35
28	7:10	37	7:35
31	7:20	38	7:35
32	7:25	48	8:00
33	7:25	49	8:00
39	7:40	50	8:05
40	7:45	51	8:05
41	7:45	52	8:05
42	7:50	53	8:05
43	7:50	54	8:05
44	7:50	55	8:10
45	7:55	56	8:10
46	7:55	57	8:10
47	7:55	58	8:10

10.2 Given two vehicle blocks with their relief points. Block 1 starts at relief point *a* and continues through points *b*, *c*, *d*, *c* and *b*, in which the piece from *b* to *c* is a deadheading trip. Block 2 starts at relief point *a* and continues through points *b*, *c*, *a* and *d*. The following table lists the segments of consecutive blocks with their associated internal piece cost (ignore cost unit). Find and list the best sets of duty pieces (minimum cost) for each block.

Block 1		Block 2	
Feasible pieces	Internal piece cost	Feasible pieces	Internal piece cost
$a - b$	3	$a - b$	3
$a - c$	5	$a - c$	7
$a - d$	9	$a - a$	12
$a - c$	14	$a - d$	24
$a - b$	20	$b - c$	3
$b - d$	7	$b - a$	8
$b - c$	8	$b - d$	16
$b - b$	12	$c - a$	7
$c - d$	5	$c - d$	12
$c - c$	5	$a - d$	8
$c - b$	9		
$d - c$	2		
$d - b$	4		
$c - b$	4		

10.3 Use Procedure Roster to determine the minimum number of drivers required to operate the following given scheduling data: Two constraints are imposed – one is 10 hours as the minimum rest period required between two consecutive working days, and the second concerns the maximum allowed roster hours for each roster type (see Table below). Each roster can contain one duty per day for days marked ‘V’.

		Mon.	Tue.	Wed.	Thur.	Fri.	Sat.	Sun.	
Duties distributed		d_1^1	d_1^2	d_1^3	–	d_1^5	d_1^6	–	
		d_2^1	d_2^2	d_2^3	d_2^4	d_2^5	–	–	Maximum roster hours
		–	d_3^2	d_3^3	d_3^4	d_3^5	d_3^6	–	
		d_4^1	–	d_4^3	d_4^4	–	–	d_4^7	
Type of roster	R_1	V	V	V	V	V	–	–	43
	R_2	V	V	–	V	V	–	–	36
	R_3	–	–	–	–	–	V	V	16

Duty for day q , $q = 1, 2, 3, 4$	Start and end times	Duty length (hours), L_j
d_1^q	6:00–16:00	10
d_2^q	12:00–20:00	8
d_3^q	16:00–23:00	7
d_4^q	18:00–midnight	6

References

Ahuja, R. K., Magnanti, T. L. and Orlin, J. B. (1993). *Network Flows*. Prentice Hall.

Ball, M., Bodin, R. and Dial, R. (1981). Experimentation with a computerized system for scheduling mass transit vehicles and crews. In *Computer Scheduling of Public Transport: Urban Passenger Vehicle and Crew Scheduling*. (A. Wren, ed.), pp. 313–334, North-Holland Publishing Co.

Ball, M., Bodin, R. and Dial, R. (1983). A matching based heuristic for scheduling mass transit crew and vehicles. *Transportation Science*, **17**, 4–31.

Banihashemi, M. and Haghani, A. (2001). A new model for the mass transit crew scheduling problem. In *Computer-Aided Scheduling of Public Transport*. Lecture Notes in Economics and Mathematical Systems, **505** (S. Voss and J. R. Daduna, eds), pp. 1–15, Springer-Verlag.

Beasley, J. E. and Cao, E. B. (1996). A tree search algorithm for the crew scheduling problem. *European Journal of Operational Research*, **94**, 517–526.

Beasley, J. E. and Cao, E. B. (1998). A dynamic programming based algorithm for the crew scheduling problem. *Computers & Operations Research*, **25**, 567–582.

Bianco, L., Bielli, M., Mingozzi, A. Ricciardelli, S. and Spadoni, M. (1992). A heuristic procedure for the crew rostering problem. *European Journal of Operational Research*, **58**, 272–283.

Bodin, L., Golden, B., Assad, A. and Ball, M. (1983). Routing and scheduling of vehicles and crews: The state of the art. *Computers and Operation Research*, **10**, 63–211.

Caprara, A., Toth, P., Fischetti, M. and Vigo, D. (1998) Modeling and solving the crew rostering problem, *Operations Research*, **46**, 820–830.

Carraraesi, P. and Gallo, G. (1984). A multilevel bottleneck assignment approach to the bus driver’s rostering problem. *European Journal of Operational Research*, **16**, 163–173.

Carraraesi, P., Nonato, M. and Girard, L. (1995). Network models, Lagrangean relaxation and subgradients bundle approach in crew scheduling problems. In *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **430** (J. R. Daduna, I. Branco and J. M. P. Paixao, eds), pp. 188–212, Springer-Verlag.

Catanas, F. and Paixao, J. M. P. (1995). A new approach for the crew rostering problem. In *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **430** (J. R. Daduna, I. Branco and J. M. P. Paixao, eds), pp. 267–277, Springer-Verlag.

- Ceder, A. (1978). *Network Theory and Selected Topics in Dynamic Programming*. Dekel Academic Press (in Hebrew).
- Clement, R. and Wren, A. (1995). Greedy genetic algorithms, optimizing mutations and bus driver scheduling. In *Computer-Aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **430** (J. R. Daduna, I. Branco and J. M. P. Paixao, eds), pp. 213–235, Springer-Verlag.
- Daduna, J. R. and Wren, A. (eds) (1988). *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **308**, Springer-Verlag.
- Daduna, J. R., Branco I. and Paixao, J.M.P. (eds) (1995). *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems **430**, Springer-Verlag.
- Desrochers, M. and Rousseau, J. M. (eds) (1992). *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **386**, Springer-Verlag.
- Desrochers, M. and Soumis, F. (1989). A column generation approach to the urban transit crew scheduling problem. *Transportation Science*, **23**(1), 1–13.
- Dijkstra, E. (1959). A note on two problems in connexion with graphs. *Numeriche Mathematics*, **1**, 269–271.
- Ernst, A., Jiang, H., Krishnamoorthy, M. and Sier, D. (2004). Staff scheduling and rostering: A review of applications, methods and models. *European Journal of Operational Research*, **153**, 3–27.
- Fores, S., Proll, L. and Wren, A. (1999). An improved ILP system for driver scheduling. In *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **471** (N. H. M. Wilson, ed.), pp. 43–61, Springer-Verlag.
- Fores, S., Proll, L. and Wren, A. (2001). Experiences with a flexible driver scheduler. In *Computer-aided Scheduling of Public Transport*. Lecture Notes in Economics and Mathematical Systems, **505** (S. Voss and J. R. Daduna, eds), pp. 137–152, Springer-Verlag.
- Freling, R., Wagelman, A. P. M. and Paixao, J. M. P. (1999). An overview of models and techniques of integrating vehicle and crew scheduling. In *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **471** (N. H. M. Wilson, ed.), pp. 441–460, Springer-Verlag.
- Freling, R., Huisman, D. and Wagelmans, A. P. M. (2001a). Applying an integrated approach to vehicle and crew scheduling in practice. In *Computer-aided Scheduling of Public Transport*. Lecture Notes in Economics and Mathematical Systems, **505** (S. Voss and J. R. Daduna, eds), pp. 73–90, Springer-Verlag.
- Freling, R., Lentink, R. M. and Odijk, M. A. (2001b). Scheduling train crews: A case study for the Dutch railways. In *Computer-aided Scheduling of Public Transport*. Lecture Notes in Economics and Mathematical Systems, **505** (S. Voss and J. R. Daduna, eds), pp. 153–165, Springer-Verlag.
- Haase, K., Desaulniers, G. and Desrosiers, J. (2001). Simultaneous vehicle and crew scheduling in urban mass transit systems, *Transportation Science*, **35**(3), 286–303.
- Hickman, M, Voss, S. and Mirchandani, P. (eds) (2007). *Computer-aided Scheduling of Public Transport*. Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin (forthcoming).
- Huisman, D., Freling, R. and Wagelman, A. P. M. (2005). Models and algorithms for integration of vehicle and crew scheduling. *Transportation Science*, **39**, 491–502.
- Kroon, L. and Fischetti, M. (2001). Crew scheduling for Netherlands railways ‘destination: customer’. In *Computer-aided Scheduling of Public Transport*. Lecture Notes in

- Economics and Mathematical Systems, **505** (S. Voss and J. R. Daduna, eds), pp. 181–201, Springer-Verlag.
- Kwan, A. S. K., Kwan, R. S. K. and Wren, A. (1999). Driver scheduling using genetic algorithms with embedded combinatorial traits. In *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **471** (N. H. M. Wilson, ed.), pp. 81–102, Springer-Verlag.
- Lourenco, H. R., Paixao, J. P. and Portugal, R. (2001). Multiobjective metaheuristics for the bus-driver scheduling problem, *Transportation Science*, **35**(3), 331–343.
- Mingozi, A., Boschetti, M. A., Ricciardelli, S. and Bianco, L. (1999). A set partitioning approach to the crew scheduling problem. *Operations Research*, **47**, 873–888.
- Mitra, G. and Darby-Dowman, K. (1985). CRU-SCHED: A computer-based bus crew scheduling system using integer programming. In *Computer Scheduling of Public Transport 2* (J. M. Rousseau, ed.), pp. 223–232, North-Holland Publishing Co.
- Paias, A. and Paixao, J. M. P. (1993). State space relaxation for set-covering problems related to bus driver scheduling. *European Journal of Operational Research*, **71**, 303–316.
- Paixao, J. M. P. (1990). Transit crew scheduling on a personal workstation (MS/DOS). In *Operational Research '90* (H. Bradley, ed.), pp. 421–432, Pergamon Press.
- Rousseau, J. M. (ed.) (1985). *Computer Scheduling of Public Transport 2*. North-Holland Publishing Co.
- Shen, Y. and Kwan, R. S. K. (2001). Tabu search for driver scheduling. In *Computer-aided Scheduling of Public Transport*. Lecture Notes in Economics and Mathematical Systems, **505** (S. Voss and J. R. Daduna, eds), pp. 121–135, Springer-Verlag.
- Smith, B. M. and Wren, A. (1988). A bus crew scheduling system using a set covering formulation. *Transportation Research*, **22A**, 97–108.
- Sodhi, M. and Norris, S. (2004). A flexible, fast, and optimal modeling approach applied to crew rostering at London Underground. *Annals of Operations Research*, **127**, 259–281.
- Tykulska, R. J., O'Neill, K. K., Ceder, A. and Sheffi, Y. (1985). A computer rail crew assignment/work rules model. In *Computer Scheduling of Public Transport 2*. (Rousseau, J. M., ed.), pp. 233–246, North-Holland Publishing Co.
- Voss, S. and Daduna, J. R. (eds) (2001). *Computer-aided Scheduling of Public Transport*. Lecture Notes in Economics and Mathematical Systems, **505**, Springer-Verlag.
- Wilson, N. H. M. (ed.) (1999). *Computer-aided Scheduling of Public Transport*. Lecture Notes in Economics and Mathematical Systems, **471**, Springer-Verlag.
- Wren, A. (ed.) (1981). *Computer Scheduling of Public Transport: Urban Passenger Vehicle and Crew Scheduling*. North-Holland Publishing Co.

Appendix 10.A

The Shortest-path problem

The following description is based on Ceder (1978). Further reading can be found in almost any OR book. This appendix supplements Sections 10.3 and 10.5, as well as material in Chapters 12–15.

Consider a connected network $G = \{N, A\}$ with set of nodes N , set of directed arcs A , and given ‘distances’ (travel time, cost, etc.), $d(i, j) \forall (i, j) \in A$. The number of nodes and arcs are $|N|$ and $|A|$, respectively. In a connected network, the shortest path may have four categories:

- (a) From one node to all other nodes with $d(i, j) \geq 0, \forall (i, j) \in A$.
- (b) From one node to all other nodes with $d(i, j)$ any.
- (c) From all nodes to all nodes with $d(i, j) \geq 0, \forall (i, j) \in A$.
- (d) From all nodes to all nodes with $d(i, j)$ any.

The shortest u - v path, termed $\bar{p}$, is defined as an u - v path, $u, v \in N$, such that $d(\bar{p}) = \sum_{(i,j) \in \bar{p}} d(i, j) = \text{Min}_{u-v \text{ paths } p \text{ on } G} \sum_{(i,j) \in p} d(i, j)$.

10.A1 Dijkstra algorithm (one to all, $d(i, j) \geq 0$)

Dijkstra’s algorithm (Dijkstra, 1959) produces a spanning tree of G containing shortest paths from a given node to all other nodes in G , with non-negative arc lengths. It is described as follows.

Step 0: Initially, assign the origin node u with a permanent label, $\pi^*(u) = 0$.

Let the set of nodes with permanent labels be $N(u) = \{u\}$. All other nodes

$$i, i \in N, i \neq u, \text{ have temporary labels: } \pi(i) = \begin{cases} d(u, i), & (u, i) \in A \\ \infty, & \text{otherwise} \end{cases}$$

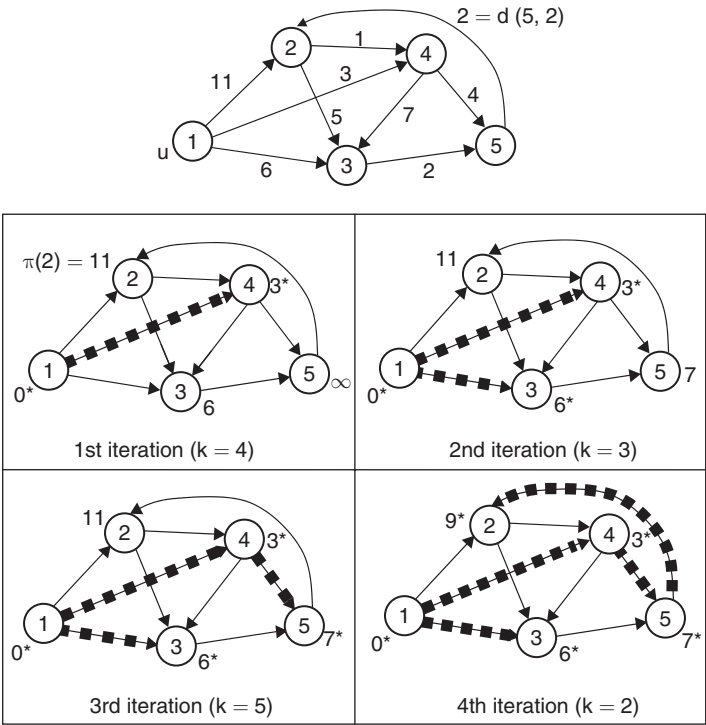
Note that a permanent label $\pi^*(i)$ represents the shortest distances from node u to node i . The symbol ‘:=’ means ‘is replaced by’ in the following steps.

Step 1: Find the node with a temporary label, say node k , such that $\pi(k) = \text{Min}_{i \in N(u)} \pi(i)$. Set $\pi^*(k) = \pi(k)$ and mark/store the arc that leads from $N(u)$ to k . Add k to $N(u)$.

Step 2: If $N(u)$ contains all nodes $i, i \in N$, terminate; otherwise, continue.

Step 3: For each arc (k, j) such that k is a node determined in **Step 1** and $j \notin N(u)$, $\pi(j) := \text{Min}[\pi^*(k) + d(k, j), \pi(j)]$, go to **Step 1**.

The following example illustrates the Dijkstra algorithm with four iterations. The tree emphasized (i.e. a connected paths without cycles) after the 4th iteration is the solution with all nodes permanently labelled. The arcs emphasized represent the desired spanning tree. The minimum spanning tree from node 1 to all other nodes is provided.



Theorem 10.A1: (a) Dijkstra algorithm terminates in, at most, $|N| - 1$ iterations.
(b) Dijkstra algorithm produces a spanning tree whose arcs correspond to the shortest paths between u and all other nodes $i, i \in N$ (the $\pi^*(i)$ values are the lengths of these shortest paths).

Proof: (a) Almost trivial, because one node is permanently labelled per iteration, and the network considered is connected.
(b) By induction. The assumption is that at the m -th iteration, the tree emphasized represents the shortest paths from u to the nodes belonging to that tree. From this point, it is not difficult to complete the proof, and therefore it is left as an exercise.

Remarks: (1) If the interest is in the shortest u - v path (between only two nodes), then terminate once $\pi(v)$ becomes permanently labelled; i.e. $\pi^*(v)$.

- (2) If the interest is in finding the shortest paths between all nodes $i \in N$ and a node u (from all to one), simply reserve all arcs $(i, j) \in A$; i.e. (i, j) will become an (j, i) arc; then apply the Dijkstra algorithm (from node u to all other nodes).
- (3) For a complete graph (there is an arc between each two pairs of nodes), the Dijkstra algorithm requires approximately $|N|^2/2$ additions and $|N|^2$ comparisons, or $3|N|^2/2$ operations. By using the complexity notation in Section 5.3 (Chapter 5), the algorithm runs in $O(|N|^2)$ time for a dense (in arcs) networks. For sparse networks, other algorithms may improve the running time; e.g. see Ahuja *et al.* (1993).

10.A2 Algorithm rearrange-node-numbers (one to all, $d(i, j)$ any, no cycles)

The major limitation of the Dijkstra algorithm is that it can run only on non-negative arc lengths. There are other algorithms that handle optimally any $d(i, j)$ that can be found; for example, in Ahuja *et al.* (1993). This section introduces an algorithm aimed at special cases in which $G = \{N, A\}$ contains no cycles. The algorithm, which is called rearrange-node-numbers (RNN) because of its first step, is described as follows:

- Step 0:** Initially, let the origin node be $u = 1$ and number all nodes $i, i \in N$, such that arc $(i, j) \in A$ if and only if $i < j$. Remove all nodes with assigned numbers less than 1 (since they cannot be reached from u). If this step cannot be done, a cycle is found. Set $j := 2, \pi(1) := 0, A_T = \Phi$.
- Step 1:** $\pi(j) = \min[\pi(i) + d(i, j)]_{(i, j) \in A/i < j} := \pi(i^*) + d(i^*, j)$. Add (i^*, j) to A_T . If there is no arc, connecting A_T with j assume $\pi(i) = \infty$, and continue.
- Step 2:** Let $j := j + 1$. If $j = |N|$ terminate. The tree $T = \{N, A_T\}$ is the required spanning tree, and $\pi(i)$ are the shortest paths. For $\pi(i) = \infty$, there is no path between u and i on G . If $j < |N|$ go to **Step 1**.

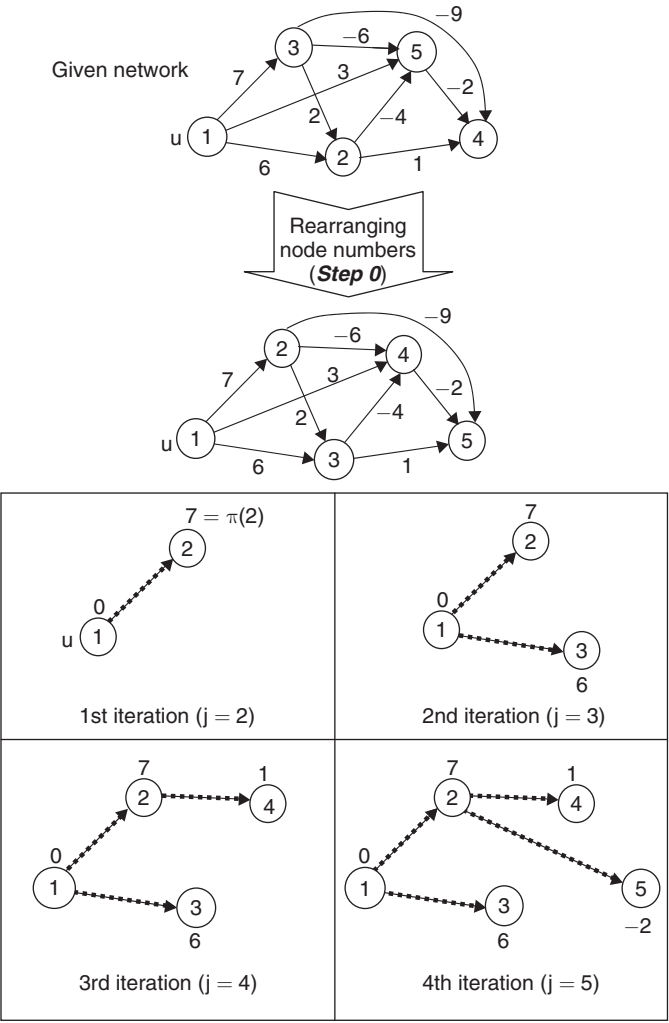
Theorem 10.A2: (a) Algorithm RNN terminates in $|N| - 1$ iterations.
 (b) Algorithm RNN produces the required spanning tree; the $\pi(i)$ value is the shortest path from u to node i (for $\pi(i) = \infty$, there is no path from u to i).

Proof: (a) Self-evident, because one node is added per iteration.
 (b) By induction. The assumption is that at the m -th iteration, the tree emphasized represents shortest paths from u to the nodes on the tree. The rest is left as an exercise.

Remarks: (1) If the interest is in the shortest u - v path (between only two nodes), then terminate once i equals the node number of v .
 (2) If the interest is in finding the shortest paths between all nodes i and a node u (from all to one), simply reserve all arcs; i.e. (i, j) will become (j, i) arc; then apply algorithm RNN.

- (3) The RNN algorithm is basically applied to known project-scheduling methods, such as PERT (program evaluation and review technique), CPM (critical path method), and others. Small changes in RNN enable one to introduce earliest-starting and latest-end times of each task (described by an arc).

The example that follows, in which some $d(i, j) < 0$, shows the RNN algorithm in four iterations. The solution is represented by the last tree, with shortest paths from $u = 1$ indicated as $\pi(i)$ on the nodes.



11 Passenger Demand



Chapter 11 Passenger Demand

Chapter outline

- 11.1 Introduction
- 11.2 Transit demand, its factors and elasticity
- 11.3 Example of a demand forecasting method and process
- 11.4 Multinomial logit (MNL) model
- 11.5 Literature review and further reading (O-D estimation)

Exercises

References

Practitioner's Corner

One of the main measures of the improvement of transit services is the increase in passenger demand. However, there is a trade-off between the cost of improvement and the extra benefit gained by the additional demand. It is common to find transit managers who are puzzled about the relationship (if any) between service changes and their impact on changes in demand. In fact, the following humoristic aphorism readily applies: “The more unpredictable the transit demand becomes, the more we rely on predictions”. Although it is easier to try out service changes for bus services (with conclusions drawn after a learning period), the rail industry requires more assurances.

The chapter contains four main parts following the introductory section. Section 11.2 presents and discusses the basic attributes and tools used in passenger-demand analysis; among these attributes are fares, travel time, service frequency, walking time, routing and transferring, and comfort and inconvenience elements. Section 11.3 provides an example of a transit-demand forecast methodology that is adapted to predict, for a given transit service, the future patronage of a specific set of routes in certain required years. Section 11.4 exhibits a known share model that divides passengers among various travel modes according to each mode's relative desirability; such a technique has been commonly used to determine modal split, although its accuracy depends heavily on the underlying mathematical model. The last main part, Section 11.5, presents a literature review on estimating origin–destination matrices, which constitute an essential input for most transit-planning and design procedures. The chapter ends with exercises.

Practitioners are advised to read the entire chapter, rather than just Section 11.4. It should be noted, however, that demand-prediction models, no matter how good they are, should be treated carefully because of the vast number of assumptions (behavioural and others) that are inherent in them. In other words, one should not believe that the models are reality (“don't eat the menu”), but should properly weight practical issues against modelling precision. The following story may ‘teach’ us about life and precision (squareness).

During the French Revolution, three professionals were sentenced to be executed by guillotine. The first person had been collecting practical data, the second analysing practical data, and the third modelling the precision of data. For the third professional, precision was above all else (a square person). The first person was asked if he wanted his head facing up or down. He requested down, and the guillotine dropped, stopping two centimetres before his neck; there was an old rule that if the guillotine stopped, you're sent free, and indeed he was released. The second person was asked about his head position; he asked face up, and the guillotine again dropped – and, unbelievable, for this had never happened before, stopped one centimetre before his throat, allowing him, too, to go free. The third person, for whom precision was all important, replied to the same question: "I don't want to go under that damn machine before you fix it!"

11.1 Introduction

Chapter 10 ends a main subject of the book, transit scheduling. Chapters 13–16 are more design oriented, focusing on service, network, route and shuttle design. Before any discussion of transit-design elements, however, it would be only natural to present an overview of the tools and models used in analysing passenger demand (Chapter 11) and the related prediction of the way the demand will flow on the transit network (Chapter 12). The first question confronting us in designing a transit service concerns the size, composition and distribution of passenger demand; it is the dominant input parameter for any new, improved, or redesign undertaking. The sort of sandwich position of this and the next chapter between the subjects of the book is displayed in Figure 11.1.

The purpose of transit-demand studies is to estimate and evaluate passenger demand by collecting and analysing data and models pertaining to current and future transit needs. Transit-demand studies and models form an essential part of any transit-planning process.

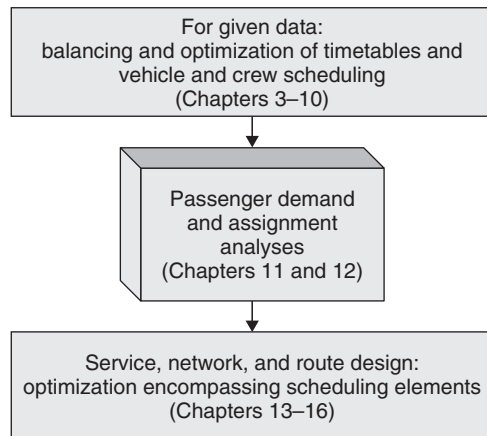


Figure 11.1 Position of Chapters 11 and 12 in the sequence of main subjects of the book.

The venerable transportation-planning practice is based on a four-step sequential process: *Trip generation* (number of trip ends associated with a zone of land), *Trip distribution* (distribution of trips among zones), *Modal split* or *Mode choice* (choice among travel modes for personal trips), and *Traffic assignment* (choice among available O-D routes and the resulting accumulated traffic). A thorough overview of this four-step process and its variants appears in Ortuzar and Willumsen (2001). This planning practice for trip forecasting contains trips associated with transit modes, thereby calling for the utilization of some of its features in analysing passenger demand.

The four-step sequential planning process is discussed in Boyce and Daskin (1997) in conjunction with solving optimal travel-choice models. Those researchers explained that the four-step process had to be solved with feedback in order to forecast equilibrium traffic conditions. In their iterative procedure involving the four-step process, the solution of the trip-assignment (route choice) model is required to solve the trip-distribution (O-D choice) model – that is, equilibrium travel times and costs. A variant of the four-step process (Ortuzar and Willumsen, 2001) indeed considers route choice as the step preceding the O-D choice step. Boyce and Daskin (1997) believe that the four-step process can at best be seen as a heuristic for solving a traffic-network equilibrium problem because it lacks the mechanism by which the process could converge to the desired equilibrium solution. Following this explication of the conceptual issues and limitations surrounding the four-step process, the sections below will provide examples of analyses of passenger-demand forecasting.

11.2 Transit demand, its factors and elasticity

The analysis and estimating of transit demand are complex mechanisms mainly because they involve the aggregate behaviour of individuals. Thus, in estimating demand, it is better to doubt what is true than accept what isn't.

The known attributes affecting transit demand are: changes in fare, travel time, service frequency, walking time, routing and transferring, stop location, comfort and inconvenience elements, information, socio-economic factors (e.g. income), external factors (e.g. land-use, security), and competition from other transportation and transit services. Usually passengers (travellers) have a travel choice among alternatives; it is reasonable to assume that their selection process is based on maximizing *utility* or finding the best way to realize their travel objectives while complying with constraints. To some extent, this selection process varies among individuals; thus, economic theories describing the behaviour of a customer searching for services or goods can be helpful. A practical guide on transit demand appears in a TRL report by Balcombe *et al.* (2004).

11.2.1 Factors affecting demand

The choice between public and private transportation is an individual decision that is influenced by government/community decisions. These various decisions often send mixed signals to transit and potential transit passengers while failing to recognize their more system-wide and integrated implications. Generally speaking, most large cities have effectively encouraged the use of private cars through planning (dispersed land-use in the suburbs), infrastructure (available parking and circulation traffic flow), pricing and financial

decisions. Consequently, there is growing confusion in many of those cities about what to do. It is therefore important to identify the factors affecting transit demand and to use them for improvements.

For the most part, increasing the quality of service (e.g. in terms of service coverage, service frequency and fare reduction) has been found to have a significant positive outcome on transit demand. Two practical guides that cover the factors affecting demand are Balcombe *et al.* (2004), and TranSystems *et al.* (2006).

The greatest impacts from increasing the transit demand are extracted from a combination of strategy-related actions and initiatives. TranSystems *et al.* (2006), based on a large survey of transit agencies in the USA, categorized these actions and initiatives in the following list (in decreasing order of effectiveness):

1. **Service adjustments or improvements** (increased route coverage; route restructuring; improved schedule/route coordination; increased service frequency; increased span of service; improved reliability/on-time performance; improved travel speed and reduced stops; targeted services; passenger-facility improvements; new/improved vehicles; increased security and safety).
2. **Partnerships and coordination** (university and school passenger programmes; travel-demand management strategies; privately-subsidized activity-centre service; consistent regional, and inter-agency, operating policies; coordination with other transportation agencies; promotion of transit-supportive design).
3. **Marketing, promotional and information initiatives** (targeted marketing and promotions; general marketing and promotions; improved informational materials; improved customer information and assistance).
4. **Fare-collection and fare-structure initiatives** (improved payment convenience; regional payment integration; fare-structure simplification; fare reduction).

In addition both Balcombe *et al.* (2004) and TranSystems *et al.* (2006) indicated that there are **external factors** influencing transit demand. The factors listed significant by TranSystems *et al.* (2006) in the US are as follow:

1. **Population characteristics and changes** (general growth in the region; high and increasing immigration; high and increasing number of elderly; high and increasing tourism; high number of college students).
2. **Economic conditions** (employment/unemployment levels; per-capita income levels; household auto-ownership levels).
3. **Cost and availability of alternative modes** (fuel and toll pricing; parking pricing and availability; taxi fares; fuel taxes; auto purchase and ownership costs; availability of a commuter-benefits programme for employers).
4. **Land-use and development patterns and policies** (density of development; relative locations of major employers and residential areas as from, e.g., increasing suburbanization; land-use and zoning controls and incentives).
5. **Travel conditions** (climate and weather patterns; traffic-congestion levels and highway capacity; traffic disruptions owing to, e.g., major construction projects).
6. **Public policy and funding initiatives** (air quality mandates; auto emission standards; Federal and state operating capital and transit-funding levels).

Another indirect set of factors affecting transit demand is related to **mode choice elements** extracted from various types of public policies. Examples of these elements given by TranSystems *et al.* (2006) are as follow:

- 1. **Price and availability of each mode** (cost of auto use; cost of transit; parking cost).
- 2. **Quality of service of each mode** (travel time; convenience; comfort; service reliability; perceived personal security and safety; perceived overall ‘image’ of each mode).
- 3. **Trip characteristics for each particular trip** (trip length and purpose; number of people to be making the trip; multiple destinations).
- 4. **Socio and demographic characteristics of the traveller** (income; origin and destination locations; status).

Retaining a high level of passenger satisfaction while fully maintaining the protection of access to less-affluent citizens is another way to look at factors for improving transit demand. For instance, TranSystems *et al.* (2006) reported on large-scale surveys conducted for the Washington Metropolitan Area Transit Authority. The results of these surveys, shown in Table 11.1, indicate significant differences between the most desired transit-service improvements

Table 11.1 *Transit-service improvements desired by bus riders and non-riders in the Washington, DC region (percentage opting for each item)*

Improvement item	Riders	Non-riders
Better reliability	49%	6%
More frequent service	31%	16%
More convenient stops	13%	18%
Better information	9%	30%
Better shelters	10%	21%
Faster service	10%	16%
Better vehicle condition	8%	13%
Service to destination	12%	13%
Longer hours*	25%	N/A
Less crowding*	22%	N/A
Lower fares*	8%	N/A
Better costumer service*	7%	N/A

* Not applicable for non-riders
Source: TranSystems *et al.* (2006)

for riders and non-riders. It was concluded that targeting new riders warrants a different set of actions and initiatives than those used to influence existing riders.

Often, desired service-improvement items can be interpreted by the willingness of riders and potential riders to pay for an increase in their satisfaction (Ceder, 2004). In one study, Kottenhoff (1998) employed the stated-preference method to ascertain Swedish train passengers' willingness to pay. The main question was how to increase the attractiveness of the rail system (the willingness to pay) and, simultaneously, to decrease its costs?

To perform these quantifications of willingness to pay, train passengers were asked to answer computer-assisted self-interviews. Table 11.2 presents the results by percentage of fare levels for which passengers are willing to pay extra in order to improve the specified

Table 11.2 *Passengers' willingness to pay for train service in Sweden
(by % of fare level to be paid as extra)*

Improvement item	Fare level
Timetable factors	
Change of train: one less transfer	19%
Speed: 20% less travel time	15%
Frequency: 2-hour to 1-hour	5%
Comfort	
Shaking and vibrations: 'a little less'	11.5%
Noise: 'a little less noisy'	10%
Climate: better ventilation	10%
Seat adjustment: reclining seat backs	8%
Seat orientation: face-to-face or -back	8%
Leg-room: 10 cm less or more	6.5%
Seat size: 5 cm wider	4%
On-board services	
Restaurant with hot food	13%
Bistro with some food	10%
Division: reading salon/quiet salons	9%
Division: play areas	9%
Coffee: free coffee and tea in the car	6%
Entertainment: video or cinema on board	5%
Entertainment: music/radio outlets	3%
Office: service	2%
Quality satisfaction	
On-time arrivals: from 80% to 90%	16%
Lavatories: 'modern and roomy'	14%
Modernity: modern coach	9%
Reservation: optional seat booking	9%
Cleanliness: clean inside	5%
Mode: going by double-decker	3.5%

Source: Kottenhoff (1998).

item. These results were also checked (and agreed with in general) against the results from other European countries.

One example of a train service for which the conditions were improved and the payment was increased is the Danish IC/3, marketed under the name Kustpilen, which replaced simple rail-bus services. The number of travellers rose by almost 200 per cent in the first four years as reported by Kottenhoff and Lindh (1996). The travelling time from Karlskrona to Malmö was reduced by 15 minutes (7%), and passengers travelling all the way no longer had to make a transfer. The frequency was increased from 6 to 11 trains per day in each direction. Although the fare level was reduced by about 10–30 per cent, the enhanced comfort offered by the new IC/3s was found to be the main attribute for the increase in patronage.

11.2.2 Demand function and elasticity

Theoretically a demand function can be mathematically expressed as a function of the attributes (explanatory variables) mentioned above. That is,

$$D_f = f(y_1, y_2, \dots, y_m) \quad (11.1)$$

where D_f is transit demand (in passengers) and y_i , $i = 1, 2, \dots, m$, are the attributes.

Basically no preference is given to any functional form of Equation (11.1); the form is determined empirically for the type of application under question, using statistical techniques.

The responsiveness (sensitivity) of demand to changes in y_i attributes is known as elasticity. This is to say, demand elasticity concerning y_i is the ratio of the percentage change in demand to the percentage change in y_i . Assuming the existence of a demand function D_i for a given attribute y_i :

$$D_i = f(y_i), i = 1, 2, \dots, m \quad (11.2)$$

then there are two main direct elasticity definitions as explicated by Balcombe *et al.* (2004): *point* elasticity and *arc* elasticity. The direct point elasticity of y_i is calculated as the slope of the demand curve times the ratio of the values of y_i to D_i :

$$\varepsilon_i^{\text{point}} = \left(\frac{\partial D_i}{\partial y_i} \right) \frac{y_i}{D_i} \quad (11.3)$$

The direct arc elasticity is calculated between two points on the demand curve, using a logarithmic form extracted from the differential $(dD_i/D_i)/(dy_i/y_i) = d(\ln D_i)/d(\ln y_i)$:

$$\varepsilon_i^{\text{arc}} = \frac{\log D_{i1} - \log D_{i2}}{\log y_{i1} - \log y_{i2}} = \frac{\Delta(\log D_i)}{\Delta(\log y_i)} \quad (11.4)$$

where D_{i1} and D_{i2} exhibit the values of the demand before and after the change of the attribute from y_{i1} to y_{i2} , respectively.

In a competitive environment, a change in an attribute (e.g. fare) for one transit service may affect the demand for another transit service. The sensitivity of these cross-effects is indicated by the phenomenon of cross-elasticity, expressed mathematically as:

$$\varepsilon_i^{uv} = \left(\frac{\partial D_i^u}{\partial y_i^v} \right) \frac{y_i^v}{D_i^u} \quad (11.5)$$

where ε_i^{uv} is the (point) cross-elasticity of demand for service u concerning the change in attribute y_i of service v ; D_i^u is the value of the demand function for attribute y_i , using service u ; and y_i^v is the value of attribute y_i , using service v .

In practice, it may be difficult to measure cross-elasticity (since direct effects frequently outweigh cross-effects). The following relationship shows how to calculate it from direct elasticities:

$$\varepsilon_i^{uv} = \left| \varepsilon_i^u \right| \frac{D_i^u}{D_i^v} \left(\frac{\partial D_i^v}{\partial D_i^u} \right) \quad (11.6)$$

where ε_i^u (an absolute value) is the direct elasticity of demand for service u for the change in attribute y_i of (the same) service u ; D_i^u/D_i^v is the ratio between the competing volumes of services, u and v ; and $\partial D_i^v/\partial D_i^u$ is the proportion of the demand change in services v and u as a result of the change in y_i^v . More on elasticity functions and approaches can be found in Balcombe *et al.* (2004).

11.2.3 Overview of elasticity results

Litman (2004) and Balcombe *et al.* (2004) summarized and compared numerous studies on elasticity values for several attributes, with an emphasis on transit fares. Litman (2004) concluded the research with the recommendation to use ranges rather than point values because of the inevitable uncertainty inherent in elasticity analysis.

Litman's (2004) recommendations appear in Table 11.3 for short- and long-run predictions. The short run refers to periods of less than two years; the long run, to periods of

Table 11.3 Recommended direct-transit elasticity values

	Market segment	Short run	Long run
Transit demand in regard to the attribute of fares	Overall	−0.2 to −0.5	−0.6 to −0.9
	Peak hours	−0.15 to −0.3	−0.4 to −0.6
	Off-peak hours	−0.3 to −0.6	−0.8 to −1.0
	Suburban commuters	−0.3 to −0.6	−0.8 to −1.0
Transit demand in regard to the attributes of service quality	Overall	0.5 to 0.7	0.7 to 1.1
Transit demand in regard to the attribute of car operating cost	Overall	0.05 to 0.15	0.2 to 0.4
Car users in regard to transit costs	Overall	0.03 to 0.1	0.15 to 0.3

Source: Litman (2004), Table 11.

five–ten years. The transit demand considered is elastic for elasticity values approaching or above 1.0 (or approaching or below -1.0). Measurable attributes influencing the quality of service are included in Table 11.3 under ‘attributes of service’; that is, frequency of service, access and egress times, waiting time, in-vehicle time and transfer time. According to Litman, it is assumed in practice that demand is inelastic for the short run; a fare increase and/or service reduction will have a marginal effect on the demand, thus increasing net revenue.

Both Balcombe *et al.* (2004) and Litman (2004) indicated that fare elasticities were about twice that for off-peak and leisure trips as for peak and commuting trips. In addition, short-distance trips and trips made in small towns have higher fare elasticities than do longer trips and trips made in large cities.

Balcombe *et al.* (2004) separated the attributes influencing the quality of service into those that can be quantified and those that can be only indirectly estimated. The latter include attributes related to transit reliability, comfort, convenience and safety. In the first group of attributes, Balcombe *et al.* (2004) used both elasticity values and attribute weighing in terms of equivalent in-vehicle time. They estimated that the elasticity of demand to vehicle-km for buses was between 0.1 and 0.7 for short runs, and between 0.2 and 1.0 for long runs; for rail, it is roughly 0.6 to 0.9 for short runs. The in-vehicle time elasticity for (local) buses ranges is -0.4 to -0.6 , and for rail -0.6 to -0.8 .

Both Balcombe *et al.* (2004) and Litman (2004) indicated that fare elasticities were about twice that for off-peak and leisure trips as for peak and commuting trips. In addition, short-distance trips and trips made in small towns have higher fare elasticities than do longer trips and trips made in large cities.

Balcombe *et al.* (2004) separated the attributes influencing the quality of service into those that can be quantified and those that can be only indirectly estimated. The latter include attributes related to transit reliability, comfort, convenience and safety. In the first group of attributes, Balcombe *et al.* (2004) used both elasticity values and attribute weighing in terms of equivalent in-vehicle time. They estimated that the elasticity of demand to vehicle-km for buses was between 0.1 and 0.7 for short runs, and between 0.2 and 1.0 for long runs; for rail, it is roughly 0.6 to 0.9 for short run. The in-vehicle time elasticity for (local) buses ranges is -0.4 to -0.6 , and for rail -0.6 to -0.8 .

The use of attribute weighing was applied to access/egress walking time, waiting time and service headway. Balcombe *et al.* (2004) estimated that access/egress walking time can be on average 1.68 times the value of in-vehicle time (depending on trip length and walking time); when transfers are involved, this average is around 1.81, which may also reflect a transfer penalty. Waiting time was valued at 1.76 times that of in-vehicle time (varying by trip length). Service headway was valued at 0.77 times in-vehicle time.

An additional attribute considered by Balcombe *et al.* (2004) was income, including car-ownership effects. Based on the national travel survey in the UK (1985–1997), they recommended values of -1.08 and 0.34 for long-run elasticities for this attribute for commuting bus and rail trips, respectively; and values of -0.33 and 0.42 for long-run elasticities for leisure trips for bus and rail, respectively. They noted that the bus-income elasticity, including car-ownership effects, was negative, compared to positive rail-income elasticities.

Hensher (2001), using mixtures of revealed-preference and stated-preference data, demonstrated an example of direct and cross-elasticities in Sydney. Revealed-preference applies to observed data (e.g. by agencies in regard to ticket sales, direct home surveys); stated-preference

applies to questionnaires inquiring into the relative importance of factors that are not easily appraised through revealed-preference methods. Table 11.4 presents Hensher's results for commuters; each column has one direct (in *italic*) and six cross-elasticities. For instance, concerning the change in the single-fare bus ticket (fourth column in Table 11.4), an increase of 10% in this ticket's fare tends to reduce its sales by 3.57% and to increase the sale of single-fare train tickets by 0.57%, and to increase the number of car users by 0.66%. Similarly this fare change can apply to the other types of tickets.

Another study on fare (or trip cost) direct and cross-elasticities for commuting trips in Sydney was presented by Taplin (1997). This study, whose findings are summarized in Table 11.5, furnished adjusted fare (or cost) elasticities for trips performed by different transit modes and cars. The direct-elasticities are emphasized and have negative values. For example, an increase of 10% in train tickets in Sydney tends to reduce train-passenger demand (sale of this ticket) by 1.56%, and to increase bus and ferry riders and car users by 0.63%, 0.39% and 0.16%, respectively.

Table 11.4 *Direct and cross-elasticities for Sydney commuters.*

Type of ticket	Mode						
	Train			Bus			Car
	Single fare	Weekly	Pass	Single fare	Ten fare	Pass	–
Single fare (train)	−0.218	0.001	0.001	0.057	0.005	0.005	0.196
Weekly (train)	0.001	−0.093	0.001	0.001	0.001	0.006	0.092
Pass (train)	0.001	0.001	−0.196	0.001	0.012	0.001	0.335
Single fare (bus)	0.067	0.001	0.001	−0.357	0.001	0.001	0.116
Ten fare (bus)	0.020	0.004	0.002	0.001	−0.160	0.001	0.121
Pass (bus)	0.007	0.036	0.001	0.001	0.001	−0.098	0.020
Car	0.053	0.042	0.003	0.066	0.016	0.003	−0.197

Source: Hensher (2001), Table 3.

Table 11.5 *Adjusted fare (or trip cost) direct and cross-elasticities for commuting trips in Sydney.*

Mode of travel	Train	Bus	Ferry	Car
Train	−0.156	0.032	0.003	0.037
Bus	0.063	−0.070	0.006	0.046
Ferry	0.039	0.037	−0.195	0.003
Car	0.016	0.011	0.000	−0.024

Source: Balcombe *et al.* (2004), Table 9.22.

Finally, elasticity may possibly be modelled through the functional relationship between demand changes and fare (or trip cost) changes. For example, Hong Kong transit agencies, according to Zhou *et al.* (2005), are more concerned with the transit-fare structure than with frequency of service in maximizing their profits. This concern derives from the relatively large sensitivity of passenger demand to fares. Zhou *et al.* (2005) mention that for a 1 per cent reduction in transit fares, there is a consequent 1.33%–1.45% increase in transit patronage in Hong Kong ($\varepsilon < -1.0$), compared with only a 0.3% ($\varepsilon = -0.3$) increase in Canada. They illustrated the demand-elasticity effects on revenues through changing fares in the following assumed exponential function:

$$D_f = B e^{\alpha p} \quad (11.7)$$

where D_f is the transit demand, p is route fare, and B and α are constant parameters (estimated from observed data). The *point* elasticity of Equation (11.5) for fares is:

$$\varepsilon = \frac{dD_f}{dp} \cdot \frac{p}{D_f} = \alpha p \quad (11.8)$$

which is typically negative. Moreover, one obtains $\alpha = \varepsilon/p$. Zhou *et al.* (2005) also referred to a denoted revenue R , in which $R = D_f \cdot p$. By using the above relationships, the revenue has the following derivative:

$$\frac{dR}{dp} = B e^{\alpha p} (1 + \alpha p) \quad (11.9)$$

For elasticity ε values less than -1.0 (thus, $\alpha < -1/p$), an increase in fare will result in a decrease in revenue. Consequently, Hong Kong transit agencies can increase their revenue by actually decreasing their fares.

11.3 Example of a demand forecasting method and process

This section presents an overview of and describes the rationale for a transit-demand forecast methodology adapted to predict, for a given transit service, which we will call Transit-A, the future patronage of a specific set of routes in certain required years. This example of a forecasting method also considers relevant survey results of passenger demand for competitor transit services associated with the same underlying set of routes. Taking into account all the parameters that affect transit-service demand is an exhaustive undertaking. Thus, simple but useful models addressing the major attributes of the passengers' decision process will be shown in this section. The method presented follows Ceder (2006) in a study of passenger demand for Hong Kong ferries.

11.3.1 Framework

We assume that Transit-A patronage relies not only on certain attributes, but also on the changes in the levels of service of the competitive transit services. Therefore, whenever there is a change in the level offered by these other services, an alternative model, such as modal split, should be employed for a coarse first estimation.

In brief, the applied approach of the demand forecast covers the following four components: (1) identification of main attributes and their weights in affecting potential Transit-A passengers; (2) demand prediction using a calibrated growth factor from past Transit-A demand figures for any projected changes in the established Transit-A routes; (3) transit modal split between specific O-D pairs on any new Transit-A routes and/or changes in fares and/or journey times in competitor routes; and (4) development of Transit-A O-D demand matrices for each alternative network in a required year.

The forecast approach principally relates to the changes in Transit-A-relevant attributes when there are changes in passenger demand; whenever changes in competitive transit links and/or new Transit-A routes arise, a modal split is necessitated for establishing the O-D demand. The relationship between change in attribute values and the corresponding change in Transit-A demand is estimated by conducting a survey (see Chapter 2 for methods) to capture the different behaviour of existing and potential Transit-A users. A demand forecast is usually network-based. The existing Transit-A network provides the base framework from which other networks can be established. Specifically, each network contains a definite set of routes and fares, an operational strategy, and different extents of competition (from other transit services). In addition to the network scenario, a route scenario can also be established to analyse individual routes when it is reasonable to assume that they are independent of one another. The latter was the case with the ferry study (Ceder, 2006).

In order to integrate the demand forecast for Transit-A into transportation-planning studies of the local area being considered, the O-D matrices for all transit services should be derived directly from the local area studies, except for those of the existing Transit-A routes, which are to be revealed by the survey. Combining all this information, the future demand for existing routes can be predicted with regard to any change in the level of service. For new routes and/or changes in other transit services, the common attributes of all relevant services between specific O-D pairs need to be identified before a multinomial logit (MNL) model (see Ortuzar and Willumsen, 2001) can be used for modal split purposes. The coefficients in the MNL model adopted the values of the local transportation planning studies. The following section explicates this MNL model.

The possible input/output lists for a transit-demand forecast methodology are shown in Table 11.6. The framework for this forecasting methodology is presented in Figure 11.2, which summarizes the flow of the forecasting tasks with the use of two models. It is worth noting that the left-hand side of Figure 11.2 relates not only to changes in the existing Transit-A routes but also to changes in the candidate routes following the establishment of their O-D demand on the right-hand side of this figure.

Another important input in Figure 11.2 is the growth factor for a required year. This factor should be updated continuously to allow for more accurate forecasting. The growth factor is extracted from percentage changes in patronage for a given design year over the past few years. It contains all the changes incurred, including changes in travel behaviour. Extrapolation (e.g. utilizing regression models) of the growth factor, which can be negative, exhibits this ongoing patronage change.

11.3.2 Opinion survey (input) interpretation

This section explains how to apply the opinion survey information to demand forecasting. Usually the main attributes affecting transit-passenger demand are as follow: fare, trip time,

Table 11.6 *Input and output lists for forecasting models*

Input	Output
1. O-D survey	1. Base-year Transit-A patronage and potential increase in Transit-A patronage from competitors
2. Opinion survey	2. Attribute weights in linear regression models
3. Estimation of transit-passenger demand from local transportation planning studies	3. Expansion of the base-year Transit-A patronage to the design-year patronage at low, medium and high levels
4. Local new-route service assumptions	4. Patronage estimate of new Transit-A routes
5. O-D matrices, fare matrices, travel time matrices and calibrated split-model coefficients taken from local transportation-planning studies	5. O-D demand forecast by modes given by MNL modal split for a design year as a base for any further changes in new Transit-A routes

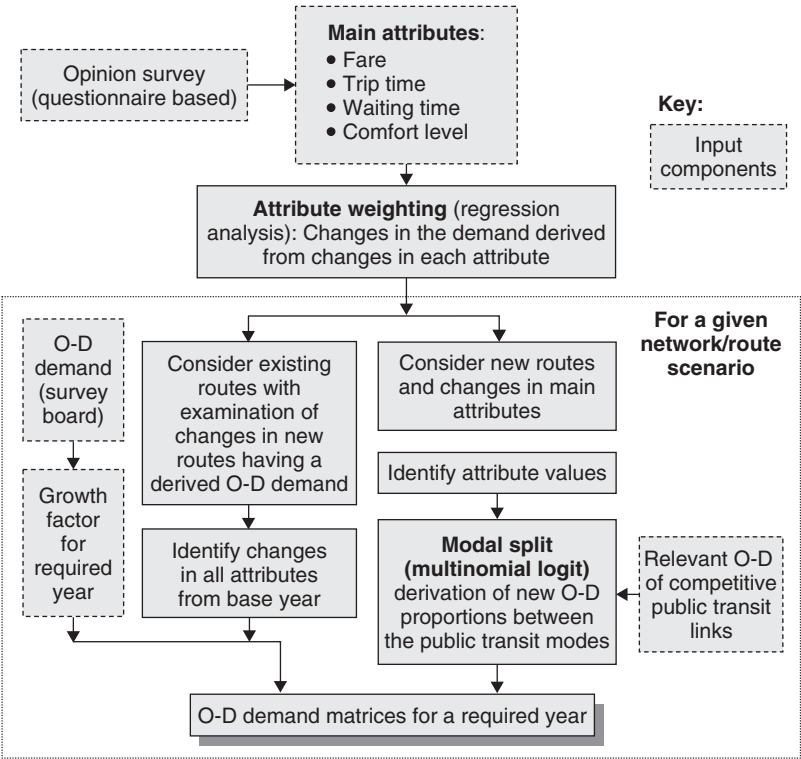


Figure 11.2 *Framework of a given forecasting method for a given transit mode*

waiting time, walking time and comfort level. Two sets of questionnaires can be designed to obtain information on the aforementioned attributes, one set for the Transit-A survey (possibly on-board) and a second for the competitor(s). The first part of the questionnaires can collect O-D information on commuters, which would help to develop the base-year Transit-A demand matrices. The second half helps in establishing the commuter's choice under different scenarios of changed services. Lastly, basic questions on the trip purpose and personal details can be included for the sake of acquiring more information. For a better understanding of people's behaviour and, hence, for more accurate future passenger-demand predictions, a stated-preference type of questionnaire can be designed for fares. This will enable a direct deduction of the relationship between a fare level and a change in demand. A similar arrangement and rationale holds for the item of comfort level, which requires indirect transformation.

The attribute values can then be converted into equivalent monetary values in order to refine the prediction mechanism by making use of the direct relationship between the fare level and a change in demand. For developing a global model for all Transit-A routes, regardless of their O-D characteristics, tailor-made questionnaires, addressing differences in the fare and the other attributes, need to be constructed for individual routes. A list of comfort-level items, from which a respondent can choose some but not all of the most important ones, will be included in the questionnaires. By combining this information with other fare-related information from other questions, an indirect deduction of the monetary value of each comfort level can be achieved, and consequently used in the prediction model.

Example of stated-preference questions:

1. Will you still ride on a Transit-A vehicle if the fare level is increased to: (i) \$7, (ii) \$6.50, (iii) \$6, (iv) \$5.50, (v) \$5.30?
2. If the Transit-A trip travel time is five minutes longer, how much do you think the fare should be reduced? (Mark an amount.)
3. If the Transit-A service waiting time is three minutes longer, how much do you think the fare should be reduced? (Mark an amount.)
4. Please mark up to three comfort items that are most important to you for improving the service? (Choose from a given list.)
5. Given improvements in the comfort items you marked in (4), are you willing to pay for an increase in the Transit-A fare to: (i) \$10, (ii) \$7.50, (iii) \$6?

A linear regression model can be utilized as a tool to relate changes in the attribute value to corresponding changes in demand. Because of different bases in most of the attributes associated with different routes, a global method can be established by expressing the percentage change for both the attribute values and the demand. For simplicity, each attribute can be assumed to be independent of all other attributes in that they can be examined separately while developing the model. For different levels of changes in fare, a corresponding percentage change in Transit-A patronage can be deducted from the survey result. This information can then be combined, using a linear regression, to determine the relative weight of the fare attribute from the slope of the regression line. Similar procedures may be used for other attributes after they are converted to monetary values, thereby making use of the fare-demand relationship developed. A procedure to combine existing riders and potential riders from a competitor within Transit-A appears in Figure 11.3.

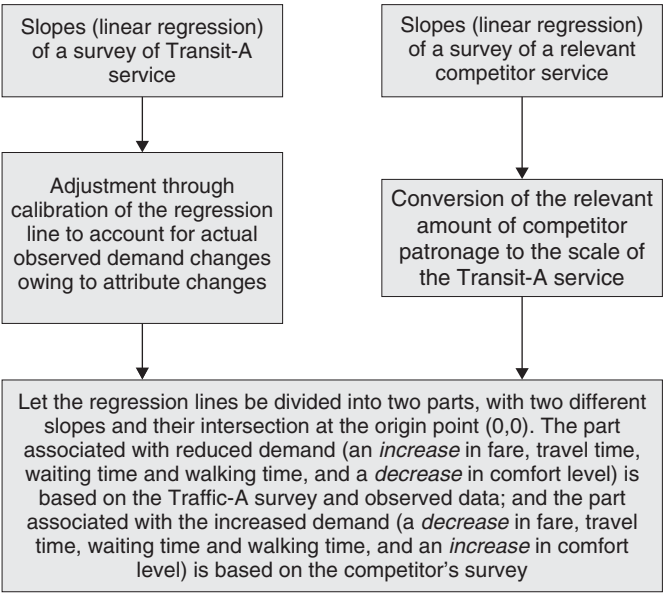


Figure 11.3 Procedure for combining the survey results of the existing and potential users of Transit-A service

The regression lines are supposed to intersect the origin (0,0), assuming that if there is no change in the attribute, there will be no change in the demand. However, it is known that the answers of survey respondents are biased, reflecting riders’ strong wishes to improve the ride without additional cost and/or the natural resistance to increase Transit-A fares and/or worsen its other attributes. As a result, the best fit for the regression line does not intersect the origin; it shows that demand is dropped even if there is no change in the attribute. One basic way to correct this bias, over and above its calibration with actual data, is to ignore the free constant element in the linear regression and to allow the regression line to intersect the origin.

11.3.3 Example of attribute derivation

A detailed example will now be provided as a useful device for understanding the underlying procedures. This example explains how to calibrate the weights of attributes from the raw survey data. It is assumed that 20 passengers were interviewed, and their choices for each question randomly assigned.

First, let us consider the question: Will you still ride on a Transit-A vehicle if the fare level is increased to: (i) \$7, (ii) \$6.50, (iii) \$6, (iv) \$5.50, (v) \$5.25? These five options correspond to five levels of fare increase from \$5 (by 5%, 10%, 20%, 30% and 40%), with ‘1’ being assigned to the option selected. In case a passenger is unwilling to pay for any fare increase, no option is assigned. The existing fares range between \$2.20 and \$5.30. The acceptable increase in fare equals the existing fare times the increase option chosen by the passenger interviewed. For example, the acceptable fare increase for one passenger was

$\$5.00 \cdot 10\% = \$0.50$; that is, for \$5.50 this passenger will still ride on Transit-A. The demand change for the fare is calculated from the number of passengers still willing to ride Transit-A given a fare increase. For instance, for a fare increase of 30%, five (out of twenty) still will ride on Transit-A; this means a 75% patronage decrease. Plotting the % change in demand versus the % change in fare for all routes and cases will construct a data set suitable for a regression analysis.

Second, we will now present a full analysis of the question of an increase in travel time. Similarly, this analysis can be performed, for example, for waiting time, walking time/distance and comfort items. Table 11.7 presents the raw survey data output for the question: if the Transit-A trip travel time is x minutes longer, how much do you think the fare should be reduced? (Interviewees need to mark an amount.)

The value of travel time is calculated by the ratio of the first two columns in Table 11.7. A set of travel-time changes is tested, whether or not passengers still ride Transit-A. For example, for the first passenger interviewed (first raw data in Table 11.7), an increase in travel time of 40% is equivalent to a fare increase of $7 \cdot 40\% \cdot 0.33$ (existing travel time *times* travel-time change *times* the value of travel time) = \$0.924. Because the acceptable increase in fare for the first passenger was calculated (in the first question) to be 0.44, this passenger will not ride on Transit-A when the travel time increases 40% or more, and thus '0' is

Table 11.7 Interpretation of survey results concerning a change in travel time

Value of x more minutes (\$)	x (minute)	Value of travel time (\$/minute)	Existing travel time (minute)	Change in travel time					
				40%	30%	20%	10%	5%	1%
1	3	0.33	7	0	0	0	1	0	0
0.5	3	0.17	7	1	0	0	0	0	0
1	3	0.33	8	0	0	0	1	0	0
0.5	3	0.17	8	0	0	0	0	1	0
1	5	0.20	13	0	0	0	1	0	0
2	5	0.40	13	0	0	0	1	0	0
2.5	3	0.83	7	0	0	1	0	0	0
1.5	3	0.50	7	0	0	0	0	0	0
3	3	1.00	5	0	1	0	0	0	0
3	3	1.00	5	0	0	0	0	0	1
2	3	0.67	7	0	0	0	0	1	0
3	3	1.00	7	0	0	1	0	0	0
2.5	5	0.50	11	0	0	0	0	0	1
1.5	5	0.30	11	0	0	1	0	0	0
2	5	0.40	12	0	0	0	0	1	0
2.5	5	0.50	12	0	0	0	0	1	0
1	5	0.20	15	1	0	0	0	0	0
2	5	0.40	15	0	0	0	0	0	0
2.5	5	0.50	10	0	0	1	0	0	0
3	5	0.60	10	0	0	0	0	0	1
Demand change for travel-time change				-90%	-85%	-65%	-45%	-25%	-10%

assigned. If, however, the travel time should increase by only 10%, this would be equivalent to a fare increase of $7 \cdot 10\% \cdot 0.33 = \$0.231$, which is less than the maximum acceptable increase, and therefore ‘1’ is assigned in Table 11.7. Thus, the demand change for the travel-time change can be calculated with this conversion; this is plotted in Figure 11.4 with the resultant fitted regression line. The R-squared value (0.816) is the fraction of the variance in the data that is explained by the regression.

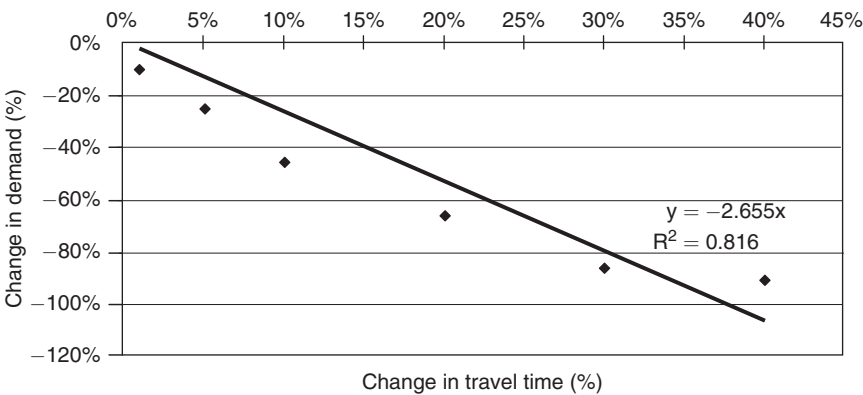


Figure 11.4 Regression line of change in travel time against change in demand

11.4 Multinomial logit (MNL) model

The MNL model (see, for example, Ortuzar and Willumsen, 2001) is a share model that divides the individuals among various travel modes according to each mode’s relative desirability for any given trip. Such a technique has been commonly used to determine modal split in a number of studies, although its accuracy depends heavily on the given data and underlying mathematical format of the logit model.

This modal split model is usually used when there are some factors that the linear regression model cannot fully address; for example, the introduction of new Transit-A routes and changes in the degree of competition with other public transit modes (e.g. via improvements, fare changes). Four public transit modes are commonly included and examined in the modal split process along with the private vehicle. These modes are rail, bus, ferry and taxi.

The input attributes for the MNL model should be consistent and valid for all the available transit modes and services. We will consider here the two main attributes affecting modal split – fare and travel time. Because of the limited number of attributes employed in the model, it may be necessary to fine-tune Transit-A demand.

The MNL model takes the following mathematical form:

$$P_m = \frac{e^{u(m)}}{\sum_{i=1}^n e^{u(i)}} \quad \text{for } 1 \leq m \leq n$$

(11.10)

where P_m = the probability of an individual's choosing mode m , or the proportion of trips using mode (or service) m ; $u(i)$ = the utility of the i -th mode, which includes the modal parameters and coefficients for a certain trip among the n available modes. For instance, the utility function with reference to buses can take this form:

$$u(i) = T(t_i - t_{bus}) + C(c_i - c_{bus}) + B_1 + B_2 \quad (11.11)$$

where t_i = the travel time incurred by mode i ; c_i = the cost (fare) by mode i ; T, C = calibrated coefficients; B_1 and B_2 are biases for the mode and movement limitations, respectively, varying according to whether the mode is accessed by walking or by another transit mode and according to the degree of access difficulty.

The following are typical MNL assumptions used in the forecast: (1) given zoning system; (2) fares and travel times for rail, bus, ferry and taxi taken from given matrices for peak and non-peak periods, as well as the cost and travel time of private vehicles; (3) total transit patronage taken from given local transportation-planning studies (aimed at capturing the right mixture of population and economy growth affecting transit demand) and based on self-zones selection; (4) given number of peak and off-peak hours for the local area considered; and (5) specific zone selection depending on the proportion of transit users.

Following is an example of the use of the MNL model. Suppose a new railway line is introduced in a transit network consisting of rail, bus and passenger ferry. In order to arrive at an estimate of daily patronage for this new line, the corresponding zones of the line's end points should initially be identified; say, Z1 and Z2. The corresponding transportation-planning-studies figures for this particular O-D pair for a designated year are as follows:

- Total daily transit demand: 164,000 trips (Z1 to Z2) and 161,000 trips (Z2 to Z1).
- Travel times (includes the trip, waiting, walking and transfer times), and fares for each of the three transit modes:

Mode	Travel time (minutes)		Fare (\$)
	Z1 to Z2	Z2 to Z1	
Rail	27.1	27.1	7.22
Bus	51.9	51.5	7.91
Ferry	49.2	48.2	5.37

The travel times and fares are inserted into the utility function (11.11) with the following coefficients:

Mode	T	C	B ₁	B ₂
Rail	-0.00373	-0.00153	0.202	0.267
Ferry	-0.00373	-0.00153	0.491	-1.1

The utility value of each of the transit modes is then calculated by Equation (11.11):
 $u(\text{rail}) = 0.567409$, $u(\text{bus}) = 0$, $u(\text{ferry}) = -0.595789$.

Utilizing the MNL model in Equation (11.10) yields the following modal split results:

$$\text{Modal split by rail} = \frac{e^{0.567409}}{e^{-0.595789} + e^0 + e^{0.567409}} = 53.2\%$$

$$\text{Modal split by bus} = \frac{e^0}{e^{-0.595789} + e^0 + e^{0.567409}} = 30.2\%$$

$$\text{Modal split by ferry} = \frac{e^{-0.595789}}{e^{-0.595789} + e^0 + e^{0.567409}} = 16.6\%$$

Applying these results to the new rail line between Z1 and Z2 will determine the estimated patronage level: $(164,000 + 161,000) \cdot 53.2\% = 172,900$ daily trips.

11.5 Literature review and further reading (O-D estimation)

Origin-destination (O-D) matrices constitute an essential input for most transit planning and design procedure. The literature describes many methods of matrix estimation, but most studies refer to the generation of car-trip matrices based on data from traffic counts. This section will review methods of estimating transit-trip O-D matrices based on passenger counts or surveys.

Simon and Furth (1985) describe the method developed by Tsygaltzky in 1977 for creating a route O-D without O-D survey data. A good representation of transit demand can usually be achieved if the trip matrix is based on a survey, in which passengers are asked directly about their precise origin and destination; however, such surveys are expensive and time consuming; therefore, several methods deal with situations in which data from surveys are not available. Tsygaltzky's method uses only passenger on-off counts; the model's passenger flow from one stop to another along the route is given in terms taken from fluid mechanics. The matrix is generated using a recursive algorithm. Simon and Furth analyse results statistically from the implementation of Tsygaltzky's method and compare them to actual trip demand. They conclude that this method yields reasonable results in cases in which the route structure is not too complicated.

Ben-Akiva *et al.* (1985) review and test several techniques for generating route-level trip matrices. The common practice of expanding on-board survey results by means of total boardings is compared with expanding results by iterative proportional fitting, constrained generalized least squares, and constrained maximum-likelihood methods. An intervening-opportunity model, which does not use the on-board survey, is also tested. The more complex methods achieve better accuracy and reduced bias by combining the survey data with passenger counts. An empirical case study demonstrates that under the assumption of error-free ride-check data, the proportional fitting technique is preferred because of its computational ease without any loss of accuracy. Ben-Akiva (1987), in a different study, discusses methods of combining transit data from different sources of partial information to create a transit-trip matrix. The same

matrix-estimation techniques described by Ben-Akiva *et al.* (1985) are examined. The use of alternative data sources, such as ticket-sales information, is illustrated. Aspects concerning the level of accuracy of combined data, biases and computational difficulties are discussed.

Nguyen *et al.* (1988) propose a method for the dynamic estimation of transit-trip matrices, using passenger counts. The matrix period is divided into several time-intervals, and formulations that connect the link flow at a given time to link flows over time segments are developed. The paper also presents an assignment model that is used for the estimation. The optimization is based on maximum entropy and maximum likelihood approaches.

Furth and Navick (1992) analyse the relationship between Tsygalnitsky's recursive method, which is commonly used for creating trip matrices from counts, and the bioproportional method, which combines data from counts and from an O-D survey. In the bioproportional method, the survey data are used to create an initial seed matrix, which is then adjusted iteratively by balancing row and columns to match on-off totals. The authors show that the Tsygalnitsky method is, in fact, a special case of the bioproportional method. As a seed matrix, it implicitly uses a null matrix, which contains information about trip directionality and minimum length. The paper illustrates why the recursive method is inappropriate when there is significant competition among routes. In addition, the authors offer a correction for cases in which the on-off data have been aggregated to the segment level.

Navick and Furth (1994) discuss possible sources of the seed matrix, which is often used as an initial basis for adjustments, and propose a model for generating this matrix. According to this model, the number of trips in each matrix cell is influenced by two factors: popularity, which is a feature of the destination zone and is not explained by models, and propensity, which is a feature of each O-D pair and depends on the distance between them. The researchers develop a propensity function that is the product of a power term and an exponential term and is equivalent to a gamma distribution. The power-function exponent is estimated by maximum likelihood, based on survey data. The gamma seed combined with the bioproportional method to match O-D totals is shown to be effective in generating O-D matrices.

Gur and Ben-Shabat (1997) develop another model that uses passenger counts for building or improving trip matrices. Their model is formulated as a nonlinear programming problem; for each pair i - j of stations, the model calculates the probability that a passenger who boards at i will alight at j . The problem takes advantage of information that is embedded in counts of individual vehicles and in the differences between different counts. The least-squares method is used for adjusting boarding totals to counts, and a minimum information model (Fratar) then generates a detailed trip table.

Wong and Tong (1998, 2003) propose a methodology for creating transit-trip matrices from passenger counts, using a maximum entropy programming formulation. The matrix-generation process depends on more detailed time data than do most previous methods; it uses a schedule-based dynamic assignment model to determine least-cost paths between all O-D pairs and clock-arrival times. Cases with and without a given seed matrix are examined. The authors also present a solution algorithm in which the matrix reduces computer-memory requirements in an efficient way.

Friedrich *et al.* (2000) present a technique for the continuous updating of demand matrices that is based on fuzzy set theory. According to this theory, counts are treated as intervals with lower and upper bounds, not as precise values. Matrix values are estimated such that all totals fall within the proper bandwidths. The method, formulated as a programming problem, may be implemented for estimating either car or transit-passenger matrices.

Nuzzolo and Crisalli (2001) and Nuzzolo *et al.* (2003) develop another method for estimating time-varying matrices from time-varying counts. The reference period is divided into several sub-periods, and detailed timetable information is considered. A dynamic schedule-based model of a stochastic type is used to predict passengers' path choices. The method is formulated as a maximization problem without constraints. The match between matrix and counts is achieved by minimizing generalized least squares.

The main characteristics of the methods reviewed above are summarized in Table 11.8.

Table 11.8 Summary of characteristics of the O-D methods reviewed.

Source	Method requires seed matrix?	Mathematical approaches used	Time segmentation (static/dynamic model)
Simon and Furth (1985) (method by Tsygalnitzky)	No		Static
Ben-Akiva <i>et al.</i> (1985)	Yes	Iterative proportional fitting, constrained generalized least-squares, constrained maximum likelihood	Static
Ben-Akiva (1987)	Yes		Static
Nguyen <i>et al.</i> (1988)	No	Maximum entropy, maximum likelihood	Dynamic
Furth and Navick (1992)	Methods that do/don't require a seed matrix are compared	Bioproportional method	Static
Navick and Furth (1994)	A method is presented for creating a seed matrix	Bioproportional method, maximum entropy	Static
Gur and Ben-Shabat (1997)	Optional	Least squares, minimum information model (Fratar)	Static
Wong and Tong (1998, 2003)	Cases with/without a seed matrix are examined	Maximum entropy	Dynamic
Friedrich <i>et al.</i> (2000)	Yes	Fuzzy set theory	Static
Nuzzolo and Crisalli (2001), Nuzzolo <i>et al.</i> (2003)	No	Generalized least squares	Dynamic

Exercises

- 11.1 Assuming that a rail company has the following linear demand relationship: $p = 20 - 0.04q$, where p is the route fare and q is the number of hourly sold tickets. Find the total revenue associated with each (p, q) pair and indicate when its maximum value is reached. In addition, specify the p - q zones under which the point elasticity is elastic and inelastic.
- 11.2 Given the following power demand function: $N = p^{-0.3} \cdot t^{-0.3} \cdot a^{0.2} \cdot c^{-0.3}$, where N is the number of transit trips, p is the fare (in dollars), t is the travel time (in hours), a is the cost of automobile trip (in dollars), and c is the average income (in dollars). (a) Given 20,000 passengers per hour who currently use the transit system at a flat fare of \$1.20 per trip, what will be the change in N for a flat fare of \$0.80 and what will be the transit agency gain? (b) Given the cost of a car trip (including parking fee) of \$4.00, what will be the change in N if the parking charges increases by \$0.60?
- 11.3 Given a calibrated utility function, $u = b - 0.04C - 0.02t$, where C is the cost of travel (in cents), t is the travel time (in minutes), and the MNL model is as described in the chapter. (a) What will be the modal split (percentage demand travelling by each mode) for the data given below? (b) Given rising gasoline price so that C for a car increases by \$1.20, what impact will this change have on the modal split?

Mode	b	C	t
Bus	-0.30	85	30
Light rail	-0.35	100	50
Car	-0.25	110	35

References

- Balcombe, R., Mackett, R., Paulley, N., Preston, J., Shires, J., Titheridge, H., Wardman, M. and White, P. (2004). *The Demand for Public Transport: A Practical Guide*. TRL Report, TRL593, TRL Limited.
- Ben-Akiva, M. (1987). Methods to combine different data sources and estimate origin-destination matrices. In *Transportation and Traffic Theory* (N. H. Gartner and N. H. M. Wilson, eds), pp. 459–481, Elsevier Ltd.
- Ben-Akiva, M., Macke, P. P. and Hsu, P. S. (1985). Alternative methods to estimate route-level trip tables and expand on-board surveys. *Transportation Research Record*, **1037**, 1–11.
- Boyce, D. E. and Daskin, M. S. (1997). Urban transportation. In *Design and Operation of Civil and Environmental Engineering Systems* (C. ReVelle and A. E. McGarity, eds), pp. 277–341, John Wiley & Sons.
- Ceder, A. (2004). New urban public transportation systems: Initiatives, effectiveness and challenges. *ASCE Journal of Urban Planning and Development*, **130**(1), 56–65.

- Ceder, A. (2006). Planning and policy of ferry passenger service in Honk Kong. *Transportation Journal*, **33**, 133–152.
- Friedrich, M., Mott, P. and Noekel, K. (2000). Keeping passenger surveys up to date: A fuzzy approach. *Transportation Research Record*, **1735**, 35–42.
- Furth, P. G. and Navick, D. S. (1992). Bus route O-D matrix generation: Relationships between bioproportional and recursive methods. *Transportation Research Record*, **1338**, 14–21.
- Gur, Y. J. and Ben-Shabat, E. (1997). Estimating bus boarding matrix using boarding counts in individual vehicles. *Transportation Research Record*, **1607**, 81–86.
- Hensher, D. (2001). Modal diversion. In *Handbooks of Transport Systems and Traffic Control*. (D. Hensher and K. Button, eds), pp. 107–123, Elsevier Ltd.
- Kottenhoff, K. (1998). Passenger train design for increased competitiveness. *Transportation Research Record*, **1623**, 144–151.
- Kottenhoff, K. and Lindh, C. (1996). The value and effects of introducing high standard train and bus concepts in Blekinge, Sweden. *Transport Policy*, **2**(4), 235–241.
- Litman, T. (2004). Transit price elasticities and cross-elasticities. *Journal of Public Transportation*, **7**(2), 37–58.
- Navick, D. S. and Furth, P. G. (1994). Distance-based model for estimating a bus route origin-destination matrix. *Transportation Research Record*, **1433**, 16–23.
- Nguyen, S., Morello, E. and Pallottino, S. (1988). Discrete time dynamic estimation model for passenger origin/destination matrices on transit networks. *Transportation Research*, **22B**, 251–260.
- Nuzzolo, A. and Crisalli, U. (2001). Estimation of transit origin/destination matrices from traffic counts using a schedule-based approach. In *Proceedings of the AET European Transport Conference Held at Homerton College*, PTRC. Cambridge, UK.
- Nuzzolo, A., Russo, F. and Crisalli, U. (2003). *Transit Network Modelling – The Schedule-Based Dynamic Approach*. FrancoAngeli, Italy.
- Ortuzar, J. de D. and Willumsen, L. G. (2001). *Modelling Transport*. 3rd Edition, John Wiley & Sons.
- Simon, J. and Furth, P. G. (1985). Generating a bus route O-D matrix from on-off data. *Journal of Transportation Engineering*, **111**, 583–593.
- Taplin, M. (1997). A world of trams and urban transit. *Light Rail and Modern Tramway*, **60**(718), 1–8.
- TranSystems Corp., Planner Coll., Inc., and Crikelair, T. Assoc. (2006). Elements needed to create high ridership transit systems: Interim guidebook. *TCRP Report 32*, Transportation Research Board, Washington, DC.
- Wong, S. C. and Tong, C. O. (1998). Estimation of time-dependent origin-destination matrices for transit networks. *Transportation Research*, **32B**, 35–48.
- Wong, S. C. and Tong, C. O. (2003). The estimation of origin-destination matrices in transit networks. In *Advanced Modeling for Transit Operations and Service Planning* (W. H. K. Lam and M. G. H. Bell, eds), pp. 287–315. Elsevier Ltd.
- Zhou, J., Lam, W. H. K. and Heydecker, B. G. (2005). The generalized Nash equilibrium model for an oligopolistic transit market with elastic demand. *Transportation Research*, **39B**, 519–544.

12 Route Choice and Assignment



Chapter 12 Route Choice and Assignment

Chapter outline

- 12.1 Introduction
- 12.2 Route choice using waiting-time strategy
- 12.3 Proportion of passengers boarding each route
- 12.4 Proportions derived for regular vehicle arrivals
- 12.5 Passenger assignment based on route choice
- 12.6 Literature review and further reading

Exercise

References

Practitioner's Corner

Once we have an idea of how many people want to travel between two points, by time of day and transit mode, the next phase is to determine the routes (direct or via transfers) they should take to reach their destination. This determination relies on passenger behaviour in regard to choosing a route if there are alternatives. What is the typical (average) choice strategy if there is a slower or less direct route, but one that does not involve long waiting times? This and other route-choice issues are discussed in this chapter in order to prepare the ground for assigning demand on the transit network. Utilizing an assignment procedure will allow for assessing/predicting changes in transit-service design at the network level.

The following riddle may indirectly indicate the complexity of the analysis exhibited in this chapter. A passenger arrives at an intersection with two bus stops in opposite directions of a desired route. He doesn't know which direction to take, since there is no indication of any destination. The passenger meets two people who know the required direction, but one tells only the truth and the other always lies (both these people, who are friends, know that). The passenger has no idea who is telling the truth and who is lying. How can the passenger, by receiving an answer from one of these two people to only one question, obtain the information required? (See the end of this Practitioner's Corner for the answer, but try first to think of the answer.)

The chapter consists of five main parts, following an introductory section. Section 12.2, using probabilities, analyses the alternatives facing a transit passenger at a stop who is deciding whether to board an arrived vehicle or to wait for a later vehicle that will have a shorter in-vehicle time. This route-choice dilemma is viewed as a decision between two categories of routes, slow and fast, in Section 12.3, from which the proportion of each category is derived. Section 12.4 crystallizes the proportions between the two categories for the case of regular vehicle arrivals. Section 12.5 presents some of the transit-assignment features that are related to route-choice modelling; emphasis is placed on estimating the accumulated demand on each segment of the network. The last part, Section

12.6, provides a literature review on transit-assignment studies, with guidance for further reading.

Practitioners may find this chapter too technical, especially the mathematical handling of probability formulas. However, it is recommended that the introduction to each section be read in order to capture the essence of the subject. For instance, knowing the rationale behind passengers' route choices may help overcome planning and operational problems. One humorous example concerns passenger behaviour in situations characterized by a high inflation rate; then it will be better to ride a taxi than a bus because in the bus you pay upfront and in the taxi at the end of the trip.

As for the question to be put to one of the two friends (it doesn't matter which one): "If I'll tell your friend that this direction (points to a direction) is the correct one, what will be his answer?" The passenger will then choose the opposite direction of the answer given.

12.1 Introduction

In constructing or revising a transit network of routes, it is not sufficient to know only the demand, its elasticity and factors affected it. What is needed in addition is a comprehension of and insight into passenger behaviour when traffic flows and disperses along the network of routes. What, if any, are the passenger paths (use of more than one route with transfers) through the network? What factors influence a passenger's choice of routes? The answers to these questions represent route-choice behaviour (modelling), which is then used in assigning passenger demand for prediction purposes (of traffic volume, capacity analysis, fleet size required, design and control elements needed, and more). This chapter will review the subject of passenger route choice and assignment, placing some emphasis on passenger waiting-time strategies.

One of the most crucial characteristics of a transit network is the existence of overlapping routes that share the same transit stops while running on common segments. Emanating from this characteristic is the fact that more than one transit route can serve the demand between a certain O-D pair. We will focus in this chapter on the decision-making process of the individual passenger who has to select the most efficient route serving a given stop. This process is one of the main obstacles in developing transit-assignment procedures.

An additional complication of transit-assignment methods in real-world problems relates to the possibility that several transit modes (bus, train, light-rail, metro, etc.) and a walking mode exist, each with its own characteristics. Moreover, each mode is perceived differently by passengers. To simplify the presentation of transit-assignment approaches, this chapter includes the simple case of a single transit mode. The analysis follows Marguier and Ceder (1984) and Israeli and Ceder (1996). The literature review of this theme appears in Section 12.6. Summaries of the main contributions of this (transit-assignment) theme can be found in Bell (2003), Bell and Schmöcker (2004), Nuzzolo (2003a, 2003b), and Nuzzolo and Crisalli (2004).

12.2 Route choice using waiting-time strategy

The decision problem facing a transit passenger at a stop with alternative choices is whether to board an arrived vehicle or to wait for a later vehicle that will have less in-vehicle time.

This issue reflects real choice situations in which the passenger can distinguish between two categories of overlapping routes: slow and fast. The objective is to minimize total travel time (waiting time and in-vehicle time), which would result from adopting an optimal strategy in the choice process. This action depends on certain parameters: time between each O-D pair, vehicle regularity, distribution of passenger arrivals at the stop, and the structure of possible paths constituting the network.

The analysis shown in this section deals with the problem of passenger waiting strategies by using mathematical formulations in a probabilistic fashion in order to derive the proportion of passengers for each category (slow and fast). This calculation depends on the evaluation of a probabilistic function that considers such affecting parameters as vehicle-frequency share, in-vehicle time difference, passenger arrivals, and vehicle-headway distribution. In addition, theoretical evaluations are presented in order to achieve correct implementation of: (a) mean waiting-time assumptions, and (b) the intuitive rule circumstances (according to which, passengers select routes in proportion to their vehicle frequencies). The outcome of the analysis can be integrated into transit-assignment models to create realistic situations.

A practical passenger-assignment procedure on transit networks consists of the following characteristics: (1) the exact structure of the network between each origin and destination (direct routes, parallel routes, sequential routes and transfer paths); (2) in-vehicle time (direct routes and transfer-path times); (3) passenger waiting time at a stop; (4) passenger waiting strategies at a stop served by overlapping routes; (5) the 'circular problem' (because frequencies are not always known at this stage): passenger flows depend on route frequencies and, conversely, the frequencies rely on the load profiles of the routes, which are the outcome of the flows; (6) network size (treating each O-D pair separately and, subsequently, considering the entire network simultaneously can help to combat computational complexities of large networks); (7) failure-to-board probabilities. This section treats the second, third and fourth characteristics, and their relationship to the first, fifth and sixth; the seventh characteristic, as well as all the others, are discussed in the literature review in Section 12.6.

12.2.1 *Passenger waiting-time dilemma*

The objective of this section is to study the problem of route choice encountered by a passenger who is able to select one of several routes (alternatives) in order to travel from stop A to stop B. The common characteristic of all the transit routes considered is that they stop at both points A and B. The central idea is that a passenger wishing to travel from A to B may disregard some of the vehicles arriving at A because of their association with relatively long travel times.

The analysis, following Marguier and Ceder (1984), is performed in a probabilistic context, in which the passengers, based on their experience, have a fairly good sense of the characteristics of each route. That is, they have some information about the headway distribution (route frequency) and expected in-vehicle time from A to B. The passengers are also influenced in the route-selection process by the amount of time they have already waited since their arrival at the transit stop, which can be considered additional information available to the passengers. Total travel times, consisting of waiting time and in-vehicle travel time for each route, are compared in the analysis. The waiting time, with its formulation shown below and in Chapter 17, clearly plays an important role in total travel time because of its

relationship to and dependency on vehicle-interval reliability and the passenger-arrival process at stop A. Before proceeding with the main thrust of the analysis, let us introduce some related formulations and assumptions.

12.2.2 Basic relationships

Chapter 17 (on transit reliability) will show how to arrive at the following, commonly used formula for mean passenger waiting time under the assumption of random passenger arrivals:

$$E(w) = \frac{E(H)}{2E(H^2)} = \frac{E(H)}{2} \left[1 + \frac{\text{Var}H}{E^2(H)} \right] \quad (12.1)$$

where w is the waiting time; $E(w)$ is the mean waiting time; $E(H)$ and $\text{Var}H$ are, respectively, the mean and variance of the time headway H between vehicles.

Underlying this relationship is the following assumption: the average waiting time of passengers arriving at an interval of length t is $\frac{1}{2}t$, and the average number of passengers arriving during such an interval is proportional to t . In addition, from data on off-peak bus intervals, the following relationship was found:

$$\text{Var}H = \frac{AE^2(H)}{A + E^2(H)} \quad (12.2)$$

where A is a constant (with time square dimension) between 0 and infinity.

$A = 0$ corresponds to the deterministic headway case (regular vehicle arrivals). $A \rightarrow \infty$ corresponds to the completely random case (Poisson vehicle arrivals; i.e. exponential headways). Actual values found for A are between 15 and 35.

The transit-vehicle interval irregularity is actually characterized by the headway-distribution coefficient of variation C , which is defined as the ratio of the headway standard deviation to the headway mean; i.e. its square is

$$C^2 = \frac{\text{Var}H}{E(H)^2} \quad (12.3)$$

For transit-vehicle (especially buses) headway distributions, C^2 ranges between 0 and 1, where $C^2 = 0$ corresponds to perfectly regular vehicle arrivals (deterministic headway) and $C^2 = 1$ to the completely random case (exponential headway).

Combining Equation (12.2) and the intrinsic definition of C^2 given by Equation (12.3) yields:

$$C^2 = \frac{A}{A + E(H)^2} \quad (12.4)$$

It may be noted that the parameter A lies between 0 (for which $C^2 = 0$) and infinity ($C^2 \rightarrow 1$). Using, respectively, the notations $f_H(t)$ and $F_H(t)$ for the probability density function

and the cumulative distribution function of the headway H , defining $\bar{F}_H(t) = 1 - F_H(t)$, and denoting the vehicle frequency as $F = 1/E(H)$ yields the following:

$$(i) \text{ for } A = 0, \quad H \text{ (deterministic)} = \frac{1}{F} \quad \text{and} \quad F_H(t) = \begin{cases} 1, & t \leq \frac{1}{F} \\ 0, & t \geq \frac{1}{F} \end{cases};$$

$$(ii) \text{ for } A = 1, \quad f_H(t) = F \cdot e^{-Ft} \quad \text{and} \quad \bar{F}_H(t) = e^{-Ft}.$$

Equation (12.4) may also be rewritten as $A = C^2/F^2(1 - C^2)$, which shows how, for a given frequency F , the parameter A and the coefficient of variation are uniquely related.

The idea, then, is that for a given A (or C^2 when using the relationship above), the headway distribution belongs to a family of functions that can approach the two extremes, the deterministic and the exponential cases, whereas Equations (12.2) and (12.4) are imposed. Two such families are considered here. The first one, developed by Marguier and Ceder (1984) and referred to as 'power' distributions, verifies the following relationships

$$f_H(t) = \begin{cases} \frac{2F \cdot C^2}{1 + C^2} \left(1 - \frac{1 - C^2}{1 + C^2} \cdot F \cdot t \right)^{\frac{3C^2 - 1}{1 - C^2}}, & t \leq \frac{1 + C^2}{1 - C^2} \cdot \frac{1}{F} \\ 0, & t \geq \frac{1 + C^2}{1 - C^2} \cdot \frac{1}{F} \end{cases} \quad (12.5)$$

$$\bar{F}_H(t) = \begin{cases} \left(1 - \frac{1 - C^2}{1 + C^2} \cdot F \cdot t \right)^{\frac{2C^2}{1 - C^2}}, & t \leq \frac{1 + C^2}{1 - C^2} \cdot \frac{1}{F} \\ 0, & t \geq \frac{1 + C^2}{1 - C^2} \cdot \frac{1}{F} \end{cases}$$

Underlying the definition of this family of distributions is the fact that for such distributions at the limit for $A = 1$ (see the relationship between A and C^2 above), the distributions approach the exponential case in a fashion corresponding to the way the binomial processes approach the Poisson process.

The second family is made up of gamma distributions and verifies the following relationship:

$$f_H(t) = \frac{\left(\frac{F}{C^2} \right)^{\frac{1}{C^2}}}{\Gamma\left(\frac{1}{C^2}\right)} \cdot t^{\frac{1 - C^2}{C^2}} \cdot e^{-\frac{Ft}{C^2}} \quad (12.6)$$

where Γ is the gamma function; i.e. $\Gamma(C^{-2}) = \int_{x=0}^{\infty} x^{C^{-2}-1} \cdot e^{-x} \, dx$.

The idea of selecting gamma distributions is that the exponential is a particular case of gamma and the generic gamma distribution has two independent parameters related to the mean

and variance, which allows the constraint given by Equation (12.2) to be met. equations (12.5) and (12.6) are used here in the route-choice problem in order to illustrate the generality of the results.

Certainly, both families give identical curves for $C^2 = 0$ and $C^2 = 1$, since these values correspond to the deterministic and the exponential cases, respectively. Real-life data show headway histograms with increasing irregularity (C^2) along a transit route; the histograms have a maximum point, as observed for the gamma situation, and an intersection with the ordinate axis (different from the origin), as observed for the 'power' situation distributions. Therefore, realistic situations fall somewhere between the 'power' and the gamma situations.

The general assumptions of this analysis are the following:

1. The (between days) vehicle-arrival variability is high enough so that passengers cannot identify any minimum wait in order to time their arrival at the transit stop. This assumption is particularly appropriate for urban routes without published timetables in which the values of A and C^2 , or F , are relatively large. This assumption can be relaxed in accordance with the study by Jolliffe and Hutchinson (1975).
2. The proportion of passengers arriving coincidentally with a vehicle ('see and rush') is fixed and independent of other parameters. Observations in urban areas (in London and Paris) found this proportion to be up to 16% (Marguier and Ceder, 1984). Together with assumption (1), we assume then that a major proportion (e.g. 84%–100%) of the passengers arrives randomly.

Assumptions related to the route-choice problem are the following:

1. Passengers are assumed to know the routes and to have a general knowledge, from experience, of the headways and waiting time. Thus, they select a strategy of choice among routes by attempting to minimize the sum of expected waiting and in-vehicle times. This assumption could be extended by considering, in a similar fashion to Jolliffe and Hutchinson's (1975) methodology, that only a certain proportion of the passengers follows the principle of minimization of total waiting and in-vehicle times, whereas the others will take the first bus that arrives and then continue on to point B.
2. At point A, the routes are statistically independent of each other; this is true if the routes do not share another common section, upstream of point A.

12.3 Proportion of passengers boarding each route

The route-choice dilemma can be viewed as a decision between two categories of routes. When there are more than two routes to travel from point A to point B, it is not irrational to assume that passengers will not consider each route individually, but rather will tend to group them into categories. Two likely categories are fast routes and slower ones. Passengers will then not choose between routes, but only between categories, and thus take the first vehicle to arrive in the category selected. The routes belonging to one category can then be viewed as forming one 'equivalent' route. For simplicity's sake, refer below only to the faster route, route 1, and to the slower route, route 2. The following analysis is suitable for individual routes 1 and 2, but may be extended for the concept of two-route categories.

Using the above formulations, the estimated proportion of passengers that travel from points A to B by route 1, and the complementary proportion that uses route 2 can be deduced. This proportion represents the quantity of interest in the travel-assignment process for different transit routes. The estimated proportion depends on two factors: the frequency of each route, the headway coefficients of variation. The latter can be replaced by Equation (12.4).

Following Marguier and Ceder (1984), use of the optimal strategy will result in route 1 users being either those who board the first vehicle, which is a route 1 vehicle, or those who allow route 2 vehicles to pass and wait for the route 1 vehicle. Conversely, route 2 users are those for whom the first arriving vehicle is from route 2 and board that vehicle. With the subscripts 1 and 2 used for routes 1 and 2, respectively, for all quantities, the latter are those for which

$$w_2 \leq w_1$$

$$\text{and} \quad \tau_2 - \tau_1 \leq RW_1(w_2), \quad (12.7)$$

where w is the waiting time, τ is the in-vehicle time between points A and B, and $RW(t^*)$ is the expected time that remains before the arrival of a route 1 vehicle, given that the passenger has already waited time t^* .

The first inequality in (12.7) corresponds to the fact that a route 2 vehicle arrives first (the time W_2 for the first vehicle on route 2 to arrive is smaller than the time W_1 for route 1). The second inequality corresponds to the criterion for boarding the route 2 vehicle instead of waiting for a faster vehicle (route 1). The proportion P_2 of route 2 passengers can thus be written as the probability that both inequalities in (12.7) exist:

$$P_2 = \text{prob} [w_2 \leq w_1 \quad \text{and} \quad \tau_2 - \tau_1 \leq RW_1(w_2)]. \quad (12.8)$$

Equation (12.8) can be expressed in a function of the distributions of the waiting times W_1 and W_2 (see Marguier and Ceder, 1984) as

$$P_2 = \begin{cases} 0, & \tau_2 - \tau_1 > \frac{1 + C^2}{2F_1} \\ \int_0^{RW_1^{-1}(\tau_2 - \tau_1)} \bar{F}_{w_1}(t) f_{w_2}(t) dt, & rw_1 < \tau_2 - \tau_1 \leq \frac{1 + C^2}{2F_1} \\ \int_0^\infty \bar{F}_{w_1}(t) f_{w_2}(t) dt, & \tau_2 - \tau_1 \leq rw_1 \end{cases} \quad (12.9)$$

where RW_1^{-1} is the reciprocal of the RW_1 function (RW_1 can be reciprocated, since it decreases monotonically); and rw_1 is the asymptotical value of the function RW_1 . The value of rw_1 depends on the type of distribution on C_1^2 and on F ; for example, $rw_1 = 0$ for the 'power' distribution family and $rw_1 = C_1^2/F_1$ for the gamma distribution family. In the limit case of *exponential headways*, $C_1^2 = 1$, $RW_1(t) = 1/F_1 = rw_1$, and

$$\int_0^\infty \bar{F}_{w_1}(t) \cdot f_{w_2}(t) dt = \int_0^\infty e^{-F_1 t} \cdot F_2 \cdot e^{-F_2 t} dt = \frac{F_2}{F_1 + F_2}$$

Equation (11.20) in this case is reduced to

$$P_2 = \begin{cases} 0, & \tau_2 - \tau_1 > \frac{1}{F_1} \\ \frac{F_2}{F_1 + F_2}, & \tau_2 - \tau_1 \leq \frac{1}{F_1} \end{cases} \quad (12.10)$$

If in this exponential case $\tau_2 - \tau_1$ is large, the proportion of route 2 users is 0. Otherwise, the share of each route's users is equal to its frequency share.

The proportion P_2 certainly does not depend on the time scale (i.e. the time unit selected). Therefore, the quantities F_1 , F_2 and $\tau_2 - \tau_1$ can be grouped within the following two variables: (a) the route 2 frequency share: $FS_2 = F_2 / F_1 + F_2$ and (b) the ratio of the in-vehicle time difference to the headway of route 1: $TH_1 = \tau_2 - \tau_1 / H_1$. The proportion P_2 is plotted against FS_2 and TH_1 in Figures 12.1 and 12.2, respectively, for each type of headway distribution ('power' or gamma) and several values of C_1^2 and C_2^2 . The curves in Figure 12.1 are based on $TH_1 = 0.54 = (31/57)$, and in Figure 12.2 on $FS_2 = 1/2$ (equal frequency share).

Figure 12.1 represents the market share-frequency share curves, in which P_2 increases with FS_2 . In Figure 12.1 for $C_2^2 = 1/3$, it should be noted that the proportion P_2 for $C_1^2 = 0$ remains fixed,

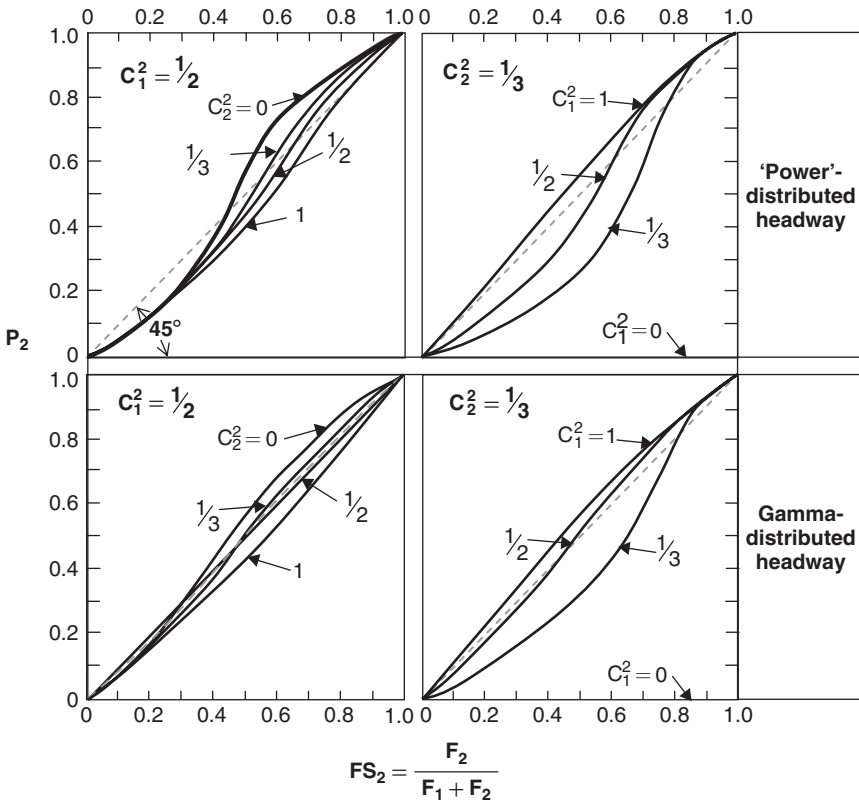


Figure 12.1 Proportion of passengers boarding the slow route 2 (complementary to the fast route 1 proportion) as a function of the slow route 2 frequency share of $(\tau_2 - \tau_1) / H_1 = 0.54$

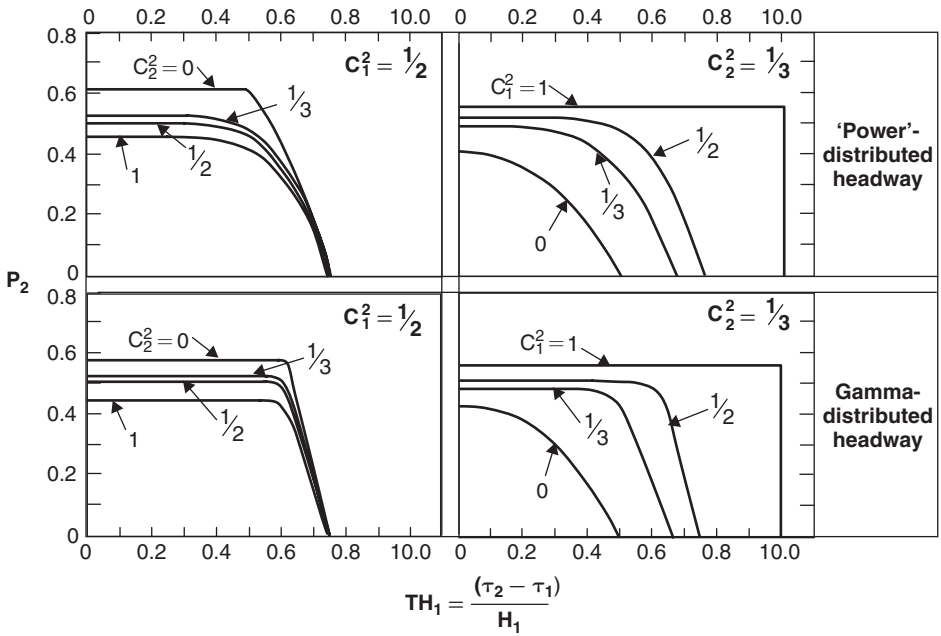


Figure 12.2 Proportion of passengers boarding the slow route 2 (complementary to the fast route 1 proportion) as a function of $(\tau_2 - \tau_1)/H_1 = 0.54$ for $FS_2 = 0.5$

equal to 0. This can be seen from Equation (12.9) because $TH_1 = 0.54 > 0.5$ for $C_1^2 = 0$. The curves in Figure 12.2 show that P_2 is either approximately or exactly constant with respect to small values of $\tau_2 - \tau_1$. For higher values, however, the value of P_2 drops rapidly and reaches 0. This drop signals the possibility of a threshold for the passengers' route-choice decision.

All in all, this analysis shows that the general intuitive rule $P_2 = FS_2$ (i.e. passengers board the routes proportionally to their frequencies) does not provide a good approximation in all cases. In addition, Marguier and Ceder (1984) present a revised formulation of Equation (12.9) for a case in which the proportion of passengers whose arrival at the transit stop coincides with the arrival of the first vehicle is not zero. The next section constructs a full description of the proportion boarding each route, but only for regular vehicle arrivals.

12.4 Proportions derived for regular vehicle arrivals

This section follows Israeli and Ceder (1996). Based on the random arrival of passengers and regular transit-vehicle arrivals (deterministic headways), the waiting time at stops associated with $i = 1, 2$ (referring to routes 1 and 2) is uniformly distributed:

$$f_{w_i}(t) = \begin{cases} F_i, & 0 \leq t \leq \frac{1}{F_i}, \\ 0, & t \geq \frac{1}{F_i} \end{cases}, \quad i = 1, 2 \quad (12.11)$$

The relevant component of interest is the distribution of the difference between the waiting times of routes 1 and 2, $w_1 - w_2$. Israeli and Ceder (1996) showed that this distribution could be simplified in the following uniform manner:

$$f_{w_1-w_2}(t) = \begin{cases} 0, & t \leq -\frac{1}{F_2} \\ \frac{F_1 F_2}{F_1 + F_2}, & \frac{1}{F_2} \leq t \leq \frac{1}{F_1} \\ 0, & t \geq \frac{1}{F_1} \end{cases} \quad (12.12)$$

with

$$E(w_1 - w_2) = \frac{1}{2} \left(\frac{1}{F_1} - \frac{1}{F_2} \right) \quad (12.13)$$

The cumulative distribution (choice probability), therefore, is given by

$$\text{Prob}(w_1 - w_2 \leq t) = F_{w_1-w_2}(t) = \int_{-\infty}^t f_{w_1-w_2}(t) dt \quad (12.14)$$

to obtain:

$$F_{w_1-w_2}(t) = \begin{cases} 0, & t \leq -\frac{1}{F_2} \\ \frac{F_1}{F_1 + F_2} (1 + F_2 \cdot t), & -\frac{1}{F_2} \leq t \leq \frac{1}{F_1} \\ 0, & t \geq \frac{1}{F_1} \end{cases} \quad (12.15)$$

The complementary cumulative distribution $F_{w_2-w_1}(t)$ can be obtained by $1 - F_{w_1-w_2}(t)$.

The probability (proportion) P_1 of selecting route 1 (fast) vehicle can be obtained by assigning $t = \tau_2 - \tau_1$ to Equation (12.15). Because $\tau_2 > \tau_1$, then $t > 0$ and Equation (12.15) becomes

$$P_1 = F_{w_1-w_2}(t) = \begin{cases} \frac{F_1}{F_1 + F_2} (1 + F_2 \cdot t), & 0 \leq t \leq \frac{1}{F_1} \\ 1, & t \geq \frac{1}{F_1} \end{cases}, \quad \begin{matrix} t = \tau_2 - \tau_1 \\ t \geq \frac{1}{F_1} \end{matrix} \quad (12.16)$$

Based on Equation (12.16), the complementary probability $P_2 = 1 - P_1$ can be derived with the following interesting property: P_2 can never bear the entire passenger demand. That is,

$$P_2 = F_{w_2 - w_1}(t) = \begin{cases} \frac{F_2}{F_1 + F_2} (1 - F_1 \cdot t), & 0 \leq t \leq \frac{1}{F_1} \\ 0, & t \geq \frac{1}{F_1} \end{cases}, \quad \begin{matrix} t = \tau_2 - \tau_1 \\ t \geq \frac{1}{F_1} \end{matrix} \quad (12.17)$$

Because $t = \tau_2 - \tau_1 \geq 0$ and $F_1, F_2 \geq 0$, that the result is that $P_2 \neq 1$.

Figure 12.3 presents P_1 as a function of the frequency proportion FS_1 , $FS_1 = F_1/F_1 + F_2$, where $F_1 + F_2 = 10$ vehicles/hour, and $\tau_2 - \tau_1$ is in hours. For those cases in which the passenger is unaware of the in-vehicle travel-time difference ($\tau_2 - \tau_1 = 0$), the proportion rule can be applied. Otherwise, Equations (12.16) and (12.17) apply. It should be noted that for $\tau_2 - \tau_1 = 0$ passengers select the first transit vehicle that arrives.

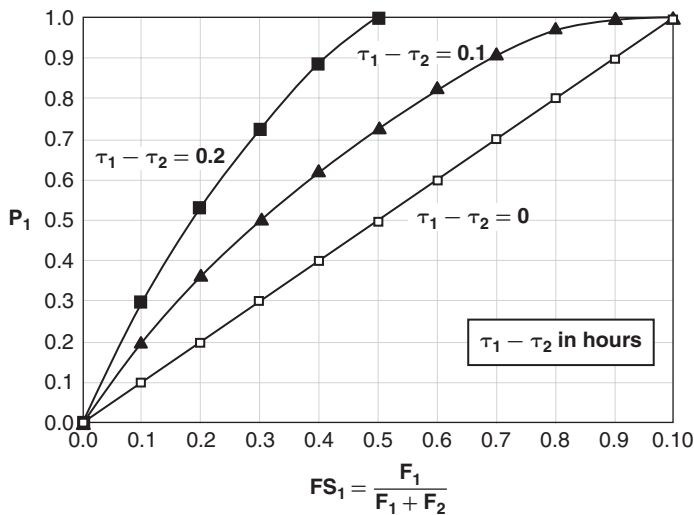


Figure 12.3 Proportion of passengers boarding route 1 (fast) vehicle as a function of the frequency share

12.5 Passenger assignment based on route choice

The beginning of this chapter indicated that vehicle and traveller assignment on a network of choices (paths or routes) is an essential part of estimating (predicting) the accumulated demand on each segment of the network. It is, therefore, an integral tool of transportation planning, in which route-choice modelling forms the base of any assignment algorithm.

In transit planning, the assignment algorithms for passengers warrant (at the planning stage) changes in routes, location of stops, selection of operational strategies, priority schemes for transit vehicles, traffic and parking arrangements, environmental impact studies and more. Section 12.6 provides a literature review of transit-assignment studies. The section presents some of the transit-assignment features that are related to route-choice modelling.

Passenger path may involve transfers and, hence, the use of more than one transit route. An arc in the transit network is a road segment that can have more than one crossing route, thus creating the possibility of overlapping transit services. Figure 12.4 illustrates the differences between a road network and a transit network. A passenger travelling from node A to node D on the transit network in Figure 12.4 has two possible paths: routes I and II with a transfer at B, or routes I and II with a transfer at C. For instance, in the network in Figure 12.4, we can examine routes A-C-D and A-B-C, in which route A-C-D replaces route B-C-D, thus making node C a single transfer point from B to D. This examination (as well as other sets of alternative routes) can use passenger-assignment procedures for given frequencies; the latter will be derived from the O-D demand. Based on given criteria or objectives, different routing recommendations can be made by the assignment algorithms. This simple description is further analysed by Guan *et al.* (2006) in combining transit assignment and route configuration.

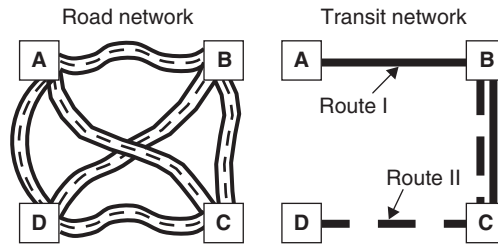


Figure 12.4 Road and transit network illustrations of four locations

The proportion of passengers boarding each transit route (Figures 12.1–12.3) can be integrated into an assignment model for an entire transit network. The use of such a procedure can provide realistic considerations to some extent of passenger behaviour in the assignment procedure. Note that the previous section assumes that the routes operate independently and do not share another common segment upstream of the transit stop with a waiting-time dilemma. Here this assumption is extended to all segments of the transit network.

Passenger assignment on the transit network is interpreted as intermitted flow of passengers: waiting to board or transferring at stops and riding on the vehicles. At node u , the total flow of arriving passengers who want to obtain service (boarding or transferring), and the frequency parameters of the service, will determine the total waiting time W_u at that node. The total expected waiting time, using Equation (12.1), can be formulated as

$$E(W_u) = \sum_r \sum_v p_{uv}^r \frac{E(H_r)}{2} \left[1 + \frac{\text{Var}H_r}{E^2(H_r)} \right] \quad (12.18)$$

where p_{uv}^r is the number of arriving passengers at u from node v who seek to board/transfer to a route (or group of routes) r ; and H_r is the headway (or combined headway for a group) of r . One objective (e.g. Spiess and Florian, 1989) can be to minimize the sum of the cost of passengers and their expected waiting times, subject to constraints for all nodes in the network. The constraints ensure conservation-of-flow conditions (what goes into a node equates with what comes out) and that the flow on each arc is greater than or equal to 0 and less than or equal to a function of the frequencies and waiting times on that arc.

Passenger-choice strategy was defined as being between fast and slow route categories but in a network context. As the proportion boarding a vehicle in each category is dependent on the frequency share and the time difference between the two categories, it is necessary to define a 'combined' frequency and 'equivalent' in-vehicle time for each category. The network structure contains different possible paths between each origin and destination; i.e. direct routes and transfer paths. The latter can consist of overlapping (parallel) routes, successive routes, or both. Although calculating the 'equivalent' in-vehicle time for each category is simple enough (passengers tend to make an average of all possible paths connecting the origin with the destination), the 'combined' frequency is more complicated, as will be explained. The complex network structure accounts for why researchers have tried to avoid this problem by presenting non-realistic simplifications, such as using only the frequencies of routes that connect directly to the transit stop or employing different frequencies for each arc of the network. The latter requires 'smoothing' techniques in order to find the correct frequency even for direct routes, a process that leads to unrealistic results.

At the outset of the process, as has been seen, passengers have to choose a category (slow or fast routes); then, the route within this category needs to be chosen. Those passengers who do not distinguish among route lengths (in-vehicle times) within each category board the first vehicle to arrive. That is, passengers are assigned to the routes in proportion to the routes' frequencies.

Each category, which refers to a sub-network between the corresponding origin and destination, is part of a 'modified' transit network: the nodes are transit stops with overlapping routes (i.e. enabling transfers) and the arcs are direct route segments between these nodes. The number of arcs in a possible path between an origin and a destination denotes the number of vehicle changes, minus one.

Figure 12.5 shows the 'combined' frequency calculation and, consequently, the demand assignment within a category, or the routes between i and j . The 'combined' frequency for parallel routes is their sum, and for successive routes it is their minimum. In such a manner, passengers between k and j in Figure 12.5, part (a), experience a 'combined' frequency of $3 + 1 = 4$ vehicles/hour, while their effective frequency along $i-k-j$ is $\min(6, 4) = 4$ vehicles/hour. The combined frequency of all categories between i and j is $6 + 4 = 10$ vehicles/hour; which basically is the minimum cut of the small network in part (a) (for the definition of minimum-cut frequency, see Section 7.4 in Chapter 7). This result (which can be either F_1 or F_2 , depending on the type of category) is then utilized in Equations (12.9), (12.16) and (12.17) in order to predict the proportion of passengers between i and j for the two categories.

Further analysis of the network shown in Figure 12.5(a) is illustrated in Figure 12.5(b) to (e) and called network synthesis. It is based on recursive calculations of the combined frequencies in a sub-network, until a single 'combined'-frequency arc is reached. Network synthesis is analogous to an analysis made for capacitor (electric) networks, with parallel and

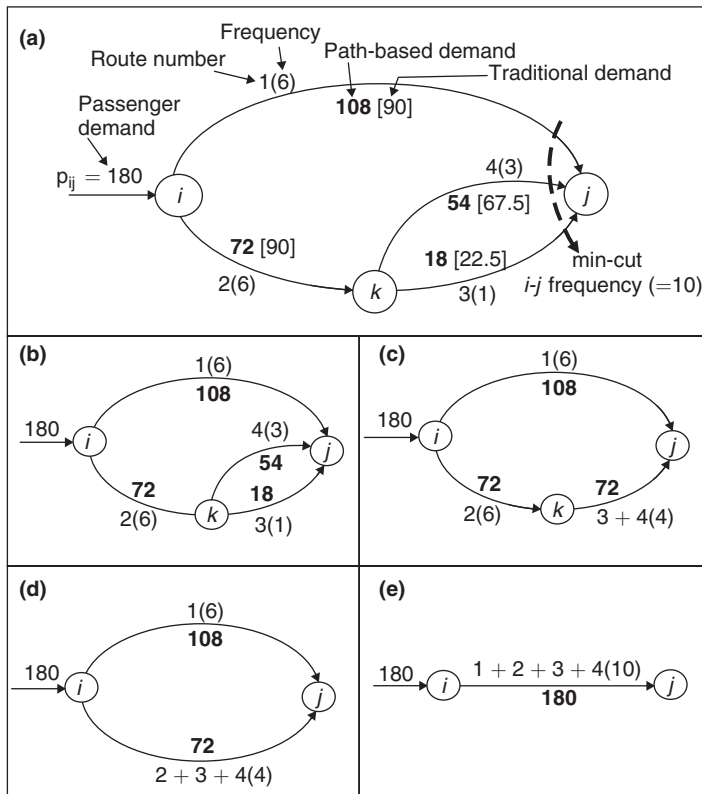


Figure 12.5 Passenger assignment on an hourly basis, in which part (a) shows that the traditional assignment is a path-based assignment; parts (b) to (e) describe in steps the network synthesis process

successive summations. A parallel summation of frequencies is equal to the sum of the frequency associated with the cut of the parallel arcs; a successive summation of frequencies equates with the minimum among the frequencies considered. The calculation of the divided demand can be viewed in a reversed process. The demand of $p_{ij} = 180$ passengers in Figure 12.5(e), which is a single-frequency arc, is divided (going backwards) in Figure 12.5(d) on the basis of the proportional frequency-share rule; e.g. see Equation (12.10): that is, $180 \cdot (6/10)$ and $180 \cdot (4/10)$. The latter demand of 72 is kept along i - k - j in Figure 11.9(c), and again (going backwards) is divided proportionally in part (b) to 54 and 18.

The separation into two route categories for each i - j pair (if passengers notice a time difference between them) creates two separate sets (sub-networks) for each route; this means that a route or path serving the demand between i and j (directly or via transfer) will be included in the sub-network of either the fast or slow category in this pair, but not in both. Thus, assignment equations can be formulated in such a manner that 'combined' frequencies and demand shares will be formulated separately for the two route categories, while the relation between variables in each category will be given in the frequency calculation constraints.

Considering the interrelationship of all O-D pairs in the network enables more manageable handling of large, complex networks. This interrelationship is affected by the mutual

dependence of the unknown frequencies and demand flows (serving as variables) in an assignment procedure. While actual frequencies are based on the methods of Chapter 3 (consisting, via ride and point-check counts, of all O-D demands for a route), the demand flows are usually based on the frequency share (with or without considering route travel times). More on transit-assignment studies and considerations appears below in Section 12.6.

12.6 Literature review and further reading

One of the vital ingredients in transit planning is the prediction of the paths that passengers choose for transit routes to take them from origin to destination. This prediction relies on the use of transit-assignment models that have appeared in the welter of professional papers on the subject in the past 40 years. Much progress was achieved in past research in adding more realistic features of passenger behaviour and operational planning elements to these models. Among the added features are waiting times at stops, transfer times between routes, preferred set of routes for an O-D pair from a passenger's perspective, preferred passengers' strategies, distributed travel and waiting times, crowding level on vehicles, and failure-to-board probabilities. A summary of some of the papers can be found in Bell (2003), Bell and Schmöcker (2004), Nuzzolo (2003a, 2003b), and Nuzzolo and Crisalli (2004). This section reviews briefly and chronologically the main contributions to transit-assignment modelling.

Early methods of transit-route choice and assignment are works by Dial (1967) and Le Clercq (1972), who used heuristic rules in considering both waiting and travel times when computing the shortest paths on a network. Passengers take the first vehicle to arrive, and a transfer penalty is equal to the expected waiting time. Chriqui and Robillard (1975) introduced, for the first time, a choice-behaviour feature by which passengers can select, from a set of alternative routes, a subset of routes from which they will board the first vehicle to arrive. This subset of routes, called 'attractive routes', represents routes that have a better chance to offer shorter travel times (based on prior knowledge). The researchers presented their idea with the use of a simple network and a single origin and destination.

Chriqui and Robillard's direction was extended in a doctoral study by Spiess in 1984. Nguyen and Pallotino (1988) introduced a graph interpretation for a strategy of selecting a set of attractive routes at a boarding point; this graph representation was denoted a *hyper-path*. Part of Spiess's original study was incorporated into a work that became known (Spiess and Florian, 1989). The latter developed a two-part algorithm to assign passenger demand from the user's perspective. Spiess and Florian assumed that passenger behaviour reflected a minimization of the expected value of access; i.e. of waiting and in-vehicle times or a weighted sum of these time elements. The first part of their algorithm computes the total expected time elements between origin and destination, including transfers; the second part assigns demand according to the strategy of choosing a set of attractive routes. Both Spiess and Florian (1989) and Nguyen and Pallotino (1988) considered the effect of the inconvenience of crowded vehicles through discomfort functions in their equilibrium-assignment models.

De Cea and Fernández (1993), inspired by Spiess and Florian, introduced a transit-equilibrium assignment model in which waiting times on access links depended on passenger flows; that is, they applied congestion functions to passengers requesting to board the first arrived vehicle. De Cea and Fernández (1993) incorporated heuristically the discomfort

effect of crowded vehicles and crowded stops. Their model was solved by Jacobi's method of using a similar diagonalization algorithm to that used by Florian (1977).

Wu *et al.* (1994), using the concept of hyperpaths (strategies) proposed by Nguyen and Pallotino (1988), introduced a network consisting of road-based and transit route-based arcs. Their passenger-related arcs included the time elements of walking, waiting, boarding, in-vehicle, transferring and alighting. Their assignment model considered that the time required to board a vehicle increased with flow (of transit vehicles); the distribution of flows across the set of attractive routes is proportional to the minimum frequency share.

The foregoing articles may be grouped under frequency-based models, in which within-day dynamics (in transit-vehicle headways) are not taken into account. In the mid-1990s, some route-choice and transit-assignment models started explicitly to consider different headways (timetable) as an input. These studies were grouped under schedule-based models. In most such studies, passenger-choice behaviour is interpreted by a *utility* function. That is, passengers are assumed to assign a utility value to each alternative route from a given choice set (of routes) and to select the one with maximum utility. Usually the utility values of the alternatives are treated as random variables, thus converting the choice process (among routes) into a search for maximum (utility) probability. Each utility function is dependent on a set of passenger attributes. Hickman and Bernstein (1997) used a sequential deterministic path-choice approach with a utility function consisting of stochastic travel-time attributes and passenger information. They studied high-frequency service, including congested networks, using equilibrium and dynamic processes.

Tong and Wong (1999) showed the differences between frequency-based and schedule-based approaches and constructed a stochastic schedule-based dynamic model using simulation. In their model, passengers either move on a walking segment, wait (queue) on a network segment, or travel on a transit vehicle along a selected route. They employ a time-dependent branch-and-bound method for the least-cost path (shortest-path); the path's costs contain the time elements of walking, waiting, in-vehicle and transferring penalty.

Lam *et al.* (1999) introduced a frequency-based, stochastic user-equilibrium model for the transit-assignment process. Their model considers capacity constraints on each transit route; the route on which the capacity constraints are not fulfilled is excluded from the set of attractive routes. For an overcrowded service, some passengers (who could already be on-board) may choose alternative services, which in practice is the case only with boarding passengers who see a loaded vehicle. Bell and Schmöcker (2004) concluded that the model by Lam *et al.* (1999) could fit situations with spare capacity, such as seat-reservation or high-fare systems.

Cominetti and Correa (2001) and Bouzaïene-Ayari *et al.* (2001) advanced the consideration of limited vehicle capacities with more realistic waiting-time functions at stops. Their frequency-based models used the effect of changes in frequencies (because of overcrowded vehicles), rather than minimum-frequency flow share. Their equilibrium-assignment models take into account queuing processes at transit stops to reflect more realistic changes in establishing the set of attractive routes. However, no specific algorithm was provided for computing their models.

In a schedule-based study, Nuzzolo *et al.* (2001) used a random utility function with passenger information for a transit-assignment procedure. They included departure-time choice and stop choice in their stochastic model. Stop choice was defined as the probability of selecting the boarding point from a set of stops located within a given access distance. They

developed the model for high-frequency service with a sequential choice process for both within-day and day-to-day dynamics.

Kurauchi *et al.* (2003) introduced a frequency-based model in which a cost is assigned to the probability of failing to board the transit vehicle. They considered the separation between passengers on-board a crowded vehicle and passengers at stops seeking to board. The latter have a reduced priority to flow on the network. Kurauchi *et al.* used Markov chains, in which the boarding probability is dependent on the leftover capacity of the transit vehicle. They also discussed the use of the Markovian approach in stochastic and deterministic user-equilibrium route-choice processes.

Finally, Cepeda *et al.* (2006) provided a frequency-based formulation built on the model by Cominetti and Correa (2001). The former introduced a transit-assignment algorithm intended for large-scale networks and tested it on real-size networks. Cepeda *et al.* (2006) discussed the assumptions made, the difficulties in reaching a satisfactory solution, and some possible future improvement directions.

A summary of the articles reviewed appears in Table 12.1, in which separate attention is given to frequency- and schedule-based models.

Table 12.1 *Summary of features concerning the articles reviewed*

Advance in transit-assignment research	Frequency-based features	Schedule (timetable)-based features	Route-choice features
Dial (1967), Le Clercq (1972)	Heuristic consideration of waiting time (and travel time) in shortest-path approach		Boarding the first transit vehicle to arrive
Chriqui and Robillard (1975)	Probabilistic selection of a subset of routes to minimize the expected sum of [wait + travel] time		Boarding the first vehicle from a set of attractive routes
Nguyen and Pallotino (1988)	Origin to destination is interpreted on an acyclic directed graph (called <i>hyperpath</i>)		Boarding the first vehicle (from a set of routes) with a planned strategy of path movement
Spiess and Florian (1989)	Choosing a set of routes to minimize expected sum of [access + wait + travel] time; using equilibrium model with linear programming		Boarding the first vehicle using a strategy of choosing only among attractive routes

(Continued)

Table 12.1 Summary of features concerning the articles reviewed (continued)

Advance in transit-assignment research	Frequency-based features	Schedule (timetable)-based features	Route-choice features
De Cea and Fernández (1993)	Incorporating a limited capacity for each route (of an attractive set) at stops in which waiting time depends on passenger flow; using asymmetric equilibrium model with Jacobi method		Boarding the first vehicle (from a set of routes), given that passenger flows do not exceed a route's capacity
Wu <i>et al.</i> (1994)	Hyperpaths are used for walk, wait, board, in-vehicle, transfer and alight time elements; boarding time increases with flow, using equilibrium model with Jacobi method		Boarding the first vehicle using a strategy of choosing only among attractive routes
Hickman and Bernstein (1997)		Model for high-frequency service using sequential choice approach	Deterministic utility path-choice model with passenger information
Tong and Wong (1999)		Simulation model consisting of a network of routes with a given number of departure times; using shortest-path of weighted [walk + wait + travel] time and route-change penalty	Random utility path-choice model for frequent service
Lam <i>et al.</i> (1999)	Stochastic user equilibrium with explicit route-capacity constraints		Boarding the first vehicle; for overcrowded service, some passengers may choose alternative services
Cominetti and Correa (2001), Bouzaïene-Ayari <i>et al.</i> (2001)	Congestion functions at stops obtained from queuing theory to increase waiting time and affect passenger-flow share		Boarding the first vehicle (from a set of routes) with available capacity

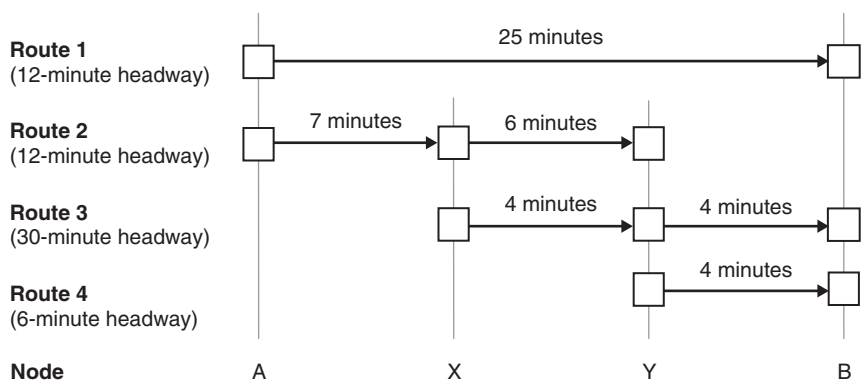
(Continued)

Table 12.1 Summary of features concerning the articles reviewed (continued)

Advance in transit-assignment research	Frequency-based features	Schedule (timetable)-based features	Route-choice features
Nuzzolo <i>et al.</i> (2001)		Introducing a set of departure-time choices and a set of stop choices for a given access distance	Random utility path-choice model for frequent service with possible passenger information
Kurauchi <i>et al.</i> (2003)	Introducing fail-to-board probability in which the demand exceeding capacity remains on the platform; using Markov chains and user-equilibrium approach		Boarding the first vehicle of a chosen single route in which route choice depends on risk of failing to board
Cepeda <i>et al.</i> (2006)	Congestion functions at stops with formulation for large-scale networks		Boarding the first vehicle (from a set of routes) with available capacity

Exercise

Given the following 4-route schematic transit network adapted from the EMME/2 user’s manual (reference: INRO Consultants Inc. EMME/2 User’s Manual, Release 9.6, May 2005); this network contains the following headways and average travel times.



Assume that walking times and transfer times are zero.

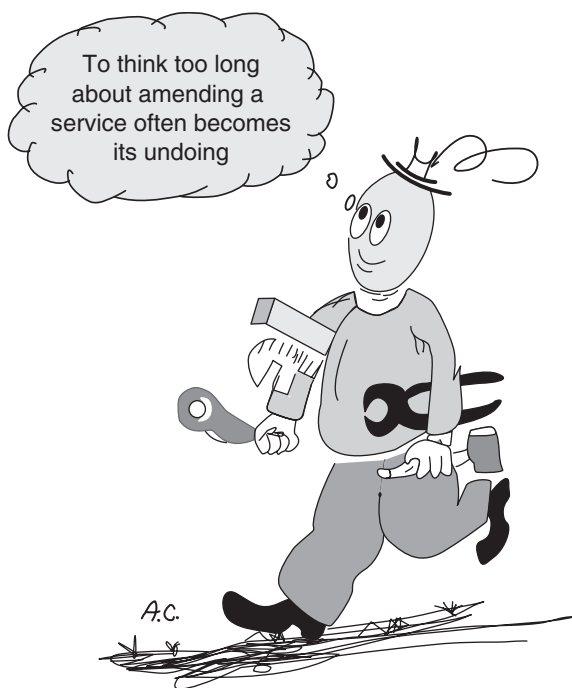
- (a) Describe the possible travel options from A to B; include the possibility that a passenger may choose among different combinations of routes.
- (b) Calculate expected travel times for each option.
- (c) Which set of options gives the minimum expected travel time?

References

- Bell, M. G. H. (2003). Capacity constrained transit assignment models and reliability analysis. In *Advanced Modeling for Transit Operations and Service Planning* (H. K. Lam and M. G. H. Bell, eds), pp. 181–199, Elsevier Ltd.
- Bell, M. G. H. and Schmöcker, J.-D. (2004). A solution to the transit assignment problem. In *Schedule-based Dynamic Transit Modeling. Theory and Applications* (N. H. M. Wilson and A. Nuzzolo, eds), pp. 263–279, Kluwer Academic.
- Bouzaïene-Ayari, B., Gendreau, M. and Nguyen, S. (2001). Modeling bus stops in transit networks: A survey and new formulations. *Transportation Science*, **35**, 304–321.
- Cepeda, M., Cominetti, R. and Florian, M. (2006). A frequency-based assignment model for congested transit networks with strict capacity constraints: characterization and computation of equilibria. *Transportation Research*, **40B**, 437–459.
- Chriqui, C. and Robillard, P. (1975). Common bus lines. *Transportation Science*, **9**, 115–121.
- Cominetti, R. and Correa, J. (2001). Common-lines and passenger assignment in congested transit networks. *Transportation Science*, **35**, 250–267.
- De Cea, J. and Fernández, J. E. (1993). Transit assignment for congested public transport systems: An equilibrium model. *Transportation Science*, **27**, 133–147.
- Dial, R. B. (1967). Transit pathfinder algorithm. *Highway Research Record*, **205**, 67–85.
- Florian, M. (1977). A traffic equilibrium model of travel by car and public transit modes. *Transportation Science*, **11**, 166–179.
- Guan, J. F., Yang, H. and Wirasinghe, S. C. (2006). Simultaneous optimization of transit line configuration and passenger line assignment. *Transportation Research*, **B 40(10)**, 885–902.
- Hickman, M. and Bernstein, D. H. (1997). Transit service and path choice models in stochastic and time-dependent networks. *Transportation Science*, **31**, 129–146.
- Israeli, Y. and Ceder, A. (1996). Public transportation assignment with passenger strategies for overlapping route choice. In *Transportation and Traffic Theory* (J.-B. Lesort, ed.), pp. 561–588, Elsevier Ltd.
- Jolliffe, J. K. and Hutchinson, T. P. (1975). A behavioural explanation of the association between bus and passenger arrival at a bus stop. *Transportation Science*, **9**, 248–282.
- Kurauchi, F., Bell, M. G. H. and Schmöcker, J.-D. (2003). Capacity constrained transit assignment with common lines. *Journal of Mathematical Modeling and Algorithms*, **2(4)**, 309–327.
- Lam, W. H. K., Gao, Z. Y., Chan, K. S. and Yang, H. (1999). A stochastic user equilibrium assignment model. *Transportation Research*, **33B**, 351–368.
- Le Clercq, F. (1972). A public transport assignment method. *Traffic Engineering and Control*, **14(2)**, 91–96.

- Marguier, P. H. J. and Ceder, A. (1984). Passenger waiting strategies for overlapping bus routes. *Transportation Science*, **18**, 207–230.
- Nguyen, S. and Pallotino, S. (1988). Equilibrium traffic assignment for large-scale transit networks. *European Journal of Operational Research*, **37**, 176–186.
- Nuzzolo, A. (2003a). Transit path choice and assignment models. In *Advanced Modeling for Transit Operations and Service Planning* (H. K. Lam and M. G. H. Bell, eds), pp. 93–124, Elsevier Ltd.
- Nuzzolo, A. (2003b). Schedule-based transit assignment models. In *Advanced Modeling for Transit Operations and Service Planning* (H. K. Lam and M. G. H. Bell, eds), pp. 125–163, Elsevier Ltd.
- Nuzzolo, A. and Crisalli, U. (2004). The schedule-based approach in dynamic transit modeling: A general overview. In *Schedule-based Dynamic Transit Modeling. Theory and Applications* (N. H. M Wilson and A. Nuzzolo, eds), pp. 1–24, Kluwer Academic.
- Nuzzolo, A., Russo, F. and Crisalli, U. (2001). A doubly dynamic schedule-based assignment model for transit networks. *Transportation Science*, **35**, 268–285.
- Spiess, H. and Florian, M. (1989). Optimal strategies: A new assignment model for transit networks. *Transportation Research*, **23B**, 83–102.
- Tong, C. O. and Wong, S. C. (1999). A stochastic transit assignment model using dynamic schedule-based network. *Transportation Research*, **33B**, 107–121.
- Wu, J. H., Florian, M. and Marcotte, P. (1994). Transit equilibrium assignment: A model and solution algorithm. *Transportation Science*, **28**, 193–203.

13 Service Design and Connectivity



Chapter 13 Service Design and Connectivity

Chapter outline

- 13.1 Introduction
 - 13.2 Service-design elements
 - 13.3 Scheduling-based solution for operational parking conflicts
 - 13.4 Optimum stop location: theoretical approach
 - 13.5 Connectivity measures and analysis
 - 13.6 Literature review and further reading
 - Exercises
 - References
-

Practitioner's Corner

After reviewing and investigating transit-scheduling components and passenger demand and assignment, this chapter opens with a description of transit service and network and route-design components, subjects that will also be covered in the three following chapters. Service-design and connectivity issues, in the subjects of this chapter, facilitate an understanding of the importance of transit planning in enhancing existing or new transit services. The planning table is the mechanism with which to start changing the prevailing opinion that transit service and problems are tied to each other as in the adage: urban areas with transit facilities are exciting places where something is almost always happening, mostly unsolved.

This chapter consists of four principle parts: following an introductory section. Section 13.2 outlines the predominant design elements, service strategies, and possible actions for both existing and new transit services. In addition, it provides standards and measures for system and connectivity performance. Section 13.3 pinpoints and solves a specific design problem in which transit vehicles need to park at route-departure/controlled/holding stops located on-street; because of the lack of parking spaces, they block the traffic lane adjacent to the parking bay/lane. The solution approach employs a surplus function model representing the surplus of parking vehicles required at a particular departure point in a multi-terminal transit system. Section 13.4 furnishes a framework and theoretical analysis for finding the smallest number of public transit stops and their locations so that no passenger is further away from a stop than a pre-selected distance. An optimal algorithm with an example is provided. Section 13.5 describes an initial methodological framework and concepts for quantifiable transit-connectivity measures, as well as directions and tools for detecting weak segments in inter-route and inter-modal transit chains (paths), for possible revisions/changes. The chapter ends with a literature review and exercises.

Practitioners are encouraged to visit thoroughly all the sections besides Sections 13.3 and 13.4. They should inspect the practical problems portrayed in these two sections and the examples with their solution.

Finally, among the important transit-service and connectivity attributes are waiting time and passenger information. When designing the service, it is worthwhile to put yourself into the shoes of waiting passengers, who wish the wait had a fast-forward button. This may lead the planner to recommend hanging the following sign at all stops: “Waiting times perceived are larger than they are”. In addition, useless information should be avoided, such as that given to a passenger who asked where to get off the bus: “See where I get off, and get off two stops before”.

13.1 Introduction

Profound service-design approaches, including well-coordinated service elements, comprise the foundation of any successful public transit service. In fact, an adequate, well-designed service is a present the transit agency gives to itself.

The importance of service-design and connectivity elements may be better comprehended after reading the following selections of a newspaper article, written by its editorial writer. Although the writer colours his personal experience with metaphors, it is hard to argue with the basic facts of the article.

Transit connectivity frustration

From the *Los Angeles Times*, 5 February, 2006 (‘Taking the rapid out of transit’ by Dan Turner).

Like many epic journeys of exploration, mine began not out of necessity but out of curiosity – the ancestral human urge to test the boundaries of endurance and knowledge. My quest: to get from my house in the Hollywood Hills to LAX, using only public transportation. “I had not anticipated that the work would present any great difficulties,” said Sir Ernest Shackleton after surviving his harrowing, failed attempt to reach the South Pole in 1915, his icebound ship by that time at the bottom of the sea. . . .

Trains ferry passengers in and out of most big airports across the country, including Atlanta’s Hartsfield, Chicago’s O’Hare and even San Francisco International. But not at Los Angeles International Airport. It is the fifth-busiest airport in the world, with more than 60 million passengers a year, and more people start their flights there than anywhere else – yet it is not served by any rail line. Like reaching the Pole, getting to the airport using only public transit is a feat requiring courage, fortitude and very bad judgment. . . .

Twenty-four minutes later the bus arrives. Knowingly, I put \$1.25 in the slot and take a seat. Leaving the bus at Hollywood Boulevard, I ask the driver for a transfer. He fixes me with a fishy stare . . . buses do not issue transfers. You have to buy a day pass, which is \$3. I hold out a \$5 bill. The driver looks at it as if it’s a used tissue. He does not give change. So begins the 1.5-mile trek to the Hollywood and Highland Red Line station, with not a sled dog or Sherpa to lead the way. Yes, I could take another bus, but I’m still steamed about the day-pass snub. . . .

At least I didn’t have to contend with bus drivers anymore. Riding the escalator into the bowels of Hollywood, I enter the Mercedes of L.A. public transit, the \$4.5-billion Red Line

subway. The 17.4-mile system is fast, semi-clean, quiet – a wonder of efficiency with nearly 120,000 boardings a day. It would attract many thousands more if only it went somewhere. Originally planned to run all the way down Wilshire Boulevard, the city's densest corridor, it instead ends with a whimper at Wilshire and Western Avenue, its spine hacked off by community opposition and weak-kneed politicians. . . .

From here it's 17 minutes to the 7th Street station in downtown L.A., where I transfer to the Blue Line. Eight minutes later, I'm flashing through downtown at 25 mph. . . .

Several days later – or maybe it's 24 minutes – I'm at the Imperial/Wilmington station in Lynwood, prepared to transfer to the Green Line. Thirteen minutes later, I'm on the train heading toward LAX. At last I can see it up ahead – the LAX/Aviation Boulevard station. But my adventure isn't over.

The Green Line from Norwalk was originally planned to end inside the airport, but in 1995, after the money ran short, so did the line. An \$11-billion plan to remodel LAX, approved in 2004, called for a people mover that would carry passengers to the terminals from a big transportation center connected to the Green Line, but when most of the plan was recently scrapped to settle a lawsuit with airport neighbors, so was the people mover. Instead, there is a shuttle bus from the Aviation station. . . .

My expedition from home to LAX takes two hours and 47 minutes, yet I am flushed with the thrill of accomplishment when the shuttle finally arrives. I am footloose and free, untied to a vehicle in a long-term parking lot. Records are sketchy, but I'm confident that no one else from the Hollywood Hills has ever attempted this journey. After all, they could drive or take a cab to LAX in about 40 minutes. Unlike Shackleton, I have reached my Pole.

Along the line of the article's metaphors, we can conclude that when it is dark enough, one can see the stars; when the transit service is poor, the transit agency can see how good it can be, and proceed to attain this objective.

13.2 Service-design elements

A discussion of service and evaluation standards and guidelines appears in Section 1.3 in Chapter 1. These standards and guidelines establish service needs as seen by the local authority/government and the transit agency, but they do not cover the entire spectrum of required service-design elements. This section outlines the predominant design elements for both existing and new transit services.

The main service-design elements are these:

- Potential markets
- Network size and coverage
- Network structure, followed by route structure
- Route coordination (intra- and inter-agency)
- Route classification
- Span of service
- Service frequency, followed by public timetables
- Schedule coordination (intra- and inter-agency)
- Vehicle scheduling, followed by vehicle types and fleet size
- Crew scheduling, followed by rostering

- Fare policy
- Passenger amenities and information systems
- Data-collection systems
- Determination of measures of performance
- Setting service and evaluation standards
- Ridership, cost and revenue estimation.

13.2.1 Description of service-design elements

The elements in the foregoing list will now be described briefly.

According to TranSystems *et al.* (2006), each **potential market** element requires an analysis of demographics and travel patterns, as well as market research. The analysis is used to identify areas with the potential to support transit services, and locate current and projected travel markets. The market research is intended to identify market segments, along with passengers' degree of satisfaction from the transit service; the research focuses on service preferences and the inclination to use/increase the use of the service, given specific improvements.

The **network size and coverage** element applies to the set of all routes – i.e. the entire system – and serves to fix the recommended spacing (distance) between routes for varying residential area densities. TranSystems *et al.* (2006) provides a few examples of a coverage measure. For instance, it suggests the provision of transit service within walking distance (defined as 400 metres) of all residents living in areas with population densities greater than 2,000 people per square kilometre.

The **network structure, followed by route structure** element can take one or a combination of the four common forms: multimodal, radial, grid and time transfer, and pulse. A multimodal network of routes coordinates short and long trips through different transit modes; e.g. short trips by buses that feed long trips by rail. The network of radial routes aims at providing a considerable amount of service to central points; e.g. the central business district (CBD). A grid structure network of routes, on the one hand, allows easy access to the transit system, but on the other hand requires many transfers; thus, the transfers in this network are timed and preferably synchronized on-line. The last network structure is based on a pulse system, in which routes are initiated at the same (central) point, which becomes a transfer point; usually this suits small urban areas.

The **route coordination (intra- and inter-agency)** element applies at the intra-agency front to a route-design system with coordinated meeting points in terms of convenient passenger-transfer facilities. At the inter-agency front, coordination is manifested in terms of operating policies and transit-support promotion programmes.

The **route classification** element exhibits various route-type needs for different geographical areas. Route classification schemes usually comprise one or a combination of the following: line haul, local, express, feeder/distributor, branching, radial line haul, zonal, commuter, circulator, cross-town, short-turn and limited stop.

The **span of service** element indicates the length of time that a service should be provided, by time of day and day of week. For example, a span of service guideline can suggest that the first trip should arrive no later than, and the last trip should depart no earlier than, the (specified) times shown for weekdays, Saturdays and Sundays.

The **service frequency, followed by public timetables** element is described and discussed explicitly in Chapters 3 and 4. Essentially, this element sets minimum service-frequency and

maximum vehicle-occupancy thresholds to guarantee a basic level of service for different geographical areas. For the timetable, this design element sets the type of vehicle headway to be utilized; e.g. even headway, clock headway, even-load headway, by time of day and day of week.

The **schedule coordination (intra- and inter-agency)** element applies at the intra-agency front to a design of timetables that will maximize simultaneous arrivals at transfer points (see Chapter 6). At the inter-agency front, schedule coordination is expressed in terms of jointly designed timetables by different agencies and the use of different transit modes to allow for easy transfers between transit routes, with minimum waiting time.

The **vehicle scheduling, followed by vehicle type and fleet size** element is described and discussed in Chapters 7, 8 and 9. Basically, this element ensures design efficiency in terms of the fleet size required, balancing deadheading trips and shifts in departure times and minimizing the cost involved in purchasing vehicles.

The **crew scheduling, followed by rostering** element is described and discussed in Chapter 10. Fundamentally this element focuses on minimizing the crew wages involved and provides a satisfactory working schedule from the crew's perspective.

The **fare policy** element aims at improving payment convenience and integration between different transit services. For instance, a fare policy can be established with free transfers, discount options, and the elimination of fare zones and special surcharges.

The **passenger amenities and information systems** element aims at improving passenger facilities and vehicle amenities, and improving passenger information for pre-trip planning, en-route riding, and waiting at terminals and stops. This element is also concerned with improving passenger safety and security.

The **data-collection system** element is described and discussed in Chapters 2, 3 and 17. This vital element constructs the foundation for effective and efficient transit-operations planning. The key components of this element are suitability, accuracy and an adequate amount of data.

The **determination of measures of performance** element aims at quantifying tools for measuring the quality of service. The measures determined show whether the service is appropriate, convenient and reliable, especially from the passengers' perspective. Two basic types of measures of performance are described below in this section: system performance and connectivity performance.

The **setting service and evaluation standards** element is described and discussed in Chapter 1. On one hand, standards maintain and improve the service; on the other hand, they present a source of fiscal pressure on the transit agencies. Evaluation standards aim at improving the efficiency, effectiveness and productivity of the transit service.

The last design element on the list is **ridership, cost and revenue estimation**. This element provides the tools for forecasting the three linked components of passenger demand, cost of service and revenue. Some of these tools are described in Chapter 11, which emphasizes ridership prediction.

13.2.2 Service strategies and possible actions

A good summary of transit-service strategies appears in the US TCRP H-32 report (TranSystems *et al.* 2006). The five following lists provide new design and design-adjustment examples by purpose and possible implementation actions (in parenthesis).

New forms of service

- Improve travel speed (introduction of express/zonal/rail/bus rapid transit [BRT] service)
- Attract new passengers (circulator shuttles; dial-a-ride service)

Area coverage service

- Increased route coverage (service expansion; integrated circulator and line-haul services)
- Increased span of service (late-night, weekends, holidays)

New and adjusted routing

- New routing (linking routes; splitting routes; feeder-based; crosstown-based)
- Routing adjustment (route extension/shortening/realignment; express/zonal/local)
- Coordination (intra- and inter-agency transfer centres/points)

New and adjusted scheduling

- Introduce interlining (area-based; trip-based; vehicle type-based)
- Introduce/improve coordination (intra- and inter-agency synchronized/timed transfers)
- Change frequency (increase/decrease; even headways; clock headways)
- Change departure times (shifts within tolerances; even-load departure times)
- Modify time elements (average travel time; layover time; recovery time)
- Improve reliability (see Chapter 17 for lists of possible actions)

Improved amenities

- Introduce/improve passenger facilities (stops/station; transfer points; transit centres; park-and-ride amenities)
- Introduce/improve vehicles (amenities; new vehicle type)
- Increase safety (introduce/improve safety features on vehicles and at stops; increase awareness and preparedness)
- Increase security (introduce/improve security agents and features on vehicles and at stops; increase awareness and preparedness)

13.2.3 Standards and measures of system and connectivity performance

This section presents service standards and measures of performance. The service standards, accompanied by the required data, follow the list given in Figure 1.4 in Chapter 1; these standards appear in Table 13.1. The measures are divided into measures of system performance (MOSP) and measures of connectivity performance (MOCP).

The MOSP and MOCP measures, which should be selected at the design stage, quantify how well transit routes are used. Table 13.2 lists ten MOSP measures and three MOCP measures, along with their notation, interpretation and the data required. Measures of reliability of service are not included in Table 13.2; they will be dealt with in Chapter 17.

The measure of emissions E_{rt} in Table 13.2 can be extracted from real data. For instance, Table 13.3 shows calculated data from the US EPA's (Environmental Protection Agency) MOBILE5 and PART5 software-based models. MOBILE5 is a vehicle-emission modelling software, and PART5 is a software model for estimating particulate emissions from highway

Table 13.1 *Service standards and data required*

Standard item	Data required
Route Length	Average running time; distance
Stop Spacing	Population density; type of service (e.g. local, express); no. of passenger using a given stop
Route Directness**	Average running time; distance; pass. O-D count
Short -Turn**	Average passenger counts by stop
Route Coverage	Population data; land-use data; public view
Route Overlapping	Scale maps of the network of routes
Route Structure**	Average pass. O-D counts; population data; public opinion
Route Connectivity	Maps of network of routes and feasible transfer points
Span of Service	Timetables by route and zone of operation
Load (Crowding) Level	Average passenger counts by stop or at max-load-points
Standees**	Passenger counts by stop
Headway Upper Limit**	Average passenger counts; permits between operator and authority
Headway Lower Limit	Averages passenger counts; no. of vehicles available
Transfers	Averages passenger counts; transfer counts; waiting time
Passenger Shelters**	Average passenger boardings by stop; no. of elderly and handicapped
Schedule Adherence**	On-board counts of departure and arrival times (manually or automatic device)
Timed Transfer	Passenger O-D counts by time-of-day; trip timetables
Missed Trips**	Dispatcher log and maintenance records (agency data)
Passenger Safety**	Accident reports, combined with average passenger counts and km driven
Public Complaints	Ordinary mail, e-mails, fax, telephone, visits (by category of complaint)

* Reflects data mainly from the US.

** Standards commonly in use.

vehicles. More information can be found in the EPA’s national emission inventories air pollutant trend website: www.epa.gov/ttn/chief/trend. A more updated model by EPA is MOBILE6, which can be found at www.epa.gov/otaq/m6.htm. Table 13.3 (using MOBILE5 and PART5 software) provides data by vehicle type, per single vehicle, and on a per-passenger basis; the last is based on an occupancy of 35 passengers/bus and 150 passengers/train.

Table 13.2 *Notations and interpretation of transit measures of performance*

Type of measure	Measure	Designated notation	Interpretation and data required
Measure of system performance (MOSP)	Number of passengers	P_{rt}	Number of passengers by route r and time-of-day t
	Revenue	R_{rt}	Agency revenue by route r and time-of-day t (\$)
	Vehicle trips	V_{rt}	Number of vehicle trips, timetable-based, by route r and time-of-day t
	Vehicle-kilometres of travel	VKT_{rt}	Vehicle-kilometres of travel, based on V_{rt} , by route r and time-of-day t , (veh-km)
	Passenger travel times	PTT_{rt}	Passenger travel times, based on trip times and load profiles, by route r and time-of-day t (hours)
	Emissions	E_{rt}	Vector of average emission of four pollutants (CO, NO _x , VOC, PM ₁₀), based on VKT_{rt} , by route r and time-of-day t (kg)
	Vehicle-hours of travel	VHT_{rt}	Vehicle-hours of travel, based on V_{rt} , by route r and time-of-day t (veh-hr)
	Passengers per veh-hr	PVH_{rt}	Ratio of passengers per vehicle-hour, based on P_{rt} and VHT_{rt} , by route r and time-of-day t (pass/veh-hr)
	Passengers per veh-km	PVK_{rt}	Ratio of passengers per vehicle-kilometre, based on P_{rt} and VKT_{rt} , by route r and time-of-day t (pass/veh-km)
Measure of connectivity performance (MOCP)	Revenue per passenger	RP_{rt}	Agency revenue per passenger, based on R_{rt} and P_{rt} , by route r and time-of-day t (\$)
	Passenger waiting times	PWT_{rt}	Passenger waiting times, based on P_{rt} and service frequency, by route r and time-of-day t (hours)
	Passenger transfers	PTR_{rt}	Number of passenger transfers, within and between modes of travel, by route r and time-of-day t
	Connectivity-production cost	CPC_{rt}	Cost of passenger waiting time, transfer penalty and on-board travel time, plus the combined average cost of vehicle-hour, based on VHT_{rt} , PWT_{rt} and PTR_{rt} , by route r and time-of-day t (\$)

Table 13.3 Average emission of pollutant per vehicle (in parenthesis, per passenger) in gram/mile

	CO	NO _x	VOC	PM ₁₀
Diesel bus	23.2 (0.66)	22.1 (0.63)	4.2 (0.12)	0.63 (0.02)
Automobile	23.0 (19.17)	3.9 (3.25)	3.7 (3.08)	0.09 (0.075)
Rail	0.03 (0.0002)	0.47 (0.003)	0.02 (0.0001)	0.009 (0.0001)

13.3 Scheduling-based solution for operational parking conflicts

One of the operational problems for urban transit, especially with buses, occurs in a situation in which vehicles need to park at route-departure/controlled/holding stops located on-street; however, because of a lack of parking spaces, the buses block the traffic lane adjacent to the parking bay/lane. Such operational scenarios are commonly observed at school dismissal times. Although a special arrangement can be made for school/factory dismissal times – e.g. using side streets – the blocking of a traffic lane in other situations can result in severe traffic congestion. One way to solve the problem is to construct more parking spaces at these conflict points; however, this is a costly and time-consuming solution. This section presents a schedule-based design solution to this problem in which lane-blocked situations are caused at route-departure points. That is, eliminating or reducing the impact of lane-blocked situations through the use of shifting departure times and/or inserting DH trips in timetables.

13.3.1 Surplus-function model

The model constructed to solve the operational parking conflicts is basically a mirror image of the deficit function (DF) model described extensively in Chapters 7, 8 and 9. This model is called a **surplus function** (SF) model because it represents the surplus of parking vehicles required at a particular departure point in a multi-terminal transit system. SF is a positive step function that basically increases by 1 at the time of each trip arrival and decreases by 1 at the time of each trip departure; its starting value depends on the starting and maximum DF values. This definition shows a strong mutual dependency between DF and SF.

To construct the SF model, the only information needed is a timetable of required trips and the number of available parking spaces at each departure point. The main advantage of SF, as with DF, is its visual nature.

Let K be the set of all departure points; $p(k, t, S)$ denote the SF for departure-point k for all $k \in K$ at time t for schedule S ; and PS_k be the number of available parking spaces at k . The maximum value of $p(k, t, S)$ over the schedule horizon $[T_1, T_2]$ is designated $P(k, S)$. For simplicity, S will be deleted when it becomes clear which underlying schedule is being considered. The DF notations defined in Sections 7.5.1, 7.5.2, 8.3.1 and 8.3.3, in Chapters 7 and 8, will also be utilized for the SF analyses.

SF, unlike DF, cannot be negative, because each departure must be preceded by an arrival, thereby creating the need for a parking space between the arrival and the departure epochs. DF is defined as the total number of departures minus the total number of trip arrivals at terminal k , up to and including time t . SF is defined as the total number of arrivals minus the total number of trip departures at terminal k , up to and including time t ,

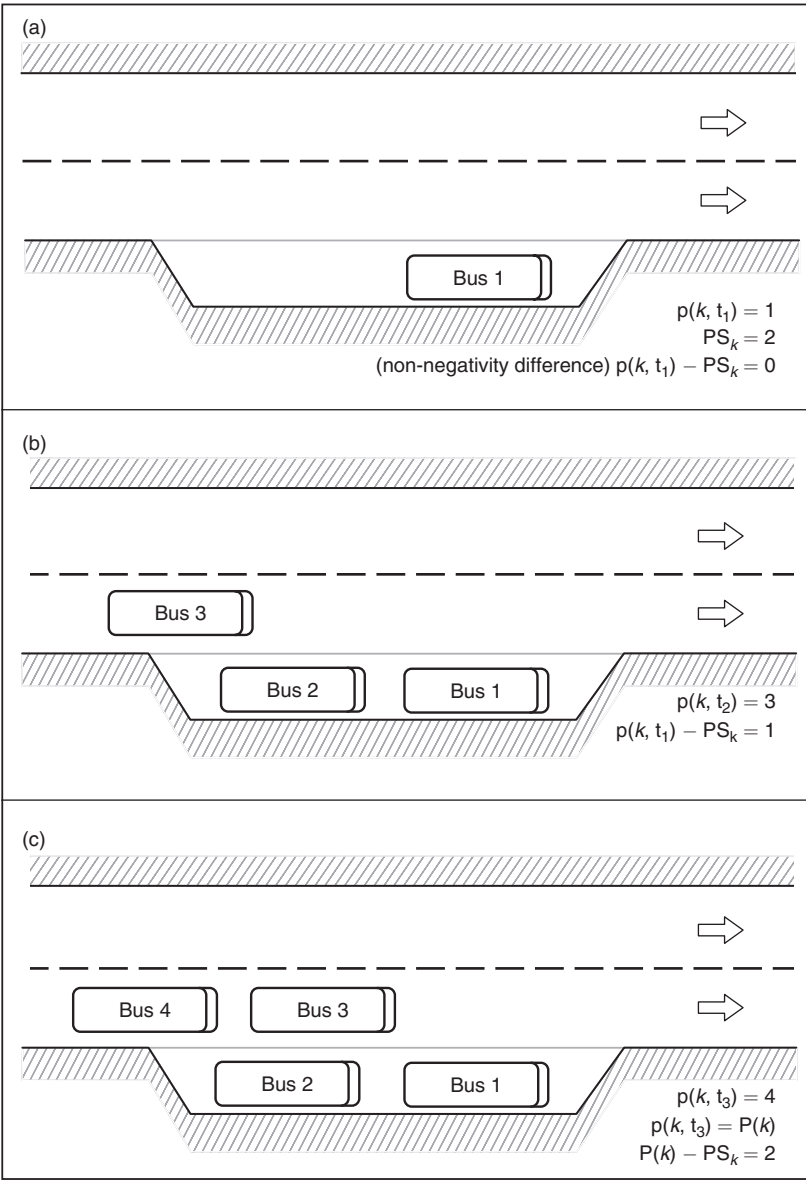


Figure 13.1 Three situations of operational parking at $k =$ route departure stop (the value of the surplus function and its difference from the two available spaces are shown)

in which the number of arrivals at T_1 is $D(k)$ minus the number of departures at T_1 . The next section further describes these relationships and definitions in a more formal way.

Figure 13.1 depicts three operational parking situations with $PS_k = 2$, at t_1 , t_2 , and t_3 . Part (a) of Figure 13.1 shows a single parked bus, hence $p(k, t_1) = 1$; because of the non-negativity

of $p(k, t)$, the excess number of buses, $p(k, t) - PS_k$, is zero (not -1). Parts (b) and (c) of Figure 13.1 exhibit lane-blocked situations; in part (c), $p(k, t_1) = P(k) = 4$, representing the maximum number of parking vehicles over $[T_1, T_2]$.

13.3.2 Minimum parking spaces required

Following the definition of SF, we can establish a few formulas and rules that will constitute the basis of an algorithmic-based solution. The first of these is Theorem 13.1.

Theorem 13.1: The minimum number of parking spaces required at k , $N_p(k)$ is equal to the maximum surplus function at k .

$$N_p(k) = P(k) = \max_{t \in [T_1, T_2]} p(k, t) \quad \forall k \in K \quad (13.1)$$

Proof: Using the notation of Section 7.5.2, F_k is the number of vehicles present at k at the start of the schedule horizon T_1 ; $s(k, t)$ and $e(k, t)$ yield the cumulative number of trips starting and ending at k from T_1 up to and including time t . The number of vehicles parking at k at time $t \geq T_1$ is $F_k - s(k, t) + e(k, t)$, which by definition is $p(k, t)$. This expression must be both non-negative and less than or equal to $N_p(k)$; i.e. $0 \leq p(k, t) \leq N_p(k)$, $T_1 \leq t \leq T_2$. The minimum number of parking spaces required at k , then, is equal to the maximum surplus function $P(k)$.

Consequently the objective of the operational parking conflict is to eliminate or minimize cases in which $P(k) > PS_k$; in other words, to make sure that the minimum number of parking spaces required at k is less than or equal to the number of spaces available. The following *Lemma* 13.1 and Theorem 13.2 facilitate the inter-dependency of SF and DF through the maximum surplus function $P(k)$.

Lemma 13.1: Compliance with $p(k, t) \geq 0$ for all $k \in K$ is attained by shifting up the mirror image of $d(k, t)$ by $+D(k)$ to obtain $p(k, t)$.

Proof: The basic concept of the SF is that it is the mirror image of the DF. The mirror image of $d(k, t)$ is simply $-d(k, t)$ for all $k \in K$; thus, the minimum value of the mirror image is $-D(k)$. The compliance with $p(k, t) \geq 0$ is attained, therefore, by shifting up $-d(k, t)$ by $+D(k)$.

Theorem 13.2:

$$P(k) = D(k) - \min_{t \in [T_1, T_2]} d(k, t), \quad \forall k \in K \quad (13.2)$$

Proof: Based on *Lemma* 13.1, $p(k, t) = -d(k, t) + D(k)$ for all $k \in K$; thus, the following holds: $\max p(k, t) = P(k) = D(k) - \min d(k, t)$ for all $k \in K$.

Figure 13.2 illustrates a nine-trip fixed schedule with two terminals/departure points. Both $d(k)$ and $p(k)$ for $k = a, b$ are shown, with an emphasis on $PS_a = 2$ and $PS_b = 1$ available parking spaces. The value of $p(a, T_1) = 2$, $T_1 = 5:00$ is determined by shifting the mirror image of $d(a, t)$ three units up because of $D(a) = 3$; thus, because the mirror image of the DF, $-d(a, t)$, starts with $-d(a, T_1) = -1$, its shift will result in $p(a, T_1) = -1 + 3 = 2$. The same applies to $p(b, T_1) = 3$.

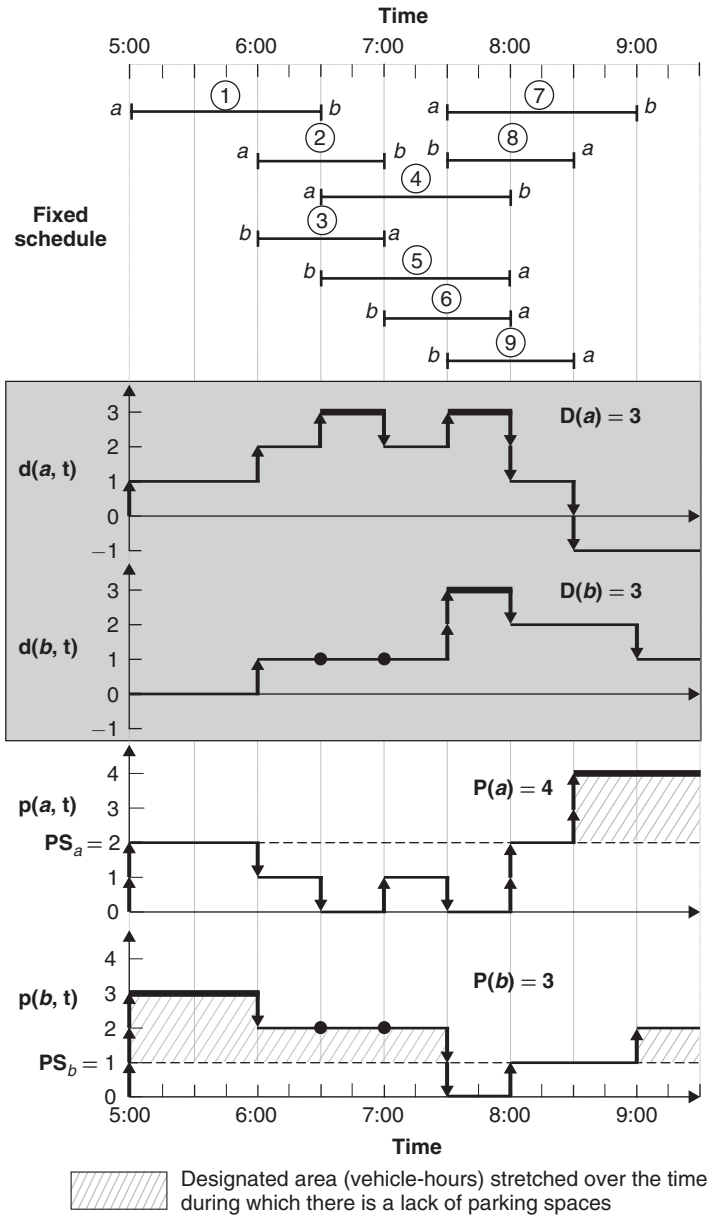


Figure 13.2 Construction of two surplus functions dependent on their deficit functions (grey background), for a nine-trip example

Note that in order to have a sufficient number of parking spaces, $P(k)$ must be less than or equal to PS_k . If it is not, attempts could be made to attain this constraint. The following section utilizes the scheduled-based tools to reduce $P(k)$ for all $k \in K$, in terms of shifting departure times and inserting DH trips.

13.3.3 Scheduled-based reduction of $P(k)$ procedure

The deficit-function theory explicated in Chapters 7 and 8 is used for a heuristic procedure to reduce required parking spaces for situations in which $P(k) > PS_k$. Ostensibly this procedure involves shifting trip-departure times and deadheading (DH) trip insertion so as to reduce $P(k)$, if necessary, without increasing $D(k)$ for all relevant $k \in K$.

The mirror-image configuration of the DF, including shifting up this configuration by $D(k)$ to attain $p(k, t)$, has properties similar to the original DF. As a result, the two following rules are applied for the construction of a heuristic procedure:

1. Shifting departure times (left, right, or in both directions), within their tolerances, for reducing $P(k)$ cannot increase $D(k)$, mainly because the maximum intervals of $P(k)$ and $D(k)$ do not overlap; this rule applies to both a single shift and a chain of shifts.
2. Inserting DH trip to reduce $P(k)$ is feasible only if the following two conditions are fulfilled: (a) the DH trip departs at or before the start of the first maximum interval of $P(k)$ and, at the same time, this DH departure must come at or after the last maximum interval of $D(k)$; (b) the DH trip must arrive, from k to k' , at or after the last maximum interval of k' , this rule applying both to a single DH trip and to a chain of DH trips.

Note that these two rules can also be formally stated and proved.

The schedule-based reduction of the $P(k)$ procedure appears in flow diagram form in Figure 13.3. This diagram contains similar features to the DF procedures described in Figures 7.9(a), 7.9(b), 8.11 and 8.12 in Chapters 7 and 8. The procedure described is designed for an interactive person-computer system. The selection of a departure point/terminal u can be made by the scheduler by inspecting the SFs on a graphical display. The search for a reduction of $P(k)$ at u (see Figure 13.3) can be performed manually by the scheduler or by requesting procedures that are based on Theorems 13.1 and 13.2 and the two rules above. If a unit-reduction shifting chain of $P(k)$ is found, all the SF information is updated for this new schedule, and a new iteration initiated. In the following step, the procedure seeks to reduce $P(k)$ by DH trips incorporated with shifting trip-departure times.

The modified URDHC (mixed with SDT) sub-routine, described in Section 8.5.3 in Chapter 8, is then applied, but with changes to account for rule (2) above; that is, inserting a DH trip from (or before) the start of the maximum interval of $P(k)$, compared to bringing a DH trip to the start (or before) of the maximum $D(k)$ interval in the DF procedures. In this DH procedure, the feasibility requirement for DH trip insertion is based on Equation (8.3) in Chapter 8. Another point to be mentioned is that DH trips that are added to the schedule S should include shifting tolerances for the next iterations of the SDT and modified URDHC procedures. Finally, if a URDHC has been found, the DH chain cost involved is compared with the saving of the cost of a single parking space. If the DH cost is higher than the saving cost, the URDHC is cancelled. Otherwise, the set of DH trips and required shifts (if any) $\{DH + SH\}$ is added to the previous schedule; all SF information, including shifting, is updated for a new schedule. The modified URDHC sub-routine continues, with updated S , until all terminals are examined.

A complete example of the required parking-space reduction procedure is illustrated in Figure 13.4. This example consists of a nine-trip schedule with three terminals, a single DH travel time of 75 minutes, and shifting tolerances of 15 minutes (both directions). The three DFs of the example appear with a grey background after going through the minimum-fleet-size

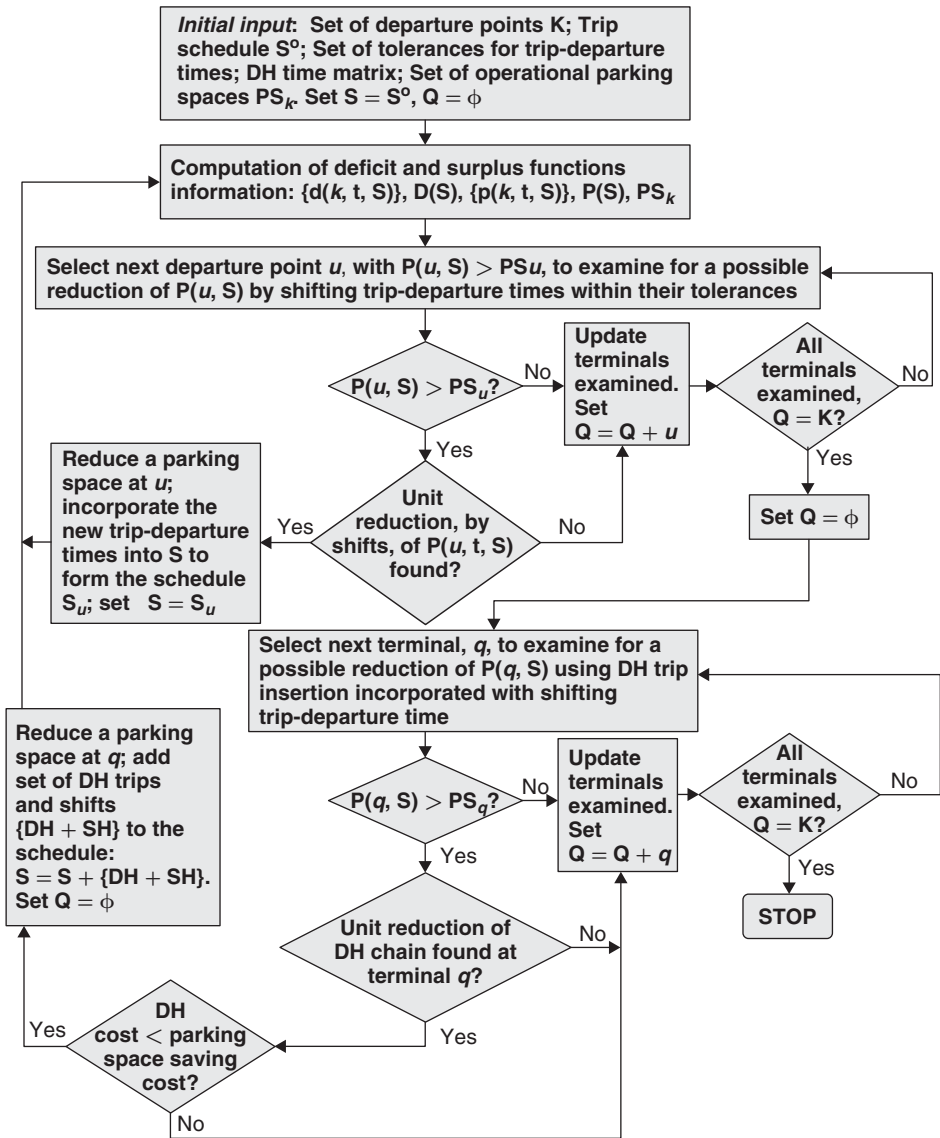


Figure 13.3 Flow diagram of the required parking-space reduction procedure, involving the shifting of departure times and DH insertions (modified URDHC sub-routine of Chapter 8)

reduction procedures; the optimum number of vehicles required is four. The SFs are illustrated below the DFs, and the only complete mirror-image configuration (multiplied by -1) is that of departure point c, where the shifting up is not performed, because $D(c) = 0$. The procedure to reduce $P(k)$ first checks which departure points are characterized by $P(k) > PS_k$; in fact, this is the case for all $k = a, b, c$.

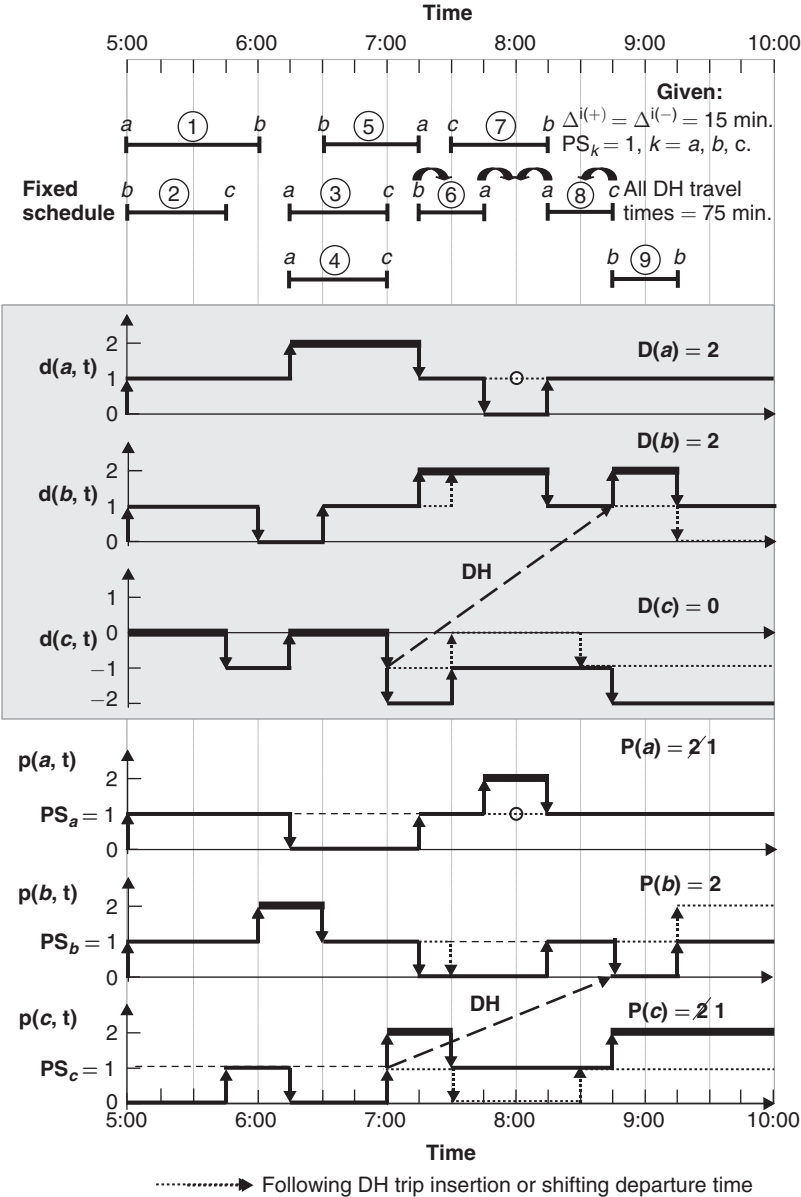


Figure 13.4 Nine-trip example of maximum reduction of the difference $P(k) - PS_k, k = a, b, c$ of the surplus functions, utilizing shifting and DH insertion (the impact of changes on the deficit functions (grey background) is also shown)

Departure point/terminal *a* is then selected in Figure 13.4 to search for possible shifting departure-time opportunities; indeed, for this search, two shifts of trips 6 and 8 are found feasible to reduce $P(a)$ from 2 to 1, thus making it equal to PS_a . These 15-minute shifts are shown in the upper part of Figure 13.4. The second step is the attempt to reduce $P(b)$ by

shifting trips 1 and 5 (15 minutes each) in opposite directions, but this cannot be realized. The reason is that by shifting trip 1 to the right, the value of $p(b, T_1)$ is changed from 1 to 2, which is the value of $D(b)$; the value reduces to 1 only at $p(b, 5:15)$. Thus, the two shifts cannot reduce $P(b) = 2$. The next step examines the reduction of $P(c)$ by shifts, but here, too, it cannot be performed. The procedure in Figure 13.3 proceeds to look for a DH trip insertion while fulfilling the two conditions of rule (2). Although it is impossible to insert a DH trip from departure points a and b , it is found that it is feasible to do so from c to b ; this results in reducing $P(c)$ from 2 to 1, making it equal to PS_c . The example ends up with only one excess-parking situation in b , between 5:45 and 6:30, but with the elimination of this undesirable circumstance in a and c .

13.4 Optimum stop location: theoretical approach

Network coverage is among the service-design elements presented in Section 13.2, which contains the criterion for providing transit service within a defined walking distance of all residents living in a certain area. The present section furnishes a framework and theoretical analysis for finding the smallest number and the locations of public transit stops so that no passenger is further away from a stop than a pre-selected distance. An optimum algorithm is described and discussed for a general road network in which the nodes are community locations and the stops are to be located along the arcs (streets) or on nodes. The algorithm follows the method described in Ceder *et al.* (1983).

In order to ensure a high level of service for public transit users, walking distances to stops should be as short as possible. Farewell and Marx (1996) state that people consider walk time to be much less convenient than in-vehicle travel time and proposed a maximum walking distance of 400 metres to a transit stop. Ceder *et al.* (1983) exhibit convenient ratios between in-vehicle travel time and walking time: 2:0 in the Netherlands, 3:5 in Chicago, and 6:2 in San Francisco. For instance, people in San Francisco consider one minute walking to be 6:2 times less convenient than one minute of in-vehicle travel time; this high value may be attributed to the difficult terrain in that city. Therefore, by reducing the length of the walking distance to stops in that city, public transit agencies will make their service more attractive.

13.4.1 Framework of analysis

The problem under consideration is to find the smallest number and the location of transit stops in a general network, so that no passenger is further away than a pre-selected distance, assuming that demand is generated at specific locations/nodes along the arcs/streets/roadway segments. The stops could be located either on the nodes or on any point along each arc. That is, the points lying on the arc are admissible as is each node. This problem, called the ' m -centre' problem in the operations research (OR) field, will be reviewed briefly here.

Minieka (1970) first proposed the m -centre problem as an optimal algorithm; this was followed by an independent study by Christofides and Viola (1971), which also presented an iterative optimal algorithm. Handler (1973) developed an improved algorithm for Minieka's approach and showed that his method was preferable to that suggested by Christofides and Viola, particularly as the problem size increases. Summaries of the algorithms developed

appear in Christofides (1975) and Minieka (1978). Several network examples serve as a stimulus for further examination of this m -centre problem, since both the Handler and the Christofides and Viola approaches might require too many steps (computational time) for the optimal solution. It should be noted, however, that the available procedures were constructed with the intention of finding the optimal locations of m -centres in a given network, rather than locating centres according to a critical distance constraint. Both problems can be solved by the same method, although more computational time is usually required for a 'given m -centre' than for a 'given-distance' problem.

Consider a connected network $G = \{N, A\}$ with a set of nodes N , a set of directed arcs A , and given distances $d(i, j)$, $\forall (i, j) \in A$. The number of nodes and arcs are $|N|$, and $|A|$ respectively. More notations are as follow:

- $D(k, q)$ = shortest distance between k and q ; k, q are two points anywhere on G (shortest-path algorithms appear in Appendix 10.A in Chapter 10)
- $SDM(i, j)$ = shortest-distance matrix for all pairs i and j ; $i, j \in N$
- SDM_d = $SDM(i, j)$ matrix information indicating the shortest paths by directions from node i to node j , $\forall i, j \in N$
- ℓ = critical distance (i.e. for each node i , $i \in N$ must be no further away than ℓ units from its closest stop)
- $S_{u,v}$ = $\{u, v \in N: D(u, v) \leq 2\ell\}$; i.e. set of all node pairs u, v that have a value equal to or less than 2ℓ in $SDM(i, j)$
- $D_i(u, v)$ = $\text{Min}[D(u, i), D(v, i)]$; $i \in N$; $u, v \in S_{u,v}$
- s = candidate stop
- $(i, j)_{u,v}$ = arc on which a candidate stop related to u, v can be located; $(i, j)_{u,v} \in A$
- ST = $\{s \in G: (i, j)_{u,v} \text{ exists}\} \cup \{s = i, \forall i \in N\}$; i.e. a set of all candidate stops, including each node as a candidate stop
- ST^* = set of stops in the optimal solution
- $d(i, s)$ = distance between node i and a candidate stop point s lying on $(i, j)_{u,v}$
- $P_{u,v}$ = $\{\text{potential active path: } (i, j)_{u,v} \text{ exists, and the critical distance criterion is satisfied for a given pair } u, v\}$; i.e. for the set of all possible paths in G between u and v , $u, v \in S_{u,v}$, so that: (1) if the candidate stop is on either i or j , the shortest distance between u and v through i or j , respectively, is equal to or less than 2ℓ ; (2) if the candidate stop lies on $(i, j)_{u,v}$, excluding i and j , the shortest distance (simple path) between u and v through $(i, j)_{u,v}$ is equal to or less than 2ℓ
- P_a = $\bigcup_{u,v \in S_{u,v}} P_{u,v}$; i.e. the set of all potential active paths
- SCP = Set Covering Problem; the corresponding problem is related to a matrix in which each row represents a node (there are n rows), and each column a candidate stop. If the distance between node i and the candidate stop is less than or equal to ℓ , the entry of 1 is indicated. The SCP is to find the least number of columns such that every row contains an entry 1 under at least one of the selected columns; that is, finding the minimum number of columns to 'cover' all the rows.

The objective function is to find the minimum number of stops, $s \in ST^*$, so that $D(i, ST^*) \leq \ell$ for all $i \in N$, where $D(i, ST^*)$ = the shortest distance between node i and its closest stop, $s, s \in ST^*$. Certainly, the SCP solution, where ST is represented by all the

columns, is the solution required for locating the transit stops. Note that there is always a feasible solution, since each node is also a candidate stop; that is, for large ℓ values, each node will be its own stop.

13.4.2 Set-ST algorithm and explanation

The algorithm to find the set ST is constructed as follows:

Step 1: Construct $\text{SDM}(i, j)$ and its related matrix information, SDM_d .

Step 2: Identify $S_{u,v}$ from $\text{SDM}(i, j)$.

Step 3: Select $u, v \in S_{u,v}$ and search for $(i, j)_{u,v} \forall (i, j) \in A$ through the procedure to determine ST (see the flow diagram in Figure 13.6); complete the procedure for ST for all $u, v \in S_{u,v}$.

Step 4: Store $\text{SDM}(i, j)$ and call SDM_d .

Step 5: Identify P_a (based on SDM_d).

Step 6: Construct SCP matrix for all $s \in \text{ST}$ (known from *Step 3*), based on P_a .

Step 7: Solve the SCP.

For *Step 1* of the Set-ST algorithm, various methods can be implemented; one of them is the Dijkstra's algorithm described in Appendix 10.A to Chapter 10. More algorithms can be found, for example, in Ahuja *et al.* (1993). *Steps 2, 4, 5 and 6* in the algorithm are self-explanatory by utilizing the abovementioned notations. *Step 3* deserves clarification and *Step 7* is a comment.

In *Step 3*, the search for $(i, j)_{u,v}$ is performed on any $(i, j) \in A$ for each u, v pair. This search enables one to check whether a candidate stop (i.e. whether the shortest distance to u and to v is less than or equal to ℓ) can be located at any point on G . The interpretations of $(i, j)_{u,v}$ point to the existence of a simple path on G (with no repetitions of either nodes or arcs along this path) between u and v , so that: (a) its length is less than or equal to 2ℓ ; (b) its mid-length point, p^* , lies on $(i, j)_{u,v}$; and (c) the shortest distance between u and p^* and between v and p^* is that along the simple path considered. In that last case, $p^* = s \in \text{ST}$. The search for $(i, j)_{u,v}$ is based on *Lemma 13.2*, below.

Before specifying *Lemma 13.2*, let us refer to a *special diagram* that was first demonstrated by Hakimi (1964) and further discussed by Handler (1973) and Minieka (1978). The x -axis in this *special diagram* represents the distance from node i to node j along (i, j) , and the y -axis represents the shortest distance from u and v to each point along (i, j) . The distances are represented by 45° lines that can be monotonically decreased or increased or can have a unique broken point of 90° along (i, j) .

Lemma 13.2: If k is a point lying on (i, j) and $D(u, k)$ and $D(v, k)$ are described by the *special diagram*, there are only four possible cases for $(i, j)_{u,v}$:

- (a) If the lines for $D(u, k)$ and $D(v, k)$ intersect, a candidate stop is located on (i, j) ;
- (b) If the lines for $D(u, k)$ and $D(v, k)$ do not intersect, a candidate stop cannot be located on (i, j) ;
- (c) If the lines for $D(u, k)$ and $D(v, k)$ are the same at one node, say i , and split toward the other node j , then i becomes a candidate stop;
- (d) If the lines for $D(u, k)$ and $D(v, k)$ are the same along each point $k, k \in (i, j)$, then both i and j are candidate stops.

Proof: The proof of this *Lemma* is almost straightforward with the use of the *special diagram* configurations, such as the one illustrated in Figure 13.5. If we follow the line (triangle shape) in part (a) of Figure 13.5 for $D(u, k)$, it can be seen that the shortest path from u to a point between i and k_1 is through i , whereas the shortest path from u to a point between j and k_1 is through j . It is also clear that $D(u, i) + d(i, k_1) = D(u, j) + d(i, j) - d(i, k_1)$ because this is the peak point (equal to the shortest distance from i and from j) of $D(u, k)$. Based on the interpretation for $(i, j)_{u,v}$, it becomes evident that point s satisfies all the requirements for a candidate stop. Part (b) of Figure 13.5 shows schematically that the solid line is the desirable simple path, whereas the dashed lines exhibit the direction of the shortest distance from u and v to other points than s that lie on (i, j) . The critical distance criterion can be checked by the length of the heavy solid line in part (a) of Figure 13.5. It is postulated that the length of this line should be less than or equal to 2ℓ (note that point s is the mid-point of the heavy solid line). The explanation for Figure 13.5 can be applied to all the other *special diagram* configurations; that is, all the possible combinations of the lines for $D(u, k)$ and $D(v, k)$ categorized by the cases of *Lemma* 13.2. For cases (c) and (d), these configurations will include equal lengths of $D(u, i)$ and $d(v, i)$ or $d(u, j)$ and $d(v, j)$.

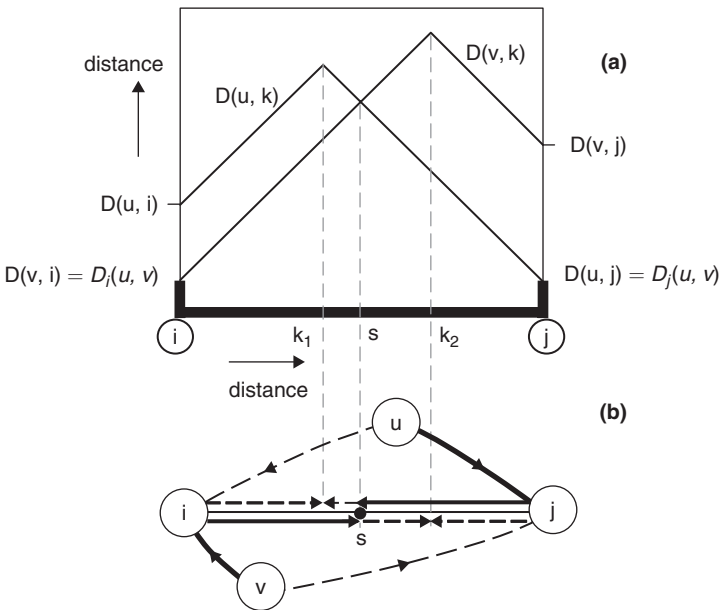


Figure 13.5 Example of a configuration for case (a) of *Lemma* 13.2 in order to search for $(i, j)_{u,v}$ and its corresponding candidate stop

The basic rules concerning the search for candidate stops (Set ST) constitute *Lemma* 13.2. These rules are integrated into a procedure to determine the set ST in Figure 13.6; this procedure exhibits *Step 3* of the Set-ST algorithm and examines each (i, j) , $(i, j) \in A$ for all u, v pairs. The set of arcs $(i, j) \in A$ should be arranged such that each node (except those connected by a single arc) is considered at least once as node i of one arc (i, j) . Note that if a node j is connected by a single arc (i, j) and can serve as a candidate stop for i (and perhaps

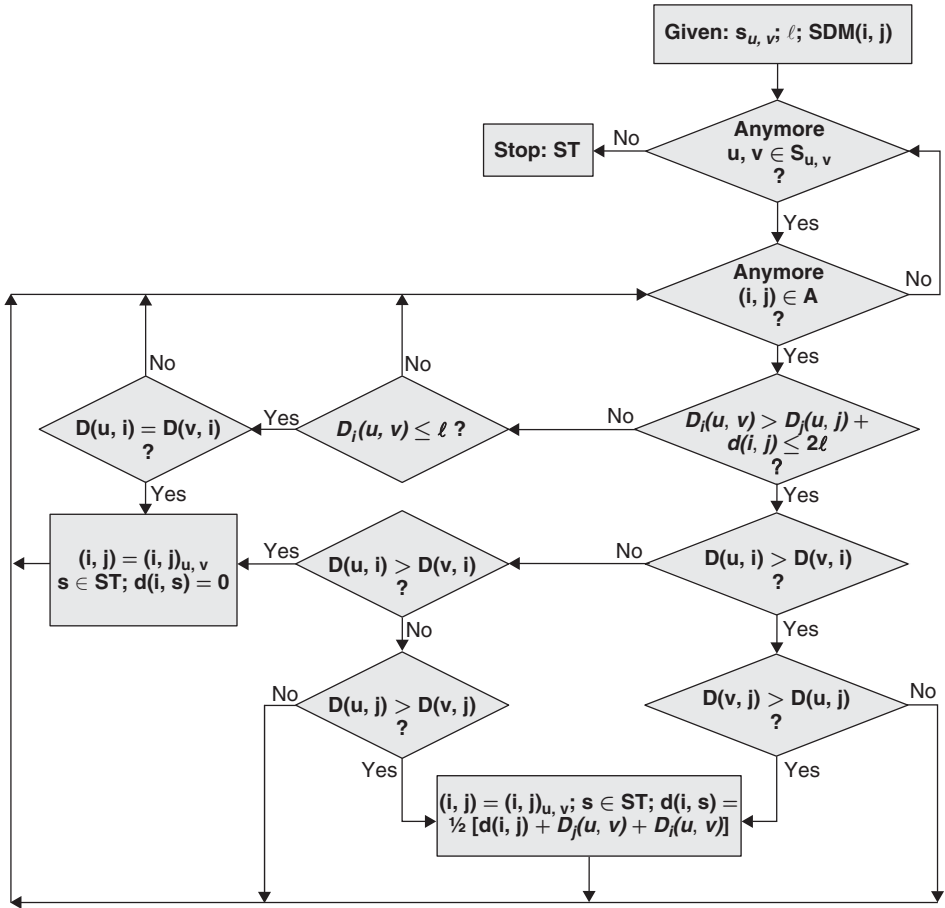


Figure 13.6 Determination of a set of candidate-stop locations

for other nodes), then it can be replaced as a candidate stop by node i . This efficient arrangement allows the omission of a special check of node j as a potential candidate stop, reduces the computation time and ensures a determination of all $s, s \in ST$, on G . Finally, in Figure 13.6, the critical distance is examined before the determination of $(i, j)_{u,v}$, an order of steps that can be changed if the value of ℓ is relatively small; e.g. $(i, j)_{u,v}$ can be determined first and then the critical distance examined.

Following are remarks pertain to *Step 7* of the Set-ST algorithm. The solution to large-scale SCPs often involved a considerable amount of computation time (Cristofides, 1975). SCPs have attracted intensive research attention, particularly because of their wide applicability; e.g. for airline crew scheduling. Rubin (1973) developed an effective heuristic algorithm for SCPs, and Balas and Padberg (1972) identified special properties of an SCP matrix that improve its solution through the Simplex method of linear programming (by avoiding degeneracy difficulties). Thus, *Step 7* can be based on known heuristics procedures in the OR field. At this stage, it is possible to set the following theorem:

Theorem 13.3: The Set-ST algorithm achieves an optimal solution for the problem of *minimum stops for a given critical distance* in a finite number of steps.

Proof: (1) *Optimality* is attained by construction; the SCP solution guarantees the selection of the minimum number of stops satisfying the distance constraint. The procedure in Figure 13.6, through *Lemma 13.2*, establishes that all candidate stops are to be considered for the analysis of the SCP. Handler (1973) showed that the optimal solution for the problem combines only candidate sources. These simple arguments suffice.

(2) *Finite convergence* is attained because an iteration of the Set-ST algorithm, which is performed only through the procedure in Figure 13.6, clearly indicates a finite operation ($S_{u,v}$ – see Figure 13.6 – is based on a finite number $|A|$).

13.4.3 Numerical example

Let us consider a small network of roads for further explanations of the Set-ST algorithm. This example is illustrated in Figure 13.7. The critical distance is $\ell = 500$ metres. The shortest-distance matrix is given in Table 13.4, with the corresponding path (list of sequential nodes) in parentheses; those distances forming $S_{u,v}$ are in bold.

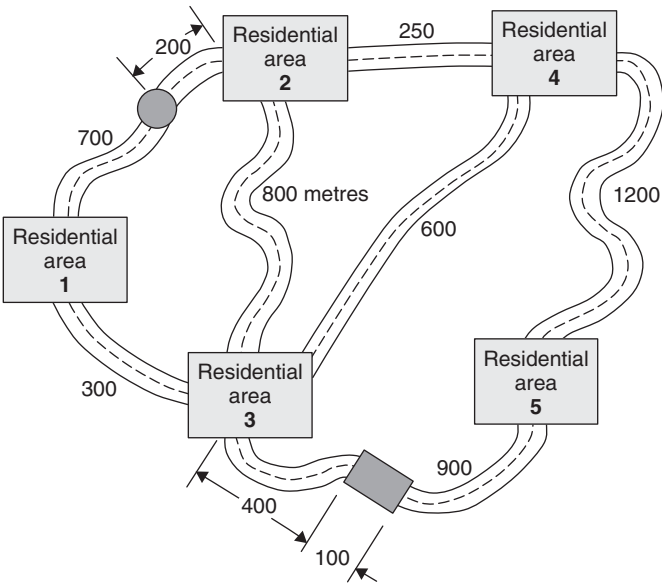


Figure 13.7 Example of a network of roads consists of five nodes (residential areas) and seven road segments, with lengths in metres; two stops (dark grey coloured) are recommended at a specific location along a stretch of 100 metres

The procedure (Step 3 of the algorithm) for Figure 13.6 is shown in Table 13.4. For each u, v pair, there is an X-shaded symbol for those $(i, j) \neq (i, j)_{u,v}$; for each $(i, j)_{u,v}$, the value of $DL = D_i(u, v) + D_j(u, v) + d(i, j)$, which should be less than or equal to $2\ell = 1,000$ metres, is indicated in the upper part of each filled cell in Table 13.5. The path $P_{u,v}$ is shown in the lower cell part for $(i, j)_{u,v}$. In this example, none of the candidate stops is located on a node.

Table 13.4 Shortest-distance matrix (in metres) of the road network example in which the corresponding path (by node number) appears in parentheses

$i \backslash j$	1	2	3	4	5
1	0	700 (1–2)	300 (1–3)	900 (1–3–4)	1200 (1–3–9)
2		0	800 (2–3)	300 (2–4)	1500 (2–4–5)
3			0	600 (3–4)	900 (3–5)
4		Symmetrical		0	1200 (4–5)
5					0

Table 13.5 Results of the procedure in Figure 13.6 for the road network example

$(i, j) \in A$ $u, v \in S_{u,v}$	(1, 2)	(1, 3)	(2, 3)	(3, 4)	(3, 5)	(4, 2)	(5, 4)
1, 2	700 (1–2)	X	X	X	X	X	X
1, 3	X	300 (1–3)	X	X	X	X	X
1, 4	1000 (1–2–4)*	X	X	900 (1–3–4)	X	X	X
2, 3	1000 (2–1–3)*	X	800 (2–3)	900 (3–4–2)	X	X	X
2, 4	X	X	X	X	X	300 (4–2)	X
3, 4	X	X	X	600 (3–4)	X	X	X
3, 5	X	X	X	X	900 (3–5)	X	X

* This $P_{u,v}$ is not the shortest-distance path between u and v

The information and intermediate results related to $s \in ST$ are shown in Table 13.6, including the SCP matrix. The optimal solution combines the two circle columns in Table 13.6, in which $(i, j)_{u,v} = (3, 5)_{3,5}$ and $(1, 2)_{1,4}$. It is interesting, and important for planning purposes, to note that the difference between 2ℓ and the DL distance indicated in Table 13.6 allows for flexibility of stop location. That is, the optimal stop lying on arc (3, 5) can be located 50 metres closer towards either node 3 or node 5 without violating the optimal solution. On the other hand, the optimal stop on (1, 2) can be located only at one point (200 metres away from node 2 on that arc).

Table 13.6 Information and results required for the SCP analysis of the road network example (first four rows) and the SCP matrix (last five rows)

u, v	1, 2	1, 3	(1, 4)	1, 4	2, 3	2, 3	2, 3	2, 4	3, 4	(3, 5)	1,1	2,2	3,3	4,4	5,5
DL*	700	300	1000	900	1000	800	900	300	600	900	0	0	0	0	0
(i, j)_{u,v}	(1, 2)	(1, 3)	(1, 2)	(3, 4)	(1, 2)	(2, 3)	(3, 4)	(2, 4)	(3, 4)	(3, 5)	0	0	0	0	0
d(i, s)	350	150	500	150	200	400	450	150	300	450	0	0	0	0	0
1**	1	1	1	1	1						1				
2	1		1		1	1	1	1				1			
3		1		1	1	1	1		1	1			1		
4			1	1			1	1	1					1	
5										1					1

* $DL = D_i(u, v) + D_j(u, v) + d(i, j)$, in metres; if s is on node I, then $DL = 2D_i(u, v)$

** Node number for the rows of the SCP matrix

In addition, it is possible to observe – or to obtain from the SCP matrix in Table 13.6 – the optimal solution for $\ell < 500$. For example, for $150 \leq \ell < 500$, three stops will be required. Practically speaking, it makes sense to have the objective of maximum location flexibility; such an objective yields the solution $(i, j)_{u,v} = (1, 3)_{1,3}, (2, 4)_{2,4}$ and $(5, 5)_{5,5}$, where the last is node 5.

This section approached the problem of transit-stop location in a network in which the nodes are located in concentrated population areas. A more challenging problem distributes the origins of passenger demand along the arcs (the continuous case) rather than in specific nodes. Nonetheless, the Set-ST algorithm can also be used as a preliminary tool to assess the location of transit stops in such a continuous case. For instance, artificial nodes (creating a new set of arcs) could be added, such that the distance between two adjacent nodes on the same arc would not be more than a given value.

13.5 Connectivity measures and analysis

Improving transit connectivity is one of the most vital tasks in transit-operations planning. A poor connection can cause some passengers to stop using the transit service. Service-design criteria always contain postulates to improve routing and scheduling coordination (intra- and inter-agency transfer centres/points and synchronized/timed transfers). Ostensibly the lack of well-defined connectivity measures precludes the weighing and quantifying of the result of any coordination effort. This section provides an initial methodological framework and concepts for (i) quantifying transit connectivity measures, and (ii) directions and tools for detecting weak segments in inter-route and inter-modal chains (paths) for possible revisions/changes.

Before proceeding with the main themes of this section, let us review a perceived concept of transit connectivity. One possible definition of a prudent, well-connected transit

path is this: *an advanced, attractive transit system that operates reliably and relatively rapidly, with smooth (ease of) synchronized transfers, part of the door-to-door passenger chain.* Interpretation of each component of this definition follows.

Attractiveness: Available information (by telephone, Internet, newspaper, radio, TV, mail leaflet), simple communication (abbreviated telephone number, automatic storage of users' telephone number and address), clear service meeting-point characteristics (clear stop sign, vehicle colour and logo), boarding/alighting/riding comfort (low-floor, extra space next to driver, comfortable seats, features for physically challenged people, low noise), on-board service (news-papers, magazines, free coffee/tea, TV/video display of timetable, weather, etc.), simple payment (electronic ticketing, pre-paid, transfers, smart-card ticketing).

Reliability: Small variance of measures of concern to passengers (total travel time, waiting time, in-vehicle time, seat availability), small variance of measures of concern to transit vehicles (schedule adherence, headways, on-time pullouts, missed trips, breakdowns, load counts, late reports), small variance of measures related to pre-trip information using telephone communication (on-line timetable, travel time to caller); for more on reliability features, see Chapter 17.

Rapidly: Easy access/egress and comfortable stops (fixed stops with shelters and information, transit-vehicle bays at timepoints, with an extra approach lane at signalled intersection), transit-vehicle preference at unsignalled intersections ('yield' or 'stop' not according to traffic procedures, special bypass arrangements at strategic points), transit vehicle preference at signalled intersections (passive priority by extending or shortening green, active priority using actuated transit-vehicle signals – e.g. radio, inductive loop), purchase and validation of tickets (electronically, ordinary) on transit vehicles (one-way, round trip, transfer, daily, weekly, monthly).

Smoothness (ease): Comfortable routing (maximum criterion for walking distance, round-trip deviation from designated route in bad weather, evolution of flexible routing and scheduling), special train-station entrance (transit-vehicle special gate, passengers' special entrance door with comfortable stairs/escalator to/from the train platform), special train exit (exit door next to the train platform for transit-vehicle ticket holders, transit vehicle waiting at exit or under shelter with vehicle-arrival announcement on variable message signs – VMS).

Synchronized: On-line communication between all modes of transit vehicles (vehicle equipped with arrival information for the relevant stations and time difference, positive or negative, for synchronization), passenger subscriptions with serial numbers (adding a variable scheduling element to suit subscribers, planning the fixed-scheduling component with subscriber information), short-turn and short-cut routing strategies (computerized suggestions for the transit-vehicle driver on short-turn and short-cut, VMS on-board information on meeting time with another or the same transit mode); for more on routing strategies and timed transfers, refer to Chapters 16 and 18, respectively.

13.5.1 Developing connectivity measures

The common denominator for all transit services are the following quality-of-connectivity attributes:

e_1 = Average walking time (for a connection)

e_2 = Variance of walking time

- e_3 = Average waiting time (for a connection)
 e_4 = Variance of waiting time
 e_5 = Average travel time (on a given transit mode and path)
 e_6 = Variance of travel time
 e_7 = Average scheduled headway
 e_8 = Variance of scheduled headway.

These eight attributes, which can be measured, will be termed *quantitative attributes*.

However, there are other important attributes that cannot easily be quantified and measured. Three of these are:

- e_9 = Smoothness (ease)-of-transfer (on a given discrete scale)
 e_{10} = Availability of easy-to-observe and easy-to-use information channels (on a given discrete scale)
 e_{11} = Overall intra- and inter-agency connectivity satisfaction (on a given discrete scale).

These hard-to-quantify attributes will be termed *qualitative attributes*. Nonetheless, it is true that the value of all 11 attributes may be perceived differently by different passengers or even by the same passenger in different situations. These different perceptions are captured in the average weighting of each attribute. The weight of each attribute is survey-based and/or based on the results of a mode (path)-choice model. The analysis framework proposed in this section distinguishes between quantitative and qualitative attributes, although both can be combined with some agreeable weighting factors. The reason for this separation is to make it easy for decision-makers to evaluate improvements and changes in the transit-connectivity chains.

As noted above, measuring transit connectivity involves various parameters and components. Therefore, the following notations are introduced in order to ease the explicit construction of connectivity measures.

Notations

For a given time window (e.g. peak-hour, average week-day):

- $O = \{O_i\}$ = set of origins O_i
 $D = \{D_u\}$ = set of destinations D_u
 $P_{Dk} = \{P\}$ = set of inter-route and inter-modal paths to D_k
 $P_{Ok} = \{P_i\}$ = set of inter-route and inter-modal paths from O_k
 $M_p = \{m\}$ = set of transit routes and modes included in path p
 t = index of quantitative attributes
 ℓ = index of qualitative attributes
 $E_t = \{e^t\}$ = set of quantitative attributes suitable for connectivity measures
 $E_\ell = \{e^\ell\}$ = set of qualitative attributes suitable for connectivity measures
 e^j_{mp} = the value of attribute e^j , $j = t, \ell$, related to mode m on path p
 α_e = weight/coefficient for each attribute e^j , $j = t, \ell$
 c^j_p = quantitative and qualitative ($j = t, \ell$) connectivity measure of path p
 F_p = average number of passengers using path p
 $c_p(i, j)$ = capacity (flow of passengers) of arc (i, j) between route and mode i , and between route and mode j ; each i can also be an origin O_i or destination D_i ; (i, j) is contained in path p and is part of a network-flow model.

Based on the notations above, the following equation-based notations are established:

$$C_p^j = \sum_{m \in M_p} \sum_{e^j \in E_j} \alpha_e e_{mp}^j, \quad j = t, \ell \quad (13.3)$$

$$C_{Dk}^j = \sum_{p \in P_{Dk}} c_p^j, \quad j = t, \ell \quad (13.4)$$

$$C_{Ok}^j = \sum_{p \in P_{Ok}} c_p^j, \quad j = t, \ell \quad (13.5)$$

$$C_D^j = \sum_{D_k \in D} C_{Dk}^j, \quad j = t, \ell \quad (13.6)$$

$$C_O^j = \sum_{O_k \in O} C_{Ok}^j, \quad j = t, \ell \quad (13.7)$$

$$c_p^{jF} = c_p^j \cdot F_p, \quad j = t, \ell \quad (13.8)$$

$$C_{Dk}^{jF} = \sum_{p \in P_{Dk}} c_p^{jF}, \quad j = t, \ell \quad (13.9)$$

$$C_{Ok}^{jF} = \sum_{p \in P_{Ok}} c_p^{jF}, \quad j = t, \ell \quad (13.10)$$

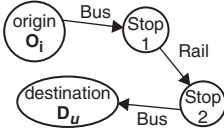
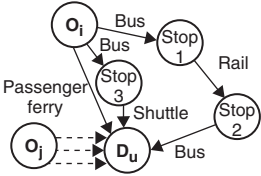
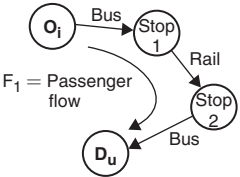
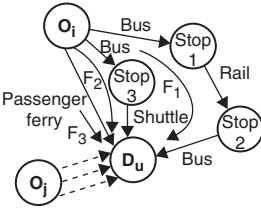
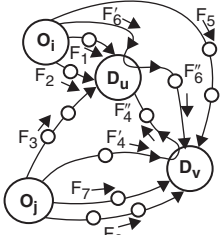
$$C_D^F = \sum_{D_k \in D} C_{Dk}^{jF}, \quad j = t, \ell \quad (13.11)$$

$$C_O^F = \sum_{O_k \in O} C_{Ok}^{jF}, \quad j = t, \ell \quad (13.12)$$

Note that the required weight/coefficient α_e in Equation (13.3) for measuring the level/quality/goodness of connectivity in Equations (13.3)–(13.12) can be estimated by the results of both passenger surveys and the path/mode-choice model (some of which may need to be constructed, being site-specific).

The interpretation, significance, and application of each of the proposed connectivity measures appear in Figure 13.8. This figure also presents graphical examples of the relevant inter-modal path for each measure and, under the applications, the purpose of each measure.

The purpose of the first connectivity measure in Figure 13.8, Equation (13.3), is to compare paths; that is, chains of trips, each from an origin to a destination, including transfers. Usually this comparison takes place following an improvement or change in one or more paths. Otherwise, this comparison, using the evaluation tool proposed, categorizes the different paths by their access/egress connectivity quality. The purpose of the second connectivity

Connectivity measure	Interpretation	Significance	Graphical example	Application
$c_p^j = \sum_{m \in M_p} \sum_{e \in E_j} \alpha_e \cdot e_{mp}, j = t, \ell$	Sum of all connectivity-component measures along a given path k	Overall connectivity value or quality for a given path k		Evaluation of path k access-connectivity value. Purpose: comparison among paths
$C_{Dk}^j = \sum_{p \in P_{Dk}} c_p^j, j = t, \ell$	Sum of all connectivity-component measures along all access paths to a given destination D_u	Overall connectivity values for all paths related to D_u		Evaluation of destination D_u access-connectivity value considering only existing paths (for new paths, C_{Dk}^{jF} are applicable). Purpose: comparison among destinations
$c_p^{jF} = c_p^j \cdot F_p, j = t, \ell$	Sum of all connectivity-component measures along a given path k, weighted by the average amount of passengers using this path	Overall exposure-connectivity measure (for all passengers) for a given path k		Evaluation of path k people-access-connectivity value (considering amount of passengers exposed). Purpose: comparison among paths, considering passenger flow
$C_{Dk}^{jF} = \sum_{p \in P_{Dk}} c_p^{jF}, j = t, \ell$	Sum of all connectivity-component measures along all access paths to D_u , weighted by the average amount of passengers using these paths	Overall exposure-connectivity measure for all paths related to terminal D_u		Evaluation of terminal D_u people-access-connectivity value, considering exiting and new paths. Purpose: comparison among destinations, considering passenger flow
$C_D^F = \sum_{D_k \in D} C_{Dk}^{jF}, j = t, \ell$	Sum of all connectivity-component measures along all access paths to a set of destinations, weighted by the average amount of passengers using these paths	Overall exposure-connectivity measure for all paths in a given set of destinations		Evaluation of a set of destinations, considering exiting and new paths. Purpose: comparison among airport destinations in a given set (e.g. zone-based, purpose-based)

Note: The measures considered can also apply to qualitative attributes

Figure 13.8 Interpretation, significance and application of quantitative O-D connectivity measures at different levels and for a given time period

measure, Equation (13.4), is to compare destinations. The purposes of the third and fourth measures, Equations (13.8) and (13.9), are the same as for the first and second measures, respectively, but with the consideration of passenger flow; that is, a determination of the average number of passengers exposed to the calculated level of connectivity. The purpose of the fifth measure in Figure 13.8, Equation (13.11), is to compare groups of destinations in regard to overall existing connectivity quality.

All connectivity measures that consider passenger flows should be updated, following routing and/or scheduling and/or service improvements or changes. When referring to a group of destinations (zonal-based, purpose-based) in the last measure in Figure 13.8, paths can have a stop at one destination and continue to others. For example, in the last-row graphical representation in Figure 13.8, the access path with F'_4 and F''_4 flows has a stop at D_v for F'_4 flow between O_j and D_v , and for F''_4 flow between D_v and D_u .

The process for the determination of transit connectivity measures, using the notations in Equations (13.3)–(13.12), is shown in Figure 13.9. This figure, constructed by an input-component-output form, shows the systematic decision sequence and process of the analysis. The output of each component positioned higher in the sequence becomes an important

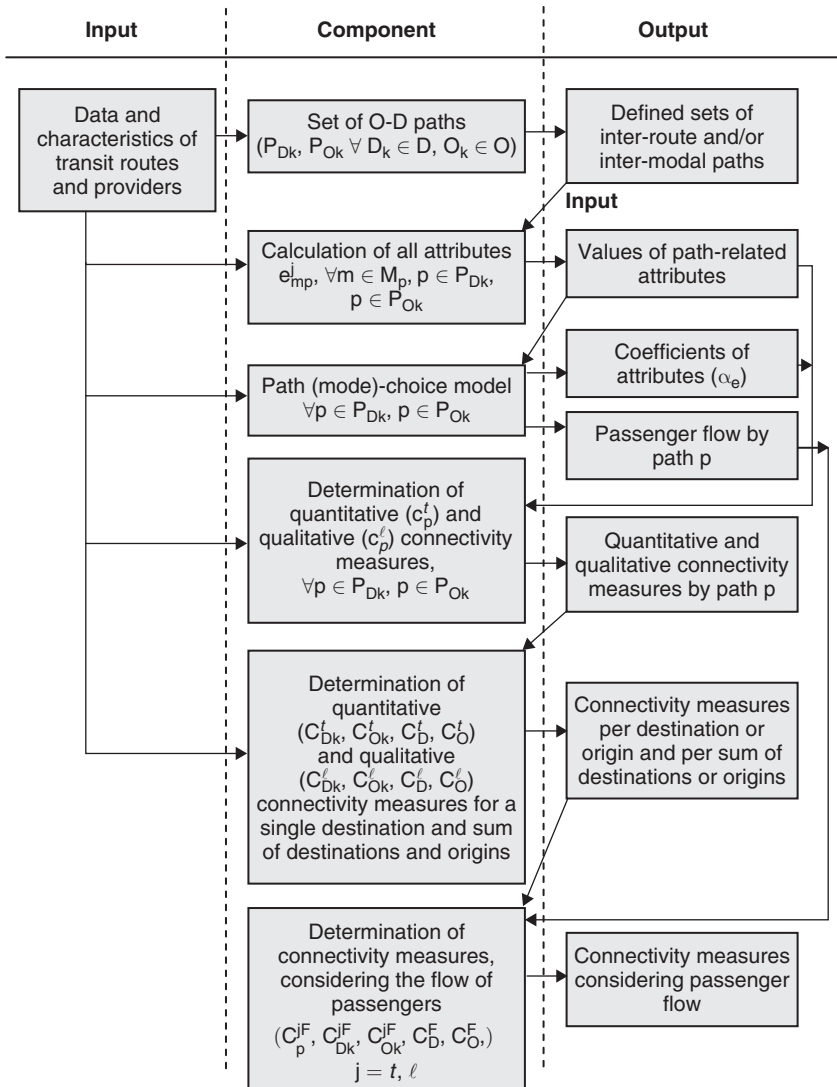


Figure 13.9 Analysis framework for the determination of connectivity measures

input into lower-level decisions. Clearly, the independence and orderliness of the separate components exist only in the diagram.

13.5.2 Detecting weaknesses on inter-route and inter-modal paths

An important aspect of a tool for transit-connectivity improvements is the ability to detect current or anticipated weaknesses in inter-route and/or inter-modal chains/paths. Figure 13.10 portrays a possible process identifying these weaknesses, based on the defined notations. Whereas the first two lists (output in Figure 13.10) do not consider passenger flow, the third and fourth lists do take this flow into account. The fourth list attempts to identify passenger-flow

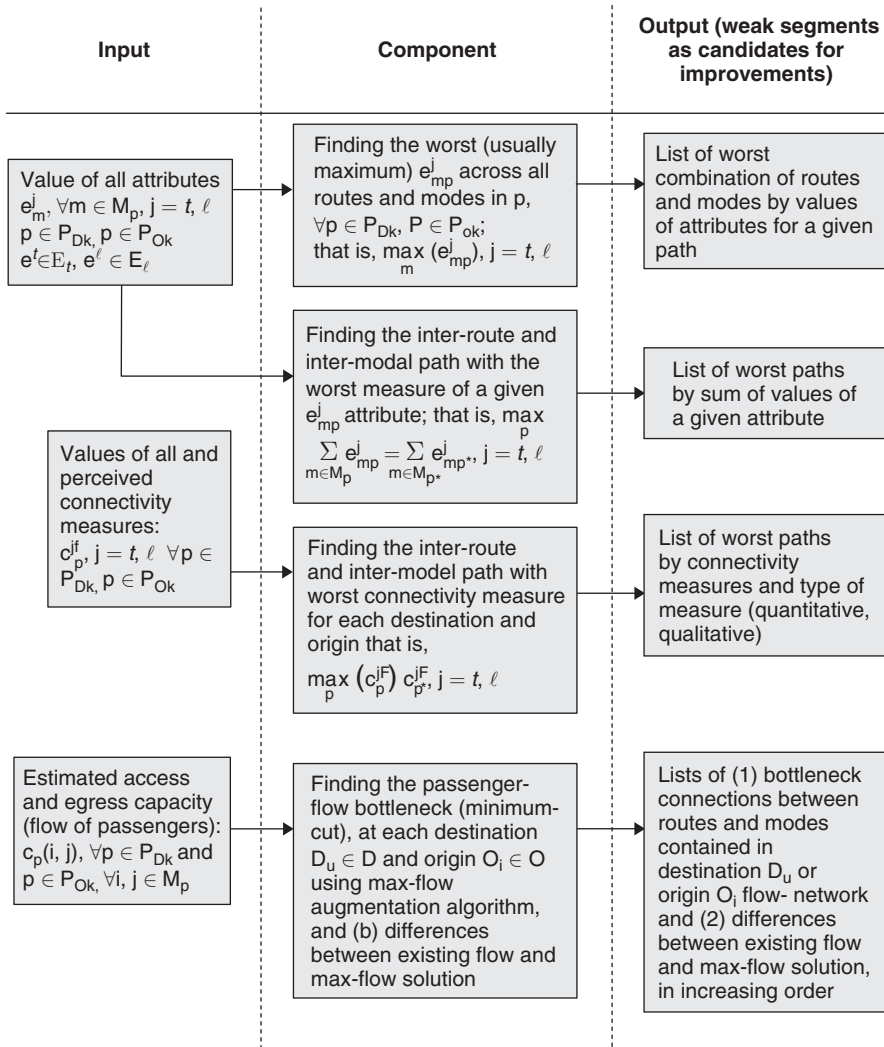


Figure 13.10 Analysis framework for the detecting weak segments on inter-route and inter-modal paths

bottlenecks at the network level. The process of attaining the fourth list is explained in the next section.

The first list of weak segments in Figure 13.10 is constructed by means of the value of each attribute across all routes and modes on a given path. That is, the worst route (operated by a certain mode) for each attribute on a given path is identified and can automatically be considered a candidate for improvement. If improvement cannot be made, the second worst route is examined and so on. This process does not consider/weigh the amount of passenger flows on each path. It simply identifies weak segments (owing to their inherent features), regardless of how many passengers use them.

The second list for detecting weak segments in Figure 13.10 considers the sum of the value of each single attribute for all routes on a given path; for example, the sum of all waiting times along a given path. The worse path, p^* , is identified and can serve as a candidate for improvements or changes in the specific attribute. In practice, passenger flow could also be considered in order to capture behavioural elements, although the inherent connectivity features of each path are not dependent on this flow.

The third list in Figure 13.10 considers the exposure of passenger flow to all inter-route and inter-modal connectivity features on each access/egress path. A list of worst paths can then be established and treated accordingly. The next section explains and interprets this third list.

13.5.3 Finding bottlenecks in inter-route and inter-modal transit networks

The last process in Figure 13.10 intends to find weak connection links in the sense that those links represent potential bottlenecks for passenger-access/egress flows. The basic idea behind this approach is to review all inter-modal activities and paths on a specially constructed flow (of passengers) network. In any flow process (vehicles, people or freight) on a network, problems arise when the flow demand exceeds the capacity. These traffic problems, well known in road traffic, are rooted in the bottlenecks of the system, some of which are dynamic in nature (e.g. a slow truck in road traffic) and are not necessarily dependent on physical characteristics. Our problem is to find the bottleneck obstructing passenger flow in a connected transit network with required passenger transfers.

A specially constructed passenger-flow network is illustrated in Figure 13.11. This network contains two sets of elements: set N for nodes i , $i \in N$, and set A for arcs (i, j) , $(i, j) \in A$. The network-flow model in Figure 13.11 consists of three subsets of nodes: first, origin or destination nodes; second, nodes related to transit routes and modes; third, a node related to a single origin or destination. Each arc of this network has a capacity value for either access or egress flow; the arc is associated with a certain inter-route and inter-modal path. In the first subset of nodes, there is a source (origin) or sink (destination) node connected artificially to all the origin/destination nodes. The process is to load this special network with either access – or egress – passenger flows while having a capacity constraint on each arc.

Passenger-access flows originate in the artificial node AO in Figure 13.11, and move to reach the destination D_u ; passenger-egress flows move from AD to O_i . Finding bottlenecks requires maximizing the flow created at AO/AD and absorbed in D_u/O_i . In the maximum-possible-flow solution, those arcs having a flow equal to capacity naturally form the bottlenecks of the network.

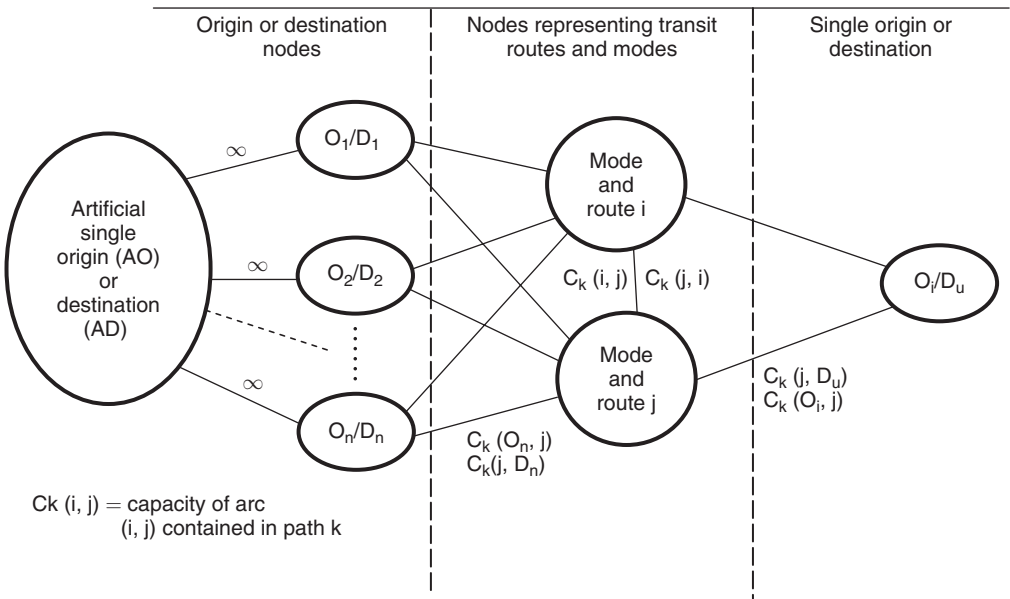


Figure 13.11 Conceptual O-D connectivity network-flow model

A vital element of the network-flow model is the capacity assigned to each arc; in other words, the maximum amount of passengers, for a given time period (peak-hour, average day), that can traverse each arc. Generally, there is a finite number of arcs in an actual transit network of routes and modes. For each path, there is need to search the capacity (passengers/hour) that can be offered by each route and mode; for instance, the maximum feasible rail/bus/passenger ferry frequency times the vehicle’s passenger capacity.

The solution for the maximum-flow (max-flow) problem, corresponding to the network-flow model in Figure 13.11, appears in Appendix 7.A of Chapter 7; it is based on the max-flow augmentation algorithm. An example of a passenger-flow transit network is shown in part (a) of Figure 13.12. This is an example of access inter-modal paths from a given origin to two destinations, D₁ and D₂. There are five transit modes, and the capacities (passengers/hour) are the numbers on the arcs. Based on the max-flow min-cut theorem in Appendix 7.A, the bottlenecks of the network-flow model are the arcs along the minimum cut. That is, the solution is a maximum produced flow of 1,500 passengers/hour. The max-flow augmentation algorithm is based on the seven following paths (in each, more pass/hr flow is added):

1. Origin - Bus - D ₁ - AD,	augmented (aug.) flow = 400
2. Origin - Bus - D ₂ - AD,	aug. flow = 200
3. Origin - Bus - Passenger ferry - D ₂ - AD,	aug. flow = 100
4. Origin - Shuttle bus - Rail - D ₁ - AD,	aug. flow = 300
5. Origin - Shuttle bus - Cable car - Rail - D ₂ - AD,	aug. flow = 250
6. Origin - Shuttle bus - Cable car - Passenger ferry - D ₁ - AD,	aug. flow = 150
7. Origin - Shuttle bus - Cable car - Passenger ferry - D ₂ - AD,	aug. flow = 100
Total flow = 1500	

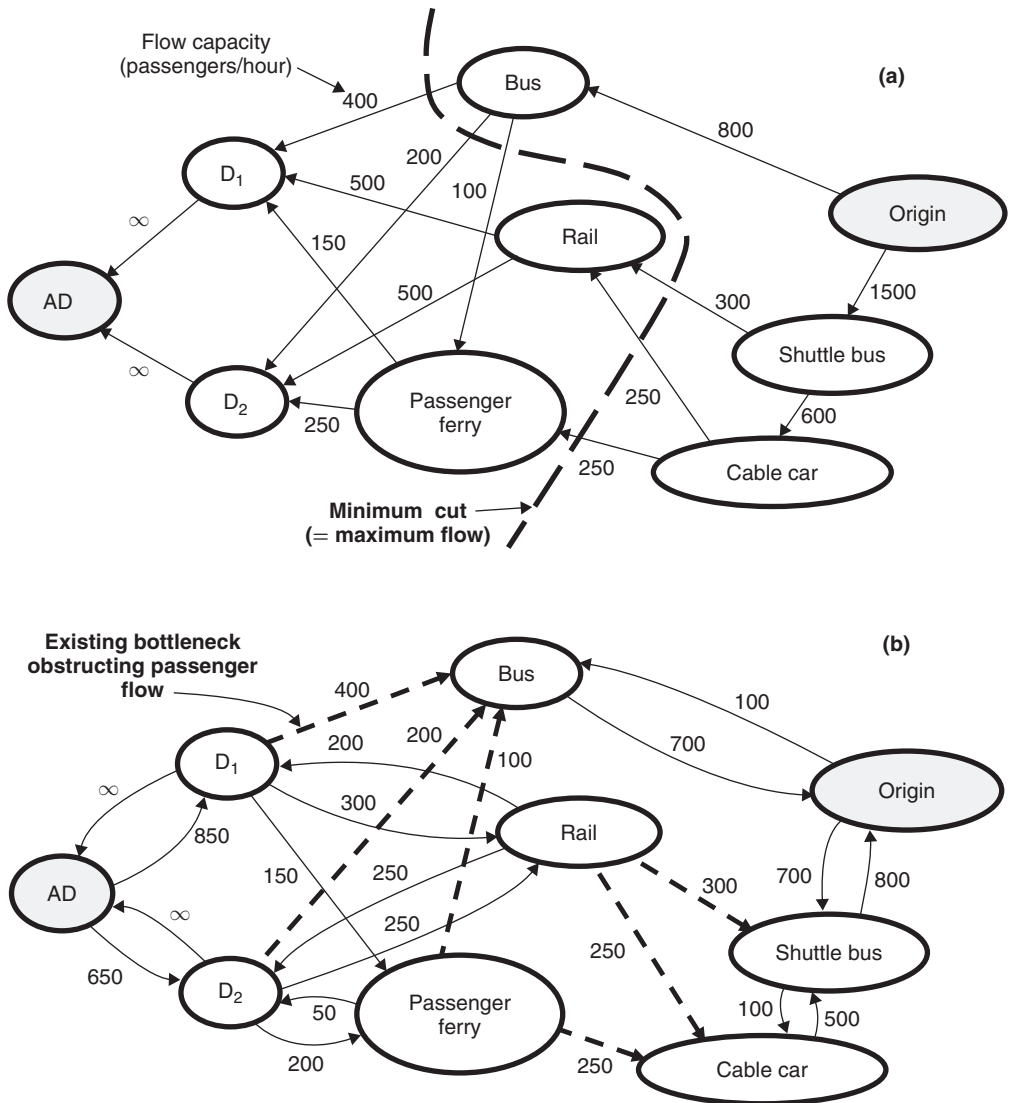


Figure 13.12 Example of the network-flow model and its minimum cut in part (a), and of the last augmented network flow in part (b)

The last augmented network-flow, based on the procedure in Appendix 7.A, appears in part (b) of Figure 13.12. The sum of augmented flows on (i, j) in Figure 13.12(b) is represented by a reverse arc (j, i) . The difference in flow between the capacity on (i, j) and this sum of flows is represented by a forward arc. Only those arcs with a flow equal to the capacity do not have forward arcs in Figure 13.12(b). Those arcs that are emphasized are the bottleneck segments.

Following is an additional procedure for the best exploitation of the results of the max-flow algorithm.

1. Solve the max-flow algorithm for maximum capacity ranges.
2. Find differences between the max-flow solution(s) and existing flow (representing existing demand).
3. Check for any current bottleneck, in which existing demand is higher than the optimal max-flow solution; if such exists, suggest immediate improvement.
4. Examine weak segments in the existing inter-route and inter-modal paths (see Figure 13.10).
5. Propose service modifications for actual and perceived attributes.
6. Update relevant attributes using path (mode)-choice model.
7. If no changes are made, continue; otherwise, go to 1.
8. Set final conclusions.

Practically speaking, it is recommended that the following data be collected from each relevant transit agency. These data constitute the base for capacity calculation, estimation of attributes, and the creation of inter-route and inter-modal connectivity proposals.

1. *Frequencies* – existing, minimum (estimated), maximum (estimated).
2. *Vehicle sizes* – existing (number of seats, allowed standees); which (and when, if not continuously) sizes are used.
3. *Passenger load* – by time-of-day.
4. *Average travel time* – between each two nodes connected by time-of-day.
5. *Routing* – existing; flexibility for routing adjustments.
6. *Routing strategies* – existing; flexibility for adding strategies (e.g. skip-stop, short-turn, short-cut).

A good example of transit-connectivity paths is an airport terminal. Table 13.7 demonstrates access and egress paths of transit service to an airport terminal.

Table 13.7 Access/egress transit modes and paths to airport terminal

Path	Access/egress mode	Path description	Notes
k ₁	Train service	Auto- <i>wait</i> -train- <i>wait</i> -shuttle-terminal	
k ₂	Public transit bus	Walk- <i>wait</i> *-public bus-terminal	
k ₃	Scheduled airport bus	Auto- <i>wait</i> -airport bus-terminal	
k ₄	Shared-ride door-to-door van**	Van-terminal	No waiting, but not a straight ride
k ₅	Charter bus	Auto- <i>wait</i> -charter bus-terminal	Special ride

* There is an additional wait for each transfer between two buses (if any) in the path
**Additional distance and time for passengers using van, by time of day
Note: Private vehicle access/egress to and from train, airport bus, and charter bus is assumed; passengers are assumed to walk to and from a public bus.

13.6 Literature review and further reading

This section reviews research works that discuss various indicators of the level of transit service (LOS). Such indicators enable a measurement of the service quality of both existing transit systems and proposed schemes and, therefore, serve as important planning tools. Note that transit coordination/connectivity studies are reviewed in Chapter 6, and reliability studies in Chapter 17.

Alter (1976) introduces a measure of transit LOS whose value is affected by accessibility, travel time, reliability, service directness, service frequency and on-board density criteria. A given set of tables and conversion tables assigns grades, from A to F, to each criterion. The accessibility grade is based on the walking time or distance to and from a stop. The reliability grade is based solely on frequency. Travel time is evaluated according to the ratio of transit travel time to car travel time on the same route. The grade for directness depends on a combination of the number of required transfers and waiting time. Grades for peak and off-peak frequencies are weighted by population density in the area served. Grades for on-board density depend on the average area for a single passenger. All six measures are aggregated into an overall LOS indicator. Some of the criteria use qualitative considerations, but the whole process is generally quantitative. Although sensitive to various criteria, the proposed evaluation methodology seems fairly simplistic.

Horowitz (1981) proposes an indicator for bus-service quality that incorporates the main performance variables as perceived by transit riders: in-vehicle time, transfer time, walking time and waiting time. The need to wait and the need to transfer, which are known to have a significant effect on passengers' perception of LOS, are accounted for by both a numerical variable and a dummy variable. To take equity considerations into account, LOS measures are calculated separately for different ridership sectors, and then weighted and summarized. The weights used represent the equivalent travel time of a private car. The final LOS rating depends on the number of people within each population sector who can travel within a certain travel-time standard.

Polus and Shefer (1984) measure bus LOS by the ratio of average bus travel time to the average car travel time on the shortest distance between the same origin and destination. The model they propose for making forecasts of this LOS indicator is based on physical or traffic-related variables: route length, stop spacing, number of intersections on route, etc. It can, therefore, be used when detailed information about bus performance or demand is unavailable, which makes their model useful for analysing suggested route changes.

Madanat *et al.* (1994) use an ordered probit model to correlate discrete responses from a bus-passenger survey with a quantitative variable representing on-board crowding levels. The ordered probit technique assists in identifying the thresholds between successive ratings presented to the respondents. The authors compare their analysis to the different on-board crowding levels that were used in the HCM manual at that time to measure transit LOS. The results are very different from the HCM values.

Henk and Hubbard (1996) propose a measure for the availability of bus or rail services that takes into account service considerations: coverage, frequency and capacity. Various alternative measures of each of these considerations are discussed, and then preferred indicators are chosen: coverage is calculated by the number of directional route-kilometres per square kilometre; frequency is indicated by the amount of vehicle-kilometres per route-kilometre; and capacity is measured by seat-kilometres per capita. The values of the three

components in each zone of the network studied are normalized by subtracting the mean and dividing by the standard deviation; an average of the three components is then computed.

Murugesan and Rama Moorthy (1998) develop an index for bus LOS. Survey results are processed using the theory of fuzzy sets, which enables a quantitative analysis based on qualitative rating responses. Twenty LOS attributes are included in the analysis, but the authors focus on a description of the fuzzy-sets methodology, not the LOS indicators.

Friman *et al.* (1998) study a database of transit riders' complaints and incidents in order to determine which LOS attributes are perceived as the most important. The analysis is qualitative, and the authors conclude that the service features most important to passengers are a driver's behaviour and service reliability. The simplicity of the transit network and the provision of sufficient information are also found to be important.

A TCRP report (1999) includes a detailed discussion of LOS and capacity issues for any transit mode. The calculation of LOS takes account of passenger loads and the accessibility of the service. In addition to the analysis of a system-wide LOS, there is a discussion of LOS considerations for a specific stop or for a specific route segment.

Prioni and Hensher (2000) develop a LOS model based on stated-preference survey data that depict the influence of various factors on the LOS as perceived by bus users. Factors found to have the highest impact on LOS perception are the tariff, travel time and access time to the stop. Driver friendliness and the smoothness of the bus ride are also found to be quite significant. The authors examine the differences in the LOS perception of buses operated by different companies and find that there is some bias towards certain operators; they show how the index developed could be used in the contracting process to monitor bus operators and performance.

Ryus *et al.* (2000) introduce a measure for transit availability based on the concept that transit routes serve a small group of potential users. The measure is defined as the percentage of time in which transit service is available and accessible to an average person within a reasonable walking distance. Computation of this indicator requires the use of a geographical information system. Frequency, service coverage, stop location and operation hours influence the value of the indicator; population density affects it, too, but does not influence the demand for transit services.

Guttenplan *et al.* (2001) propose a methodology for determining LOS in a combined bus, cycling and walking network. The bus LOS is influenced by common factors, such as frequency, but also by inter-modal factors that are unique to this method, such as difficulty in crossing the road or the connection between the bus stop and the pavement.

Yang and Wang (2001) evaluate LOS on the traffic-zone level, based on the Fuzzy-c Means method. Using geographical information system functions, the authors compute each zone's characteristics, such as its route-network density and population density. The Fuzzy-c Means method is used to identify clustering patterns and to aggregate the detailed measurements into composite, zone-level LOS indicators.

The major features of the LOS indicators reviewed are presented in Table 13.8.

Exercises

- 13.1 Given the following timetable with 7 bus trips, 3 departure points, and shifting tolerances of ± 5 minutes and deadheading (DH) travel time of 25 minutes for all

Table 13.8 Summary of the characteristics of the models reviewed

Source	Modes	Quantitative?	Factors influencing LOS
Alter (1976)	All transit modes	No	Accessibility, travel time, reliability, service directness, service frequency, on-board density
Horowitz (1981)	Bus	Yes	In-vehicle time, transfer time, walking time, waiting time
Polus and Shefer (1984)	Bus	Yes	External factors (route length, stop spacing, number of intersections on route, etc.)
Madanat <i>et al.</i> (1994)	Bus	Partially	On-board density
Henk and Hubbard (1996)	All transit modes	Yes	Coverage, frequency, capacity
Murugesan and Rama Moorthy (1998)	Bus	Partially	(Twenty factors are discussed)
Friman <i>et al.</i> (1998)	All transit	No	Driver behaviour, reliability, network simplicity, information
TCRP (1999)	All transit modes	Yes	Accessibility, passenger load
Prioni and Hensher (2000)	Bus	Yes	Tariff, travel time, access time, driver friendliness, smoothness
Ryus <i>et al.</i> (2000)	All transit modes	Yes	Frequency, service coverage, stop location, operation hour, population density
Guttenplan <i>et al.</i> (2001)	Bus	Yes	Factors related to the pedestrian environment (difficulty in crossing road, etc.)
Yang and Wang (2001)	All transit modes	Yes	Network density, population density

trips: the number of available parking spaces in each of the three departure points is 1 space. Note that the DH time is larger than the travel time (buses need to detour the service route when empty).

- Find the minimum fleet size required by using deficit function modelling.
- While maintaining the result of (a), find the best bus scheduling solution so as to minimize the number of departure points with maximum parking spaces

required greater than the given number of available spaces; utilize surplus function modelling to arrive at the solution.

Trip number	Departure terminal	Departure time	Arrival terminal	Arrival time
1	<i>a</i>	6:00	<i>b</i>	6:15
2	<i>b</i>	6:25	<i>a</i>	6:40
3	<i>b</i>	5:40	<i>c</i>	5:55
4	<i>c</i>	5:45	<i>a</i>	6:00
5	<i>a</i>	6:30	<i>b</i>	6:50
6	<i>b</i>	6:35	<i>c</i>	6:50
7	<i>a</i>	6:20	<i>c</i>	6:42

13.2 Use the same technical data and queries as in Exercise 13.1, but with 10 trips, 4 departure points and the following timetable:

Trip number	Departure terminal	Departure time	Arrival terminal	Arrival time
1	<i>a</i>	7:00	<i>b</i>	7:20
2	<i>b</i>	7:30	<i>a</i>	7:45
3	<i>c</i>	8:15	<i>d</i>	8:30
4	<i>b</i>	7:00	<i>c</i>	7:15
5	<i>a</i>	7:25	<i>c</i>	7:40
6	<i>b</i>	7:45	<i>a</i>	7:55
7	<i>a</i>	8:05	<i>c</i>	8:15
8	<i>c</i>	7:25	<i>c</i>	7:40
9	<i>d</i>	7:30	<i>b</i>	7:45
10	<i>d</i>	7:25	<i>b</i>	7:45

- 13.3 Employ the same technical data and queries as in Exercise 13.1, but with 9 trips and this timetable:

Trip number	Departure terminal	Departure time	Arrival terminal	Arrival time
1	<i>a</i>	8:00	<i>b</i>	8:20
2	<i>b</i>	8:30	<i>a</i>	8:45
3	<i>c</i>	9:15	<i>b</i>	9:30
4	<i>b</i>	8:00	<i>c</i>	8:15
5	<i>a</i>	8:25	<i>c</i>	8:40
6	<i>b</i>	8:45	<i>a</i>	8:55
7	<i>a</i>	9:05	<i>c</i>	9:15
8	<i>c</i>	8:25	<i>c</i>	8:40
9	<i>a</i>	9:15	<i>b</i>	9:35

- 13.4 Given a network of nine two-way street segments and six residential-area nodes in which the distances between nodes, in metres, appear in the following table and are symmetrical (same distances between nodes *i* and *j*, and between *j* and *i*).

Node	2	3	4	5	6
1	900	300			
2		1100	1300		
3			1000	400	
4				700	500
5					600

- Find the minimum number of stops in the network, such that the critical distance between each node and its closet stop is less than or equal to 900 metres.
- Find the optimal locations of four stops in the street network so as to minimize the maximum distance between a node and its closet stop; what is this distance?
- Find the optimal location of a single stop in the network using the same criterion as in (b); what is the mini-max distance obtained, and from what node?

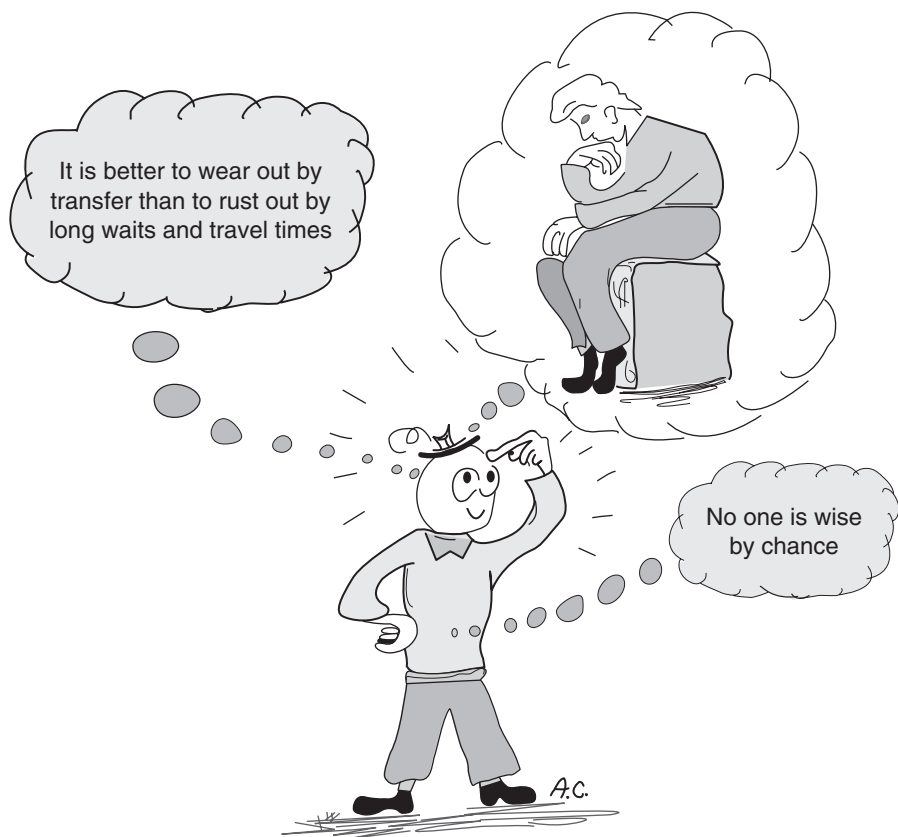
References

- Ahuja, R. K., Magnanti, T. L. and Orlin, J. B. (1993). *Network Flows*. Prentice Hall.
- Alter, C. H. (1976). Evaluation of public transit services: The level-of-service concept. *Transportation Research Record*, **606**, 37–40.
- Balas, E. and Padberg, M. W. (1972). On the set covering problem. *Operations Research*, **20**, 1152–1161.
- Ceder, A., Prashker, J. and Stern, H. I. (1983). An algorithm to evaluate public transportation stops for minimizing passenger walking distance. *Applied Mathematical Modeling*, **7**, 19–24.
- Christofides, N. (1975). *Graph Theory; An Algorithmic Approach*. Academic Press.
- Christofides, N. and Viola, P. (1971). The optimum location of multicenters on a graph. *Operations Research Quarterly*, **22**, 45–54.
- Farewell, R. G. and Marx, E. (1996). Planning, implementation, and evaluation of OmniRide demand-driven transit operations: Feeder and flex-route services. *Transportation Research Record*, **1557**, 1–9.
- Friman, M., Edvardson, B. and Garling, T. (1998). Perceived service quality attributes in public transport: Inferences from complaints and negative critical incidents. *Journal of Public Transportation*, **2**(1), 67–89.
- Guttenplan, M., Landis, B. W., Crider, L. and McLeod, D. S. (2001). Multimodal level-of-service analysis at the planning level. *Transportation Research Record*, **1776**, 151–158.
- Handler, G. Y. (1973). Minimax location of a facility in an undirected graph. *Transportation Science*, **7**, 287–293.
- Hakimi, S. L. (1964). Optimum locations of switching centers and the absolute centers and medians of a graph. *Operations Research*, **12**, 450–459.
- Henk, R. H. and Hubbard, S. M. (1996). Developing an index of transit service availability. *Transportation Research Record*, **1521**, 12–19.
- Horowitz, A. J. (1981). Service-sensitive indicators for short-term bus-route planning. *Transportation Research Record*, **798**, 36–39.
- Madanat, S. M., Cassidy M. J. and Ibrahim, W. H. W. (1994). Methodology for determining level of service. *Transportation Research Record*, **1457**, 59–62.
- Minieka, E. (1970). The m -center problem. *SIAM Review*, **12**, 138–139.
- Minieka, E. (1978). *Optimization Algorithms for Networks and Graphs*. Marcel Dekker.
- Murugesan, R. and Rama Moorthy, N. A. (1998). Level of public transport service evaluation: A fuzzy set approach. *Journal of Advanced Transportation*, **32**(2), 216–240.
- Polus, A. and Shefer, D. (1984). Evaluation of a public transportation level of service concept. *Journal of Advanced Transportation*, **18**(2), 135–144.
- Prioni, P. and Hensher, D. (2000). Measuring service quality in scheduled bus services. *Journal of Public Transportation*, **3**(2), 51–74.
- Rubin, J. (1973). A technique for the solution of massive set covering problems, with application to airline crew scheduling. *Transportation Science*, **7**, 34–48.
- Ryus, P., Ausman, J., Teaf, D., Cooper, M. and Knoblauch, M. (2000). Development of Florida's transit level-of-service indicator. *Transportation Research Record*, **1731**, 123–129.

- TCRP (1999). *Highlights of the Transit Capacity and Quality of Service Manual*. *TCRP Research Results Digest*, **35**, Transportation Research Board.
- TranSystems Corp., Planner Coll., Inc. and Crikelair, T. Assoc. (2006). Elements needed to create high ridership transit systems: Interim guidebook. *TCRP Report 32*, Transportation Research Board.
- Yang, X. and Wang, W. (2001). GIS-based Fuzzy C-means clustering analysis of urban public transit network service: the Nanjing city case study. *Road and Transport Research, ARRB Transport Research*, **10**(2), 56–65.

This page intentionally left blank

14 Network (Routes) Design



Chapter 14 Network (Routes) Design

Chapter outline

- 14.1 Introduction
- 14.2 Objective functions
- 14.3 Methodology and example
- 14.4 Construction of a complete set of routes
- 14.5 Multi-objective technique
- 14.6 Literature review and further reading

Exercises

References

Practitioner's Corner

This chapter focuses on the first activity of the operational planning process described in Chapter 1: public transit route design and evaluation at the network level. The problem addressed is that of how to design a new transit network or to redesign an existing network. Naturally, this activity would be appropriate only very infrequently because of the disruption that would be imposed on passengers if wholesale changes were made to the transit network. In addition, many transit agencies have not gone through such a reappraisal, mainly because of the lack of clear-cut, practical and measurable design criteria for evaluating the ‘goodness’ of transit routes and comparing sets of routes. Theodore Roosevelt said: “Nine-tenths of wisdom is being wise in time”. Thus, a design and evaluation tool may provide a timely boost for reconstructing transit routes, although one should bear in mind the required inter-connectivity of a transit network.

Following the introductory section of this chapter, Section 14.2 proposes a framework for the construction of the operational objective functions of the transit network design problem. This framework takes into account the passengers, as well as agency and community/government interests – the three perspectives emanating from the broad spectrum of transit activities. Section 14.3 utilizes the objective functions described for facilitating the construction of a methodology for the practical design of an efficient transit network of routes. This methodology combines the philosophy of mathematical programming approaches with decision-making techniques; it allows the transit planner to select a set of routes from a number of alternatives. Section 14.4 provides an efficient procedure for constructing sets of routes, in which each set covers the transit demand considered. The sets of routes are created in such a manner that connectivity between all nodes (in the network) is maintained and their total deviation from the shortest path minimized. Section 14.5 presents a multi-objective technique for obtaining optimal or near optimal solutions. The chapter ends with a literature review and exercises.

Practitioners are encouraged to visit Sections 14.2 and 14.3, and they may skip Sections 14.4 and 14.5. However, it is recommended that they look at the introductory parts and examples illustrated in these two sections.

It is worth noting that a new design or a redesign of a transit network needs time to be fully comprehended; this is the ‘same’ time defined as nature’s way of keeping everything from happening at once. Finally, this Danish proverb seems most apt: “Better to ask twice than to lose your way once”. In our case, it is better to present the new routes in at least two clear ways than to have the user misinterpret them even once.

14.1 Introduction

There are two main approaches to restructuring transit routes: (1) at the route level or for a small group of routes; (2) at the network level. For the first approach, Pratt and Evans (2004), in *TCRP Report 95*, suggested that restructuring was to simplify routes, accommodate new travel patterns, ease or eliminate transfers, reduce route circuitry, or otherwise alter route configuration. This approach is dealt with to some extent in Chapters 13, 15 and 16. The present chapter proffers a practical solution utilizing the second approach.

Only a few researchers have studied the interrelationship of the scheduling and the network-design planning activities shown in Figure 1.2 in Chapter 1. The interrelationship has two categories: (a) each set of routes, based on demand, yields a different set of frequencies and timetables and, ultimately, the required fleet size; (b) the operational cost derived from the scheduling activities and passenger level of service affects the search for the optimal route design while relying on a compromise solution between the agency and the passengers.

For many public transit agencies whose network of routes has not been reappraised for 20 to 50 years, it is high time to consider such an undertaking. This should provide sufficient motivation to seek an efficient network route-design method, based on certain objective functions and a set of constraints. The main purpose of the methodology presented in this chapter is to transport a given origin-destination (O-D) demand through the transit network in the most cost-effective way. The special characteristics of route-design problems are as follow: (i) passenger demand is spread throughout the entire network, where it is generated and terminated at many points along the network’s links and can be grouped in terms of an O-D matrix; (ii) the demand must be transported simultaneously (usually during peak hours); (iii) over a given planning horizon, it is impossible to reconstruct the routes (i.e. once the route network is designed, it will remain unchanged over an entire planning period).

Prior approaches to the public transport network-design problem can be grouped into those that simulate passenger flows, that which deal with ideal networks, and those based on mathematical programming. *Simulation models* are presented in Dial and Bunyan (1968), Heathington *et al.* (1968), and Vandebona and Richardson (1985). These models require a considerable amount of data, and their proximity to optimality is uncertain. *Ideal network methods* are based on a broad range of design parameters and a choice of objectives reflecting user and agency interest. Such methods appear in Kocur and Hendrickson (1982), Tsao and Schonfeld (1984), and Kuah and Perl (1988). These methods are adequate for screening or policy analyses, in

which approximate design parameters rather than a complete design are determined; thus, these methods cannot represent real situations. *Mathematical programming models* are divided into generalized network-design models and transit-specific networks models. Known generalized network models are well summarized and reviewed in Kim and Barnhart (1999); also see, as an example, the heuristics developed by Farvolden and Powell (1994). The transit-specific network-design models are inevitably heuristic because of the extremely high computational effort required. These partial optimization approaches appear in Lampkin and Saalmans (1967), Silman *et al.* (1974), Dubois *et al.* (1979), Mandl (1979), and Keudel (1988). None of the models and methods mentioned in this section has actually been applied, however. A further literature review appears below, in Section 14.6.

An overview of prior work done on this theme calls for a method that, given the availability of typical data, will be more practical and less complex than other methods and models. Such a method, described in this chapter, may increase the chances of its acceptance by most transit agencies. The method developed seeks to have transit routes as close as possible to the shortest paths. At the same time, planners often encounter a Murphy's law: the shortest distance between two points is usually under repair. This is the reason that the next two sections provide the possibility of a limited detour to the shortest path.

Finally, it is interesting to observe Figure 14.1, which is currently a distributed info-page at the taxi stand at the Washington (DC) National Airport. This page includes the following note: *Fares are based on shortest route; however, the shortest route may not be the quickest route.* (Which is also a nice way of them to remind us that laughter is the shortest distance between two people.) In our case, the shortest path intends to be the quickest one because of its time units; in real-time, however, this shortest path may not hold because of traffic congestion (e.g. caused by road accidents or road repairs, etc.).

14.2 Objective functions

This section proposes a framework for the construction of the operational objective functions of the public transit network-design problem. This framework takes into account the passengers, the agency and community interests. It follows the studies by Ceder and Israeli (1992), Israeli and Ceder (1995), Ceder (2001), Ceder *et al.* (2002), and Yin *et al.* (2005).

14.2.1 Three perspectives and four criteria

From the literature review presented above and that at the end of the chapter, there are no clear-cut, practical, or measurable criteria for evaluating the 'goodness' of transit routes and comparing sets of routes at the network level. The only comprehensible matter is that the design of transit routes should be looked at simultaneously from three perspectives: that of the passenger, the agency and the community/government. These three perspectives emanate from the broad spectrum of transit activities.

Four criteria can be considered when measuring the quality of a transit route: (1) minimum passenger waiting time; (2) minimum empty-seat/space time; (3) minimum time difference from shortest path; and (4) minimum fleet size. The first three criteria are measured

WELCOME TO RONALD REAGAN WASHINGTON NATIONAL AIRPORT

Once you have been comfortably seated in the taxicab, please note the Airport Taxi Operator's Permit hanging from the rear view mirror. For future reference, you should record the Permit Number #_____.

(NOTE: Driver is allowed to remove the Permit after exiting the Airport.)

APPROXIMATE FARES FOR THESE AREAS

Fares are based on shortest route, however, the shortest route may not be the quickest route.
Fares are not flat rates, they are APPROXIMATE fares and do not include the \$1.75 airport fee.
Your actual fare may be different than the amount shown on the map.

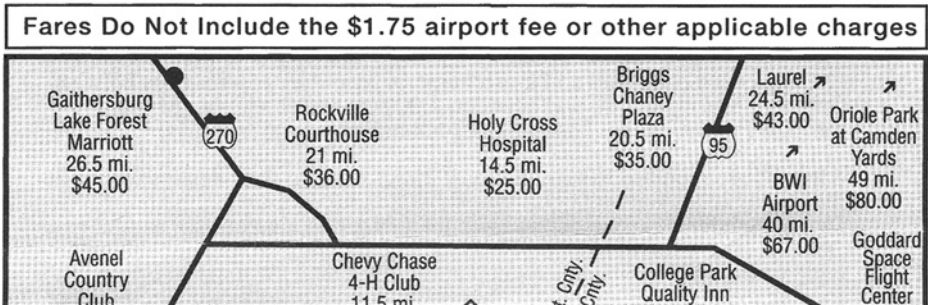


Figure 14.1 Current use of the term 'shortest route' by the taxi service at the Washington (DC) National Airport (see the marked section under APPROXIMATE FARES. . .)

in passenger-hours, and the last one in number of vehicles. Clearly, criterion (1) represents the passengers' perspective; criteria (2) and (4), the agency's perspective; and criterion (3), both the passengers' and the community's perspectives.

When the purpose of measuring is to compare sets of transit routes or different transit modes, monetary weights can be introduced to the four criteria. Optionally, criterion (3) can be replaced by the total monetary loss (or saving, if it is negative) if all the transit passengers are switched to the shortest path. For instance, when comparing a set of bus routes and a set of metro lines, in which the latter is the shortest path (in time), criterion (3) will provide the total monetary loss/saving if all the bus passengers switch to the metro lines.

The next section constructs a quantitative framework for the four criteria in order to devise tools for the design of an optimal set of transit routes. This quantitative framework furnishes a clearer picture of the four criteria, with two optional versions of criterion (3).

14.2.2 Formulation

The following established notations will be used throughout the analyses of this chapter.

Notations

Consider a connected network composed of a directed graph $G = \{N, A\}$ with a finite number of nodes $|N|$ connected by $|A|$ arcs. Define:

Route	= Progressive path initiated at a given transit terminal and terminated at a certain node while traversing given arcs in sequence
Transfer path	= Progressive path that uses more than one route
$R = \{r\}$	= Set of transit routes
$TR = \{tr\}$	= Set of all transfer paths
$S = \{sp\}$	= Set of all shortest paths (minimum average travel times)
N_r	= Set of nodes located on route r
N_{tr}	= Set of nodes located on transfer path tr
N_{sp}	= Set of nodes located on the shortest path sp
d_{ij}^r	= Passenger demand between i and j , $i, j \in N$, riding on route r
d_{ij}^{tr}	= Passenger demand between i and j along the transfer path tr
d_{ij}^{sp}	= Passenger demand between i and j along its shortest path
F_r	= Vehicle frequency associated with route r
F_{min}	= Minimum frequency (reciprocal of policy headway) required
t_{ij}^r	= Average travel time between i and j on route r
t_{ij}^{tr}	= Average travel time between i and j on transfer path tr (can include transfer penalties)
t_{ij}^{sp}	= Average travel time between i and j on its shortest path
t_r	= Overall travel time on route r between its start and end
L_r	= Maximum passenger load on route r
w_r	= Passenger waiting time on route r
d_o	= Desired occupancy on each vehicle (load standard)
a_{tr}^r	= $\begin{cases} 1, & \text{transfer } tr \text{ moves through route } r \\ 0, & \text{otherwise} \end{cases}$
α	= Maximum allowed deviation from shortest path for any O-D pair on a transit path (including transfers)
k_{tr}	= Maximum degree of transfer path tr (number of vehicle changing).

Two principal objective functions

The transit network-design problem is based on two principal objective functions, minimum Z_1 and minimum Z_2 , across the different sets of transit routes:

$$Z_1 = \begin{cases} a_1 \sum_{i,j \in N} WT(i, j) + a_2 \sum_r EH_r + a_4 \sum_{i,j \in N} DPH(i, j), & \text{for single set} \\ a_1 \sum_{i,j \in N} WT(i, j) + a_2 \sum_r EH_r + \sum_{i,j \in N} [a_3 PH(i, j) - a_4 DPH(i, j)], & \text{for comparison} \end{cases} \quad (14.1)$$

$$Z_2 = FS \quad (14.2)$$

where:

PH (i, j) = Passenger-hours between nodes i and j, $i, j \in N$ (defined as passengers' riding time in a transit vehicle on an hourly basis; it measures the time spent by passengers in vehicles between the two nodes)

DPH (i, j) = Difference in passenger-hours between PH (i, j) and total passenger-hours from i to j when using only the shortest path, $i, j \in N$

WT (i, j) = Waiting time between nodes i and j, $i, j \in N$ (defined as the amount of time passengers spend at the transit stops between the two nodes)

EH_r = Empty-seat/space hours on route r (defined as the unused seats/spaces in a transit vehicle on an hourly basis; empty-seat/space hours measures the unused capacity on vehicles)

FS = Fleet size (number of transit vehicles needed to provide all trips along a chosen set of routes)

a_k = Monetary or other weights, $k = 1, 2, 3, 4$.

Equation (14.1) contains the two options for the Z_1 objective function, which can be interpreted as minimum waiting time and maximum utilization; for given weights of 1 or without units, this equation results in units of passenger-hours. Equation (14.2) is simply the required minimum fleet size.

Objective function components

Equations (14.1) and (14.2) essentially combine five objective function components. The first straightforward objective is to minimize passengers' total waiting time. This is strictly the perspective of the transit user. The formulation of this objective takes the following form:

$$\text{Min } a_1 \sum_{i,j \in N} WT(i, j) \quad (14.3)$$

where a_1 = monetary value of one-hour's waiting time.

The second objective is to minimize the total unused seat capacity so as to allow more viable transit service. This is strictly the perspective of the agency, which wishes to see more occupation of the available seats. The following is the formulation of this objective:

$$\text{Min } a_2 \sum_r EH_r \quad (14.4)$$

where a_2 = the equivalent of one-hour's average monetary revenue divided by the average number of hourly boarding passengers. This objective is to minimize the total monetary value of unused seat capacity.

The third and fourth objectives are two versions of the same objective: (a) to minimize the total time loss (in monetary value) between riding the transit vehicle and travelling by car (assumed to be the shortest path); (b) to minimize the total loss (in monetary value) if all the passengers are switched to the shortest path. Versions (a) and (b) take the following forms, respectively:

$$\text{Min } a_4 \sum_{i,j \in N} DPH(i, j) \quad (14.5)$$

$$\text{Min } \sum_{i,j \in N} [a_3 \text{PH}(i, j) - a_4 \text{DPH}(i, j)] \quad (14.6)$$

where a_3 = equivalent of one-hour's difference in average cost/fare between riding the shortest path (car or a transit competitor) and the transit route; and a_4 = monetary value of one-hour's in-vehicle time.

The value of Equation (14.6) is the total monetary loss (or saving, if it is negative) if all the passengers are switched to the shortest path, where $a_3 \text{PH}$ = total monetary loss, with respect to cost/fare only, if all the passengers are switched to the shortest path; and $a_4 \text{DPH}$ = total monetary value of the time saved if all the passengers are switched to the shortest path. The latter fits Equation (14.5). These objectives represent the perspectives of the community/government and the passengers.

The fifth objective is to minimize the number of vehicles required for a given set of routes and frequencies (timetables). This is strictly the agency perspective, which wishes to perform all transit trips using a minimal number of vehicles. This objective takes the form:

$$\text{Min FS} \quad (14.7)$$

Objectives (14.3)–(14.6) are all in terms of passenger-hour cost; for the sake of simplicity, therefore, this can be summed up to $\text{Min } Z_1$ as shown in Equation (14.1). Objective (14.7) stands alone to some extent and is termed $\text{Min } Z_2$ in Equation (14.2).

Calculation of Z_1 functions

The objective function Z_1 is based on the passenger-load profile. Figure 14.2 presents the example of a small transit network in part (a). The input data during a given time period (usually peak hours) consist of average travel times, an estimated O-D demand in part (b), and $a_k = 1$ for all $k = 1, 2, 3, 4$. Part (c) of Figure 14.2 displays the feasible routes complying with the following five practical constraints: (i) maximum route length of 30 minutes; (ii) maximum allowed deviation of 40% from the shortest path; (iii) maximum of 1 transfer for all O-D pairs; (iv) a chosen route should not be included in another feasible route; and (v) no circular routes considered.

Because routes can start only at terminals (node 1 in the example), a mapping process is applied, with constraint (i) considered first. This results in six routes: [1-2], [1-2-3], [1-2-3-4], [1-3], [1-3-2] and [1-3-4], all of which comply with the maximum of 30 minutes average travel time. In parallel, a shortest-path algorithm (see Appendix 10.A to Chapter 10) is applied. Constraint (ii), maximum deviation from shortest path, can then be checked using the shortest-path information. Table 14.1 provides this check while describing the procedure for a determination of feasible routes; that is, making sure that a feasible route will not be included in another feasible route in order to avoid overlapping.

Once the feasible routes are established, their load profiles are constructed, based on the input data. The load profile of route [1-3-4] is presented in Figure 14.2 part (d). The load on 1-3 is the maximum load ($L_{1-3-4} = 680$ passengers) of the O-D demand: 1-3, 1-4, 2-3 and 2-4 (see part (b) of Figure 14.2). The load on 3-4 is 300, consisting of the O-D demand: 1-4, 2-4, 3-4. The shaded area of the load profile between 680 and 300 passengers along the 16-minute segment represents 101.3 empty-seat hours for $d_o = 50$ (number of seats on the vehicle) and

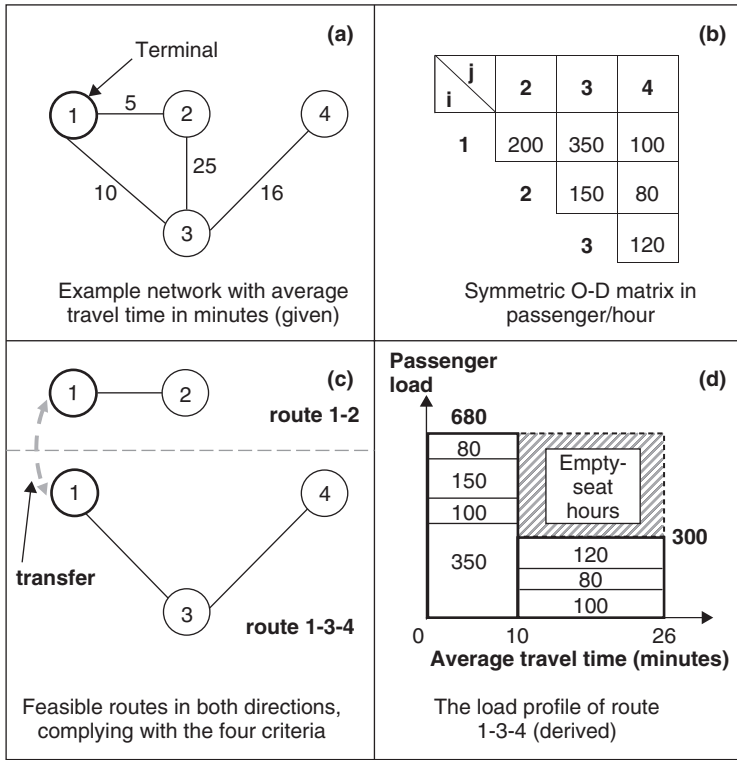


Figure 14.2 Example network, its given data and the construction of a load profile

$F_{\min} = 4$ during the associated time period of the example. In addition, the load profile of the other feasible route, 1-2, is simply rectangular with a maximum load of $L_{1-2} = 200 + 150 + 80 = 430$ passengers and $F_{1-2} = 8.6$.

The first function of Z_1 in Equation (14.1) is the total wait-time hours both at the transit stops and during transfers. Different formulations of the expected passenger waiting time appear in Sections 12.2 and 17.4 in Chapters 12 and 17, respectively. Utilizing Equation (12.1), in which passengers arrive randomly at the transit stop and headways are deterministically distributed, we obtain the expected waiting time on route r , which is half the headway:

$$w_r = \frac{1}{2F_r}, \quad \text{for all } r \in R \quad (14.8)$$

Thus,

$$\sum_{i,j \in N} WT(i, j) = \sum_{r \in R} \frac{1}{2F_r} \left(\sum_{i,j \in N_r} d_{ij}^r + \sum_{ij \in N_{tr}} d_{ij}^{tr} a_{tr}^r \right) \quad (14.9)$$

Table 14.1 *Determination of feasible routes in the example problem*

Route end points	1 → 2	1 → 3	1 → 4
Shortest-path (minutes)	5	10	26
Shortest-path route	1-2	1-3	1-3-4
Symmetric (both directions), O-D served optimally	1-2	1-3	1-3, 1-4
Other routes not complying with the maximum length criterion (40%)	1-3-2	1-2-3	1-2-3-4
Symmetric (both directions), O-D served by other routes, and their deviation (%) from the shortest path	1-2 (600%) 1-3 (0%) 2-3 (67%)	1-2 (0%) 1-3 (200%) 2-3 (67%)	1-2 (0%) 1-3 (200%) 1-4 (77%) 2-3 (67%) 2-4 (32%) 3-4 (0%)
Feasible route	1-2	1-3-4	
Symmetric (both directions), O-D served optimally via one transfer		2-3, 2-4	

Applying Equation (14.9) to the example of Figure 14.2 yields (for both directions of routes):

$$\begin{aligned} \sum_{i,j \in N} WT(i, j) &= \frac{2.60}{2.680/50} (350 + 100 + 150 + 80 + 120) \\ &+ \frac{2.60}{2.430/50} (200 + 150 + 80) = 6529 \text{ pass-hour} \end{aligned}$$

where $d_o = 50$. Note that any demand between i and j , $i, j \in N$, can split (e.g. some via a direct route and the remaining via transfers); in this case, an assignment procedure (see Chapter 12) can be applied. However, this is not the case in the example.

The second function of Z_i , in Equation (14.1), describes the total empty-space hours or empty-seat hours (when d_o equals the number of seats on the vehicle). This function represents an unproductive measure for the agency (e.g. unused seat capacity). Its formulation is:

$$\sum_r EH_r = \sum_{r \in R} [\max(L_r, F_{\min} \cdot d_o)] t_r - \sum_{i,j \in N} PH(i, j) \quad (14.10)$$

In the example of Figure 14.2, this equation yields: $EH_r = 2(101.3 + 0) = 202.6$ passenger-hours for both directions of the feasible routes.

The first version of the third function in Equation (14.1), which appears in Equation (14.5), is the total passenger-hour difference between PH (i, j) on r and PH (i, j) on the shortest path sp:

$$\sum_{i,j \in N} \text{DPH} (i, j) = \sum_{i,j \in N} \text{PH} (i, j) - \sum_{sp \in S} \sum_{i, j \in N_{sp}} d_{ij}^{sp} t_{ij}^{sp} \quad (14.11)$$

in which, for the example in Figure 14.2: $\sum \text{DPH} = 0$; the reason for this is that routes [1-2] and [1-3-4] include only the shortest paths.

The second version of the third function in Equation (14.1), which appears in Equation (14.6), has two parts, the second part being Equation (14.11). The first part is total passenger-hours in the routing system:

$$\sum_{i,j \in N} \text{PH} (i, j) = \sum_{r \in R} \sum_{i,j \in N_r} d_{ij}^r t_{ij}^r + \sum_{tr \in TR} \sum_{ij \in N_{tr}} d_{ij}^{tr} t_{ij}^{tr} \quad (14.12)$$

For the example in Figure 14.2, Equation (14.12) yields:

$\sum_{i,j \in N=1} \text{PH} = 2(10 \cdot 680 + 16 \cdot 300 + 430 \cdot 5)/60 = 458.3$ passenger-hours; in which the transfers (both directions of 2-3, 2-4) do not include penalties (estimated extra-effort cost).

Estimation of Z_2

Estimation of the minimum fleet size can utilize the deficit function modelling described in Chapters 7 and 8. Note that it may be sufficient to use the procedure to determine the stronger fleet-size lower bound (see Section 8.2.2 in Chapter 8) for this estimation. Practically speaking, the design of optimal transit routes involves a vast amount of computations of sets of routes. Thus, the lower-bound-based Z_2 calculation can ease the computation effort for each route considered.

14.2.3 Applications

The framework of the objective functions described is believed to be a useful toolset for more than designing a new transit network of routes. For instance, this framework can apply to:

- optimal design for expansion or curtailment of an existing transit network of routes;
- assessment of the performance of an existing transit network from the aspects of: (i) agency efficiency (fleet size, empty-seat hours, length of routes, number of transfers), and (ii) passenger level of service (average waiting time, deviation from shortest path, crowding level);
- sensitivity analysis of transit network performance for a variety of system parameters (such as different fleet sizes, different levels of service, changes in passenger demand, changes in frequencies, changes in travel time and more).

One application that utilized the objective functions described was presented by Yin *et al.* (2005). This application proposes a deployment-planning framework that provides, in a

sequence of steps, a general structure for the optimal deployment of buses in a rapid transit (BRT) system. The following BRT elements were considered for system enhancement:

- a. Bus signal priority (extended over existing deployment)
- b. Exclusive lanes
- c. Articulated buses
- d. Multiple door boarding and alighting
- e. Stop enhancements
- f. Electronic fare payment
- g. Precision docking.

Given these BRT elements, the deployment-planning framework was used to determine cost-efficient combinations for system enhancement. Equations (14.1)–(14.6) were then utilized to calculate performance measures for each combination of BRT elements. From these calculations, an optimal combination was selected and recommended for deployment. Table 14.2 presents seven alternative combination of BRT elements considered in the study, together with their estimated cost (in millions of \$US).

Table 14.2 Cost estimates for combinations of BRT elements

Alternative	BRT elements	Cost (\$M)
A	a, b, c, d, e, f	99.9
B	a, b, c, d, e, f, g	101.4
C	a, b, e, f	19.9
D	a, b, e, f, g	21.4
E	a, c, d, e, f	93.9
F	a, c, d, e, f, g	95.4
G	a, b, c, d	86.6

Source: Yin *et al.* (2005)

In the Yin *et al.* (2005) study, the budget limit for implementing the BRT elements was considered as a given in the amount of \$90 million. If the number was high enough, there would be no trade-off between elements. Therefore, the financially feasible alternatives are C, D and G. Moreover, by refining Alternative E through excluding element e, stop enhancement, and making it financially feasible, a new Alternative H is created that includes the elements a, c, d and f. The total cost is \$83.9 million.

The results of the Yin *et al.* analysis are shown in Table 14.3, in which the objective is to find the best alternative with minimum Z_1 and minimum Z_2 . It can be seen from Table 14.3 that Alternative H is dominated by Alternative G and that both G and D are non-dominated. Although the capital cost of Alternative D is much less than that of Alternative G, the former requires a much larger fleet size, which may lead to higher operating and maintenance costs. Therefore, transit agencies should look carefully at the trade-off between these two objectives and, based on their preferences and non-quantitative considerations, decide which of the two

Table 14.3 Evaluation results for combinations of BRT elements

	Passenger travel time (pass-hrs)	Passenger waiting times (pass-hrs)	Empty-seat hours (pass-hrs)	Z_1	Z_2
Alternative D	1999	94	1614	23966	138
Alternative G	2025	139	1639	28648	84
Alternative H	2075	139	1665	29065	86

Source: Yin *et al.* (2005)

recommended alternatives (G and D) should be deployed. Section 14.5 provides the tools for selecting the better (compromise) solution.

14.3 Methodology and example

The objective functions described facilitate the construction of a methodology for the practical design of an efficient transit network of routes. This methodology, outlined in this section together with an example, combines the philosophy of mathematical programming approaches with decision-making techniques; it allows the transit planner to select a set of routes from a number of alternatives. The methodology and the example follow Ceder and Israeli (1992) and Ceder (2003).

14.3.1 Six-element methodology

The transit route-design methodology consists of six elements, shown in Figure 14.3. The *first element* generates every feasible route and transfer (throughout the entire network) from all terminals, including the shortest-path computation. Initially, the network contains average travel times covering a time window, which is usually the peak period. These measured average travel times are then used as input for the calculation of the shortest path between each origin-destination (O-D) pair. Each candidate route determined meets the route-length-factor constraint; that is, one procedure of this element screens out routes according to given boundaries of a route length. In addition, there is a limit on the route's average travel time between each O-D pair. That is to say, a given passenger demand, usually during peak hours, cannot be assigned to a candidate route if its average travel time exceeds the shortest-path travel time by more than a given percentage. Feasible transfers (between O-D pairs without direct routes) are based on establishing additional direct routes between O-D pairs characterized by high O-D demands (predetermined O-D); feasibility is determined by the travel-time limit in comparison with the shortest path. These extra direct routes are initiated and/or terminated at non-terminal nodes and, consequently, deadheading trips are responsible for their connection to the terminals. The feasible transfers are created using a mapping algorithm (branching of routing possibilities along with a check of constraints). Finally, low O-D demand, without a direct route, may not be considered for implementing a transit service. The example below further explains this element.

The *second element* in Figure 14.3 creates a minimum set of routes and related transfers, such that connectivity between nodes is maintained and their total deviation from the shortest

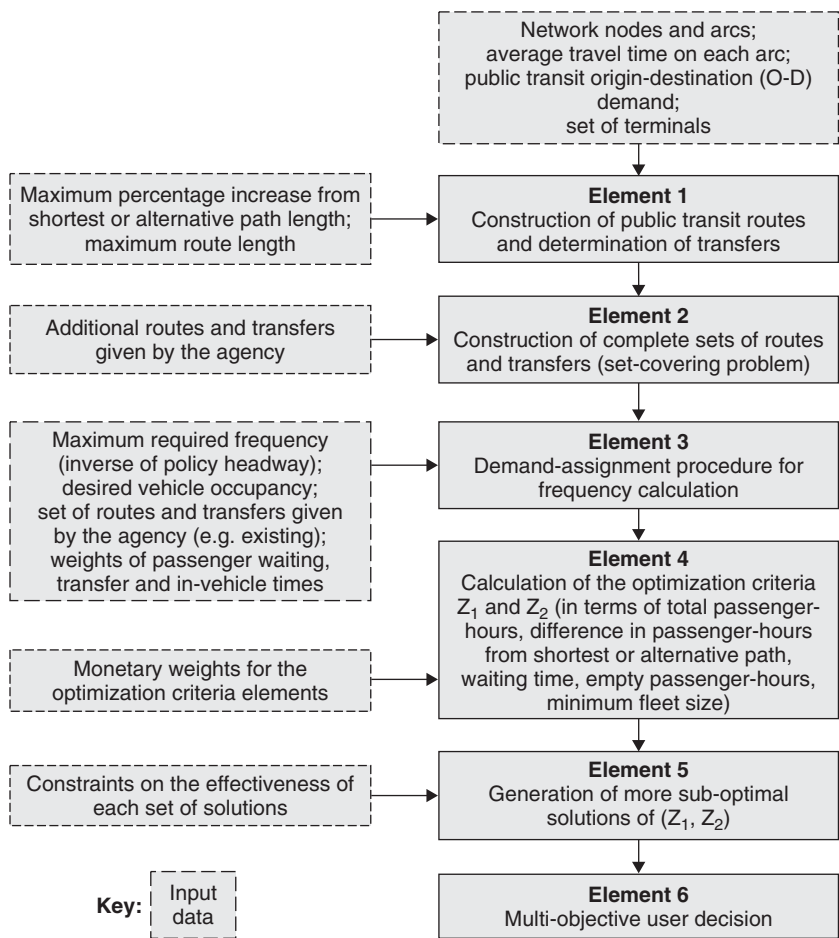


Figure 14.3 A methodology for designing public transit routes

path is minimized. This problem is defined as a set-covering problem (SCP) similar to the one in Section 13.4.3 in Chapter 13. The SCP determines the minimum set of routes from the matrix of feasible routes, in which each column represents either a feasible route or a feasible transfer. The procedures included in this element are described in Section 14.4.

The *third element* assigns the entire O-D demand to the chosen set of routes. The assignment algorithm follows the procedure described in Section 12.5 in Chapter 12, and includes steps that are related to a route-choice decision investigation. That is, the algorithm includes a probabilistic function for passengers who are able to select the transit vehicle that arrives first or, alternatively, who wait for a faster vehicle. The passengers’ strategy is to minimize the total weight of wait, transfer and in-vehicle times.

The *fourth element* represents the optimization criteria from the passengers’, agency’s, and community/government’s perspectives. It is detailed in the previous section and based on Equations (14.1)–(14.6).

The *fifth element* is responsible for constructing alternative sets of routes in order to search for additional (Z_1, Z_2) values in the vicinity of their optimal setting. The procedure for this search is based on incremental changes in the set of routes, much like the known reduced-gradient methods. Given the set of routes associated with the minimum Z_1 value, the single route that is the worst contributor to Z_1 is deleted and then the SCP is resolved, followed by the execution of the third and fourth elements. This process could continue, but there is no guarantee that a previous alternative will not be repeated. In order to overcome this problem, a new matrix is constructed with the idea of finding the minimum and worst set of candidate routes for possible deletion in each iteration; i.e. a new SCP matrix is constructed in which the candidate routes are the columns and each row represents a previous set of routes that was already identified in the vicinity of the optimal (Z_1, Z_2) setting. The solution to this new SCP matrix is a set of rejected routes, so as not to repeat a previous alternative solution. During this process, a number of unique collections of routes are termed ‘prohibited columns’, as they are the only ones that can fulfil a certain demand. These prohibited columns are assigned an artificially high cost value so as not to be included in the solution. This process also involves some bounds on the number of (Z_1, Z_2) solutions and number of iterations.

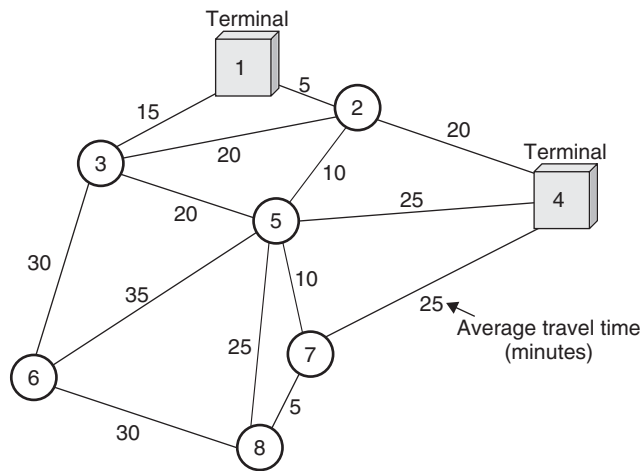
The *sixth element*, and the final one in Figure 14.3, involves multi-objective programming of the two objective functions, Z_1 and Z_2 . Given the alternative sets of routes derived in the fifth element, the purpose is to investigate the various alternatives for the most efficient (Z_1, Z_2) solution. The method selected in this element is called the compromise-set method. The outcome of the compromise-set method is the theoretical point at which (Z_1, Z_2) attains its relatively minimum value. The results can be presented in a table or a two-dimensional graph showing the trade-off between Z_1 and Z_2 . These results also indicate the optimal zone or the so-called Pareto front. The decision-maker can then decide whether or not to accept the proposed solution; for example, the decision-maker can see how much Z_1 is increased by decreasing Z_2 to a certain value, and vice versa. This element is detailed in Section 14.5 below.

14.3.2 Example

A simple 8-node network for a bus service will be used as an example to demonstrate some of the procedures discussed; it is depicted in Figure 14.4 with two terminals (from which trips can be initiated). The input of passenger demand appears in Table 14.4; in addition, it is given that $a_k = 1$ for all $k = 1, 2, 3, 4$. The aim is to find the best bus routes in the network while complying with given constraints.

The outcome of the *first element* in Figure 14.3 is presented in Table 14.5, while using the maximum deviation from the shortest path as $\alpha = 0.4$; that is, no route length or portion of it can exceed its associated shortest travel time by more than 40%. Construction of the nine feasible routes emanating from terminal 1 are shown in Figure 14.5. The first element is based on an algorithm that mainly produces feasible transfers throughout the entire network. The transfers that are created using a mapping algorithm are shown in Table 14.6 for the example problem, the numbers in parentheses being the route numbers of Table 14.5 that comprise the transfers. In the transfer-path description, the numbers outside the parentheses represent nodes while those inside the parentheses represent routes.

The *second element* of the methodology creates a minimum set(s) of routes and related transfers, defined as an SCP matrix. Each row in this matrix represents either a feasible route or a transfer. The ‘1’ and ‘2’ in the matrix are inserted whenever an O-D demand can



Input: Maximum increase (from shortest/alternative path) factor, $\alpha = 0.4$
Maximum number of transfers allowed, $k_{tr} = 2$
Desired vehicle occupancy, $d_o = 50$ passengers

Figure 14.4 Example of an eight-node network with its basic input

Table 14.4 Passenger demand between nodes (assumed to be symmetrical) for the example problem

Nodes	2	3	4	5	6	7	8
1	80	70	160	50	200	120	60
2		120	90	100	70	250	70
3			180	150	120	30	250
4				80	210	170	230
5		Symmetrical			250	40	130
6						130	120
7							70

be, respectively, feasibly and optimally (shortest path) transported by the route or transfer; or ‘0’ otherwise. The word ‘covering’ in the SCP refers to at least one column with ‘1’ in each row. Section 14.4 below provides further description of the SCP analysis. In the example problem, about 100 sets of routes were generated. An example of one set is {4, 6, 9, 11, 25, 28, 32, 46}, in which the numbers refer to those in Tables 14.5 and 14.6; five routes and two transfer paths cover all the O-D pairs in the defined feasible manner. The results of the SCP analysis for this set of routes appear in Table 14.7.

The *third element* assigns the entire O-D demand to the chosen set of routes. Vehicle frequency in this element is derived from the passenger-load profile of each route; the load on

Table 14.5 *All feasible routes of the example problem generated by Element 1*

Route no.	Description
1	1-2
2	1-2-4
3	1-2-5
4	1-2-5-6
5	1-2-5-7
6	1-2-5-7-8
7	1-2-5-7-8-6
8	1-3
9	1-3-6
10	4-2
11	4-2-1-3
12	4-2-1-3-6
13	4-2-3
14	4-2-3-6
15	4-2-5
16	4-2-5-6
17	4-5
18	4-5-3
19	4-5-6
20	4-5-7
21	4-5-7-8
22	4-5-7-8-6
23	4-7
24	4-7-5
25	4-7-5-3
26	4-7-5-6
27	4-7-8
28	4-7-8-6

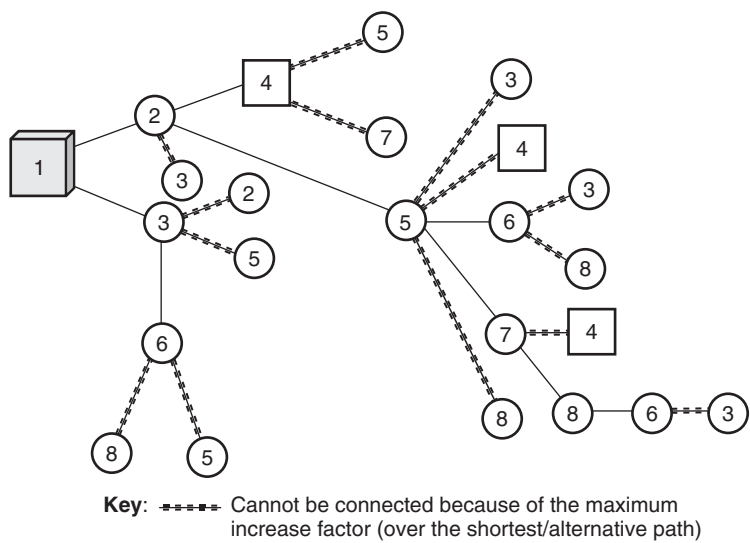


Figure 14.5 Example solution steps in generating all (nine) feasible routes from terminal 1 (see Table 14.5)

Table 14.6 All transfers for the example problem connecting nodes 3 and 8 during their 35-minute shortest path (route numbers in parenthesis*)

Transfer no.	Description
29(5, 18, 27)	3(18)–5(5)–7(27)–8
30(5, 18, 28)	3(18)–5(5)–7(28)–8
31(6, 18)	3(18)–5(6)–7(6)–8
32(6, 25)	3(25)–5(25, 6)*–7(6)–8
33(7, 18)	3(18)–5(7)–7(7)–8
34(7, 25)	3(25)–5(25, 7)–7(7)–8
35(18, 20, 27)	3(18)–5(20)–7(27)–8
36(18, 20, 27)	3(18)–5(20)–7(28)–8
37(18, 21)	3(18)–5(21)–7(21)–8
38(18, 22)	3(18)–5(22)–7(22)–8
39(18, 24, 27)	3(18)–5(24)–7(27)–8
40(18, 24, 28)	3(18)–5(24)–7(28)–8
41(18, 26, 27)	3(18)–5(26)–7(27)–8

(Continued)

Table 14.6 All transfers for the example problem connecting nodes 3 and 8 during their 35-minute shortest path (route numbers in parenthesis*) (continued)

Transfer no.	Description
42(18, 26, 28)	3(18)–5(26)–7(28)–8
43(21, 25)	3(25)–5(25, 21)–7(21)–8
44(22, 25)	3(25)–5(25, 22)–7(22)–8
45(25, 27)	3(25)–5(25)–7(27)–8
46(25, 28)	3(25)–5(25)–7(28)–8

* When there is a possibility of more than one transfer, the node in which the transfer is considered indicates the two routes in parentheses

Table 14.7 SCP matrix of one set of routes and transfers

O-D pair	Routes						Transfer path	
	4	6	9	11	25	28	32	46
1, 2	2	2	–	2	–	–	–	–
1, 3	–	–	2	2	–	–	–	–
1, 4	–	–	–	2	–	–	–	–
1, 5	2	2	–	–	–	–	–	–
1, 6	–	–	2	–	–	–	–	–
1, 7	–	2	–	–	–	–	–	–
1, 8	–	2	–	–	–	–	–	–
2, 3	–	–	–	2	–	–	–	–
2, 4	–	–	–	2	–	–	–	–
2, 5	2	2	–	–	–	–	–	–
2, 6	2	–	–	–	–	–	–	–
2, 7	–	2	–	–	–	–	–	–
2, 8	–	2	–	–	–	–	–	–
3, 4	–	–	–	2	1	–	–	–
3, 5	–	–	–	–	2	–	–	–
3, 6	–	–	2	–	–	–	–	–

(Continued)

Table 14.7 SCP matrix of one set of routes and transfers (continued)

O-D pair	Routes						Transfer path	
	4	6	9	11	25	28	32	46
3, 7	–	–	–	–	2	–	–	–
3, 8	–	–	–	–	–	–	2*	2*
4, 5	–	–	–	–	1	–	–	–
4, 6	–	–	–	–	–	2	–	–
4, 7	–	–	–	–	2	2	–	–
4, 8	–	–	–	–	–	2	–	–
5, 6	2	–	–	–	–	–	–	–
5, 7	–	2	–	–	2	–	–	–
5, 8	–	2	–	–	–	–	–	–
6, 7	–	–	–	–	–	2	–	–
6, 8	–	–	–	–	–	2	–	–
7, 8	–	2	–	–	–	2	–	–

* Optimum path: contains a single transfer
Key: – The O-D pair is not covered by the route
1 Covered but not optimally
2 Optimally covered

each route is determined by the demand and assignment method. The *fourth element* calculates the optimization parameters PH_r , DPH_r , WH_r , EH_r on a selected route basis for computing Z_1 , and determines the minimum fleet size required Z_2 to meet the passenger demand. For instance, the calculated values of these optimization parameters of the set of routes, in Table 14.7, are shown in Figure 14.6. The step-by-step derivation of these parameters is shown in Figure 14.6 for Route 4 (the 1st route in the set); that is, given the assigned passenger demand for each O-D pair (the result of the third element) a load profile is constructed. Then, the max load, frequency (from which the waiting time is derived), and empty-space hours can be calculated. The values of PH_r and DPH_r depend on the average travel time for each O-D pair as well as on its assigned demand.

The *fifth element* is responsible for constructing alternative sets of routes in order to search for additional (Z_1, Z_2) values in the vicinity of their optimal setting. In the example problem nine sets were produced by this element; these sets are as shown in Table 14.8, including the resultant Z_1 and Z_2 values.

The *sixth element* of the methodology involves multi-objective programming of the two objective functions, Z_1 and Z_2 . Given the alternative sets of routes derived in the fifth element, the purpose here is to investigate which set provides the more efficient solution. The trade-off situation regarding the example problem is depicted in Figure 14.7. The lower left corner of

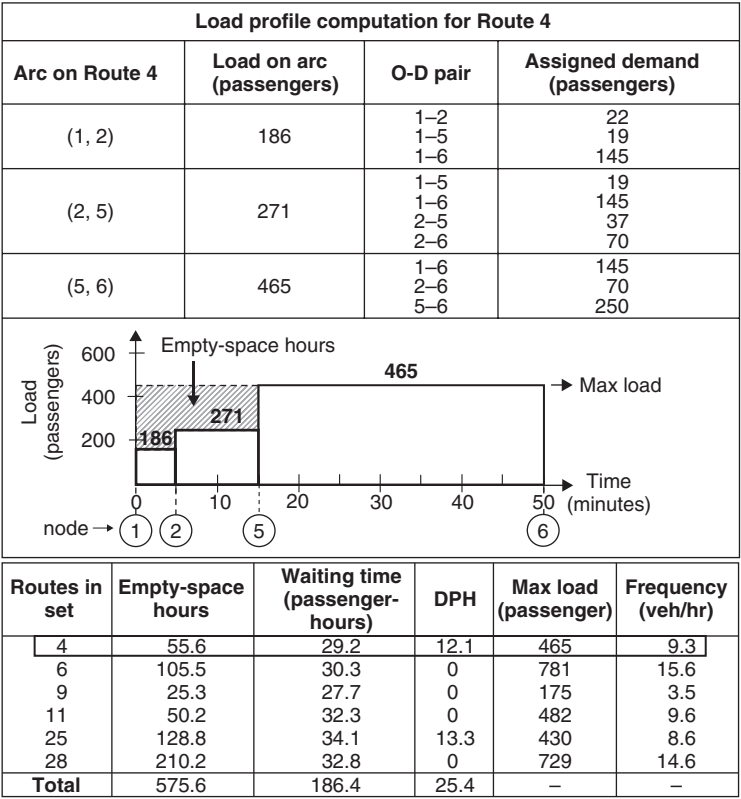


Figure 14.6 Computation of the load profile of Route 4 and the analysis required for Set 1 of routes (no. 4, 6, 9, 11, 25, 28) and transfers (no. 32, 46)

Table 14.8 All selected sets of the example problem and their objective functions values

Set	Description	Z ₁	Z ₂
1	{4, 6, 9, 11, 25, 28, 32, 46}	787	106
2	{7, 9, 11, 19, 25, 27, 34, 45}	900	109
3	{7, 9, 11, 25, 28, 34, 46}	1105	117
4	{6, 9, 11, 16, 25, 28, 32, 46}	866	102
5	{6, 12, 19, 25, 28, 32, 46}	937	103
6	{7, 12, 25, 27, 34, 45}	997	101
7	{7, 12, 25, 28, 34, 46}	1213	113
8	{4, 6, 12, 25, 28, 32, 46}	869	103
9	{6, 12, 16, 25, 28, 32, 46}	961	105

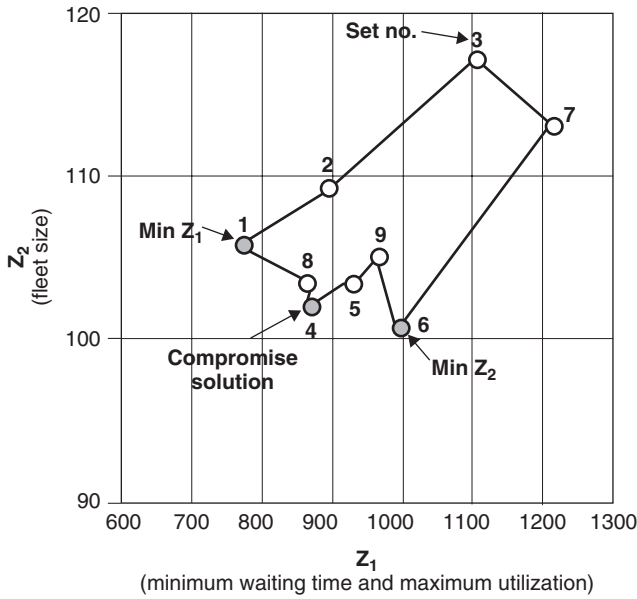


Figure 14.7 Trade-off between Z_1 and the minimum fleet size, Z_2 in the example problem

the envelope contour of the nine solutions represents the best sets; for a bi-objective problem, this boundary is called the Pareto front (Coello Coello *et al.*, 2002). The user is then able to choose a desired solution with this information in hand. In the example case, the choice is between $[Z_1 = 787, Z_2 = 106]$, $[Z_1 = 866, Z_2 = 102]$ and $[Z_1 = 997, Z_2 = 101]$ for sets 1, 4 and 6, respectively. Section 14.5 provides more details on the multi-objective analysis.

14.4 Construction of a complete set of routes

This section deals with the second element of the methodology described. The set-covering problem (SCP) is the creation of a minimum set of routes and their related transfers, such that connectivity between all nodes is maintained and their total deviation from the shortest path is minimized. An efficient heuristic search algorithm that has been tested with random networks will be described here. Its solution is compared to: (i) integer-programming optimization, without considering transfers; (ii) a nonlinear programming using relaxation methods on the integer variables for transit networks, with transfers; and (iii) a complete enumeration of all possible covering scenarios. The outcome is a set of a minimum number of routes that cover all the O-D pairs in the network, and on which demand can be assigned.

14.4.1 Formulation

The notations in Section 14.2 are extended to include:

$MA = \{A_R, A_{TR}\}$ = matrix of binary parameters, where $A_R = \{a_{ij}^r\}$, $A_{TR} = \{a_{ij}^{tr}\}$ and

$$a_{ij}^r = \begin{cases} 1, & \text{demand } i, j \in N \text{ can be handled directly by route } r \in R, \\ 0, & \text{otherwise} \end{cases}$$

$$a_{ij}^{tr} = \begin{cases} 1, & \text{demand } i, j \in N \text{ can be handled by transfer path } tr \in TR, \\ 0, & \text{otherwise} \end{cases}$$

The matrix MA, described in Table 14.9, consists of a set of columns A_R of direct routes and a set of columns A_{TR} of combined transfers. Each transfer column contains a set of routes that can simultaneously handle the demand of each $a_{ij}^{tr} = 1$ at a specific k_{tr} value. If only a subset of these routes can handle the demand $i, j \in N$, then $a_{ij}^{tr} = 0$. In this case, $a_{ij}^{tr} = 1$ either for transfers with lower k_{tr} value or for one of the routes that can directly handle the demand. The costs of a direct route and of a transfer path (each one is referred to a single column in Table 14.9) are designated c_r and c_{tr} , respectively.

Table 14.9 Configuration of feasible paths in matrix form (MA matrix)

	$r \in R$	$tr \in TR$			
		$k_{tr} = 1$	$k_{tr} = 2$	$\dots$	$k_{tr} = n$
$i, j \in N$	$\{a_{ij}^r\}$	$\{a_{ij}^{tr}\}$	$\{a_{ij}^{tr}\}$	$\dots$	$\{a_{ij}^{tr}\}$
	$\{c_r\}$	$\{c_{tr}\}$	$\{c_{tr}\}$	$\dots$	$\{c_{tr}\}$

These costs define the total deviation, in time units, from the shortest path for one passenger who applies the route or the transfer path:

$$c_r = \sum_{i,j \in N_r} (t_{ij}^r - t_{ij}^{sp}) \quad \forall r \in R \quad (14.13)$$

$$c_{tr} = \sum_{i,j \in N_{tr}} (t_{ij}^{tr} - t_{ij}^{sp}) \quad \forall tr \in TR \quad (14.14)$$

The SCP minimization formulation can now be established:

$$\text{Min} \left[\sum_{r \in R} c_r x_r + \sum_{tr \in TR} c_{tr} \pi_{r:tr} x_r \right] \quad (14.15)$$

s.t.

$$\sum_{r \in R} a_{ij}^r x_r + \sum_{f \in F} a_{ij}^{tr} \pi_{r:tr} x_r \geq 1 \quad \forall i, j \in N \quad (14.16)$$

$$x_r = \begin{cases} 1, & \text{route } r \in R \text{ in the solution} \\ 0, & \text{otherwise} \end{cases} \quad (14.17)$$

where $a_{ij}^{tr} = \{0, 1\} \quad \forall tr \in TR$, $a_{ij}^r = \{0, 1\} \quad \forall r \in R$

Note that in the covering matrix MA , a link exists for an O-D pair via either a direct route ($r \in R$) or a transfer ($tr \in TR$), and:

$$\sum a_{ij}^r + \sum a_{ij}^{tr} \geq 1 \quad \forall i, j \in N \quad (14.18)$$

The notation $r:tr$ signifies route r that contains the transfer path tr . Since the costs c_r and c_{tr} at this stage, before the assigning of demand, refer to the length of the travel path (direct routes and via transfers), the solution of the covering problem will yield the travel-path network (direct and indirect) with the minimum travel times.

It should be noted that consideration of the degree of transfer creates non-linearity, but reduces the size of the problem. While methods for solving the classic SCP (with no transfer columns; i.e. $a_{ij}^{tr} = 0, \forall i, j \in N$) that have been reviewed showed the employment of integer-programming (IP) techniques, the formulation (14.15)–(14.18) of nonlinear SCP with integer variables has not been considered.

14.4.2 Heuristic approach

The case of absence of transfer columns defines the classic SCP (linear with integer variables). The SCP is then formulated as IP, for which solution methods exist in the literature (e.g. Syslo *et al.*, 1983). These methods are incapable of encountering large matrices corresponding to real problems. In such problems, the number of variables (columns) is relatively large even for a small number of constraints. Thus, it is difficult or even impossible to use IP techniques or enumeration. Since the complexity of the SCP is of the NP-complete type (see Section 5.3 in Chapter 5), the use of known IP algorithms that are polynomials for a large problem is inefficient. The common method of solving the SCP is relaxation of the linear variable $x_r = \{0,1\}$, such that the integer constraint is replaced by $0 \leq x_r \leq 1$ (and sometimes $x_r \geq 0$). Then, in the first step, a linear programming (LP) is solved; and in the next step, regular LP techniques can be applied to solve this matrix if the non-integer solutions of the LP create a diminished matrix.

The nonlinear integer SCP formulated in (14.15)–(14.18) corresponds to practical and realistic cases. All but one of the studies mentioned in the previous section (the exception is Conely, 1980), refer to the classic IP case, whose relaxation makes it linear (LP). Thus, these approaches are not appropriate for solving the problem with transfers. Conely (1980) utilizes LP for the reduction of the matrix dimensions; however, a nonlinear problem (NLP) can be solved randomly on the next step of the random lottery. In this method, combinations of covering columns are chosen randomly while a frequency function is built in order to identify the probability of the global optimum. The drawback of this procedure is its high computation time.

It is worth noting that relaxation of the binary variable x_r entails a convex NLP, the quality of the solution to which is not clear; this is for two reasons: (1) obtaining non-integer values for the variables and rounding them off upward to an integer cannot guarantee optimality; (2) the closeness of a local optimum (obtained by integer variables in the NLP solution) to the global optimum is unknown. In addition, the type of the problem's complexity, NP-complete, can be an obstacle to solving NLP with large matrices.

Consequently, a heuristic approach is called for in which consideration is given to a compromise between efficiency and accuracy. This algorithmic-based approach provides

a number of solutions (if such exist) on a search tree. It should be noted that a single solution (one branch of the search tree) is attained in a short time.

The initial stage of the heuristic approach is a possible reduction of matrix MA by the following definitions:

- *Exclusive columns*: Columns that serve exclusively one or more O-D pairs will be chosen for the solution. If such a column is a transfer $tr \in TR$, then all the routes associated with it will be chosen. The corresponding rows will also be eliminated from the matrix.
- *Dominant columns*: If a column k covers at least all O-D pairs of column ℓ , and $c_k \leq c_\ell$, then column ℓ will be eliminated.
- *Dominant rows*: If all columns that cover row q can cover row u and more, then row u will be eliminated.

The heuristic algorithm, called CPCC (cost per covered cell), embodies two stages: *Forward* – with an ‘add column’ step; and *Backward* – with a ‘delete column’ step. At these stages, columns can be added or deleted according to the criterion of cost per covered cell. This approach, as opposed to the one that considers the total cost of each column, can lower the objective function (14.15) by choosing a column with relatively high cost, but with high coverage (number of covered rows).

These are further notations for the algorithm:

- MA^ℓ = matrix MA in iteration ℓ where $MA^0 \equiv MA$; and $MA^\ell = \{A_R^\ell, A_{TR}^\ell\}$
 P_r = columns contained in route r ; P_r is part of A_R^ℓ
 P_{tr} = columns contained in transfer tr ; P_{tr} is part of A_{TR}^ℓ
 P_k = column k in the matrix MA; k can be r or tr
 $\{P_k\}$ = set of columns (routes and transfers) generated by adding or deleting a column to/from the solution
 $i, j \in N^\ell$ = O-D pair $i, j \in N$ whose row exists in the reduced matrix MA^ℓ
 S = solution matrix (set of columns that cover matrix MA)

Forward stage

The search at this stage is based on varied upper bounds of the cost of the candidate column; this bound is decreased during the iteration. The process described is for iteration ℓ .

Step 1 (candidate column): In MA^ℓ , choose an unmarked (definition in *Step 4*) column P_k with $\text{Min}_k c_k$. The number of covered rows (cells) in column k is CV_k^ℓ . The cost per covered cell of column k is

$$\hat{c}_k = \frac{c_k}{CV_k^\ell}, \quad P_k \text{ is part of } MA^\ell, \quad c_k = \text{Min}_{k'} c_{k'} \quad (14.19)$$

Two main cases exist:

- (a) $k = r$, in which the route does not entail transfer columns (or the problem does not contain transfers); hence,

$$c_k = c_r \quad \forall k = r \in R \quad (14.20)$$

$$CV_k^\ell = CV_r^\ell = \sum_{i,j \in N^\ell} a_{ij}^r \quad \forall k = r \in R \quad (14.21)$$

(b) $k = r$, in which the route entails candidate transfer columns. Now $tr(r)$ is defined as a candidate transfer column tr and entered into the solution because of the candidacy of route r ; thus,

$$c_k = c_r + \sum_{tr(r)} c_{tr} \quad \forall k = r \in R \quad (14.22)$$

$$CV_k^\ell = CV_r^\ell = \sum_{i,j \in N^\ell} \left(a_{ij}^r + \sum_{tr(r)} a_{ij}^{tr} \right) \quad \forall k = r \in R \quad (14.23)$$

Note that the maximum value of the expression in parenthesis of Equation (14.23) must be 1 for each $i, j \in N^\ell$ because some ‘entailed’ transfers can cover the same pairs; thus if it exceeds 1, only the value 1 is considered.

Step 2: (marked columns): Indicate the candidate column P_k and its associated entailed columns according to the two cases of *Step 1*:

$$\{P_k\} = \begin{cases} P_r, & \text{case (a) of Step 1} \\ P_r & \text{in which } \{P_{tr}|tr = tr(r)\}, \text{ case (b) of Step 1} \end{cases} \quad (14.24)$$

Step 3: Calculate the uncovered rows in matrix MA and denote them as NC^ℓ ; this NC^ℓ is equal to the total number of rows in matrix MA^ℓ .

Step 4: (varied upper bound): Calculate the varied upper bound, denoted UB^ℓ , for the search of more possible columns in matrix MA^ℓ :

$$UB^\ell = NC^\ell \cdot \hat{c}_k, \quad P_k \text{ is part of } MA^\ell, \quad c_k = \min_{k'} c_{k'} \quad (14.25)$$

Step 5: If in matrix MA^ℓ a column exists with

$$c_k \leq UB^\ell, \quad P_k \text{ is part of } MA^\ell \quad (14.26)$$

then again perform (once only) *Steps 1* and *2*; next go to *Step 6*; otherwise, go to *Step 7*.

Step 6: If the chosen column fulfils the following:

$$\hat{c}_k \cdot NC^\ell < UB^\ell \quad (14.27)$$

then assign:

$$UB^\ell = C_k \cdot NC^\ell \quad (14.28)$$

A reduced new upper bound is obtained; go to *Step 5*.

Step 7: End of column search of iteration ℓ . From all the marked columns, select the one that fulfils Equation (14.28); i.e. the new upper bound obtained by this column. If more than one such column, P_k , exists, select the one with the $\max_k CV_k^\ell$ value; if more than one exists with this maximum value, select the one with $\min_k CV_k^\ell$ in order to prevent the selection of long routes that serve only a few

O-D pairs. In case several columns exist with the same such maximum and minimum values, then the solution contains more than one branch.

- Step 8:* If the chosen column (and its entailed columns) $\{P_k\}$ fulfils $S \cup \{P_k\} = S'$, in which S' is a partial solution matrix obtained in one of the previous searches (former iterations or in the same iteration), then there is a repetition of the same junction in the search tree that was previously reached; backtrack to another branch. If there are no more solutions, go to *Step 9*.
- Step 9:* For the chosen $\{P_k\}$, execute $S = S \cup \{P_k\}$, assign $x_r = 1$, and $MA^\ell = MA^\ell - \{P_k\}$; cancel the marks of columns in *Step 2*.
- Step 10:* Cancel the rows that were covered by the chosen column and its entailed columns $\{P_k\}$; in each row, a_{ij}^k in the CV_k^ℓ calculation. A new reduced set of rows, D^ℓ , is obtained.
- Step 11:* If $D^\ell = \phi$, *Forward stage* is terminated. Go to *Backward stage*; otherwise, $\ell = \ell + 1$. A reduced matrix MA^ℓ (in columns and rows) is obtained; go to *Step 1*. If $MA^\ell = \phi$, the search is completed.

Backward stage

In this stage, covered columns are deleted from the solution matrix S , which means that all their O-D demands can be carried out by other columns (routes and transfers). The size of the matrix is relatively small, thus yielding three different approaches for solving the problem: heuristic, complete enumeration, and mathematical programming. These are explained as follows.

- (1) *Heuristic approach.* Here, the covered columns in matrix S are marked. The column selected for deletion is the one with the maximum c_k and includes the associated transfers costs, similar to Equation (14.22), with $\min_k CV_k^0$ the latter means a less effective column. The last condition affirms the possibility of deleting other covered columns. The process continues sequentially, one column at a time, until the set of covered columns is empty. It should be noted that covered columns that have the same $\max_k c_k$ and $\min_k CV_k^0$ create several branches in the solution tree.
- (2) *Complete enumeration approach.* This approach creates a matrix MA^{en} of all covered columns. Since the size of such a matrix is small, it is possible to treat it manually. A complete enumeration can be performed on all combinations of columns, while each combination is covered by other columns in MA^{en} . The combination that contributes the largest decrease to the objective function (i.e. has the $\max \sum_k C_k, p_k$ is part of MA^{en} , where c_k is similar to Equation (14.22)) will be selected. In this approach all relevant covered columns will be deleted after one iteration, and the final reduced matrix S of the solution will be obtained.
- (3) *Mathematical programming approach.* In this approach, a collection of columns to remain in S will be selected from MA^{en} . These columns will contribute the least to the total cost (objective function), while affirming connectivity. The problem is formulated as nonlinear integer programming and can easily be solved, since the matrix is small. Hence, the possibility exists of obtaining an integer solution; or otherwise, it is relatively

easy to round off fractions upwards to integers. The mathematical formulation is similar to the SCP Equations (14.15)–(14.18). In Equation (14.15), only columns in MA^{en} will be considered; and in Equation (14.16), the right-hand side will include 1 minus the sum of binary parameters of the columns in $S-MA^{en}$ that serve the same $i, j \cup N$ demands.

14.4.3 Numerical example utilizing the CPCC algorithm

A simple example of a five-node network is depicted in Figure 14.8, in which there is one terminal at node 1 and the travel times are indicated on the arcs. Additional data are as follows: (i) maximum allowed 40% deviation from the shortest path, $\alpha = 0.4$; (ii) maximum of one transfer allowed, $k_{tr} = 1$.

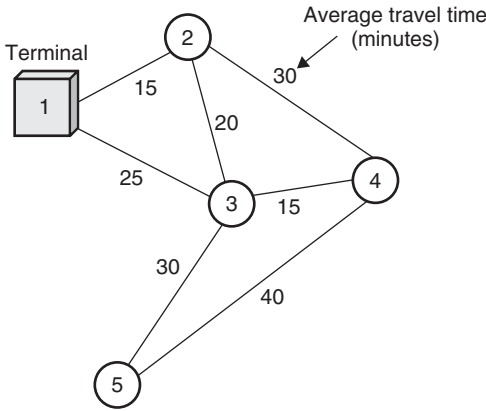


Figure 14.8 Example of a five-node network for demonstrating the set-covering analysis

The process of mapping the feasible routes results in eight routes. Following is a list, with route number and path description: (1) [1-2]; (2) [1-2-3]; (3) [1-2-3-4]; (4) [1-2-3-5]; (5) [1-2-4]; (6) [1-3]; (7) [1-3-4], and (8) [1-3-5]. For simplicity sake, transfers are allowed only between disconnected O-D pairs; in the example problem, only pair (4,5) does not have a direct connection to any of the eight routes. A possible path for the (4,5) demand is 4-3-5, using node 3 as the transfer node. Thus, four transfer paths can be constructed, listed here by transfer number and the combination of routes (route number) required: (9) [3 + 4]; (10) [3 + 8]; (11) [4 + 7]; and (12) [7 + 8].

With this information in hand, we can now construct the MA matrix for solving the SCP. Table 14.10 presents the MA matrix, in which the cost is the sum of deviations from the shortest path for each i, j pair on each route; the cost of a transfer column is assumed to be zero. Whenever a column (route or transfer) can serve the row (O-D pair), ‘1’ is assigned; otherwise, ‘0’. No distinction exists between being served by the shortest path or by a feasible longer one.

Table 14.10 Matrix MA for the five-node network example

O-D pair	Route no.								Transfer no.			
	1	2	3	4	5	6	7	8	9	10	11	12
1, 2	1	1	1	1	1	0	0	0	0	0	0	0
1, 3	0	1	1	1	0	1	1	1	0	0	0	0
1, 4	0	0	1	0	1	0	1	0	0	0	0	0
1, 5	0	0	0	1	0	0	0	1	0	0	0	0
2, 3	0	0	1	1	0	0	0	0	0	0	0	0
2, 4	0	0	1	0	1	0	0	0	0	0	0	0
2, 5	0	0	0	1	0	0	0	0	0	0	0	0
3, 4	0	0	1	0	0	0	1	0	0	0	0	0
3, 5	0	0	0	1	0	0	0	1	0	0	0	0
4, 5	0	0	0	0	0	0	0	0	1	1	1	1
Cost c_k	0	10	25	20	5	0	0	0	0	0	0	0

The following notations (some of which are repeated) are used to ease the description of the CPCC algorithm:

ℓ = iteration number

NC = number of uncovered rows

Cand = candidate column

c_k = cost of a candidate column (Cand) k

Cells = number of uncovered rows (cells) covered by Cand

$\hat{c}_k$ = cost per covered cell of Cand

UB = varied upper bound

Chosen = column chosen to enter the solution

Solution = partial solution in iteration ℓ

Z = objective function value

Remark = remark about the chosen column or a partial solution.

The CPCC algorithm is illustrated in Tables 14.11 and 14.12. The first illustration, shown in Table 14.11, includes the consideration of exclusive columns; that is, those columns that exclusively serve one or more O-D pairs and are automatically chosen for the solution. In the example case, Route 4 (see Table 14.10) is the only exclusive column chosen that covers six of the ten O-D pairs. The next iteration selects columns 7 + 11 (Route 7 and Transfer 11), both of which manage to cover three more O-D pairs. The iteration ends with Route 5, which covers only the 'leftover' uncovered O-D (2,4) pair. In this case, no *Backward stage* is needed.

Table 14.11 CPCC solution with consideration of exclusive columns

ℓ	NC	Cand	c_k	Cells	$\hat{c}_k$	UB	Chosen	Solution	Z	Remark
1	10	4	20	6	$3\frac{1}{3}$	$13\frac{1}{3}$	4	4	20	Exclusive column
2	4	7 + 11	0	3	0	0	7 + 11	4, 7, 11	20	
3	1	5	5	1	5	5	5	4, 5, 7, 11	25	End Forward
	0									No Backward

Table 14.12 CPCC solution without consideration of exclusive columns

ℓ	NC	Cand	c_k	Cells	$\hat{c}_k$	UB	Chosen	Solution	Z	Remark
1	10	1	0	1	0	0	–			
		6	0	1	0	0	–			
		7	0	3	0	0	7	7	0	High coverage
		8	0	3	0	0				Another branch
2	7	8 + 12	0	3	0	0	8 + 12	7, 8, 12	0	
3	4	1	0	1	0	0	1	1, 7, 8, 12	0	
4	3	5	5	1	5	15	5	1, 5, 7, 8, 12	5	
		2	10	1	10	–				
5	2	2	10	1	10	20				
		4 + 11	20	2	10	20	4 + 11	1, 4, 5, 7, 8, 11, 12	25	High coverage; End Forward
6	0	1	0	1			1	4, 5, 7, 8, 11, 12	25	Start Backward; low overlapping
	0	8 + 12	0	4	–					
7	0	8 + 12	0	4			8 + 12	4, 5, 7, 11	25	End Backward

The second illustration, shown in Table 14.12, implements the CPCC algorithm without any preliminary search for ‘good’ columns. In the first iteration, the procedure selects Route 7 because it covers three O-D pairs; however, this first alternative creates another branch of Route 8, which also covers three O-D pairs. The procedure continues with the *Forward stage* and ends this stage with seven columns (Routes 1, 4, 5, 7 and 8; and Transfers 11 and 12) that cover all O-D pairs. Then, the *Backward stage* is activated, in two iterations, and leads to a reduction in unnecessary overlapping columns. It ends with the same optimal (for this example) solution as in Table 14.11.

14.4.4 Evaluation of the CPCC algorithm

The common evaluation criteria of a heuristic algorithm are as follows: number of solutions, accuracy (closeness to optimality), and computing time. The CPCC algorithm was tested using different parameters on random networks. The computing time depends on the number of rows, number of columns and the density of the SCP matrix. The number of rows is more dominant than the number of columns. The density of the matrix has two impacts: (a) the computing time of one branch of the search tree is relatively small, because fewer columns are needed for the cover; (b) there are more possible branches for the tree.

The algorithm was tested for the SCP matrix with and without transfers. In the case without transfers, the algorithm was compared with an IP approach and with two known algorithms, those of Salkin (1975) and Conely (1980). The IP approach employed a relaxed LP by the MPSX software, similar to the process and package utilized in Section 10.4.2 in Chapter 10. The test results appear in Table 14.13, in which ‘single branch’ denotes the selection of the first branch of the search tree of the heuristic algorithm and ‘minimal branch’ denotes the execution of the algorithm for all possible branches affirming solutions. The comparison criteria in Table 14.13 are the degree of accuracy and the deviation in percentage; the former defines the fraction of time in which the solution attains optimality, while deviation (%) refers to the maximum difference between the best and the worst solutions. About 100 SCP matrices were created with the largest rows-times-columns size of 100×100 . In addition to the results in Table 14.13, the heuristic CPCC has the advantage of a smaller computation time than the other methods.

For the case of the SCP matrix with transfers, the two-version CPCC algorithm was compared with (a) complete enumerations of all possible covers and (b) a relaxation of the

Table 14.13 Comparison of two versions of the CPCC algorithm and other methods for the case without transfers

Algorithm	Degree of accuracy	Deviation (%)
CPCC – single branch	88.9	8.3
CPCC – minimum branch	100	0.0
MPSX	100	0.0
Conely (1980)	66.7	58.0
Salkin (1975)	88.9	21.0

nonlinear problem (NLP) together with rounding off upward of the fractions into integer values. Table 14.14 presents the results derived from the solutions of 100 SCP matrices utilizing different network sizes. Clearly, solving the CPCC for all possible branches (and selecting the minimum) yields the optimal solution, as do the results from complete enumeration. The relaxed NLP provides the least accurate results.

Table 14.14 *Comparison of two versions of the CPCC algorithm and other methods for the case with transfers*

Algorithm	Degree of accuracy	Relative deviation (%)
CPCC – single branch	91.5	7.4
CPCC – minimum branch	100	0.0
Relaxed NLP	74.6	38.5
Complete enumeration	100	0.0

In summary, the test results, shown in Tables 14.13 and 14.14, show that if the heuristic algorithm is solved for only single branches – which does not necessarily minimize the objective function – there is about a 90% probability of achieving the optimal solution. This finding, for instance, can be used for large SCP matrices intended for small computers.

14.5 Multi-objective technique

This section deals with Elements 5 and 6 in Figure 14.3 and their relationship to Elements 2, 3 and 4 (SCP, assignment procedure, and calculation of Z_1 and Z_2). Following Israeli and Ceder (1995, 1996), the section provides a general procedure and algorithm based on a given covering matrix that generates ‘promising alternative solutions’ (sets) for the multi-objective solution. Although the solution of the SCP is based on the costs derived from the assignment procedure, the assignment procedure itself is based on the solution of the SCP. This yields an iterative procedure, the outcome of which is an alternative-solutions generation process.

It may be seen from Equations (14.3)–(14.7) that there is need to minimize the four objective functions in which either Equation (14.5) or (14.6) is considered. Let us term Equations (14.3), (14.4), (14.5 or 14.6) and (14.7) as Z_{11} , Z_{12} , Z_{13} and Z_2 , respectively. In fact, because of the conflict among the objective functions, it is impossible to arrive at an ideal solution incorporating simultaneous minimization of all the functions constructing Z_1 and Z_2 . Since the four objective functions, Z_i , $i = 11, 12, 13, 2$ are complete, it is complicated to combine them into one unit scale without an accompanying loss of information, although their separation into Z_1 and Z_2 is a good practical alternative. Treating the four objective functions separately entails a trade-off among these functions. By doing so, however, there might be no optimal solution, but a variety of compromise solutions among the objective functions. The

choice of the 'best' compromise can only be facilitated by the establishment of various solutions, themselves created by due process.

The multi-objective programming problem can be classified into two differing alternative characteristic types: discrete problems and continuous problems. Discrete-type problems are based on a number of alternatives from which one is preferred. Continuous problems require a model entailing decision variables (e.g. the columns in the SCP), constraints and objective functions for creating suggested alternatives. Such variables may have any value from a given successive value structure (Cohon, 1978; Coello Coello, *et al.* 2002). The solution techniques of multi-objective programming are based on the fact that the dimension of the space of the objective functions in most practical problems is much smaller than that of the decision variables; hence, it is easier to perform the space of the objective functions while referring only occasionally to that of the decision variables. The choice of multi-objective technique requires two stages: creating efficient solutions and choosing the compromise solutions. The problem, when analysed, is of a nonlinear nature (concave) and contains integer variables, including the complex form of NP-complete. Such characteristics prevent the use of mathematical programming techniques usually inherent in small dimensional problems.

14.5.1 Iterative process

The real costs of matrix MA columns (routes and transfers in Table 14.9) are not known. The costs can result from the demand-assignment procedure, which by itself is based on the minimum set of routes; i.e. the SCP solution. That is why it is impossible to solve the SCP to minimum real costs but only to minimum travel time. This section commences the iterative process of constructing efficient alternatives for resolving the multi-objective problem. The process consists of three main stages:

- Stage 1:* Producing a minimum set of direct/indirect travel paths (routes/transfers) affirming the connectivity of the network; this is Element 2 in Figure 14.3, covered in the previous section.
- Stage 2:* The execution of the assignment process regarding the covering set created in *Stage 1* in the previous stage; the outcome of this stage consists of vehicle frequencies of the set of routes, passenger-load profiles, demand assignment across the set of routes, and the optimization parameters of Z_i , $i = 11, 12, 13, 2$. At the same time, the minimum fleet size required to meet the demand, as well as to satisfy the determined frequency on each route, is estimated.
- Stage 3:* A deletion of the detrimental variables (routes/transfers) up to the value of the acceptance of a new (reduced) set of routes and transfers; then back to *Stage 1*. The process executed in *Stage 3* guarantees two main results: (1) there will always be a cover to the reduced matrix MA; (2) previous alternatives will not be repeated.

Once the criteria reach the desired number of alternative solutions, the multi-objective problem is resolved. This section deals mainly with the relationship between *Stages 1* and *3* and with resolving the preceding multi-objective problem. Methods for *Stage 2* appear in Chapter 12.

The process separates into master problem and sub-problem. The master problem deals with building the covering matrix MA and solving it in such a manner as to create different solutions (new sets of routes transferring network demand). The initial point is the covering

sets established in Section 14.4. At this stage, the ideal solution is defined anew; that is, the point at which each of the objective functions has its minimum value. In addition, the process defines the set with the minimum distance to the ideal point. The column most detrimental to the set's cost is deleted from this set, and the process repeats itself. The minimum distance varies, since the ideal point can be changed during the iterations.

The sub-problem, termed SDP (set deletion problem), handles the production of feasible solutions for the covering problem, SCP, of the master problem. Two main elements affect the SDP: (1) If a column is deleted in the master problem (or a combination of columns) from MA, a situation could arise in which MA has no covering; in such a case, a column should be returned (or a partial combination of columns). (2) In the covering process for the master problem, it is essential that no endless loop be allowed to enter; this means that old existing sets (solutions) should not be accepted.

14.5.2 The master problem and SDP

The algorithm described examines at each stage the sets of feasible points of the five objective functions in Equations (14.3)–(14.7) while generating a new solution that can guarantee approaching the ideal point. If such solutions exist, the outcome will consist of all sets of efficient solutions (efficient points in the five-dimension description). The following are additional (to that in Section 14.4) notations for the master-problem algorithm. The algorithm then follows.

CAN = a group of all sets of routes (solutions of SCP) examined and are candidates for change

SOL = a group of all sets of routes selected to solve the multi-objective programming problem

n_c = number of sets in CAN

n_s = number of sets in SOL

$n_{\max}$ = an upper bound for the required number of sets

n_{set} = an upper lexicographic bound for the selection of a column from a re-examined set

L_s^k = distance from the exponent of s of the set (of columns) $k \in \text{CAN}$.

Step 0 (Selection of criteria): Assign initial values to $n_{\max}$, n_{set} , and s .

Step 1 (A group of candidate sets): Solve the SCP of the initial matrix MA; on the solution's set(s), execute the demand-assignment process and calculate the estimator of the minimum fleet size; a feasible point of Z_i , $i = 11, 12, 13, 2$ is created; group the sets in CAN and in SOL (at this stage $n_s = n_c$).

Step 2 (Initial conditions of the algorithm): If $n_s > n_{\max}$, take SOL as the final feasible solution; otherwise, go to Step 3.

Step 3 (Finding the ideal point in the present iteration): The sets in CAN yields $Z_i^* = \text{Min } Z_i$, $i = 11, 12, 13, 2$ and defines the point $(Z_{11}^*, Z_{12}^*, Z_{13}^*, Z_2^*)$.

Step 4 (Distance values): For all sets in CAN, calculate the distance values L_s from the ideal point at which

$$L_s = \left[\sum_{i=1}^p (Z_i(x) - Z_i^*)^s \right]^{\frac{1}{s}}, \quad \forall 1 \leq s \leq \infty \quad (14.29)$$

and $Z_i(x)$ is the optimal solution for alternative set x (see, for example, Goicoechea *et al.*, 1982, for more details).

Step 5 (Choosing the candidate set): From the sets in CAN, choose k' , which keeps a minimum 'potential distance' from the ideal solution; that is,

$$L_s^{k'} + \Delta L_s^{k'} = \min_{k \in \text{CAN}} (L_s^k + \Delta L_s^k) \quad (14.30)$$

where ΔL_s^k shows the difference in distance $L_s^j - L_s^k$ when set j is created last as a result of candidate set k in former stages (the set that has not been generated will be marked $\Delta L_s^k = 0$). This ΔL_s^k marks the production potential of distances chosen by set k . Its aim is to prevent the degeneration of the problem.

Step 6 (Candidate set's status): Check whether the set selected in *Step 5* was executed in former iterations (i.e. was a candidate from which other sets were generated). If so, go to *Step 8*; otherwise, continue.

Step 7 (A candidate set executed for the first time): Select the column of a singular route about to be deleted from the covering matrix; select from the set the column of route r' that delineates:

$$\hat{L}_s^{r'} = \max_r L_s^r \quad (14.31)$$

where $\hat{L}_s^{r'}$ is the distance value of the route (after the assignment process) from the set's ideal point for a covering unit, meaning the distance value divided by the number of pairs 'i, j' (rows in the covering matrix): those served by this route and the resultant transfers.

Step 8 (A candidate set that is re-executed): Select in the candidate set the preferred column to be deleted according to the hierarchical order of decreasing distance values, using Equation (14.31). There are two possible cases: (1) A column of a single route that was not performed in former stages (was not chosen as a candidate for deletion); (2) A combination of columns that was not performed in former stages. The hierarchical order is based on a binary number of n_{set} digits, in which '1' denotes the column to be selected and '0' otherwise.

Step 9 (A candidate set that cannot be re-executed): Examine whether one of two conditions exists for a candidate set: (a) the group of columns to be rejected is given by n_{set} with '1' in all its digits, in which case, the n_{set} bound was completely extracted; (b) the group of columns to be rejected is given by a number less than n_{set} , whose right digit is '0', the others are '1', and the set itself contains n_{set} columns. In the latter case, the set is totally extracted. If the lexicographic deletion process is continued, the set will be aborted at the next stage.

Step 10 (Deletion of columns from MA): Present the column(s) of the route(s) selected in *Steps 7 or 8* as columns in an SDP matrix (see next section).

Step 11 (Examining the SDP solution): If SDP performs a feasible solution, go to *Step 12*. Otherwise, if the candidate set is still in CAN, replace the columns that were deleted and go to *Step 8*; otherwise, go to *Step 15*.

Step 12 (Solving SCP using the reduced MA): Delete from matrix MA the columns of routes r that are the solution of SDP and the columns of transfers t_r associated with them; solve SCP.

- Step 13 (Obtaining another alternative):* In the set obtained in *Step 12* (the solution of SCP), the demand-assignment procedure will be executed and, hence, the minimum fleet size calculated. These results yield a new point for Z_i , $i = 11, 12, 13$, 2 in the four-dimension description. Add the new set to CAN and to SOL.
- Step 14 (Process terminating – first possibility):* Let $n_s = n_s + 1$; if $n_s < n_{\max}$, go to *Step 3*; otherwise, the process is terminated. $n_{\max}$ feasible alternatives are attained.
- Step 15 (Process terminating – second possibility):* If $CAN = \emptyset$, go to *Step 5*; otherwise, the process is terminated. $n_s < n_{\max}$ feasible alternatives are attained.

It should be noted that in each iteration of the algorithm described, SOL will certainly increase while CAN might (although not necessarily) decrease or even receive the value $CAN = \emptyset$ (meaning that $n_s \geq n_c$). The decrease in n_c is subject to *Steps 5 and 11*. The number of alternatives to be generated depends on the density (configuration) of matrix MA and the bounds selected in *Step 0*.

The sub-problem SDP (set-deletion problem) checks the columns about to be deleted from matrix MA in the master problem. A different matrix is defined whose columns are those deleted from sets in former iterations and from the present candidates, and its rows are the fitting sets. The dimensions of the matrix increase as the iteration progresses within the master problem, so that a row must be added in each iteration to mark the examined set (*Step 5* in the master-problem algorithm); a column may be added (*Step 7*) or replaced (*Step 8*). Thus, the covering problem is solved, resulting in a minimum amount of columns to be rejected from MA and no repetition of sets (unlike in SOL).

The SDP problem has definite covering and, according to the chosen columns, two types of solutions emerge: (1) a feasible solution, in which the columns chosen are deleted from matrix MA, and MA has coverage; (2) an unfeasible solution, in which the columns chosen are not permitted a covering for MA. The second case is derived from a possible combination of columns in SDP that exclusively cover one entire row in MA (meaning that '1' appears on at least one row of these columns, while '0' appears on the rest). Such a combination, designated a 'unique combination', if rejected from MA, will result in '0' values at least on one entire row and, as a result, will avoid coverage. The formulation and solution of SDP attempt to prevent the entrance of a 'unique combination' into the solution through a so-called UCP (unique combination problem) covering analysis. In the absence of choice, it would be possible to perform a backtrack in the algorithm of the master problem in order to transfer the rejected route columns as described in *Step 11* of the master problem. Figure 14.9 presents a diagram of the generation of 'good' sets of solution alternatives (called 'alternatives'). The complete SDP, including UCP, algorithm, and formulations, appears in Israeli and Ceder (1996).

14.5.3 Evaluation and selection of alternative solutions

An evaluation of the alternatives generated and the selection of 'the best' one are performed by the method of 'compromised programming'. This procedure, developed by Zeleny (1974), corresponds mainly to the solution of linear multi-objective problems (continuous), while a derivative variation of this method is used for solving discrete problems. The latter is employed for our four objective-functions case.

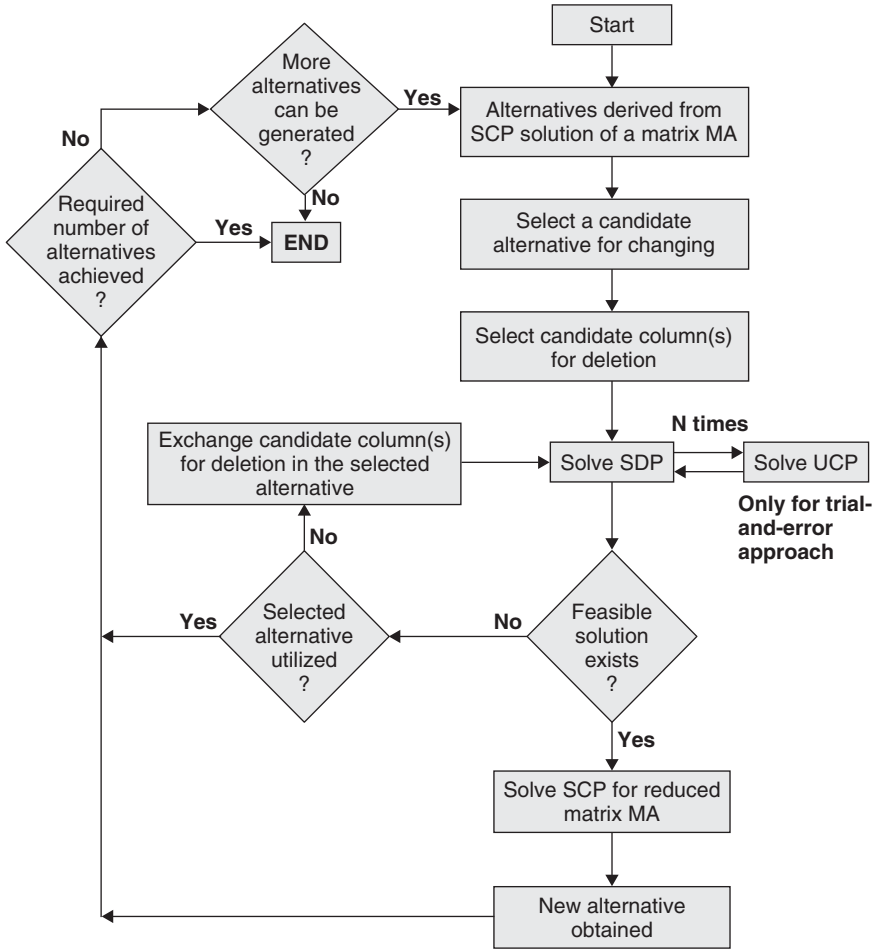


Figure 14.9 Flow diagram of the generation process of alternative solutions

From the process of generating alternative solutions, a finite number of feasible points $(Z_{11}, Z_{12}, Z_{13}, Z_2)$ are attained and can be arranged in a trade-off table. The calculation of the feasible solution points follows.

$$Z_i^* = \min_{k \in \text{SOL}} Z_i^k, \quad \forall i = 11, 12, 13, 2 \quad (14.32)$$

$$Z_i^M = \max_{k \in \text{SOL}} Z_i^k, \quad \forall i = 11, 12, 13, 2 \quad (14.33)$$

$$I_s^k = \left[\sum_{\forall i} \left[\frac{Z_i^k - Z_i^*}{Z_i^M - Z_i^*} \right]^s \right]^{1/s} \quad \forall k \in \text{SOL}, 1 \leq s \leq \infty, \quad i = 11, 12, 13, 2 \quad (14.34)$$

The compromise solutions are those that determine minimum distances from the ideal solution to each given s :

$$L_s^* = \min_{k \in \text{SOL}} L_s^k, \forall 1 \leq s \leq \infty \tag{14.35}$$

The study by Israeli and Ceder (1996) considers the values $s = 1, 2, \infty$ (s is exponent).

14.5.4 Numerical experience

The idea behind the entire heuristic process of generating alternative solutions is to provide most of the efficient points during the first stages of the process. The initial condition for choosing a candidate set was given in Equation (14.30) in Step 5 of the master problem. In order to examine this condition, it was compared to other initial conditions on an empirical basis. Two groups of 50 small-size (10 nodes) networks were produced in each condition. The first group was designed in such a fashion as to be able to produce a high number of alternative solutions from each network (depending on density, number of terminals, their location, etc.). The second group was designed to produce a small number (fewer than ten) of alternative solutions (usually small or low-density networks). The initial conditions that were compared:

- (1) the set with the minimum ‘potential distance’: $\min_{k \in \text{CAN}} (L_s^k + \Delta L_s^k)$
- (2) the set with the minimum distance: $\min_{k \in \text{CAN}} L_s^k$
- (3) the set with the maximum distance: $\max_{k \in \text{CAN}} L_s^k$

The first group of problems examines the solution quality; the results are described in Table 14.15. ‘Accuracy level’ defines the percentage of solutions that are identical to the minimum-distance solution. Closeness to the minimum solution in Table 14.15 denotes the relative (in %) distance from the minimum. For example, using condition (1), 50% of the alternative solutions that do not provide a minimum value have a distance of no more than 20% from the minimum value.

Table 14.15 Influence of different initial conditions on the solution

Initial condition	Accuracy level (%)	Closeness to minimum value (%)				
		0–20	20–40	40–60	60–80	80–100
(1)	88	50	33	17	0	0
(2)	60	25	25	45	4	1
(3)	28	6	14	28	20	32

The second group of problems examines the location (of the minimum distance) of the efficient alternative solutions during the process. Table 14.16 presents the percentage of solutions that achieve the minimum value. For instance, using condition (1), 65% of the minimum-value solutions are generated during the production of up to 25% of all possible alternative solutions. Tables 14.15 and 14.16 show that condition (1), used in the master problem, is the most effective one.

Table 14.16 Influence of different initial conditions on the location of the minimum solution

Initial condition	Location distribution of solution (%)			
	0–25	25–50	50–75	75–100
(1)	65	20	10	5
(2)	40	25	25	10
(3)	5	20	30	45

14.6 Literature review and further reading

This section contains a review of papers that propose methods of optimizing the configuration of transit-route systems. The output of such methods is a route itinerary that usually includes headways or frequencies. Papers that focus on headway determination, without a route-itinerary design, are discussed in Chapter 3. Papers in which special attention is given to unique objectives, such as the minimization of transfer times, are concentrated in Chapter 6. A review of additional network-design algorithms, mainly from the 1960s and 1970s, can be found in Axhausen and Smith (1984).

Hobeika and Cho (1979) present a method for determining the structure of a bus-route system. A heuristic algorithm partitions the existing bus stops in an urban area into sectors and seeks a way to link the stops while trying to minimize the total distance travelled by all buses. In the optimization process, each bus is subject to capacity and distance constraints. The routes developed are improved iteratively; in each iteration, a disaggregate choice model is used to examine passengers' behaviour. Equilibrium between supply and demand is reached when the proportion of passengers using buses cannot be increased by improving the bus network.

Marwah *et al.* (1984) develop a method for the simultaneous design of routes and frequencies. First, passenger flows are assigned on the road network. Then, a large set of possible bus routes that satisfy certain constraints is generated. Finally, routes that minimize the number of system-wide transfers are selected. Heuristics are used for the concentration of flows on the road network and for the initial generation of routes. Linear programming is used for the selection of optimal routes and for assigning frequencies.

Van Nes *et al.* (1988) formulate a programming problem that sets route itineraries and frequencies with the objective of maximizing the number of trips in which a transfer is not needed. The constraints include a given fleet size and budget limitations. The authors discuss the advantages of formulating the route-design problem as a programming problem, such as the ability to add further constraints. With additional constraints, it is possible to use the model in systems in which some existing routes may not be changed or there is a given limit to the number of routes, etc.

List (1990) describes a methodology for preparing optimal sketch-level service plans. The sketch-level plan does not include precise routes, but it does determine passenger-flow

values and frequencies on the road network. The model includes constraints for demand satisfaction, fleet size, minimal frequency, load factor, junction capacity, train length and crew requirements. Several sub-models are developed that help to formulate those parts of the model that stem from limited resources, such as energy consumption. The model makes sure that network flows are balanced and takes into account multi-period passenger demand.

Baaj and Mahmassani (1991, 1992, 1995) develop transit-network-design methods based on artificial intelligence (AI). The methods discussed are developed by a typical formulation of the network-design problem as a programming problem with minimum frequency, load-factor and fleet-size constraints. The first paper (1991) uses flowcharts to present a quantitative description of a three-stage design process for a route network. In the first stage, a large set of routes is generated; the second stage involves network analysis and a determination of frequencies; the third stage is network improvement. The second paper (1992) focuses on a method of representing the transportation network by using lists and arrays to make the solution procedure efficient. The third paper (1995) concentrates on the stage of creating the initial set of routes, which are supposed to be modified now and improved later on. In order to generate this initial route set, a set of basic skeletons is created along the shortest paths between nodes with high passenger demand; the skeletons are expanded, using a set of node-insertion manipulations.

Spasovic and Schonfeld (1993) introduce a method for determining optimal route lengths, route spacing, headways, and stop spacing in a radial network with one central business district (CBD). Cost functions are developed for both the operator and the users by minimizing their sum; equations are derived for the optimal values of all decision variables. A many-to-one demand pattern is assumed, and passenger density can either be uniform or decrease linearly with distance from the CBD. In addition, a solution algorithm is developed that incorporates realistic vehicle-capacity constraints.

Spasovic *et al.* (1994) extend the model described in the previous paragraph. The decision variables in this version include route lengths, route spacing, headways and fares. Two alternative optimality criteria are examined – operator profit and social welfare, the latter being the sum of user and operator surpluses. Social welfare is optimized with both unconstrained subsidies and break-even constraints. Analytical solutions are sought for a rectangular transit corridor with elastic demand, uniformly distributed passenger-trip density, and many-to-one travel patterns.

Ramirez and Seneviratne (1996), using GIS, propose two methods for route-network design with multiple objectives. Both methods involve ascribing an impedance factor to each possible route and then choosing those routes that have the minimum impedance. In the first method, the impedance factor depends on passenger flow and on the road length travelled. This method requires the use of an assignment model. In the second method, the impedance factor depends on the number of employees who have a reasonable walking distance from the route.

Patanik *et al.* (1998) present a methodology for determining route configuration and associated frequencies, using a genetic algorithm. Solutions are chosen in an iterative process from a large set of possibilities in which the chances of a solution's surviving through the iterations are higher if it yields a high value for a given fitness function. The method presented here adopts the typical programming formulation of the route-network-design problem with the objective of minimizing a weighted combination of passenger-time costs and operator-time costs; the objective function provides the basis for the calculation of the fitness-function

values. A methodology is also presented for the coding of variables as strings with a fixed or variable length.

Soehodo and Koshi (1999) formulate a programming problem for designing transit routes and frequencies. Similar to other models, the problem is solved by first creating all feasible routes and then choosing an optimal subset. In addition to some traditional components, such as minimum frequency and fleet-size constraints, the problem adds some unique elements, such as the inclusion of private car user costs, transit-passenger crowding costs, and transfer costs, to the minimized objective function. A sub-model is developed for each of these cost types. Equilibrium of network flows is another constraint. The model assumes that demand is elastic, and therefore the shift of passengers between different modes of transport plays a major role. Both transit and non-transit demand-assignment models are used.

Bielli *et al.* (2002) describe another method for designing a bus network, using a genetic algorithm. As in other genetic algorithms, each population of solutions goes through reproduction, crossover and mutation manipulations, whose output is a new generation of solutions. In the proposed model, each iteration involves demand assignment on each network of the current set of solutions and a calculation of performance indicators based on the assignment results. These indicators supply input to a multicriteria analysis of each network, leading to the calculation of its fitness-function value.

Wan and Lo (2002) develop a network-design model with an explicit consideration of inter-modal and inter-route transfers. The model has two separate phases. First, the points that are to be connected with a direct service are determined in a heuristic algorithm. This algorithm uses a network-representation approach called State Augmented Multi-Model (SAM), which involves inserting imaginary links into the actual road network where a direct service is provided. Afterwards, an actual bus-route system is built into a mixed-integer linear programming problem.

Yan and Chen (2002) present a method for designing routes and timetables that aims at optimizing the correlation between bus-service supply and passenger demand. The method is based on the construction of two time-space networks: a fleet-flow network and a passenger-flow network. Both networks are depicted in bi-dimensional diagrams in which the horizontal dimension represents bus stops and the vertical dimension represents time. While the fleet-flow network shows the potential activities of the bus fleet, the passenger-flow network illustrates trip demand. The objective of the model is to feed buses and passengers at minimum cost in both networks simultaneously. A mixed-integer, multiple-commodity, network-flow problem and a solution algorithm based on Lagrangean relaxation are presented.

Van Nes and Bovy (2002) and Van Nes (2003) investigate the influence of the definition of objective functions on the design of stop spacing and line spacing, and the preferences of different traveller groups, respectively. Van Nes and Bovy develop an analytical model, in which the objective functions are defined for the traveller, the agency and the authorities. Two alternative objectives for the traveller and six objectives for the authorities are examined, and the results of the interactions of different combinations are discussed. In addition, various alternative assumptions regarding demand elasticity are accounted for, and the difference in outcomes between different city sizes is analysed. Van Nes (2003) investigates the same model, but with different weights for different traveller groups. Van Nes found that the optimal result of network of routes was similar to the traditional single-user-class approach, thus concluding that a more realistic description of passenger groups was not necessary for the network-design theme.

Tom and Mohan (2003) continue the development of genetic methods for route-network design. In the current model, frequency is the variable; thus, it differs from earlier models in terms of the coding scheme adopted. Whereas fixed-string length coding and variable-string length coding were used in previous models, a combined route and frequency-coding model is proposed here.

Table 14.17 summarizes the main features reviewed and discussed in this section.

Table 14.17 *Summary of the characteristics of the methods reviewed*

Source	Special features	Unique objectives/ constraints	Intermodal considerations
Hobeika and Cho (1979)	Stops are divided into sectors		Yes
Marwah <i>et al.</i> (1984)		Minimizing the system-wide number of transfers	No
Van Nes <i>et al.</i> (1988)		Maximizing the number of direct trips	No
List (1990)	Sketch-level design only Multi-period demand	Junction capacity, train length, and crew-requirement constraints	Yes
Baaj and Mahmassani (1991, 1992, 1995)	Artificial intelligence		No
Spasovic and Schonfeld (1993)		Minimizing the sum of user and operator costs. Capacity constraint.	No
Spasovic <i>et al.</i> (1994)		Maximizing either operator profit or social welfare, with unconstrained subsidies and break-even constraints	No
Ramirez and Seneviratne (1996)	GIS-based. The 2nd proposed model takes into account potential demand (based on employment density), and not just existing demand		No
Pattanik <i>et al.</i> (1998)	Genetic algorithm		No

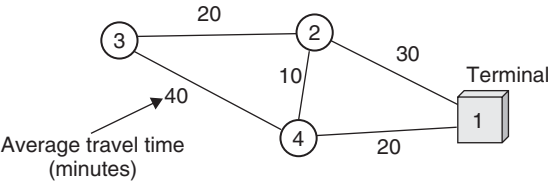
(Continued)

Table 14.17 Summary of the characteristics of the methods reviewed (continued)

Source	Special features	Unique objectives/ constraints	Intermodal considerations
Soehodo and Koshi (1999)	Elastic demand	Private car user costs, transit passenger crowding costs and transfer costs are included in the minimized objective function	Yes
Bielli <i>et al.</i> (2002)	Genetic algorithm	The fitness function is based on multi-criteria analysis	No
Wan and Lo (2002)		Inter-modal and inter-route transfers are considered through the network representation method	Yes
Yan and Chen (2002)	Network-flow problem formulation		No
Van Nes and Bovy (2002)	Elastic and fix demand examined	Multiple objectives of the traveller, the agency, and the authorities are defined and compared	No
Van Nes (2003)	Different traveller groups	Multiple objectives in which travel time and relationship between supply and demand depend on traveller groups	No
Tom and Mohan (2003)	Genetic algorithm		No

Exercises

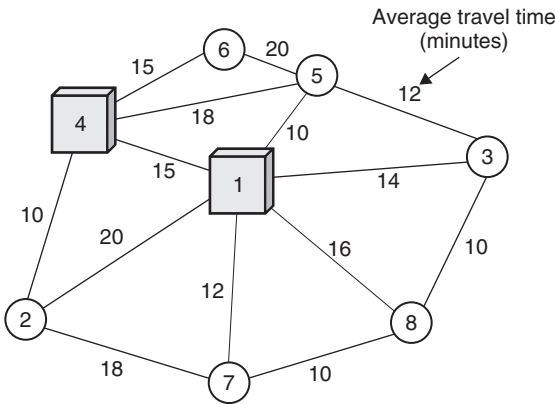
14.1 Given: (1) A bus service network consisting of four nodes (one terminal) and five arcs with average travel times (minutes) for both directions of travel.



(2) An O–D symmetrical demand matrix for a two-hour peak period; the arrival pattern of passengers is assumed to be homogeneously distributed for each cell;

From \ To			
	2	3	4
1	960	380	160
2		220	200
3	Symmetrical		240

- (3) The desired occupancy is 60 passengers per bus. (4) Bus trips are generated (start and end) only at the terminal (node 1) and deadheading trips are not allowed. (5) Two routes (round trips) serve the given network demand, Route A: 1 → 2 → 3 and backward to node 1; and Route B: 1 → 2 → 4 and backward to node 1. (6) Boarding, alighting, and transfer times are neglected.
- (a) For the two existing routes, derive the required headways based on the maximum load section method; if more than one alternative exists, select the most appropriate one and explain your decision.
- (b) What is the minimum fleet size required for Routes A and B, based on the derived schedule (headways).
- (c) Your task is to evaluate all possible single routes (round trips – same route forwards and backwards from the terminal) in the given network, visiting all nodes during the two-hour peak period. Note that the round trips may end after the peak hours.
- (c-1) Define all possible combinations of single routes (visiting all nodes).
- (c-2) Set an appropriate criterion (criteria) for your evaluation.
- (c-3) Construct the load profile with respect to travel time for each possible single route.
- (c-4) What is the minimum fleet size required for each single route?
- (c-5) Suggest a single route and compare it with the existing two routes.
- 14.2 Given: (1) The following bus-service network:



- (2) In accordance with their respective sequence of nodes, the following three routes exist: the first employs an articulated bus, 1–2–4–6, with a desired occupancy (d_o) of 75 passengers; the second, 4–5–3, uses a standard bus with $d_o = 50$; and the third, 1–8–7, employs a minibus with $d_o = 25$. (3) Symmetrical O-D matrix of a peak period, by number of passengers.

From \ To							
	2	3	4	5	6	7	8
1	80	100	120	100	180	30	40
2		100	100	140	200	50	30
3			100	80	120	20	30
4				60	140	30	20
5		Symmetrical			100	0	40
6						10	20
7							50

- (a) For each route r (one direction), calculate: difference in passenger-hours, PH_r ; empty-space hours, EH_r ; difference between passenger-hours and shortest path, DPH_r (without transfers); and passenger waiting time, w_r .
- (b) Calculate total DPH for all three routes, including transfers consideration.

- 14.3 One of two independent transit modes is considered for implementation, either fast ferry or fast train. The route examined is $A \rightarrow B \rightarrow C$. The following data are given:

Data item	Transit segment			
	Mode	A $\rightarrow$ B	A $\rightarrow$ C	B $\rightarrow$ C
Average travel time (minutes)	Fast train	40	50	10
	Fast ferry	70	90	20
Expected fare (\$)	Fast train	35	45	12
	Fast ferry	30	35	9
O-D demand (passengers)	Both modes	280	1100	700

Additional given data and information: (1) desired occupancy, in terms of number of seats, on the fast train and fast ferry is 600 and 450 passengers, respectively; (2) value of passenger waiting time is \$9/hour; and (3) value of passenger travel time is \$3/hour; (4) passengers arrive randomly at stops A and B; and (5) time-tables in both modes are based on even headways and Method 2 (see Section 3.2 in Chapter 3).

- (a) Calculate the hourly waiting-time cost of each alternative mode.
- (b) Calculate the loss cost of hourly empty seats and travel times.
- (c) Calculate the hourly income and profit of each alternative mode.
- (d) Based only on the given data, suggest the preferred mode.
- (e) List briefly actual cost elements that were neglected in this exercise.

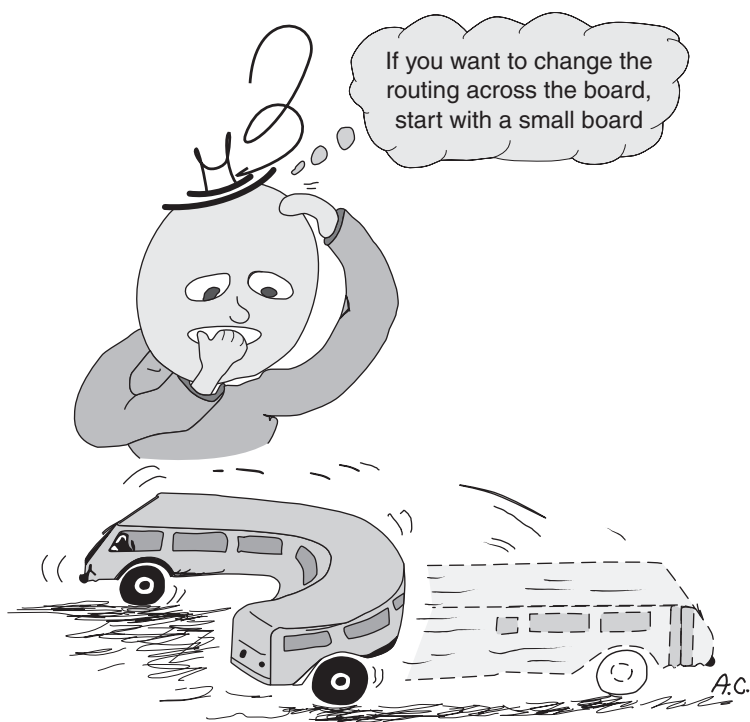
References

- Axhausen, K.W. and Smith, R.L. (1984). Evaluation of a Heuristic Network Optimization Algorithm, *Transportation Research Record*, **976**, 720
- Baaj, M. H. and Mahmassani, H. S. (1991). An AI-based approach for transit route system planning and design. *Journal of Advanced Transportation*, **25**, 187–210.
- Baaj, M. H. and Mahmassani, H. S. (1992). Artificial intelligence-based system representation and search procedures for transit route network design. *Transportation Research Record*, **1358**, 67–70.
- Baaj, M. H. and Mahmassani, H. S. (1995). Hybrid route generation heuristic algorithm for the design of a transit network. *Transportation Research*, **3C**, 31–50.
- Bielli, M., Caramia, M. and Carotenuto, P. (2002). Genetic algorithms in bus network optimization. *Transportation Research*, **10C**, 19–34.
- Ceder, A. (2001). Operational objective functions in designing public transport routes. *Journal of Advanced Transportation*, **35**, 125–144.
- Ceder, A. (2003). Designing public transport network and routes. In *Advanced Modeling for Transit Operations and Service Planning* (W. Lam and M. Bell, eds), pp. 59–91, Elsevier Ltd.
- Ceder, A., Gonzalez, O. and Gonzalez, H. (2002). Design of bus routes: Methodology and the Santo Domingo case. *Transportation Research Record*, **1791**, 35–43.
- Ceder, A. and Israeli, Y. (1992). Scheduling consideration in designing transit routes at the network level. In *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **386** (M. Desrochers and J. M. Rousseau, eds), pp. 113–136, Springer-Verlag.
- Coello Coello, C. A., Van Veldhuizen, D. A. and Lamont, G. B. (2002). *Evolutionary Algorithms for Solving Multi-objective Problems*. Kluwer Academic/Plenum.
- Cohon, J. L. (1978). *Multi-objective Programming and Planning*. Academic Press.
- Conely, W. (1980). *Computer Optimization Techniques*. Petrocelli Books.
- Dial, R. B. and Bunyan, R. E. (1968). Public transit planning system. *Socio-Economic Planning Science*, **1**, 345–362.
- Dubois, D., Bel, G. and Libre, M. (1979). A set of methods in transportation network synthesis and analysis. *Operations Research*, **30**, 797–808.
- Favolden, J. M. and Powell, W. B. (1994). Subgradient methods for the service network design problem. *Transportation Science*, **28**, 256–272.

- Goicoechea, A., Hansen, D. R. and Duckstein, L. (1982). *Multi-objective Decision Analysis with Engineering and Business Applications*. John Wiley & Sons.
- Heathington, K. W., Miller, J., Knox, R. R., Hoff, G. C. and Bruggman, J. (1968). Computer simulation of a demand scheduled bus system offering door to door service. *Highway Research Record*, **91**, 26–40.
- Hobeika, A. G. and Cho, C. (1979). Equilibration of supply and demand in designing bus routes for small urban areas. *Transportation Research Record*, **730**, 7–13.
- Israeli, Y. and Ceder, A. (1995). Transit route design using scheduling and multi-objective programming techniques. In *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **430** (J. R. Daduna, I. Branco and J. M. P. Paixao, eds), pp. 56–75, Springer-Verlag.
- Israeli, Y. and Ceder, A. (1996). Multi-objective approach for designing transit routes with frequencies. In *Advanced Methods in Transportation Analysis*. (L. Bianco and P. Toth, eds), pp. 157–182, Springer-Verlag.
- Keudel, W. (1988). Computer-aided line network design (DIANA) and minimization of transfer times in networks (FABIAN). In *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **308** (J. R. Daduna and A. Wren, eds), pp. 315–326, Springer-Verlag.
- Kim, D. and Barnhart, C. (1999). Transportation service network design: Models and algorithms. In *Computer-aided Scheduling of Public Transport*. Lecture Notes in Economics and Mathematical Systems, **471** (N. H. M. Wilson, ed.), pp. 259–283, Springer-Verlag.
- Kocur, G. and Hendrickson, C. (1982). Design of local bus service with demand equilibration. *Transportation Science*, **16**, 149–170.
- Kuah, G. K. and Perl, J. (1988). Optimization of feeder bus routes and bus-stop spacing. *Journal of Transportation Engineering*, **114**, 341–454.
- Lampkin, W. and Saalmans, P. D. (1967). The design of routes, service frequencies, and schedules for a municipal bus undertaking: A case study. *Operations Research*, **18**, 375–397.
- List, G. F. (1990). Towards optimal sketch level transit service plans. *Transportation Research*, **24B**, 324–344.
- Mandl, C. E. (1979). Evaluation and optimization of urban public transportation networks. *European Journal of Operation Research*, **5**, 396–404.
- Marwah, B. R., Umrigar, F. S. and Patnaik, S. B. (1984). Optimal design of bus routes and frequencies for Ahmedabad. *Transportation Research Record*, **994**, 41–47.
- Patnaik, S. B., Mohan, S. and Tom, V. M. (1998). Urban bus transit route network design using genetic algorithm. *Journal of Transportation Engineering*, **124**, 368–375.
- Pratt, R. and Evans, J. (2004). Traveler response to transportation system changes: Bus routing and coverage. In *TCRP Report 95, Chapter 10*. Transportation Research Board, Washington, D.C.
- Ramirez, A. I. and Seneviratne, P. N. (1996). Transit route design applications using geographic information systems. *Transportation Research Record*, **1557**, 10–14.
- Salkin, H. (1975). *Integer Programming*. Addison-Wesley.
- Silman, L. A., Barzily, Z. and Passy, U. (1974). Planning the route system for urban buses. *Computers and Operations Research*, **1**, 201–211.
- Soehodo, S. and Koshi, M. (1999). Design of public transit network in urban areas with elastic demand. *Journal of Advanced Transportation*, **33**, 335–369.

- Spasovic, L. N. and Schonfeld, P. M. (1993). Method for optimizing transit service coverage. *Transportation Research Record*, **1402**, 28–39.
- Spasovic, L. N., Boile, M. P. and Bladikas, A. K. (1994). Bus transit service coverage for maximum profit and social welfare. *Transportation Research Record*, **1451**, 12–22.
- Syslo, M. M., Deo, N. and Kowalit, J. S. (1983). *Discrete Optimization Algorithms*. Prentice Hall.
- Tom, V. M. and Mohan, S. (2003). Transit route network design using a frequency coded genetic algorithm. *Journal of Transportation Engineering*, **129**, 186–195.
- Tsao, S. and Schonfeld, P. (1984). Branched transit services: An analysis. *Journal of Transportation Engineering*, **110**, 112–128.
- Vandebona, U. and Richardson, A. J. (1985). Simulation of transit route operations. *Transportation Research Record*, **1036**, 36–40.
- Van Nes, R. (2003). Multiuser-class urban transit network design. *Transportation Research Record*, **1835**, 25–33.
- Van Nes, R., Hamerslag, R. and Immers, B. H. (1988). Design of public transport networks. *Transportation Research Record*, **1202**, 74–83.
- Van Nes, R. and Bovy, P. H. L. (2002). Importance of objectives in urban transit-network design. *Transportation Research Record*, **1735**, 25–34.
- Wan, Q. K. and Lo, H. K. (2002). 2-Phase transit network design in an integrated transit system. *Advanced Modeling for Transit Operations and Service Planning – Work-shop Proceedings*. Crousher Advanced Study Institute.
- Yan, S. and Chen, H. L. (2002). A scheduling model and a solution algorithm for inter-city bus carriers. *Transportation Research*, **36A**, 805–825.
- Yin, Y., Miller, M. and Ceder, A. (2005). Framework for deployment planning of bus rapid transit systems. *Transportation Research Record*, **1903**, 11–19.
- Zeleny, M. (1974). A concept of compromise solutions and the method of the displaced ideal. *Computers and Operations Research*, **1**, 479–496.

15 Designing Short-turn Trips



Chapter 15 Designing Short-turn Trips

Chapter outline

- 15.1 Introduction
 - 15.2 Methodology
 - 15.3 Candidate points and example
 - 15.4 Excluding departure times
 - 15.5 Maximum extensions of short-turn trips
 - 15.6 Literature review and further reading
 - Exercises
 - References
-

Practitioner's Corner

Subsequent to the overview and methods for designing routes at the network level, the next two chapters deal with specific design features at the route level. This present chapter presents a set of procedures to design transit timetables efficiently with trips that are initiated beyond the route departure point and/or terminated before the route arrival point. Such trips are called short-turn trips. In practice (see Chapter 4), transit frequency is determined at the heaviest load-route segment, whereas the operation at other segments may be inefficient because of situations, characterized by empty seats. Transit planners attempt to overcome this problem by manually constructing short-turn trips with the objective of reducing the number of vehicles required to carry out the transit timetable. The purpose of this chapter is to improve and automate this task.

The following riddle may serve as a stimulus for the need to handle a design element only after comprehending the process in which this element appears. The riddle: Given a micro-organism-biological process in which a microbe that is put into a glass splits into two after one second, the two then split into four after another second, the four into eight after another second, and so on; after one minute, the glass is full of microbes. How many seconds are required to fill half of the same glass? (Answer is given at the end of this Practitioner's Corner.)

This chapter contains four main parts, following an introductory section. Section 15.2 outlines the framework of the methodology for the efficient design of short-turn trips. Section 15.3 identifies a minimum number of candidate short-turn points and presents the example that is subsequently used throughout the other sections of the chapter. Section 15.4 provides a procedure to adjust the number of departures at each short-turn point to that required by the load data, provided that the maximum headway (associated with passenger waiting time) attained is minimized. Section 15.5 constructs another procedure to minimize the number of short-turn trips while ensuring that the minimum fleet size is preserved. The chapter ends with a literature review and exercises.

Practitioners are encouraged to visit Sections 15.1 to 15.3. In addition, they should follow all the figures, tables and corresponding paragraphs of the chapter that are related to the example problem.

It is known that an error in the premise will appear in the conclusion. Often, transit agencies avoid a profound design because of their premise that it would be costly, time-consuming and unrewarding. It is analogous to a common saying, and perhaps even belief, that good things in life are illegal, immoral or fattening. It will be shown in this chapter that a profound design (of short-turn trips) can, on the contrary, reduce cost, reduce time, and be rewarding. Lastly, the answer to the riddle is 59 seconds (in the 60-th second, the half will multiply itself); obviously this answer is strongly dependent on the process described.

15.1 Introduction

The previous chapter discussed the attainment of an efficient network of transit routes. This chapter focuses on improving the cost-effectiveness of each single route. Transit planners certainly understand the need to accommodate the observed passenger demand as well as possible. At the same time, however, their efforts are also directed at the minimization of vehicle and driver costs. The trade-off between increasing the passenger's comfort and reducing the cost of the service makes the planner's task cumbersome and complex. The design of short-turn trips is one such trade-off. A short-turn trip is initiated beyond the route's departure terminal and/or terminated before its arrival terminal. The possibility of generating short lines opens up an opportunity to further save on vehicles while ensuring that the passenger load on each route segment does not exceed the desired occupancy (load factor).

Planners in most transit agencies usually include a short-turn operating strategy in their attempts to reduce the cost of the service. The procedures commonly used are based only on visual observation of the load profile; that is, a potential turn point is determined at the timepoint (major stop) nearest to a stop at which a sharp decrease or increase of passengers is observed. Although this procedure is intuitively correct, planners do not know whether all the short-turn trips are actually needed to reduce the fleet size. Unfortunately, each short-turn trip limits the service and, hence, tends to reduce passenger level of service.

The major objectives of a short-turn operating strategy, and therefore of this chapter, are as follow:

- to identify minimum candidate short-turn points based on passenger load profile data;
- to adjust the number of departures at each short-turn point to that required by the load data, provided that the maximum headway obtained is minimized; this objective results in the maximum possible short-turn trips and the minimum required fleet size (including shifting departure times and DH trips);
- to minimize the number of short-turn trips, provided that the minimum fleet size attained (with short-turn trips) is maintained; for a given timetable, this objective results in increasing the level of service as seen by the passengers.

Several methods will be presented to meet the objectives. These methods are based on procedures and algorithms that use data commonly inventoried or collected by most transit agencies. The chapter is based on the methodology and modelling presented by Ceder (1990, 1991), with some improved and corrected elements.

15.2 Methodology

The proposed methodology relies on the following input data: (1) a complete timetable of all route timepoints; (2) passenger load profiles, by load and distance, for each time period; (3) minimum frequency or policy headway; and (4) a set of candidate short-turn points. These data are given for both directions of the route (each direction with its own data). Candidate short-turn points are usually all the major route stops (timepoints) at which the public timetable exists. In some cases, it may be limited to only those timepoints at which vehicles can actually turn back.

The comprehensive tasks needed to accomplish the objectives of the methodology are described in flow-diagram form in Figure 15.1; this methodology also appeared in Ceder (2003). It starts with a procedure to determine the set of feasible short-turn points R_j among the candidate points. Then the deficit function (DF) theory (Chapters 7 and 8) is used to derive the minimum number of vehicles required to carry out all the trips in the complete, two-direction timetable, $N_{\min}$. The required number of departures is determined at each of the feasible short-turn points, and then the so-called Minimax H algorithm is applied. The basis of the algorithm is the elimination of some departures from the complete timetable in order to obtain the number of departures required. In this procedure, the algorithm minimizes the maximum difference between two adjacent departure times (headway). At this stage, as shown in Figure 15.1, the DF method derives the minimum fleet size required with short-turns, $N'_{\min}$. If this minimum is less than the size required without short turns, then another procedure is applied. This second procedure inserts (back) the maximum possible departures among those previously eliminated, provided that the minimum fleet size, $N'_{\min}$, is maintained. The final step of the overall program is to create vehicle blocks to cover all the trips that appear in the last version of the two-direction timetable.

15.3 Candidate points and example

The short-turn points are usually route timepoints at which the vehicle can turn back without interfering with the traffic flow. It is therefore anticipated that for each route, the initial set of candidate short-turn points will be given by the transit planner.

15.3.1 Minimum candidate short-turn points

Let the set of candidate short-turn points be designated as set R_1 for one direction and R_2 for the opposite route direction. Note that R_1 does not necessarily coincide with R_2 . More specifically,

$$\begin{aligned} R_1 &= \{r_{11}, r_{12}, \dots, r_{1n}\} \\ R_2 &= \{r_{21}, r_{22}, \dots, r_{2q}\} \end{aligned} \quad (15.1)$$

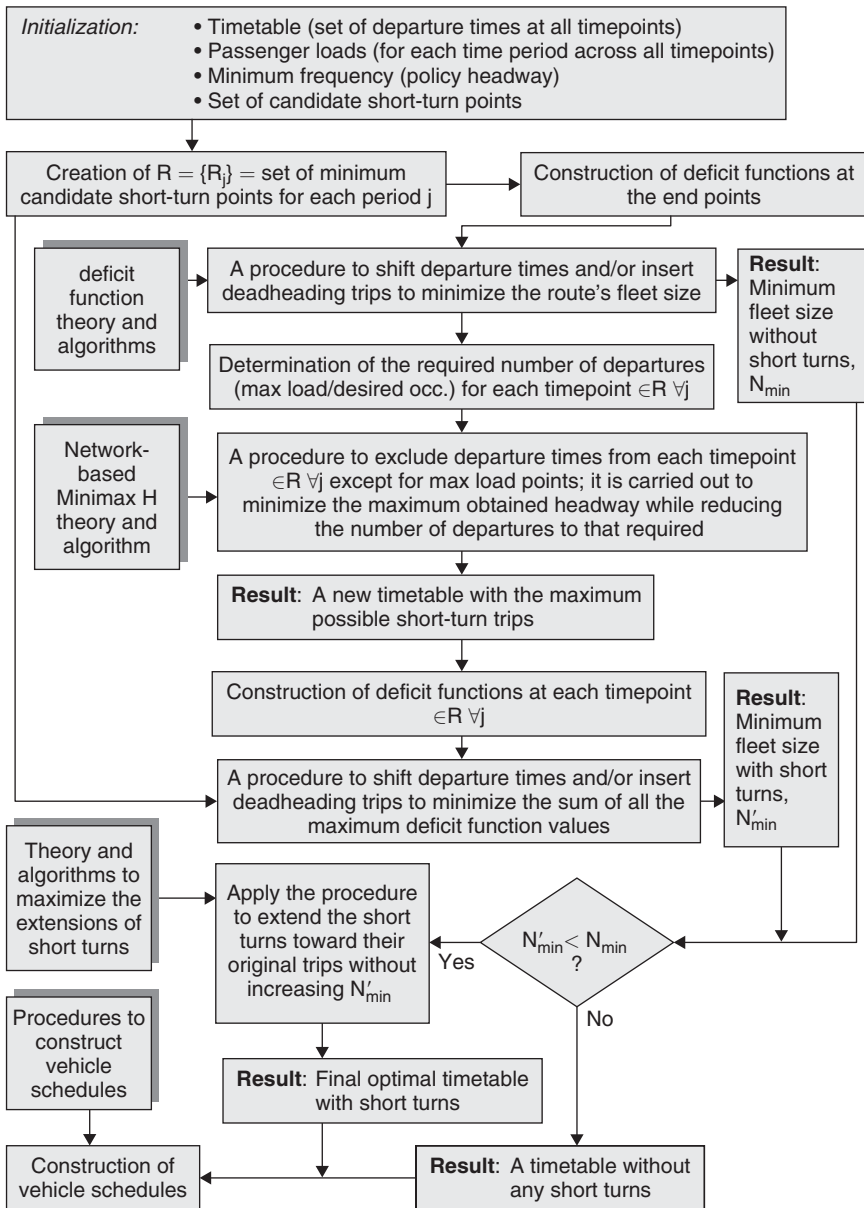


Figure 15.1 Flow diagram describing the design of efficient short-turn trips

where r_{ij} is the j -th candidate short-turn point in the j -th direction ($j = 1, 2$) and n and q are such points for directions 1 and 2, respectively.

For a given time period, the fluctuation of a passenger load along the entire route (load profile) may reveal that some short-turn points are actually redundant; thus, we can establish a set of a minimum number of candidate points. For example, say a load profile consists of

13 stops and 5 candidate short-turn points as shown in Figure 15.2. Theoretically, each segment between two adjacent short-turn points can be treated independently with respect to its required frequency. This frequency is determined by the maximum observed load on the segment, which is marked by the hatched area in Figure 15.2. In short-turn strategy, however, all the trips must serve the heaviest load segment of the route (in the example, all trips must cross the r_4 – r_5 segment). Another observation is that fewer trips are required between r_3 and r_4 than between r_2 and r_3 while both groups of trips must cross the r_4 – r_5 segment. Consequently, the point r_3 is redundant.

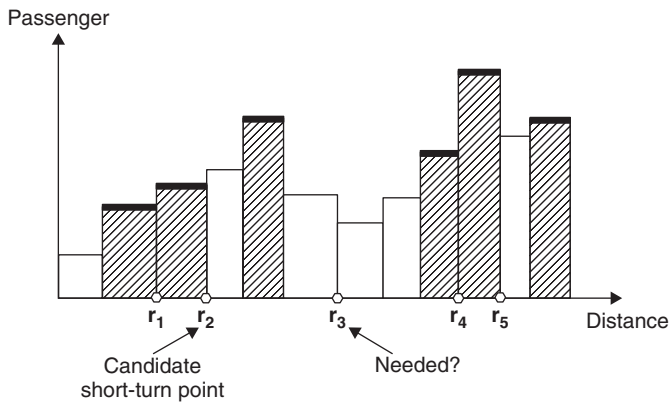


Figure 15.2 Construction of a set R of feasible short-turn points

The exclusion of the redundant points at each time period j results in a set of minimum candidate short-turn points, R_j ; this analysis is important from the viewpoint of computational time. The formal description of the algorithm to determine the set R_j includes an additional analysis of the difference between the required frequencies at the short-turn point associated with the max load segment and a considered short-turn point. If this difference is small, the considered short-turn point can be deleted. Note that the difference in the frequencies is equivalent to the difference in the load, and there are always stochastic variations of that load. Hence, if this difference is small, it is not reasonable to consider short-turn trips from the associated short-turn point. This is actually similar to the manual procedure performed in current practice in which the planner selects short-turn points only on the basis of an observed sharp increase or decrease in the load profile. For subsequent analyses, the union of all R_j for all (time periods) j is denoted as the set R .

15.3.2 Example

A simple example is used as an expository device to illustrate the DF approach and the procedures developed. This example appears in Figure 15.3 for a two-hour schedule of departure times at the max load points. The route includes three timepoints (a , b , c), which is the set R ; the average travel times for service and DH trips are given below the timetable in Figure 15.3; no shifting in departure times is allowed.

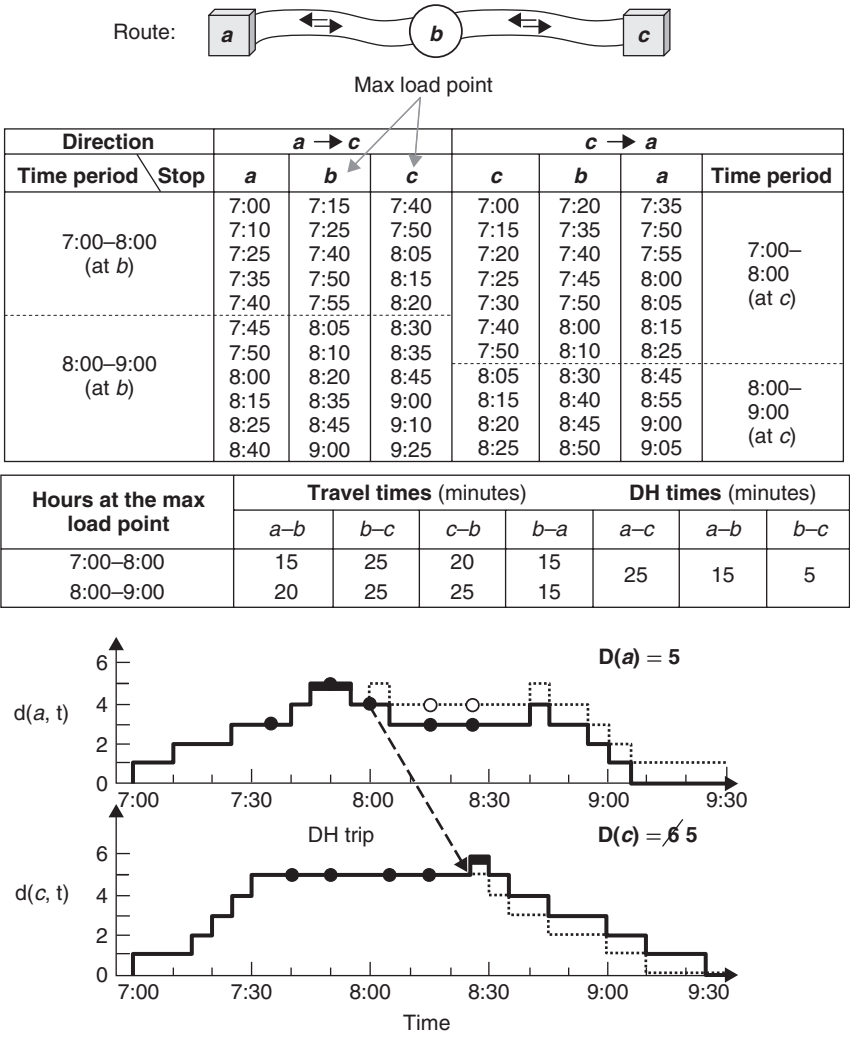


Figure 15.3 Example of a two-way route with a two-hour schedule for which 10 vehicles are required (based on the graphical deficit function method)

Construction of $d(a, t)$ and $d(c, t)$ can then take place, and the minimum number of vehicles required without DH trips is $D(a) + D(c) = 11$. However, a DH trip can be inserted from a to c , departing after the last maximum interval of $d(a, t)$ and arriving just before the start of the first maximum interval of $d(c, t)$. Both $d(a, t)$ and $d(c, t)$ are then changed, as seen by the dashed line in Figure 15.3; thus, $D(c)$ is reduced from 6 to 5, and the overall fleet size is reduced from 11 to 10. After that, it is impossible to further reduce the fleet size through DH trip insertion; hence, $N_{\min} = 10$. This condition can also be detected automatically by the lower-bound test. That is, the maximum of the combined DFs is 10, and therefore $N_{\min}$ reaches its lower bound.

15.4 Excluding departure times

The basic information required to consider short turns is the route's load profile. Based on this load-profile information, each route segment between two adjacent short-turn points can be treated separately. That is, the required number of trips between the $(k-1)$ th and k -th short-turn points for a given direction and given time periods is similar to Equation (3.2) in Chapter 3:

$$F_k = \max \left(\frac{P_k}{d_o}, F_{\min} \right) \quad (15.2)$$

where P_k is the maximum load observed between the two adjacent short-turn points; d_o is the desired occupancy; and $F_{\min}$ is the minimum required frequency. F_r is the route frequency, determined as in Chapter 14.

15.4.1 Level-of-service criterion

The manual procedure usually undertaken by the planner to create short-turn trips simply involves the exclusion of departure times in order to set the frequency at each short-turn point k to F_k instead of to F_r . This exclusion of departure times is performed without any systematic instructions in the (scheduler) belief that it is possible thereby to reduce the number of vehicles required to carry out the timetable.

The result of excluding certain departure times is that some passengers will have to extend their wait at the short-turn points. To minimize this adverse effect, it is possible to set the following (Minimax H) criterion: *Delete $F_r - F_k$ departure times at k with the objective of minimizing the maximum headway obtained.*

The Minimax H criterion attempts to achieve the minimization of maximum passenger waiting time; the result may represent an adequate passenger level of service whenever short-turn strategy is employed.

15.4.2 Minimax H algorithm

To solve the optimization problem with the Minimax H criterion, a theory based on three stages was developed by Ceder (1991): (1) representation of the problem on a directed network with a special pattern; (2) application of a modified shortest-path algorithm on the network to determine the Minimax headway; and (3) application of an algorithm to ensure that the exact number of required departures is included in the optimal solution. The Minimax H algorithm will now be outlined and then applied to the example problem presented in Figure 15.3.

Let $G_m = \{N_m, A_m\}$ be the special network consisting of a finite set of nodes N_m and a finite set A_m of directed arcs. Figure 15.4 presents a general illustration of the special network, accompanied by an example. There are n given departures from the complete timetable, and the requirement is that only $m < n$ need remain to satisfy the Minimax H criterion. The construction of G_m is based on $m - 2$ equally spaced departure times between the first and last given departures: t_1 , and t_n , respectively. These equally spaced departure times are denoted by $t'_2, t'_3, \dots, t'_{m-1}$ and have an equal headway of $t_e = (t_n - t_1)/(m - 1)$.

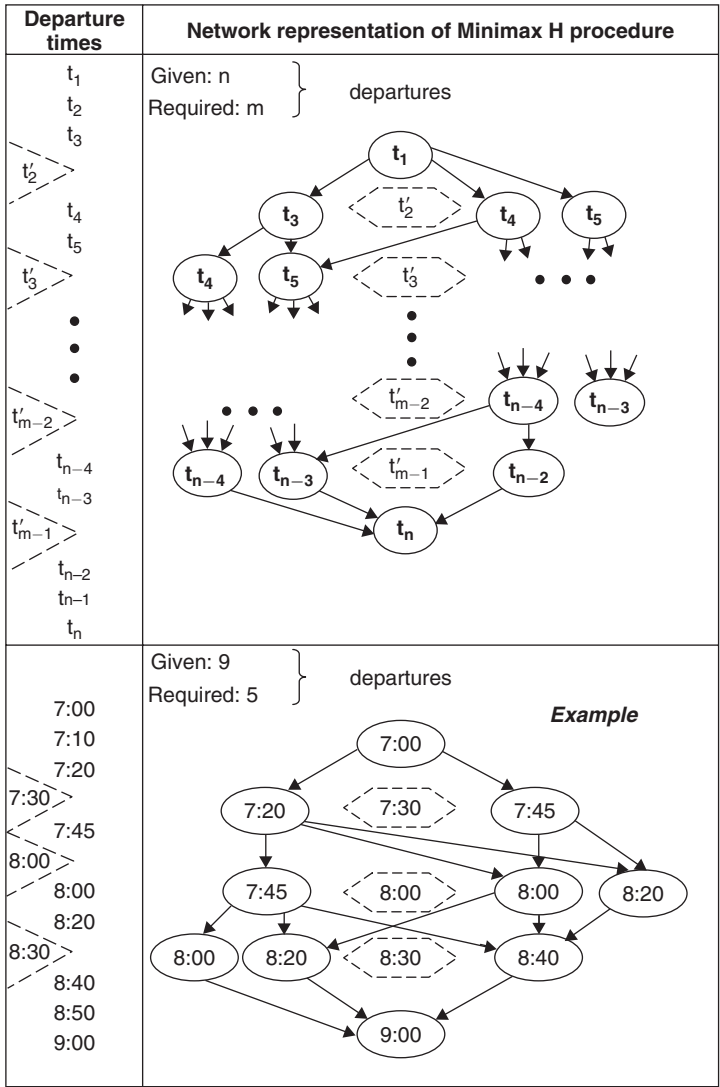


Figure 15.4 General network representation of the Minimax H procedure at one terminal, accompanied by an example

- The G_m network has these six characteristics:
1. G_m consists of m rows; the first and last rows are nodes t_1 and t_n , respectively, and there is a row for each t_j , $j = 2, 3, \dots, m - 1$.
 2. Each node in N_m represents a departure time in the given set of departures; however, it is not necessary that all given departures be included in N_m (see 7:10, 8:50 in the example of Figure 15.4), and also the same departures may be represented by several nodes (see 7:45, 8:00, 8:20 in Figure 15.4).

3. The nodes in each row are organized from left to right in increasing time order with respect to their associated t'_i . That is, all the given nodes t_i , such that $t'_k \leq t_i < t'_{k+1}$, are positioned twice, once to the right of t'_k and once to the left of t'_{k+1} , where t'_k, t'_{k+1} are two adjacent, equally spaced departure times. An exception occurs in the second and the $(m - 1)$ th rows, where only one node is positioned to the left of t'_2 and one to the right to t'_{m-1} , respectively. These single nodes, t_3 and t_{n-2} in Figure 15.4, are selected such that t_3 is the closest node to t'_2 , provided that $t_3 < t'_2$, and t_{n-2} is the closest node to t'_{m-1} , provided that $t'_{m-1} \leq t_{n-2}$.
4. The directed arcs in A_m connect only nodes from the k -th row to the $(k + 1)$ th row, $k = 1, 2, \dots, m - 1$.
5. A directed arc from t_i to t_j is included in A_m if $t_j > t_i$, and from t_i to t_i (in subsequent rows) if and only if G_m is disconnected without this arc.
6. The length of an arc from t_i to t_j is exactly $t_j - t_i$.

After constructing G_m , a modified shortest-path algorithm is applied as the second stage of the Minimax H procedure. This is a modified version of the Dijkstra algorithm described in Appendix 10.A in Chapter 10. The Dijkstra algorithm is based on assigning temporary labels to nodes, the label being an upper bound on the path length from the origin node to each node. These labels are then updated (reduced) by an iterative procedure. At each iteration, exactly one of the temporary labels becomes permanent, implying that it is no longer the upper bound but rather the exact length of the shortest path from the origin to the considered node.

The modification of the Dijkstra method takes place in the computation step, in which the labels are updated. It is modified from $\pi(t_i) = \text{Min}[\pi^*(t_k) + (t_i - t_k), \pi(t_i)]$ (see Step 3 of Dijkstra algorithm in Appendix 10.A), to

$$\pi(t_i) = \text{Min}\{\max[\pi^*(t_k), (t_i - t_k)], \pi(t_i)\} \quad (15.3)$$

where $\pi(t_i)$, $\pi^*(t_k)$ are temporary and permanent labels of nodes t_i and t_k , respectively.

This modified-Dijkstra algorithm, instead of minimizing the sum of arc values, searches for the minimum of the maximum arc value. The algorithm is applied to G_m when the origin node is t_i ; the algorithm terminates when the temporary label on node t_n becomes permanent.

Figure 15.5 exhibits the flow diagram of the three stages of the Minimax H algorithm. When the modified-Dijkstra procedure ends with $p < m$ departures, a secondary criterion, for the second largest headway, is introduced to find the difference of $m-p$ departure times; this is, basically the third stage. The third stage of the Minimax H procedure ensures that the optimal result includes exactly the required number of departures (m). Although the modified shortest-path algorithm for G_m determines the value of the Minimax headway, it does not ensure that the result will include all the m required departures.

The third stage is interpreted utilizing the eight different network configurations illustrated in Figure 15.6. Each example in Figure 15.6 is based on a different initial DT (departure times) set of five departures, from which one of the middle three is unnecessary. Cases (e) and (f) consist of a single path with three departures, whereas four departures are required. The additional departure required will be selected between t_2 and t_3 for case (e) and between t_3 and t_4 for case (f). The Minimax H = 45 minutes in both cases. If we select

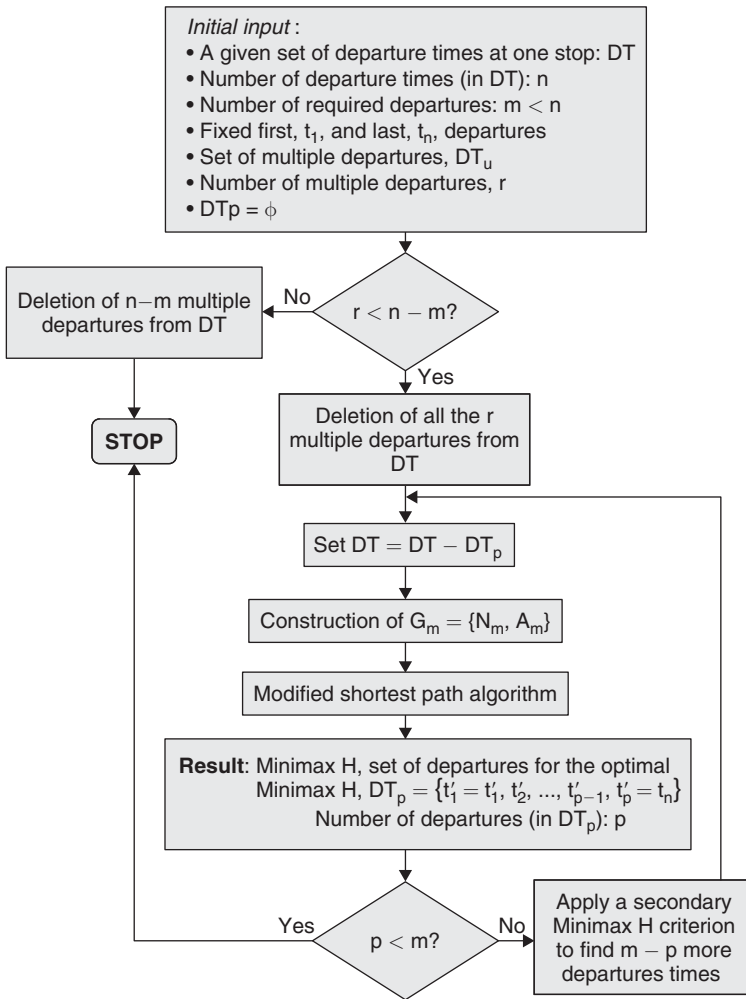


Figure 15.5 Schematic flow diagram of the Minimax H algorithm

the t_2 departure for case (e), then the second largest headway will be 11 minutes, as opposed to 10 minutes for t_3 . A similar situation exists when selecting t_3 instead of t_4 for case (f).

Cases such as (e) and (f) call for a secondary criterion of minimizing the maximum headway of a subset of DT. That is, for case (e), t_1 remains the same, t_n becomes $t_4 = 7:15$, $n = 4$, $m = 3$ and a new G_m is constructed and solved. For case (f), $t_n = t_5$ remains the same, t_1 becomes $t_2 = 7:45$, $n = 4$, and $m = 3$ of the new G_m . The detailed description of the third stage of the Minimax H algorithm appears in Ceder (1991), along with a procedure for treating multiple departures (with the same or more than one departure time in the given timetable).

The Minimax H algorithm is now applied to the example problem presented in Figure 15.3. The given and required numbers of departures for each hour and direction of travel are

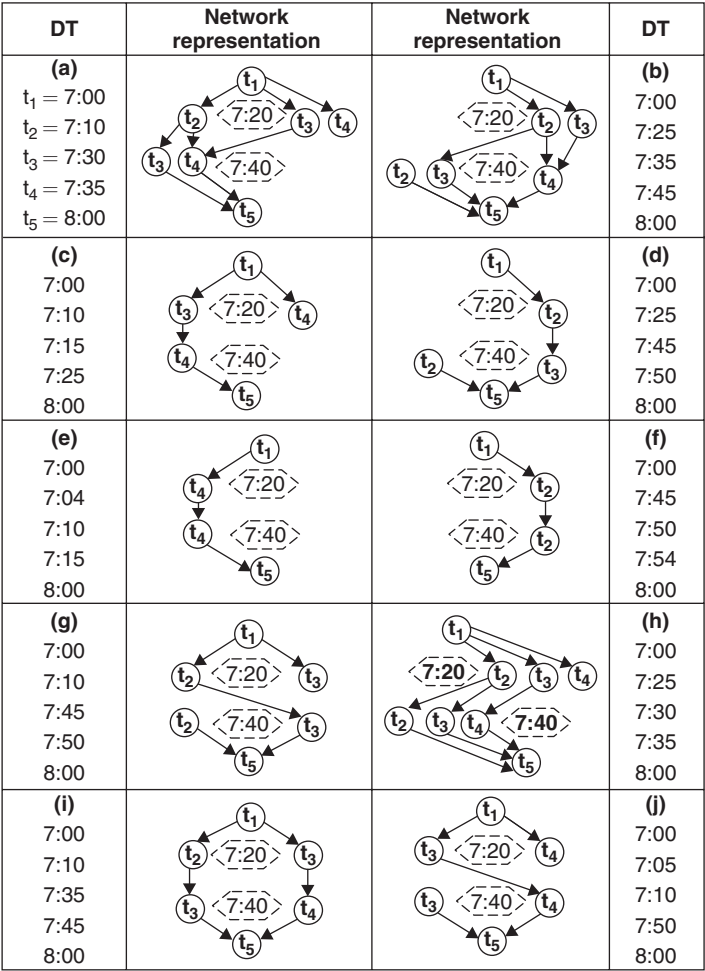


Figure 15.6 Different networks (based on a given 5-departure timetable at one stop) for a required schedule of 4 departures from 7:00–8:00

determined by the load profiles and $d_0 = 50$ in Figure 15.7. These required numbers (values of m) then undergo the Minimax H procedure in Figure 15.8. Four G_m networks are constructed to derive the Minimax headway. This derivation appears in the figure with an emphasized line indicating the optimal path of G_m and the labels of the shortest-path algorithm according to Equation (15.3). A dashed line and arrowhead indicate the direction of another optimal solution. Also note that between 7:00 and 8:00 t_n becomes t_1 for 8:00–9:00 in order to preserve the continuity of the analysis. The other parts of Figure 15.8 are self-explanatory. In all four cases, there is no need to proceed to the third stage of the Minimax H algorithm, because all the required departures are determined by the modified shortest-path algorithm.

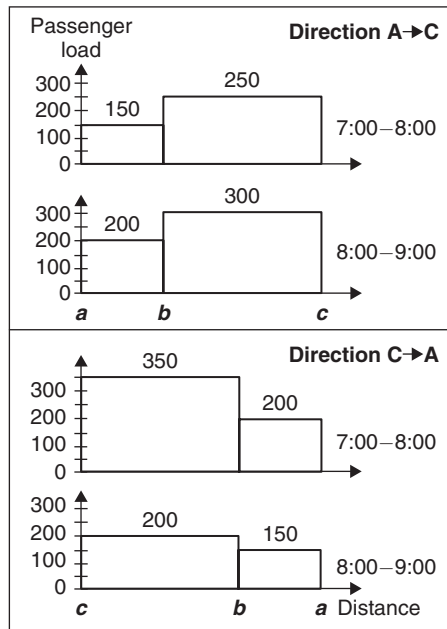


Figure 15.7 Load profiles of the example problem with a desired occupancy of 50 passengers per vehicle

The results of the Minimax H algorithm are then applied to the example problem in Figure 15.3. After the deletion of departures at timepoints a and b in directions $a \rightarrow c$ and $c \rightarrow a$, it is possible to construct the new timetable, along with the DFs. This time, however, all three timepoints (a, b, c) are involved. That is, in the modified timetable, some trips are initiated at b and some terminate at b in directions $a \rightarrow c$ and $c \rightarrow a$, respectively. Thus, point b also becomes an end/start point, and the deficit function description can be applied to it. The new timetable and DFs are presented in Figure 15.9; the resultant timetable (with maximum short turns) appears in the upper part of this figure. The corresponding DFs show that 10 vehicles are required to carry out the timetable without DH trips and that 9 vehicles ($N'_{\min}$) are required with a single DH trip from $d(c,t)$ to arrive at $d(b,t)$ before or at 8:35. Following the condition of Figure 15.1 for $N'_{\min} < N_{\min}$, where $N_{\min} = 10$ vehicles, the next step is to attempt to reduce the number and impact of short-turn trips, provided that $N'_{\min}$ is maintained.

15.5 Maximum extensions of short-turn trips

There are two similar undertakings for possible extensions of short-turn trips: (1) convert trips to points $r_i \in R$ from DH to service trips; (2) extend arrival and departure trips to points $r_i \in R$ toward their original schedule.

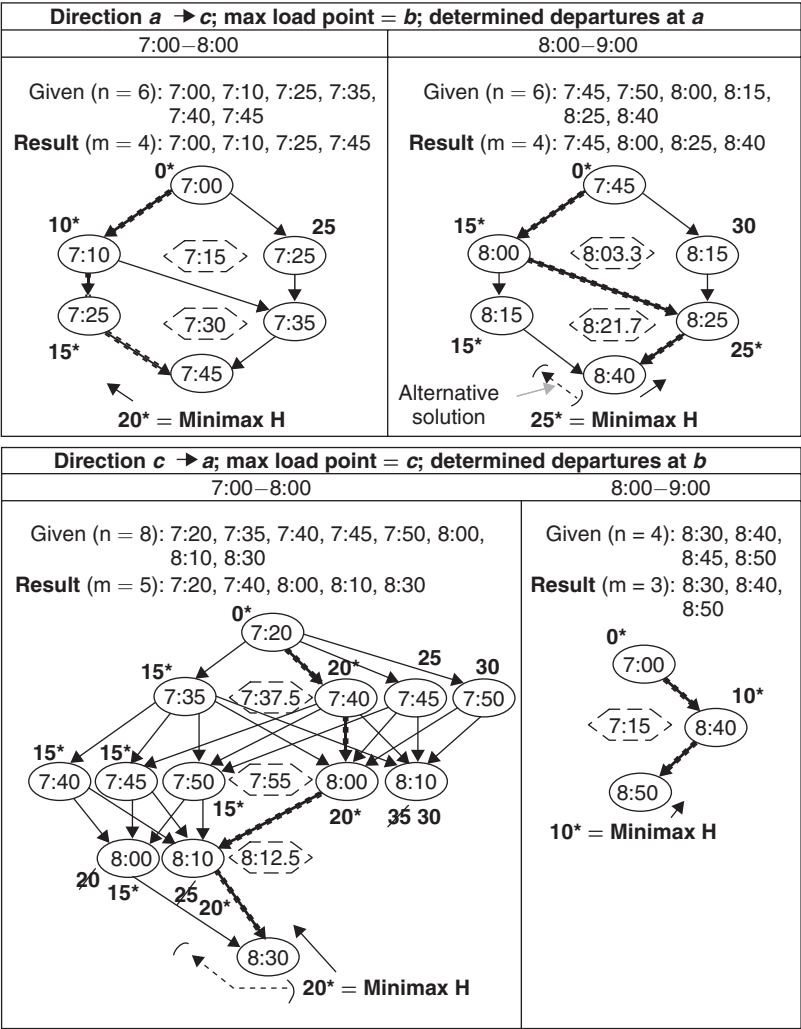
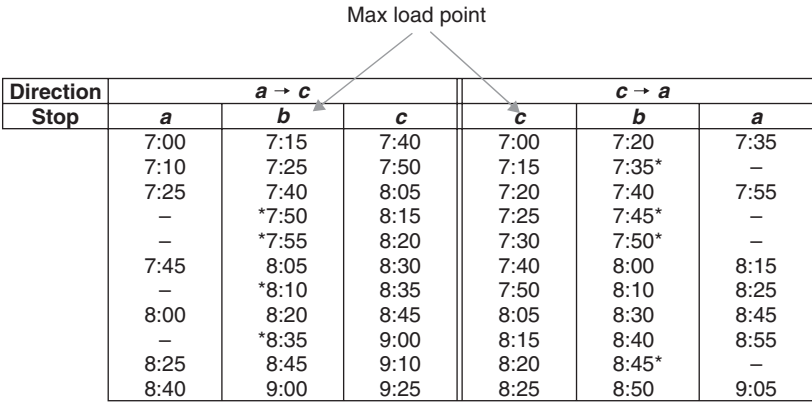


Figure 15.8 Minimax H algorithm and results for the example problem

15.5.1 Converting DH to service trips

To convert DH to service trips, start by denoting the modified timetable with maximum short turns by DT' ; the route end points by e_q , $q = 1, 2$; and the intermediate short-turn points by $r_i \in R$, $i = 1, 2, \dots, V$, where there are V short-turn points. To attain $N'_{\min}$, the overall schedule carrying out DT' might also include DH trips. This overall schedule is designated S . In this section, the DF properties will be exploited to check whether a DH trip can be interpreted as an extension of a short-turn trip in DT' .

By using DF theory, a DH trip can be inserted into a certain time window in order to reduce the fleet size by one. To simplify this possibility, a DH trip is inserted from one terminal to terminal k so that its arrival time always coincides with the first time that



* Departures and arrivals at b (creating short-turn trips)

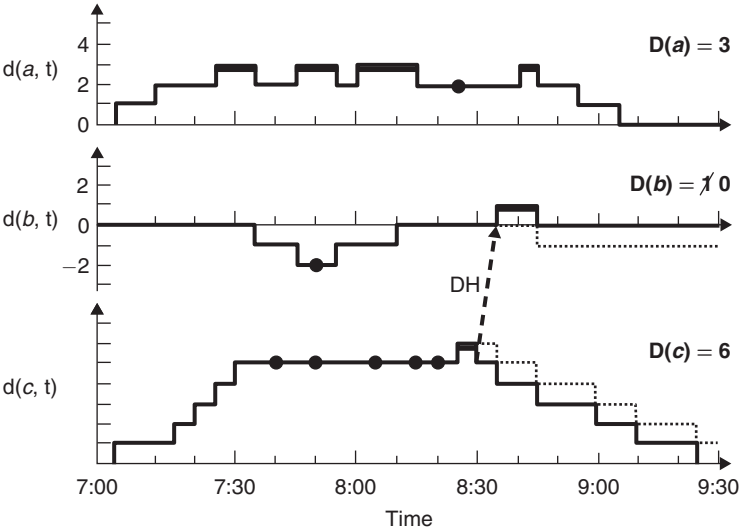


Figure 15.9 New timetable with maximum excluded departure times (following the Minimax H procedure) and associated deficit functions

$d(k, t)$ attains its maximum. The following steps attempt to describe the procedure to convert DH trips in S into service trips, used in the original timetable:

- Step 1: Select a DH trip in S and call it $\overline{DH}$; if there is no DH trip in S , stop.
- Step 2: If the $\overline{DH}$ is from r_i to e'_q ($q' = 1$ or 2 , $r_i \in R$), go to Step 3; if the $\overline{DH}$ is from e'_q to r_i , go to Step 4; and if the $\overline{DH}$ is from r_i to r_k ($r_i, r_k \in R$), go to Step 5.
- Step 3: Examine the arrival in $d(r_i, t)$ to the left of the departure time of $\overline{DH}$ (start with the one closest to that departure time) to see whether it can be extended to e'_q (by replacing $\overline{DH}$). If the arrival considered is associated with trip p_1 , the extension can be executed but if and only if the following three conditions are met: (a) the arrival time of p_1 at r_i is within the hollow that contains the $\overline{DH}$ departure time; (b) p_1 was

originally planned to continue toward e'_q ; and (c) the originally planned arrival time of p_1 at e'_q is equal to or less than the arrival time of $\overline{DH}$. If all the three conditions are fulfilled, delete $\overline{DH}$ from S , update DT' , $d(r_i, t)$, and $d(e'_q, t)$, and go to *Step 1*; otherwise, $\overline{DH}$ remains in S , go to *Step 1*.

- Step 4:* Examine the departure in $d(r_i, t)$ to the right of the arrival time of $\overline{DH}$ (start with the one closest to that arrival time) to see whether it can be extended to e'_q (by replacing $\overline{DH}$). If the departure considered is associated with trip p_2 , the extension can be executed, but if and only if the following three conditions are met: (a) the departure time of p_2 at r_i is less than or equal to the arrival time of $\overline{DH}$; (b) p_2 was originally planned to start at e'_q ; and (c) the originally planned departure time of p_2 at e'_q is within the hollow that contains the $\overline{DH}$ departure time. If all three conditions are fulfilled, delete $\overline{DH}$ from S , update DT' , $d(e'_q, t)$ and $d(r_i, t)$, and go to *Step 1*; otherwise, $\overline{DH}$ remains in S , go to *Step 1*.
- Step 5:* Set $r_k = e'_q$ and use the procedure in *Step 3*; if it is terminated successfully ($\overline{DH}$ is converted to a service trip), execute Adjustment A and go to *Step 1*. Otherwise, set $r_i = e'_q$ and use the procedure in *Step 4*; if it is terminated successfully, execute Adjustment A and go to *Step 1*. Otherwise, $\overline{DH}$ remains in S , go to *Step 1*. Adjustment A: delete $\overline{DH}$ from S , update DT' , $d(r_i, t)$, and $d(r_k, t)$.

15.5.2 Extensions at short-turn points

Following the use of DH trips from/to short-turn points, this section will now describe the principles and procedures for possible extensions of short-turn trips toward their original schedule without increasing $N'_{\min}$. To this end, let the updated timetable DT' be denoted by DT'_1 , including the conversion of DH trips to service trips by extending their associated short-turn trips. An extension of a short-turn trip can be viewed as stretching the trip toward the route's end points, e'_q . An extension does not necessarily mean that the short-turn trip is converted to a full trip along the entire route, because it can be only partially extended; that is, an extension can be performed only from r_i to r_k ($r_i, r_k \in R$). The three stages at which the extensions at $r_i \in R$ can be analysed and executed are as follow: (a) zeroing the maximum DF ; (b) stretching the maximum interval; and (c) treating the DF hollows.

Zeroing the maximum deficit function

On the basis of the DF properties, it is possible to prove that while $N'_{\min}$ is preserved, the number of extensions in each $r_i \in R$ from r_i to e'_q is greater than or equal to $D(r_i)$. This rule is based on the observation that in each $r_i \in R$, exactly $D(r_i)$ departures can be extended to their original departure point without increasing $N'_{\min}$. This procedure will eventually lead to $D(r_i) = 0$ for all $r_i \in R$. These extensions are obtained through the following basic steps. Note, for subsequent steps, as well, that t_s and t_e denote the beginning and the end of the DF maximum interval.

Step 1: (Initialization): set $R = \bar{R}$.

Step 2: Select $r_i \in \bar{R}$; if $\bar{R} = \phi$ (empty), stop.

- Step 3:* Check to see whether $D(r_i) = 0$; if so, delete r_i from $\bar{R}$ and go to *Step 2*; otherwise continue.
- Step 4:* Identify a trip (there might be more than one) whose departure time is at t_s of $d(r_i, t)$ and extend this departure to its original time at e'_q ; update DT'_1 , $d(r_i, t)$ and $d(e'_q, t)$ and go to *Step 3*.

Figure 15.10 illustrates an example of three extensions to $d(r_i, t)$. The first two (numbered 1 and 2) induce $D(r_i)$ to decrease from two to zero. Each extension in Figure 15.10 is followed by an update of $d(r_i, t)$. After the three extensions, only two (of five) short-turn trips remain at r_i . Note that these two trips could also be extended if both are associated with the same terminal k and that their arrival time in the original schedule is at or before the departure time at k .

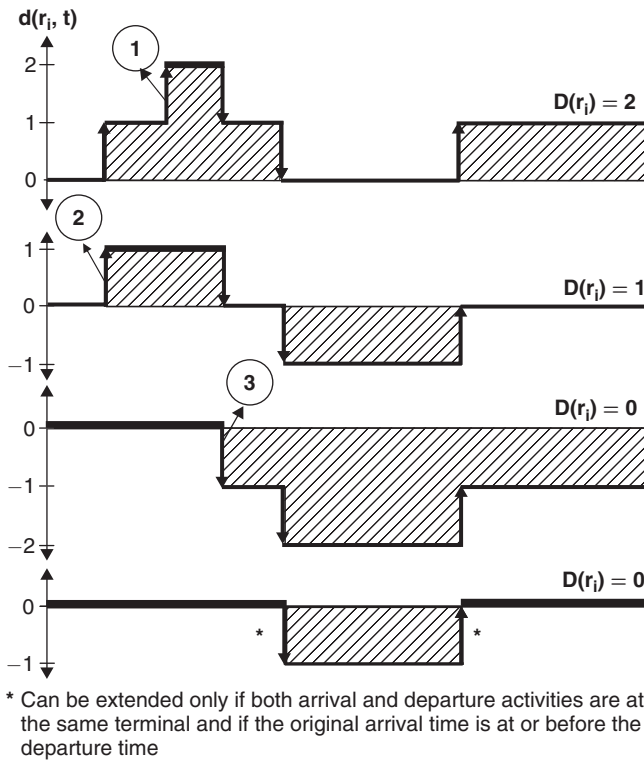


Figure 15.10 Schematic example of updated deficit functions at an intermediate short-turn point after each of the three indicated extensions

Stretching the maximum interval

After reducing all $D(r_i)$ to zero, it is possible to prove the following rule: while preserving $N'_{\min}$, further extensions can be performed from $r_i \in R$ to e'_q , up to the point at which the

maximum interval is stretched over the whole span of the schedule horizon. This rule is based on the observation that certain arrivals can be extended without increasing $D(r_i)$ above zero. The span of the schedule horizon is determined by the earliest departure and latest arrival in the original timetable.

These additional extensions to the route's end points are executed using the following steps, where DT'_2 denotes the updated timetable after the stage described previously (zeroing the DF of r_i):

Step 1: (Initialization): Set $R = R'$.

Step 2: Select $r_i \in R'$; if $R' = \phi$ (empty), stop.

Step 3: Check whether the r_i maximum interval (from t_s to t_e) coincides with the span of the schedule horizon; if so, delete r_i from R' and go to *Step 2*; otherwise, continue.

Step 4: Identify a trip (there might be more than one) whose arrival is at t_e of $d(r_i, t)$ and extend this arrival to its original time at e'_q ; update DT'_2 , $d(r_i, t)$ and $d(e'_q, t)$ and go to *Step 3*.

The above procedure is demonstrated by Extension 3 in Figure 15.10.

Treating the deficit function hollows

At this third stage, a search is made to determine more extensions at $r_i \in R$ regarding departures and arrivals in hollows. Each hollow in $d(r_i, t)$ contains the same number of arrivals as departures. The procedure developed does not treat hollows consisting of only one point. In Figure 15.10, for example, the third DF with Extension 3 has one hollow that consists of two arrivals followed by one departure. DF theory, as outlined in Chapters 7 and 8, permits the construction of the following extension search procedure, in which DT'_3 denotes the updated timetable after the stages just discussed:

Step 1: (Initialization): set $R = \bar{R}'$.

Step 2: Select $r_i \in R$; if $R' = \phi$ (empty), stop.

Step 3: Check the next (with respect to time) trip in $d(r_i, t)$; if it is the last departure, go to *Step 2*; if it is an arrival, go to *Step 5*; otherwise, continue.

Step 4: Examine this departure by extending it to its original time at e'_q ; execute this extension if $D(e'_q)$ is unchanged or if $D(e'_q)$ is increased, but it can also be reduced (back) through the unit reduction DH chain, (URDHC) procedure; update DT'_3 and all the DFs involved. Then, if t_e of $d(r_i, t)$ does not coincide with the right boundary of the schedule horizon, go to the extension procedure described in the previous stage (stretching the maximum interval); otherwise go to *Step 3*. If the extension cannot be made, repeat this extension examination toward a different short-turn point r_k (instead of e'_q) each time, selecting them backward from e'_q to r_i .

Step 5: Examine this arrival by extending it to its original time at e'_q and use the URDHC procedure to check whether $D(r_i)$ can remain the same; if so, execute the extension; update DT'_3 and all the DFs involved; otherwise, repeat this extension examination the same way as in *Step 4*.

Finally, if this procedure leads to the introduction of a new DH trip, the procedure for extensions of DH trips needs to be repeated.

15.5.3 Extensions of the example problem

The example problem, after the Minimax H procedure, is shown in Figure 15.9; this result will now be subjected to possible extensions related to the arrivals and departures at $d(b, t)$. First, the extensions of DH trips procedure will be used to determine whether or not the DH trip can be converted into a service trip. In the example it is found that it cannot be converted, and hence the single DH trip remains in the schedule.

Second, the procedures described for extensions of short-turn points are applied; their execution is shown in Figures 15.11 and 15.12. Because $D(b) = 0$ (after inserting the DH trip), the algorithm in the first stage cannot be utilized. However, because of the algorithm in the second stage, Extension 1 can be performed, increasing $d(b, t)$ at 8:45 from -1 to 0 . Then, the algorithm in the third stage is used. It can be observed that Extension 2 alone

Max load point

Direction	$a \rightarrow c$			$c \rightarrow a$		
Stop	a	b	c	c	b	a
Timetable with maximum short turns,* including DH trip from c to b (8:30–8:35)	7:00	7:15	7:40	7:00	7:20	7:35
	7:10	7:25	7:50	7:15	7:35*	–
	7:25	7:40	8:05	7:20	7:40	7:55
	–	*7:50	8:15	7:25	7:45*	–
	–	*7:55	8:20	7:30	7:50*	–
	7:45	8:05	8:30	7:40	8:00	8:15
	–	*8:10	8:35	7:50	8:10	8:25
	8:00	8:20	8:45	8:05	8:30	8:45
	–	*8:35	9:00	8:15	8:40	8:55
	8:25	8:45	9:10	8:20	8:45*	–
	8:40	9:00	9:25	8:25	8:50	9:05

* Departures and arrivals at b

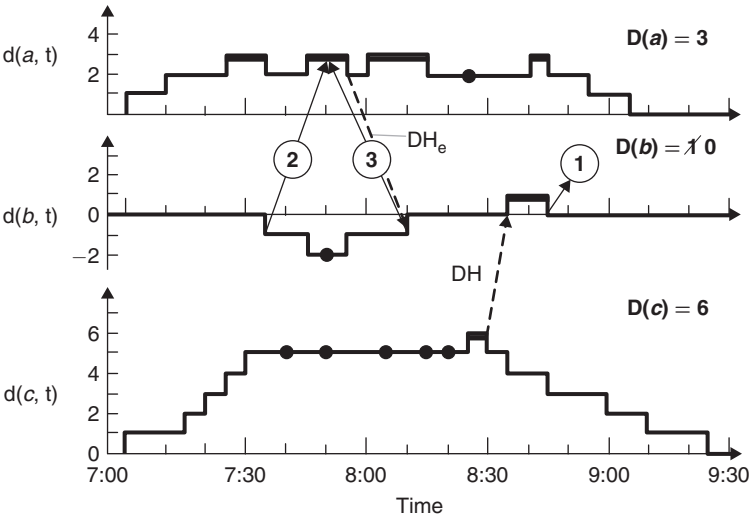


Figure 15.11 Modified timetable and deficit functions following the Minimax H procedure, along with an indication of three short-turn trip extensions

Direction Stop	$a \rightarrow c$			$c \rightarrow a$		
	a	b	c	c	b	a
Final timetable	7:00	7:15	7:40	7:00	7:20	7:35
with minimum	7:10	7:25	7:50	7:15	7:35	(7:50)*
number of	7:25	7:40	8:05	7:20	7:40	7:55
short turns, but	—	7:50	8:15	7:25	7:45	—
with the same	—	7:55	8:20	7:30	7:50	—
minimum	7:45	8:05	8:30	7:40	8:00	8:15
number of	(7:50)*	8:10	8:35	7:50	8:10	8:25
vehicles	8:00	8:20	8:45	8:05	8:30	8:45
	—	8:35	9:00	8:15	8:40	8:55
	8:25	8:45	9:10	8:20	8:45	(9:00)*
	8:40	9:00	9:25	8:25	8:50	9:05

(...)* is an extension (see below)

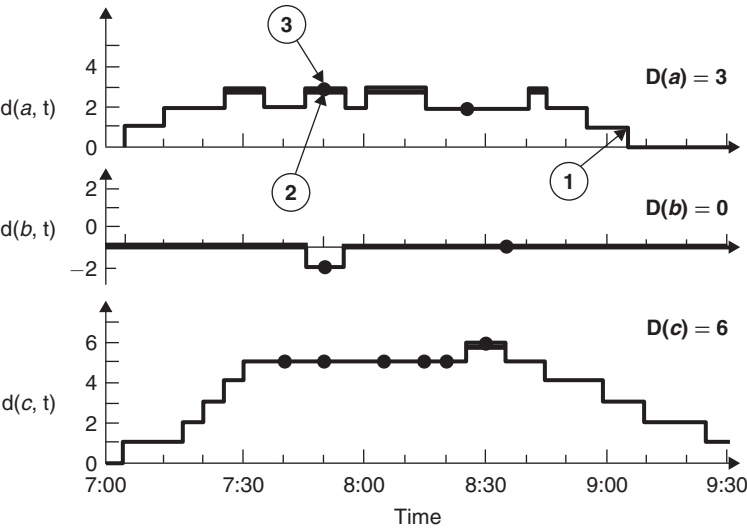


Figure 15.12 Optimal timetable for the problem and associated deficit functions and extensions

affects $D(b)$, increasing it by one at 8:10; the URDHC procedure, therefore, searches for a DH trip that can arrive at b at 8:10. Such a DH trip, designated DH_e in Figure 15.11, is inserted from a . Note that the insertion of DH_e is possible because $d(a, t)$ drops to 2 at 7:50 after resuming the original arrival time associated with Extension 2; thus, $D(a)$ remains 3.

The final step in the third stage is to check DH_e with the procedure of extensions of DH trips. This permits Extension 3 to be performed. Consequently, three of the eight short-turn trips in the timetable of Figure 15.10 were extended to their original schedule, while $N'_{\min}$ remains 9. In other words, the procedures developed identify only the minimum (crucial) short-turn trips that are required to reduce fleet size. Figure 15.12 illustrates the updated DFs after the three extensions. It can be observed that no more extensions can be made.

The timetable in Figure 15.12 achieved, for the example problem, both the minimum fleet size required (of nine vehicles) and the minimum required five short-turn trips to reach this minimum fleet size. The next step is to construct the nine blocks using either the FIFO method or the ‘within hollow’ method described in Section 7.5.5 in Chapter 7. Lastly, a

summary of performance measures of the nine blocks can be calculated for administrative and further evaluation purposes.

For instance, Table 15.1 lists the final schedule of Figure 15.12 in order of increasing departure times, when for the same departure times those at *a* come before those at *c* and

Table 15.1 *List of trips of the example problem in order of increasing departure times*

Trip number	Departure time	Departure location	Arrival time	Arrival location
1	7:00	<i>a</i>	7:40	<i>c</i>
2	7:00	<i>c</i>	7:35	<i>a</i>
3	7:10	<i>a</i>	7:50	<i>c</i>
4	7:15	<i>c</i>	7:50	<i>a</i>
5	7:20	<i>c</i>	7:55	<i>a</i>
6	7:25	<i>a</i>	8:05	<i>c</i>
7	7:25	<i>c</i>	7:45	<i>b</i>
8	7:30	<i>c</i>	7:50	<i>b</i>
9	7:40	<i>c</i>	8:15	<i>a</i>
10	7:45	<i>a</i>	8:30	<i>c</i>
11	7:50	<i>c</i>	8:25	<i>a</i>
12	7:50	<i>b</i>	8:15	<i>c</i>
13	7:50	<i>a</i>	8:35	<i>c</i>
14	7:55	<i>b</i>	8:20	<i>c</i>
15	8:00	<i>a</i>	8:45	<i>c</i>
16	8:05	<i>c</i>	8:45	<i>a</i>
17	8:15	<i>c</i>	8:55	<i>a</i>
18	8:20	<i>c</i>	9:00	<i>a</i>
19	8:25	<i>c</i>	9:05	<i>a</i>
20	8:25	<i>a</i>	9:10	<i>c</i>
21(DH)	8:30	<i>c</i>	8:35	<i>b</i>
22	8:35	<i>b</i>	9:00	<i>c</i>
23	8:40	<i>a</i>	9:25	<i>c</i>

the last departures are at *b*. The 23 trips listed are subjected to the FIFO procedure and nine blocks are constructed, grouped by trip number, as follows: [1,9,20], [2,10,21,22], [3,11,23], [4,13], [5,15], [6,16], [7,12,17], [8,14,18], [19]. A summary of the performance measures of each block is shown in Table 15.2.

Table 15.2 *Performance measures of the nine-block result of the example problem*

Block number	Service time (minutes)	DH time (minutes)	Idle time (minutes)	Service km	DH km	Block time (minutes)
1	120	0	10	72	0	130
2	100	10	10	76	14	120
3	120	0	15	72	0	135
4	75	0	0	48	0	75
5	80	0	5	48	0	85
6	80	0	0	48	0	80
7	80	0	5	52	0	90
8	85	0	0	52	0	85
9	40	0	0	24	0	40
Total	785	10	45	492	14	840

15.6 Literature review and further reading

This section reviews primarily the research on the design of short-turn trips. It ends with several studies on using short turning as a real-time control strategy.

Furth (1987) presents a methodology for schedule coordination between short-turn and full-length trips in a system consisting of several transit modes. The objective is to find an optimal trade-off between the minimum possible fleet size and the minimum passenger waiting time. The schedule offset between full and short-service patterns is determined as are the turning points of the short-turn service. Optimal service headways and vehicle size are also calculated. The frequency of the short-turn route is set to a multiple of that of the full-length route, and the ratio of frequencies is called scheduling mode; different models are developed for different scheduling modes. Vehicle capacity is given significant attention if the researcher shows that a problem of overcrowding may occur even when overall capacity exceeds volume on every road link. The timetable-design problem includes an integer-minute constraint.

Miller and Bunt (1987) describe a computer program that simulates light-rail operations on a specific route in Toronto. This program is designed to analyse the impact of a range of

operating policies on the regularity of the streetcar service and to compare alternate means of improving regularity. One of the strategies compared is the introduction of short-turn service. The model assists in determining where and when a vehicle will be directed to turn around before the end of the route. This determination is based on a set of decision rules. The basic rules recommend turning a vehicle when a gap is found that exceeds a given threshold between successive vehicles travelling in the opposite direction of the vehicle to be turned. Other rules make sure, for example, that the gap in the original direction, which occurs as a result of the turning, does not exceed a given threshold.

Vijayaraghavan and Anantharamaiah (1995) examine the possibility of reducing the required number of buses operated on a single route by using two strategies: partial service (i.e. short turning) and express service. Fleet-assignment options and their effect on the efficiency of fleet use are analysed graphically.

Dell Site and Filippi (1998) present a model for bus-service optimization under an elastic demand. The model is formulated as a programming problem in which decision variables are the locations of turning points, the time offsets between departures of the full-length and the short-turn routes, frequencies and fares. Vehicle size is also considered as a variable, but is represented indirectly. The model enables taking into account different demand patterns at different periods. A different frequency and offset are ascribed to each period; for other variables, a single optimal value is determined. A numerical procedure is presented to solve the problem.

The main characteristics of the models reviewed to this point are summarized in Table 15.3.

Table 15.3 *Summary of features of the models reviewed*

Source	Decision variables	Required demand input	Elastic/constant demand	Operation period
Furth (1987)	Offset between full and short service, turning points, headways, and vehicle size	O-D matrix	Constant	Single
Miller and Bunt (1987)	Turning points and times	None	Not considered	Single (afternoon 5-hour peak)
Vijayaraghavan and Anantharamaiah (1995)	Turning points, number of stops	Total demand	Constant	Single
Dell Site and Filippi (1998)	Offset between full and short service, turning points, headways, fares, and vehicle size	O-D matrix and elasticities	Elastic	Multi-period

The above-mentioned papers relate to the design of short-turn trips. Some other papers also discuss short turning as a real-time control strategy.

Huddart (1973) discusses short-turn trips as a strategy to avoid bus bunching. The author explains that in order to use this strategy efficiently, a bus should be taken out of a bunch in one direction and be entered into a schedule gap in the opposite direction. In addition, difficulties in real-time decision-making regarding short-turn trips are described.

Strathman *et al.* (2001) mention that a bus chosen to turn around should ideally be one with a small number of passengers, a small gap from the preceding bus and a small gap from the following bus. The authors mention that inconvenience is caused to passengers whose destination is further from the turning point, since they are forced to transfer from a short-turn trip to a full-length trip.

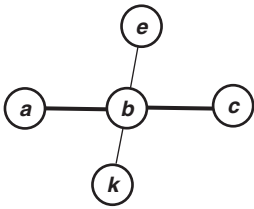
Shen and Wilson (2001) develop a disruption-control model for rail transit systems. The model, formulated as a deterministic mixed-integer program, is used to examine the introduction of several real-time control strategies, including short turning. The main conclusion is that the best system performance is achieved when holding and short-turn strategies are combined. The paper does not include a methodology for short-turn strategy design.

Exercises

- 15.1
- Given a bus route between terminals *a* and *c* (both directions of travel) with a single short-turn point *b* along the route. The input data appear in the following table for the period 6:00–8:00.
The desired occupancy of a bus is 50 passengers; there is a minimum frequency of 3 vehicles/hour, and no shifting in departure time is allowed.

Direction	<i>a</i> → <i>c</i>		<i>c</i> → <i>a</i>	
	<i>a</i> → <i>b</i>	<i>b</i> → <i>c</i>	<i>c</i> → <i>b</i>	<i>b</i> → <i>a</i>
Original departure times	6:00; 6:12; 6:17; 6:25; 6:32; 6:40; 6:46; 6:53; 7:00; 7:10; 7:18; 7:25; 7:35; 7:40; 7:48	–	6:00; 6:10; 6:18; 6:25; 6:32; 6:40; 6:50; 6:55; 7:10; 7:18; 7:25; 7:32; 7:40; 7:48; 7:53	–
Service travel time (minutes)	15	20	25	10
DH travel time (minutes)	10	17	20	8
Passenger load	6:00–7:00	400	250	150
	7:00–8:00	400	300	450

- (a) Apply the method for designing short-turn trips so as to minimize both the number of vehicles required and the number of short turns; provide the final timetable.
- (b) Establish schedules/blocks for all vehicles using the FIFO procedure.
- 15.2 Given the following 2-route network, with stop **b** as a single candidate short-turn point for both routes in both directions of travel:



The following table contains the data required.

Route and direction	$a \rightarrow c$		$c \rightarrow a$		$e \rightarrow k$		$k \rightarrow e$	
	$a \rightarrow b$	$b \rightarrow c$	$c \rightarrow b$	$b \rightarrow a$	$e \rightarrow b$	$b \rightarrow k$	$k \rightarrow b$	$b \rightarrow e$
Service travel time (minutes)	20	30	25	15	10	15	20	15
DH travel time (minutes)	18	25	20	12	7	10	15	10
$Hmin_i$ (minutes)*	5		10		8		13	
$Hmax_i$ (minutes)*	6		15		7		20	
Passenger load	6:00–7:00	150	250	250	400			
	7:00–8:00	250	350	150	200			

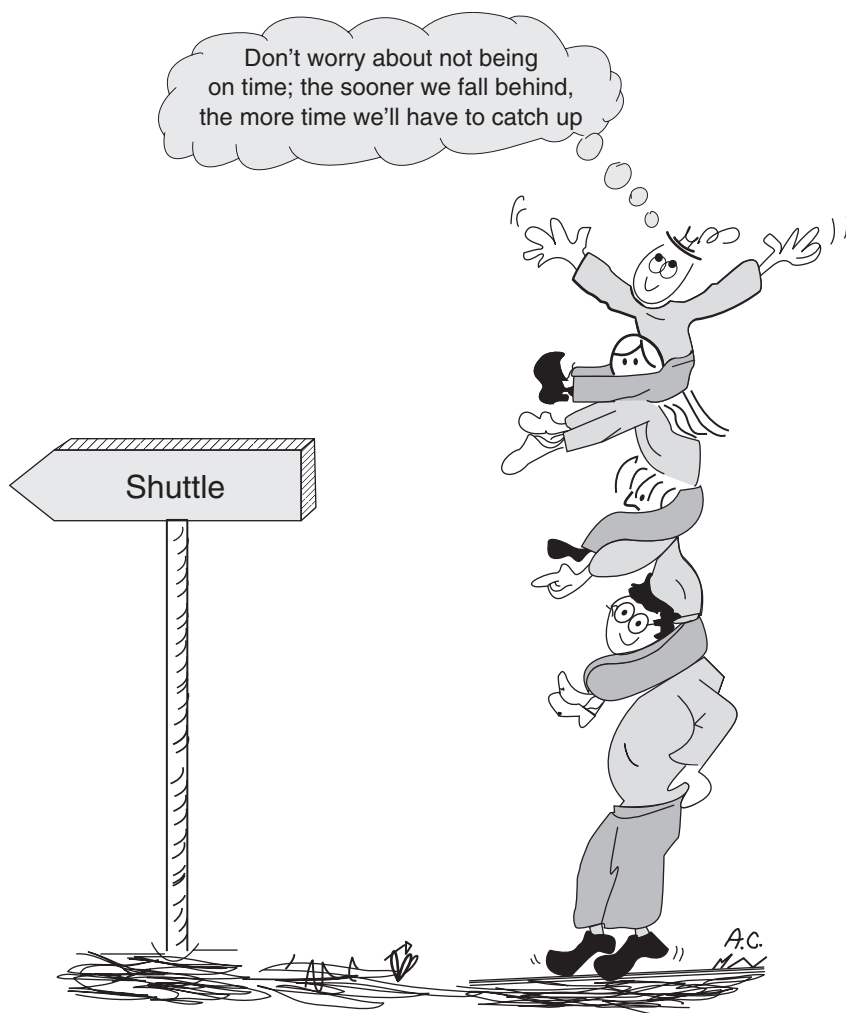
* $Hmin_i$ and $Hmax_i$ are the minimum and maximum headways permitted, respectively, between two adjacent departures on route i (see Chapter 6 for more detailed description and examples)

- In addition, it is given that the determined frequency for the two-hour period considered (departures are between 6:00 and 8:00) is 12 vehicles/period for each route and direction of travel; no shifting in departure time is allowed, and the derived timetable starts at 6:00 (the first fixed departure).
- (a) Construct the timetable for only routes $a \rightarrow c$ and $c \rightarrow a$, using the method of synchronization described in Chapter 6.
- (b) Apply the method to create short-turn trips for routes $a \rightarrow c$ and $c \rightarrow a$ so as to minimize both the number of vehicles required and the number of short turns; provide a final timetable.

References

- Ceder, A. (1990). Optimal design of transit short-turn trips. *Transportation Research Record*, **1221**, 8–22.
- Ceder, A. (1991). A procedure to adjust transit trip departure times through minimizing the maximum headway. *Computers and Operations Research Journal*, **18**, 417–431.
- Ceder, A. (2003). Designing public transport networks and routes. In *Advanced Modeling for Transit Operations and Service Planning* (W. Lam and M. Bell, eds), pp. 59–91, Elsevier Ltd.
- Dell Site, P. and Filippi, F. (1998). Service optimization for bus corridors with short-turn strategies and variable vehicle size. *Transportation Research*, **32A**, 19–38.
- Furth, P. G. (1987). Short turning on transit routes. *Transportation Research Record*, **1108**, 42–52.
- Huddart, K. W. (1973). Bus priority in greater London: Bus bunching and regularity of service. *Traffic Engineering and Control*, **14**, 592–594.
- Miller, E. J. and Bunt, P. D. (1987). Simulation model of shared right-of-way streetcar operations. *Transportation Research Record*, **1152**, 31–41.
- Shen, S. and Wilson, N. H. M. (2001). An optimal integrated real-time disruption control model for rail transit systems. In *Computer-aided Scheduling of Public Transport. Lecture Notes in Economics and Mathematical Systems*, **505** (S. Voss and J. R. Daduna, eds), pp. 335–364, Springer-Verlag.
- Strathman, J. G., Kimpel, T. J., Ducker, K. J., Gerhart, R. L., Turner, K., Griffin, D. and Callas, S. (2001). Bus transit operations control: Review and an experiment involving TRI-MET's automated bus dispatching system. *Journal of Public Transportation*, **4**, 1–26.
- Vijayaraghavan, T. A. S. and Anantharamaiah, K. M. (1995). Fleet assignment strategies in urban transportation using express and partial services. *Transportation Research*, **29A**, 157–171.

16 Smart Shuttle and Feeder Service



Chapter 16 Smart Shuttle and Feeder Service

Chapter Outline

- 16.1 Introduction
- 16.2 Minimum fleet size required for a circular (shuttle) route
- 16.3 Routing strategies
- 16.4 Simulation
- 16.5 Case study
- 16.6 Customer survey
- 16.7 Optimal routing design: base network
- 16.8 Optimal routing design: algorithm
- 16.9 Implementation stages
- 16.10 Literature review and further reading

Exercise

References

Practitioner's Corner

The choice between public and private transportation is an individual decision that is influenced by government/community decisions. These decisions often send mixed signals (e.g. 'use transit'; 'we're trying to reduce traffic congestion to ease your automobile driving') to transit and potential transit passengers while failing to recognize their more system-wide, integrated implications. Chapter 13 facilitates an understanding of the importance of service-connectivity elements in enhancing existing or new transit services. One such element is the design of an integrated, smart feeder/shuttle service. Such a service may stem, for instance, from the need to overcome the problem of an excessive number of automobiles arriving and parking at a train station, resulting in high parking demand around the station. The purpose of this chapter is to examine an innovative feeder/shuttle system that will comply with (i) passengers' needs and desires, (ii) intelligent transportation technologies, and (iii) an agency's viability. After all, a solution for being transported by a door-to-door service is the mother of attraction.

This chapter contains eight main parts, following an introductory section. Section 16.2 continues the analysis of the minimum fleet size required that was the subject of Chapter 7, but for a radial/shuttle route having a single departure and arrival point. Section 16.3 proposes ten different feeder/shuttle routing strategies with various combinations of fixed/flexible routes, fixed/flexible schedules, a uni- or bi-directional concept, and short-cut (shortest path) and/or short-turn (turn-around) concepts. Section 16.4 investigates these strategies by employing a simulation model specifically developed and constructed for this purpose. This simulation model is used in Section 16.5 for a case study of Castro Valley in California, where the feeder/shuttle service is coordinated with the San Francisco Bay Area Rapid Transit (BART) service; the ten routing strategies are compared for four fleet-size scenarios. Section 16.6 reports on a survey conducted in the

case-study area concerning: (a) willingness to use the smart-shuttle service, (b) willingness to pay for the service, and (c) the attributes that would enhance the shuttle service. Section 16.7 introduces a framework for the optimal design of circular feeder/shuttle fixed routes, including a method for estimating potential passenger demand. Section 16.8 shows the modelling of an optimal circular route that can handle any size road network. The chapter ends with a literature review and exercises.

Practitioners are advised to visit all sections of the chapter, except perhaps for Section 16.8, which is of a more mathematical nature. They are especially encouraged to follow the examples and the case study.

Lastly, an idea that has been introduced is that passengers who want to take the shuttle service should be able to use an on-line, intelligent information system through a telephone/cellular call. The system will announce the arrival time to the point closest to the caller unless this estimated time is not reliable enough. In the latter case, the system will call back once an estimation becomes reliable. It may even have a special ring for easy recognition . . . does the name Pavlov ring a bell?

16.1 Introduction

When exploring the design possibility of a transit feeder/shuttle service, we can refer to Section 13.5 in Chapter 13, in which an ideal coordinated service is portrayed. For simplicity, we will include the transit-feeder service within a shuttle service in the remaining parts of the chapter. Ideally, a smart transit shuttle system will provide an advanced, attractive service that operates reliably and relatively rapidly, and has smooth, synchronized transfers as part of the door-to-door transit-passenger chain. In order to arrive at the design of such a system, new integration and routing concepts had to be developed. This chapter integrates work done by Ceder and Yim (2002), Yim and Ceder (2006), and Jerby and Ceder (2007).

A growing concern for public transit is its inability to encourage people to switch their mode of transportation from solo to shared driving. The majority of large cities has encouraged private car use through planning (dispersed land-use in the suburbs), infrastructure (available parking and circulation traffic flow), pricing, and financial decisions. Consequently, there is growing confusion among drivers in many of those cities about what to do. One way to handle the known decline in transit use is to retain a high level of satisfaction among transit users while fully maintaining the protection of accessibility for less-affluent travellers.

What follows is a description of the motivation that was employed by research reported in this chapter. Although overall transit ridership is declining in cities, an encouraging trend is increased ridership in long-haul express bus or rail transit. When long-haul express transit systems were built in the 1970s and 1980s, parking facilities were also provided for the riders, under the rubric of 'park and ride'. The concept was readily accepted by the public, and a large number of commuters preferred to take an express bus or train to avoid rush-hour traffic and high parking costs. However, most of the arrivals at the train stations are made by private cars, hence creating traffic congestion and parking overloading in the station area.

Such a case was observed in the Bay area around San Francisco in connection with the use of the local metro system, called BART (bay area rapid transit). This observation, which

served as the motivation for this chapter, appeared on the front page of the *San Francisco Chronicle* on three consecutive days in 2001. Figure 16.1 shows the relevant sections of the newspaper, providing news about the problem and commentary in the form of suggested parking solutions (bottom right of figure might be hard to read). The main parking solution



Figure 16.1 Motivation for the study effectively appeared on the front page of the *San Francisco Chronicle* on three consecutive days

considered by BART was this: “Building reserved parking lots, where users would be charged a fee for a guaranteed space”. An obvious question arises: Why not employ a smart bus shuttle service instead?

Mark Twain said: “You cannot depend on your eyes when your imagination is out of focus”. Our eyes see what transit services are currently providing for high- and low-density communities. Our eyes can read reports covering urban transportation characteristics, the influence of transportation investment, ground transportation strategy and passenger-transportation action plans. However, we cannot depend on our eyes alone to trigger our imagination. As Einstein said: “Imagination is more important than knowledge”. Consequently, in order to design an imaginative, ‘ideal’ shuttle service (see definition in the first paragraph of this section), it will be worthwhile to start almost from scratch in order to attain such a smart service.

The purpose of this chapter is to describe the conceptual construction of an innovative shuttle system that will: (1) meet the needs and desires of end users, (2) utilize intelligent-transportation technologies, and (3) increase operational efficiency. A simulation model, a passenger survey, and a case study will help us take a practical approach to the optimal design of the shuttle route(s).

16.2 Minimum fleet size required for a circular (shuttle) route

Section 7.2 in Chapter 7 discussed the case in which interlinings are not allowed and each route is operated separately. Let T_r be the average round-trip time, including layover and turn-around times, of a circular route r (whose departure and arrival points are the same). The minimum fleet size is then equal to the largest number of vehicles that departs within T_r (Salzborn, 1972).

Although Salzborn’s modelling provides the basis for fleet-size calculation, it relies on two assumptions that, in practice, do not hold for a circular (shuttle) route: (i) vehicle departure rate is a continuous function of time; (ii) T_r is the same throughout the period under consideration. In practice, departure times are discrete (see Chapters 4–6), and average trip time is usually dependent on time-of-day. The analysis of Section 7.2 is used in the present section for the case of a circular (shuttle) route.

Let route r start and end at b . Let T_{rj} be the average trip time on route r for a vehicle departing at t_j from b , including its layover time at b . Also, let n_j be the number of departures from b from, and including, departure j at t_j until, but excluding, departure j' at $t_{j'}$. We further define that j arrives at b and continues with departure j' at $t_{j'}$, which is the *first* feasible departure from b at a time greater than or equal to $t_j + T_{rj}$. Theorem 7.1 (see Chapter 7) without n_{ia} but with the setting $n_{jb} = n_j$ yields the following:

$$N_{\min}^r = \max_j n_j \quad (16.1)$$

where $N_{\min}^r$ is the minimum fleet size required for the circular route r .

The timetable for b in the example shown in Figure 7.3 in Chapter 7 is used for the shuttle case in Figure 16.2. Both a single T_{rj} value of 30 minutes throughout the timetable and a case in which T_{rj} varies are used in Figure 16.2. On the left-hand side of Figure 16.2, the

The case with $T_{rj} = 30$ minutes for all trips j	The case with different T_{rj} for each trip j	
Timetable	Timetable	T_{rj} (minutes)
1 { 5:00	5:00 } 1	25
5:30 } 1	1 { 5:30	25
1 { 6:00	6:00 } 1	30
6:30	2 { 6:30	30
6:50 } 2	6:50	30
3 { 7:05	7:05 } 4	40
7:10	7:10 } 4	40
7:15 } 5	7:15	35
7:20	4 { 7:20 } 5	30
7:30	7:30 } 3	25
3 { 7:40	2 { 7:40	25
8:00	8:00 } 2	—
$N_{min}^r = 5$	$N_{min}^r = 6$	

Figure 16.2 Example of the derivation of radial-route fleet size when the round-trip time, including layover time, can either be the same (left side) or vary (right side) for each departure time

time windows are all with the same length (30 minutes); on the right-hand side, they are T_{rj} dependent. The maximum number of departures within T_{rj} is emphasized for each case, leading in both instances to a determination of the minimum fleet size, N_{min}^r , according to Equation (16.1).

Following the determination of N_{min}^r , the construction of vehicle chains (blocks) can be carried out by using the FIFO (first-in-first-out) rule of first-feasible connection. For example, for the case with varied T_{rj} in Figure 16.2, the first block is [5:00–5:30–6:00–6:30–7:05–8:00], in which the 7:05 departure will be ready for another departure at 7:45 ($T_{r6} = 40$ minutes). The second block, using the FIFO rule, is [6:50–7:20]. Because there is need for a minimum of six vehicles, some trips of the first block can certainly be performed by the second block, thereby opening some idle time for the vehicles associated with these blocks. For instance, the first two blocks can be [5:00– 6:00–6:50–7:20] and [5:30–6:30–7:05–8:00].

Finally, two points are worth noting: (1) When shifting in departure times is allowed, a further reduction in the fleet size can be achieved. (2) When more than one circular (shuttle) route is operated from the same point, it is recommended that the deficit-function method (of Chapters 7 and 8) be used to arrive at the minimum fleet size required for all routes.

16.3 Routing strategies

The previous section dealt with a fixed route and a fixed shuttle-service schedule; however, other routing strategies can be imagined and should be explored. These strategies represent the flexibility and part of the attractiveness of the transit system. Before embarking on a study of these other strategies, we will first present an overview of various known and relevant concepts of transit-shuttle operation.

In order to alleviate the problems encountered in traditional transit services, several flexible services were studied. Dial-a-ride and door-to-door paratransit have played a vital role in providing equitable transportation service to elderly and handicapped persons who have difficulty in accessing regular public transit systems (Cervero, 1998). Such a demand-responsive transit (DRT) system, which was investigated by Ioachim *et al.* (1995) and Borndorfer *et al.* (1999), does not, though, fulfil the needs of the entire transit population. An interesting study by Melucelli *et al.* (2001) distinguishes between two classes of users, so-called *passive users* and *active users*. The *passive users* make use of traditional transit; i.e. boarding and alighting at compulsory stops. No reservation is necessary, since vehicles are guaranteed to serve each compulsory stop within a given time window. The *active users*, who board or alight at an optional stop, must issue a *service request* and specify pickup and drop-off stops, as well as earliest departure and latest arrival times. In the Melucelli *et al.* study, transit vehicles had to be rerouted and re-scheduled in order to satisfy as many requests as possible, complying with passage-time constraints at compulsory stops while activating optional stops between two compulsory stops on demand.

Several studies have made use of simulation as a tool to arrive at satisfactory routing and scheduling DRT solutions. Two waves of simulation studies can be traced in the literature. The first wave consists of the research conducted by Wilson *et al.* (1970), who evaluated various heuristic routing rules and algorithms used in a computer-aided routing system. These rules and algorithms, developed for mainframe computers, have limitations in their handling of large-size road networks with different routing strategies. The second wave of research was conducted by Fu (1999, 2001) and Fu and Xu (2001), who considered the use of advanced technologies. Their studies present a simulation model, Sim-Paratransit, which was developed for evaluating advanced paratransit systems, such as AVL (automatic vehicle location) and CAD (computer-aided dispatch) systems. The simulation model is described by Fu (1999, 2001), and the evaluation of AVL and CAD systems by Fu and Xu (2001). The ability to track the location of a transit vehicle continuously allows for the introduction of intelligent paratransit systems, which will naturally lead to operating such systems at a significantly improved level of productivity and reliability.

This chapter investigates ten routing strategies:

1. Fixed route with a fixed schedule (timetable) and fixed direction (of travel)
2. Fixed route; flexible (demand-driven) schedule; fixed direction
3. Fixed route; flexible schedule; bi-directional
4. Fixed route; flexible schedule; fixed direction; possible short turns
5. Fixed route; flexible schedule; bi-directional; possible short turns
6. Fixed route; flexible schedule; fixed direction; possible short cuts
7. Fixed route; flexible schedule; bi-directional; possible short cuts
8. Fixed route; flexible schedule; fixed direction; possible short turns and short cuts

9. Fixed route; flexible schedule; bi-directional; possible short turns and short cuts
10. Flexible (demand-responsive) route with a flexible schedule.

Fixed direction means that the shuttle will always maintain the same direction of travel (same sequence of stops), whereas bi-directional allows for flexibility in selecting a direction based on real-time demand information. The term 'short cut' means that, based on certain loading-threshold and synchronization criteria, the shuttle will not continue its fixed route and, instead, will use the shortest path (minimum travel time) to arrive at the required point; e.g. the train station. For convenience, we will use the shuttle terminal as a train station throughout this chapter, although this connection point can also serve other transit modes.

The loading threshold is a given (input) number of passengers on board the shuttle. The synchronization criterion means matching the shuttle's new (via short cut) arrival time with an earlier train than that originally planned if the entire route were completed. The term 'short turn' means that, based on certain loading-threshold and synchronization criteria, the shuttle will not continue on its fixed route. Instead, it will turn around and arrive at the train station in the opposite direction in order to try to pick up passengers who were too late to catch this shuttle when it first passed their stops. The loading-threshold and synchronization criteria for the short-turn strategy (including the consideration of more pickups) and the short-cut strategy are similar. Each strategy allows for the flexibility of the other. In other words, the loading threshold of the short-cut strategy is initially higher than the loading threshold of the short-turn strategy. If the latter is reached and there is a possibility of picking up x passengers (after turning around), where x is equal to or greater than the difference between the two loading thresholds, then the short-turn strategy is recommended.

Figure 16.3 represents the ten strategies on a small network with two shuttle routes, one marked with a dashed line and the other with a dotted line. Strategies 2 through 10 are characterized by a flexible schedule; and, it is emphasized, by 'no timetable' in Strategy 2. Arrows in both directions of the route means a bi-directional situation. The short-cut strategy has lines with arrows in Figure 16.3 that indicate deviations from the fixed route. The short-turn strategy has arrows indicating a turn-around at a certain point in the network. Both representations appear in Strategies 8 and 9, involving a possible combination of the two. The last strategy is for a DRT-type of service, allowing for the creation of a new route every time, based on trip bookings.

The idea of covering almost every possible practical routing strategy stems from the need to satisfy user desires and understandings. Certainly, there is no intention that all strategies be used at the same time; rather, it is to examine which strategy is best for a given demand pattern and magnitude while taking into consideration the real-time traffic situation in the area of the shuttle's trips. A simulation model was devised for that purpose. This simulation tool, explained in the next section, enables a comparison of the various strategies, based on the following measures:

- sum of total time (in passenger-hours) from passenger pickup to train-departure time;
- sum of total time (in passenger-hours) riding the shuttle vehicle;
- sum of total waiting time (in passenger-hours) for the train;
- sum of total waiting time (in passenger-hours) for the shuttle vehicle;
- total number of transit vehicles (by number of seats) required to meet the demand.

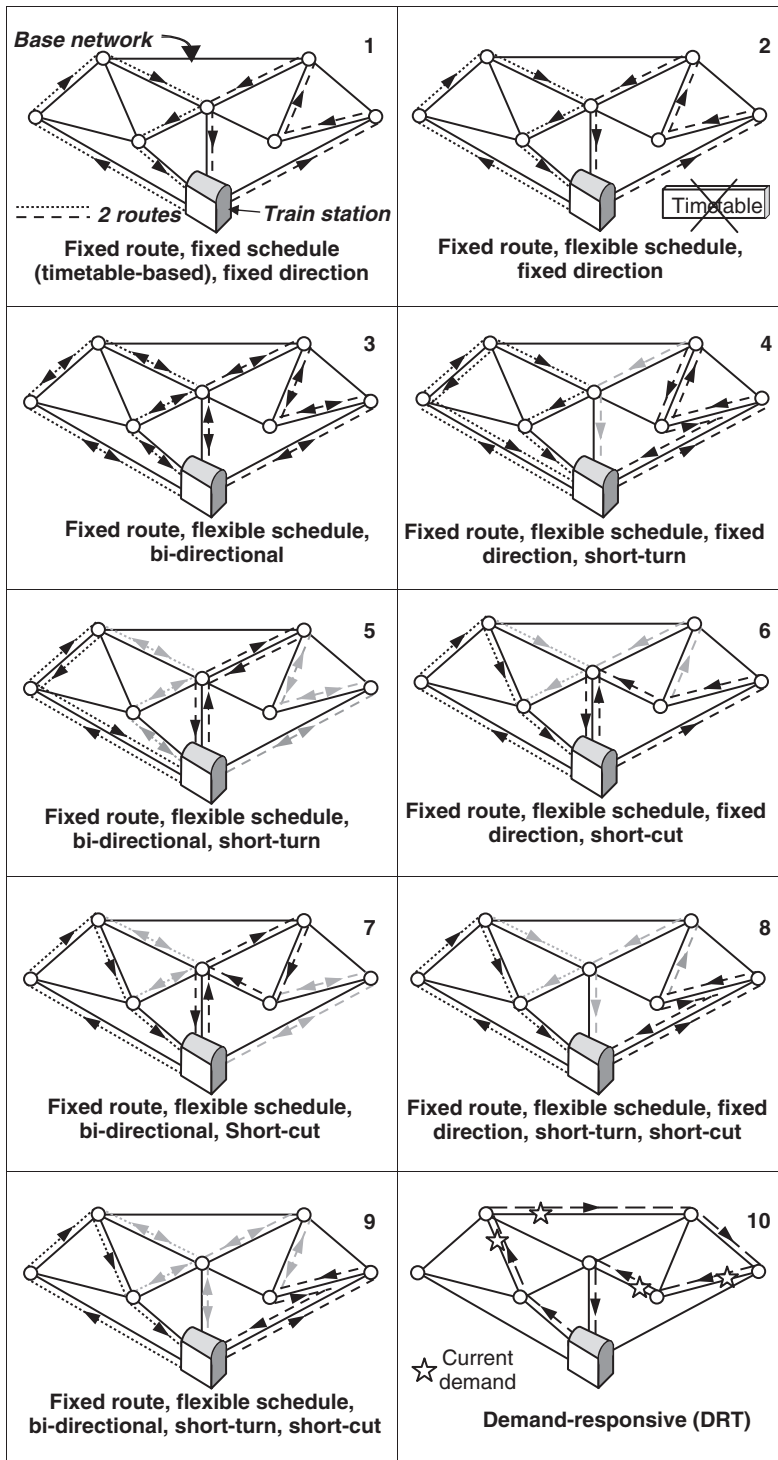


Figure 16.3 Ten routing strategies

These measures of travel and waiting times and number of vehicles characterize the effectiveness and efficiency of each strategy. The strategy selected for a given passenger demand is the one with the minimum weighted travel and waiting times (user perspective) and the minimum number of vehicles (agency perspective).

16.4 Simulation

The purpose of the simulation model is to examine the ten different routing strategies. Passenger demand can either be inserted as part of the input or be generated randomly on the network.

16.4.1 Simulation input variables

Following are the input variables of the simulation program, which was written in C++ language. Each variable is presented by its simulation name, as well as an explanation and interpretation of its substance. What is referred to here as 'bus' can be applied to any shuttle vehicle.

Bus2Train	= Time in seconds that a bus must be at the station before a train arrives in order to ensure an efficient meeting
Train2Bus	= Time (seconds) that a bus must wait after a train arrival to ensure pickup
SizeType	= Number of seats in this bus type
Quantity	= Number of vehicles of this SizeType
FixPick	= Fixed time (seconds) for one passenger pickup, including bus slow down
FixDrop	= Fixed time (seconds) for one passenger drop-off, including bus slow down
FixBoard	= Fixed (additional) boarding time (seconds) per passenger
FixAlight	= Fixed (additional) alighting time (seconds) per passenger
NodeNo	= Node index at section end points
SectionNo	= Section number between two nodes
StopNo	= Stop-number starts with SectionNo, representing an intersection, not a node
MeanDemand	= Mean number of potential travel requests per given hour and SectionNo, to the train station
MeanDestin	= Mean number of potential travel requests per given hour and SectionNo, from the train station
MeanTime	= Mean section travel time (seconds)
StDevTime	= Standard deviation of section travel time (seconds)
Min4Turn	= Minimum number of on-board passengers to allow a short turn
Min4Cut	= Minimum number of on-board passengers to allow a short cut
Min4Trip	= Minimum number of travel requests by calls to allow a non-scheduled trip
Min4Dep	= Minimum of number of waiting passengers to allow a non-scheduled trip
RouteNo	= Unique route index

RouteDir = Direction of RouteNo (start westbound or eastbound)
 TTimeTable = Fixed train timetable in hhmmss form, hh is from 00 to 24
 BTimeTable = Fixed bus timetable in hhmmss form, hh is from 00 to 24
 Layover = Fixed time for driver rest at the end of each trip

All these variables interact in each simulation iteration as part of the different strategies and other internal features.

16.4.2 Simulation procedures

The simulation model is based on events. The simulation starts with reading the input data and proceeds by arranging train-arrival events and passenger-arrival events. Figure 16.4 presents the basic event-oriented simulation logic.

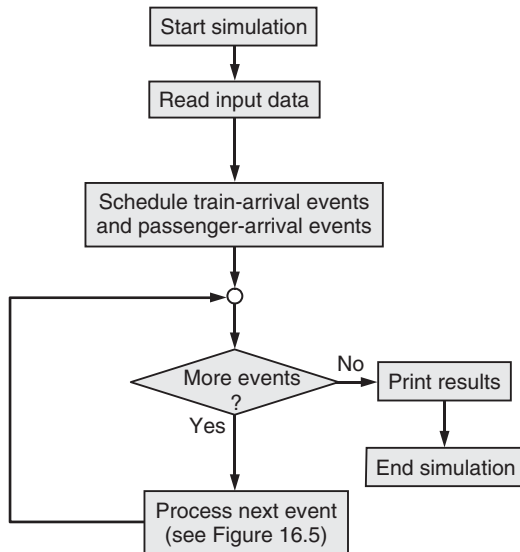


Figure 16.4 Basic simulation logic

There are eight main events, classified in Figure 16.5. Event 1 represents passengers walking to the stop to wait for the next shuttle in order to arrive at the train station. Event 2 represents passengers who have arrived on the train and are now waiting for the next shuttle. Event 3 occurs when a vehicle becomes available for the next trip. Event 4 is when the shuttle arrives at a node (intersection) on the road network being considered. Event 5 represents the arrival of passengers who want to ride the shuttle from its stop to the train station. Event 6 represents passengers who are about to arrive at the train station and will seek to ride the shuttle. Event 7 is the arrival of the train at the station, including the time for the passengers to arrive at Event 2. Lastly, Event 8 is the time when the shuttle departs, in accordance with the timetable.

The actions taken for Events 1, 2 and 3 appear in Figure 16.6. They start with enquiring whether the number of passengers who want the service has reached the minimum required for

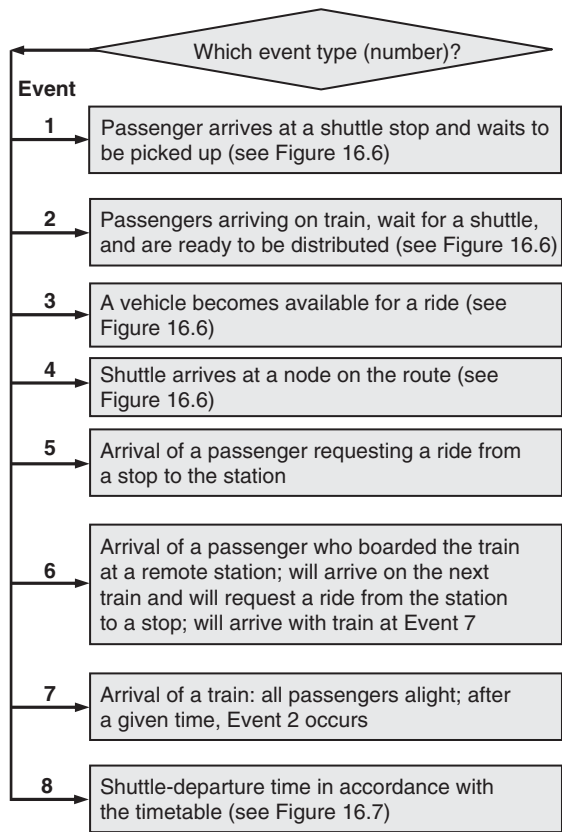


Figure 16.5 *Event classification*

dispatching a vehicle. For the DRT strategy (no fixed route), the procedure in Figure 16.6 identifies the section with current (booked) demand and the application of a shortest-path algorithm (e.g. see the appendix to Chapter 10). The dynamic-routing procedure involves moving the shuttle from the train station across all demand points to the first demand point that is within the shortest path from the station. From the last point, the shortest-path algorithm is used again for all the other demand points that were not visited until all the points have been included in the dynamic route. This DRT routing procedure has been found to be effective and convenient to use in the simulation.

Once a vehicle is available (see Figure 16.6), the next event is of type 4, described in Figure 16.7. Thus Event 4, in Figure 16.7, starts either with a station node (train station), whether at the end or at the beginning of the shuttle ride, or at an intermediate node. For any intermediate node, the procedure checks whether the minimum number of passengers on-board the shuttle has reached the threshold for either a short-turn or a short-cut strategy. The procedure then checks to see whether creating a short turn or a short cut will enable passengers to arrive at an earlier train than that which would be met by completing the entire route.

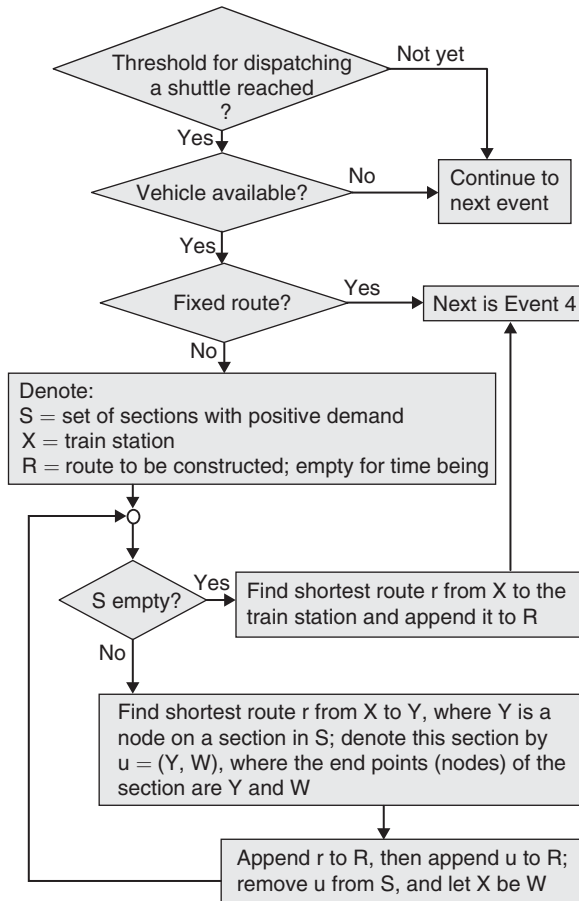


Figure 16.6 Actions taken for Events 1, 2 and 3

Finally, there are the actions taken in the simulation for Event 5 following the process of informing the passenger through the available on-line service. There are two possible actions: (a) the passenger will be able to reach the shuttle stop on time after learning of the expected arrival time; (b) the passenger will not be able to arrive on time and will ask to be notified of the next arrival time by a call-back. It is assumed that passengers will either call an automated system or look at the shuttle website to ascertain the arrival time. The passenger will then be asked to click, for instance, '1' (for wanting to use the service) or '2' (for not wanting to use it following the announcement of the expected arrival time). Only those who click '1' (OK) are taken into account in the simulation process. The simulation model can either consider a given demand figure or be used to generate a random demand based on the residential density of each section of the network. In the fixed-route case, passengers reach their closest stop on the network. The travel time is a random variable with a normal distribution, and the simulation model calculates the probability of being on-time. If this

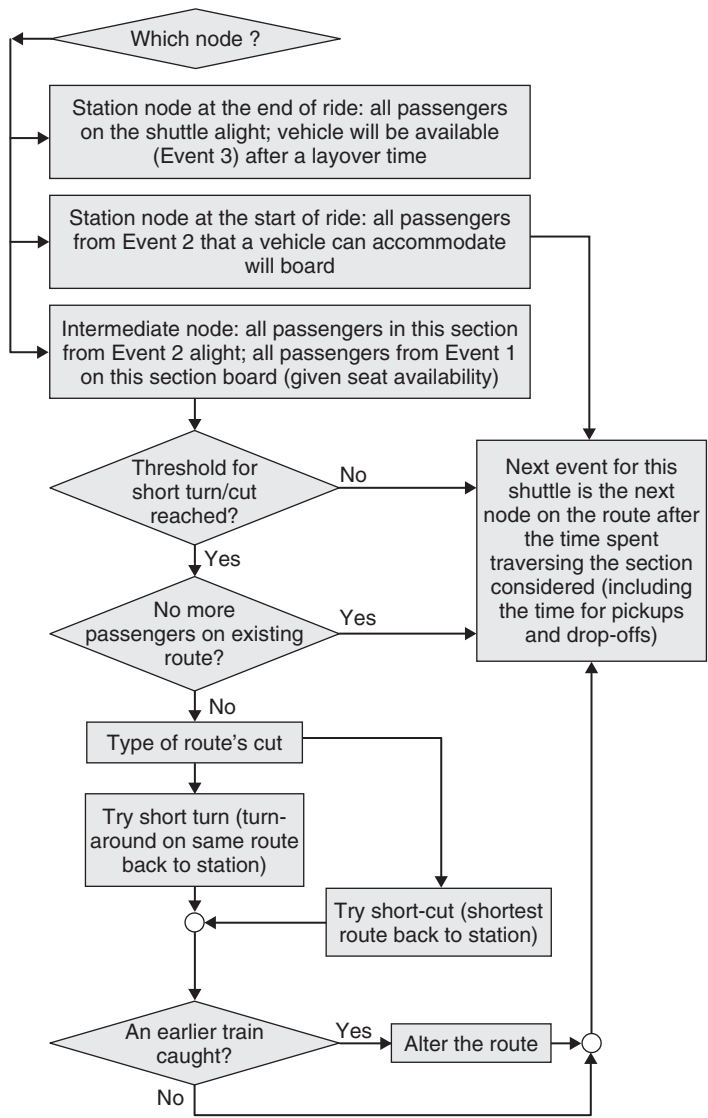


Figure 16.7 Actions for Events 4 and 8

probability is below 90%, the user is notified to wait for a call-back. In this way, the system uses the philosophy of advanced technologies and maintains a highly reliable service.

16.5 Case study

In order to test a real-life situation, the area of Castro Valley in California was selected for data collection and simulation runs. The BART station in Castro Valley is on the ‘blue’ line,

Dublin/Pleasanton to Daly City. A site observation was conducted in the Castro Valley area, from which the base network and stops were created. The optimal routing procedures given in Sections 16.7 and 16.8 were not considered in this case study; instead, two routing scenarios starting at the BART station were introduced, one with a single route and one with a two-route shuttle system. The routing decided on followed a site visit. Figure 16.8 presents the base network (described in Section 6.7.1), both the single-route case (solid line) and the two-route case (dashed lines), and shows the location of the BART station at which the routes start and end. The two circular routes (dashed lines) are extended from the station – one to the left and one to the right.

A large number of simulation runs were executed across the ten routing strategies described for four different groups of (available) shuttle buses: 1, 2, 3 and 4 buses. The estimated demand was 400 passengers daily, generated randomly. Table 16.1 summarizes the results obtained for the waiting time per passenger in 40 cases related to the single-route scenario. The minimum (best) passenger-waiting-time results are indicated in this table by an asterisk for each group of available buses; the second-best results are marked by '✓'. The waiting time in Table 16.1 is the average time per passenger, in minutes, that elapses from the time of a phone call to the shuttle bus-information centre until pickup occurs. This time period includes the walking time from the place of the call to the bus stop (assuming the shortest walking distance) and the waiting time until the bus arrives.

The results presented in Table 16.1 show that the fixed route–fixed schedule strategy (no. 1) results in the highest (longest) waiting times. At the same time, the flexible route–flexible schedule (demand-responsive) strategy (no. 10) does not always provide the best results; hence, it cannot *a priori* be superior to the other strategies. In fact, the best routing strategies observed in this real-life test are those with two asterisks and two '✓': fixed route–flexible schedule and bi-directional, with possible short turn (no. 5); fixed route–flexible schedule and bi-directional, with possible short cut (no. 7); and fixed route–flexible schedule and bi-directional, with possible short turn and short cut (no. 9). The short-turn, short-cut, and bi-directional-based routing strategies indeed prove worthwhile to consider. These three uncommon strategies reflect the current availability of on-line information and communication.

In addition, six more simulation runs were performed for the two-route scenario, using strategies no. 1, 8 and 10 with a group of four buses for both picking up passengers for the train station and distributing them from the station. An additional criterion, attention to which had to be paid at the phone location, was established for maximum waiting time on these six runs: 20 minutes (could be changed in the simulation runs). This criterion reflected the fact that the caller would not actually wait if the announced waiting time for the shuttle bus was more than 20 minutes; in such a case, the caller would cancel the request. Table 16.2 summarizes the average waiting time per passenger for these two pickup and drop-off cases, including the standard deviation determined for each simulation run.

The cumulative curves for the waiting times in these six situations appear in Figures 16.9, 16.10 and 16.11. Table 16.2 shows that the average waiting time for distributing passengers in the fixed schedule case is much higher than for the flexible schedule cases. Furthermore, the standard deviations are lower in the pickup cases than in the drop-off cases. More precise configurations for these results are shown in Figures 16.9, 16.10 and 16.11.

In these figures, the upper cumulative curve refers to the pickup case, and the lower curve to the drop-off case. Obviously the waiting time in the drop-off case depends on the bus-departure time from the train station. This is the reason that the cumulative curves for

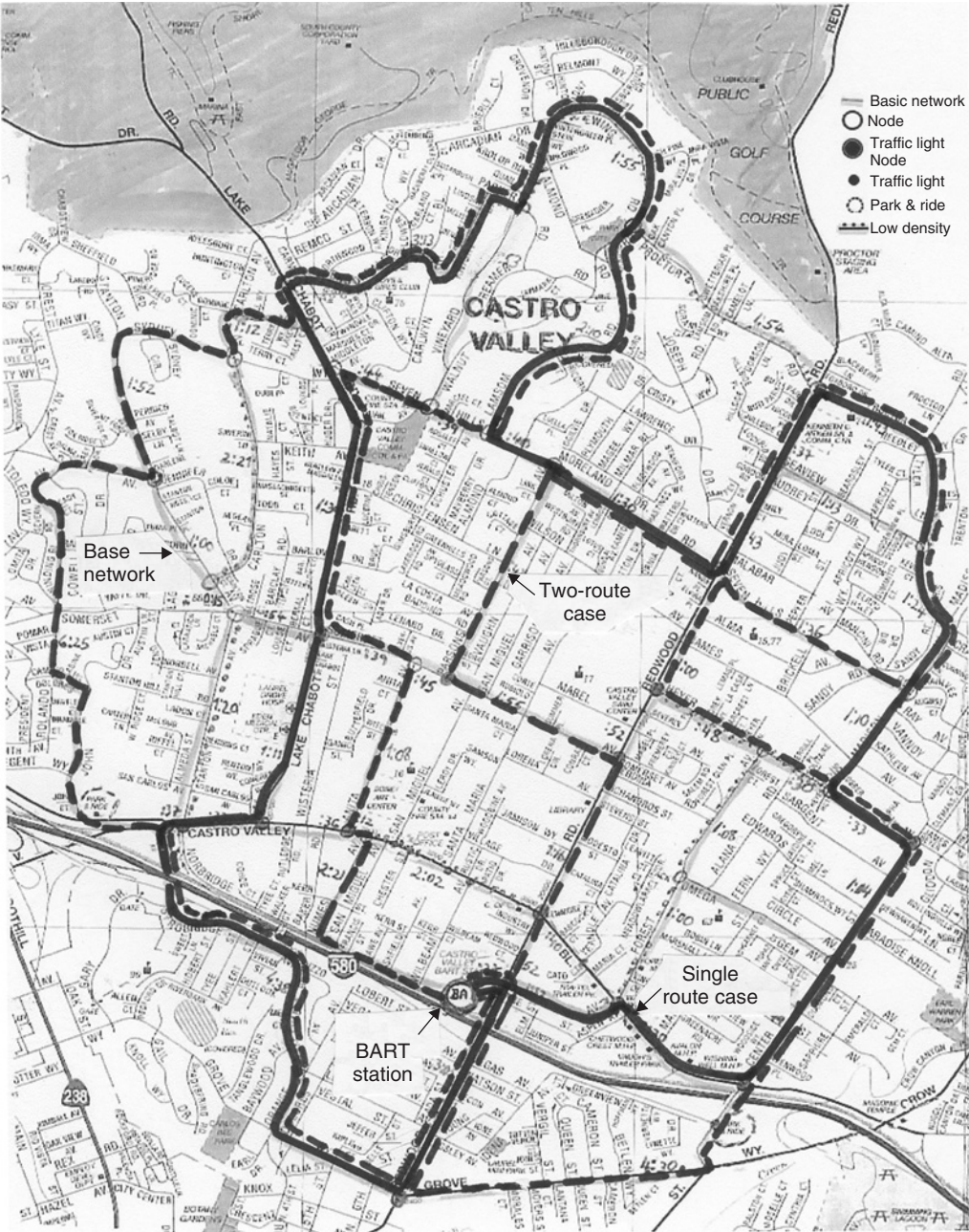


Figure 16.8 Routing map of the Castro Valley case study: Base network is defined by grey lines, single-route case by a solid line, and the two-route case by dashed lines

Table 16.1 Simulation results for waiting time per passenger (in minutes), using different combinations of strategies and numbers of buses for the Castro Valley case study (given demand: 400 buses daily)

Strategy	No. of buses			
	1 bus	2 buses	3 buses	4 buses
1	51	22	20	20
2	25	22	17	15
3	24 ✓	23	15 *	14 ✓
4	25	17 *	16 ✓	15
5	24 ✓	18 ✓	15 *	12 *
6	24 ✓	17 *	16 ✓	15
7	24 ✓	18 ✓	15 *	12 *
8	24 ✓	23	16 ✓	15
9	24 ✓	18 ✓	15 *	12 *
10	22 *	18 ✓	15 *	15

* = Best result

✓ = 2nd-best result

Table 16.2 Simulation analysis for the two-route, 4-bus case, with a 20-minute criterion (maximum waiting after call)

Strategy ⇔	Fixed-route, fixed-schedule (no. 1)	Fixed-route, flex-schedule (no. 8)	Flex-route, flex-schedule (no. 10)
Average wait from phone call to bus arrival (minutes)	13.3	8.9	6.8
Standard deviation (minutes)	3.0	3.7	2.5
Average wait at train station (minutes)	13.6	1.6	1.4
Standard deviation (minutes)	15.3	8.5	5.0

waiting at the train station have the shape of a large step function. It should also be mentioned that the X-axis scale is not the same in all cases; it simply reflects the resultant waiting-time range. A comparison between Strategies 1 and 2 (Figures 16.9 and 16.10) for the pickup case reveals that whereas the wait ranges from 5–20 minutes in the fixed schedule,

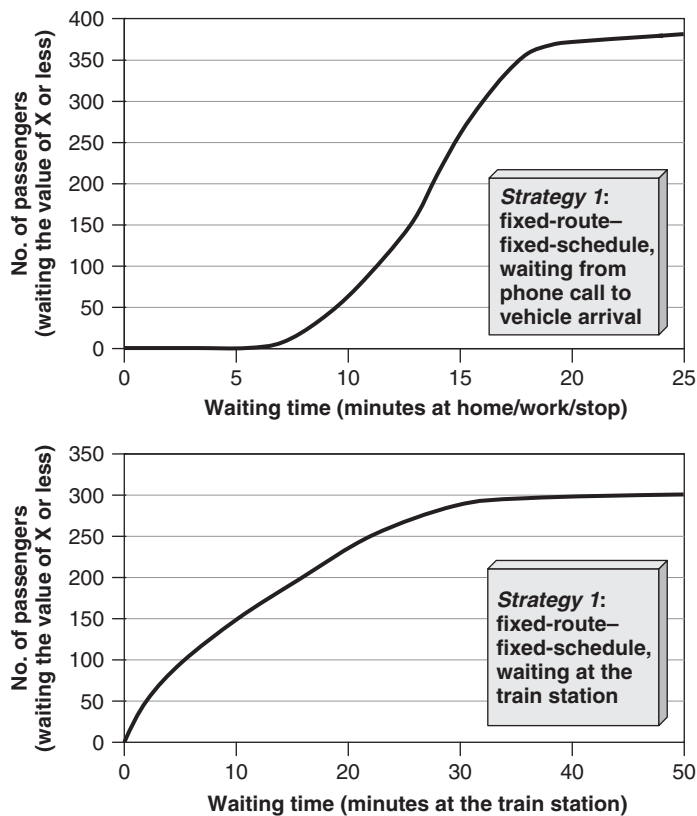


Figure 16.9 *Waiting time (phone call to bus arrival in upper curve, and at the train station in lower curve), in minutes, for Strategy 1 (fixed route–fixed schedule); note that waiting-time scales are not the same*

it ranges from 3–18 minutes in the flexible schedule. In the demand-responsive case (Strategy 10, Figure 16.11), the waiting time for the pickup case ranges from 3–13 minutes.

These simulation runs are only preliminary steps toward the examination of a smart-shuttle operation. More simulation runs are required for different numbers of fixed routes and various demand levels, along with further sensitivity analyses of the input parameters.

16.6 Customer survey

To obtain the needed consumer information, the case study used the survey-research method described by Yim and Ceder (2006). The test market was identified as being within a two-mile radius of the Castro Valley BART station. Four hundred telephone interviews were conducted in this market area, using a random-digit-dial sample and the computer-aided

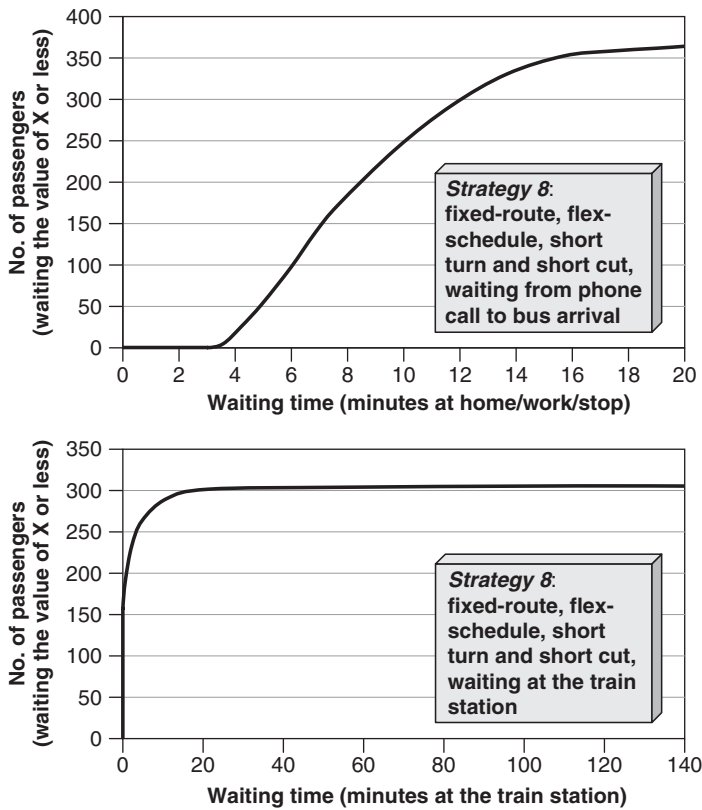


Figure 16.10 Waiting time (phone call to bus arrival in upper curve, and at the train station in lower curve), in minutes, for Strategy 8 (fixed route-flex schedule; with short turn and short-cut); note that waiting-time scales are not the same

telephone-interview (CATI) technique. The following criteria were used for screening survey participants:

- 18 years old or older;
- a permanent resident of the household surveyed;
- said that BART was a possible means of transportation for them;
- commute or make their most frequent trip away from home by some means other than walking or bicycling.

The margin of error for a 400-respondent sample is $\pm 5.0\%$ at the 95% level of confidence.

The survey questions covered the following topics: (1) trip characteristics, (2) mode of access transportation to BART, (3) willingness to use a smart shuttle, (4) willingness to pay for the service, (5) desired attributes of the shuttle service, and (6) demographic characteristics of the survey respondents. The study identified the features that would attract consumers in terms of routing characteristics (i.e. intermediate stop options, express service), travel time, waiting time, number of stops and willingness to pay for the shuttle service.

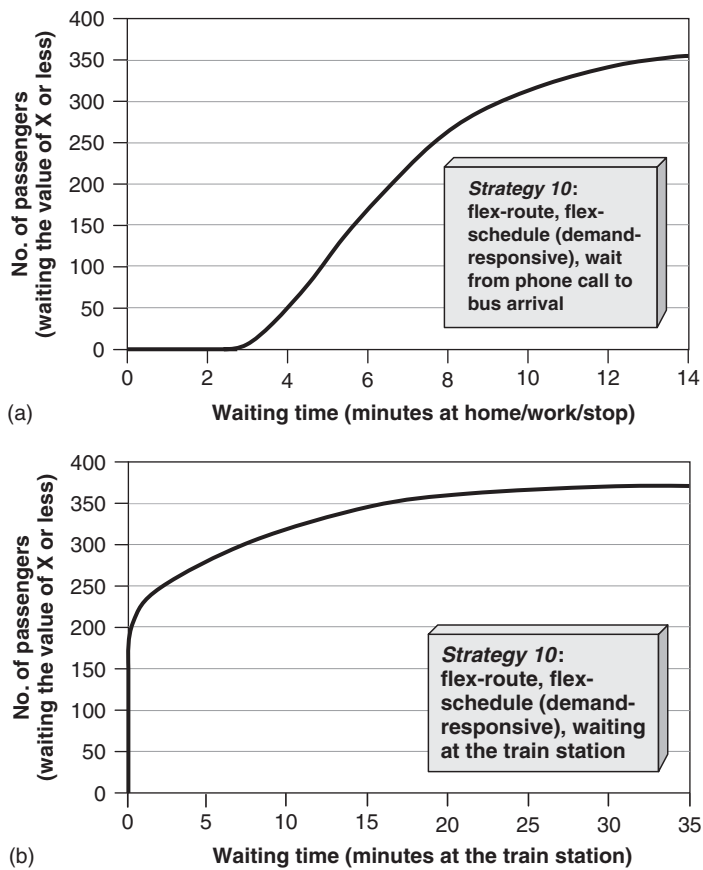


Figure 16.11 *Waiting time (phone call to bus arrival in upper curve, and at the train station in lower curve), in minutes, for Strategy 10 (demand-responsive); note that waiting-time scales are not the same*

16.6.1 Survey results

Several attributes were investigated with respect to the design of the shuttle service. Among them were the number of pickups and drop-offs, the size of a shuttle vehicle, acceptable number of riders, travel time and waiting time. The results for these attributes and for other questions are as follow.

Important attributes for shuttle design

The respondents said that the most important attribute was the cost of the shuttle service. The second most important was overall travel time, including waiting time for the shuttle either at BART or at the pickup location. The third most important was the on-time reliability of the service at the pickup location or at the BART station.

Pickups and drop-offs

Most people expected four to five pickups (median of five pickups) on the way to the BART station. Similarly, they expected four to five drop-offs on the way home (median of five drop-offs).

Travel time

The question was, If the average travel time to the BART station was slower than it currently is, do you think you might use the BART shuttle service? Nearly one-third (29%) said they would take the shuttle, whereas half (53%) said they would take the shuttle if it took about the same time. Only 12% said they would take it only if it were faster. The survey suggests that people are willing to accept a longer travel time when using a shuttle because of whatever benefits they perceive it affords.

Arrival time and schedule information

One of the reasons that people are hesitant to go by public transit is the uncertainty associated with vehicle-arrival times. The advanced transit information system (ATIS) can disseminate real-time vehicle-schedule information to those who are regular transit users, as well as to the occasional transit rider. ATIS vehicle schedules can also attract those who seldom or never used public transit in the past.

Cost

The cost question for riding a shuttle was constructed to ask about the higher price first and then lower prices. Suppose the cost of the shuttle service were \$1 for a one-way trip, how likely would you be to use this service? According to the survey, half the respondents said they would be interested if the cost were \$1; furthermore, 41% said that they would still be interested in taking a shuttle at the price of \$2 for a one-way trip. As expected, consumer interest in using the shuttle service is highly elastic with respect to the cost, as the results in Figure 16.12 illustrate. However, the survey found that price elasticity was not directly proportional to the cost of the shuttle service. The willingness to use the service is significantly (95% level of confidence) different between the low and the high cost of the shuttle service.

Waiting time

Questions about waiting time were also asked at the same time as the cost of the shuttle service. If the waiting time was 20 minutes, 15 minutes, 10 minutes and 5 minutes, how likely would respondents be to use the shuttle service? The results are shown in Figure 16.12. The survey showed that the longer the waiting time, the less willing people are to take the shuttle; however, there was not a significant difference between a ten-minute waiting time and a five-minute waiting time. This suggests that half of the shuttle users are willing to wait five minutes and up to ten minutes for a vehicle.

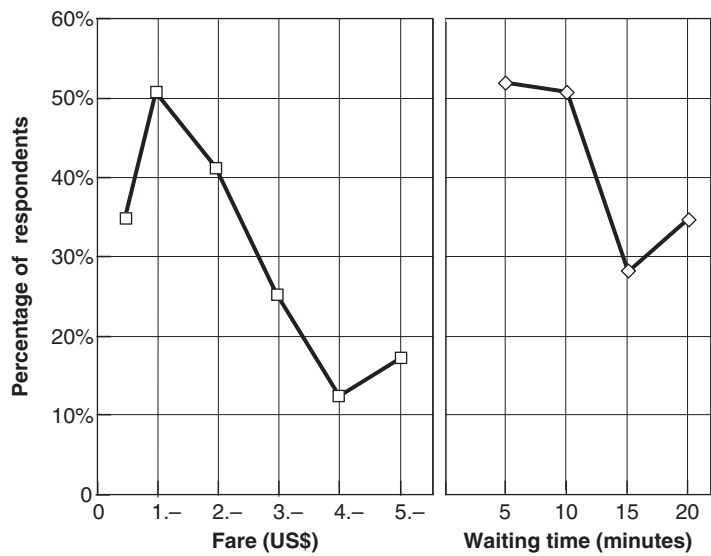


Figure 16.12 *Willingness to pay and wait for the shuttle service*

Frequency of using the shuttle service

Most respondents said that they would use the service two to three times a week.

Payment method

Over half (55.4%) of the respondents were interested in paying for the service on a per-user basis. Only 11.2% responded favourably to a weekly fee basis, and 31.2% said they could work with a monthly subscription arrangement.

Preferred means of receiving information about the shuttle

The survey showed that 62% would like to receive their information from a pamphlet. Approximately one-third (30.9%) preferred to retrieve it through the Internet, and only 5.9% would like to receive it by telephone.

The benefits of the shuttle

A question pertained to the biggest benefits derived personally from using the shuttle service. It was an open-ended question and accepted up to three responses; therefore, the percentages shown in this section are not mutually exclusive. Respondents mentioned a variety of personal benefits. Among them were these: (i) convenience, including no need to park (25%), avoid walking in bad weather (2%), avoid wear on vehicle (21%), and others (30%); (ii) safety, including reduced stress and anxiety (8%), less chance of an accident (2%), the avoidance of traffic fights (18%); (iii) travel time savings (14%); (iv) less cost (18%); (v) reduced pollution (7%); and (vi) the chance to meet people and to socialize (2%).

16.7 Optimal routing design: base network

This and the following sections present, respectively, a methodology and a procedure for designing optimal shuttle (circular) routes, following Jerby and Ceder (2007). The objective set forth for an urban road network is maximal coverage of potential passenger demand to a transportation centre (e.g. train station), with travel time along the entire route not exceeding a certain threshold value. Generally, an objective function that aspires for maximum coverage is supposed to create a long, winding route, which in this case is blocked by the route-time constraint to ensure an adequate level of service. The problem input consists of an urban network with trip-generation nodes and a single destination (main) node, average travel time between each segment on the network to the main node, and a constraint defining maximum travel time along the route.

The methodology described is based on a modular approach, enabling the partitioning of a complex problem into a series of sub-problems. Hence, each sub-problem can be referred to as an independent component. The methodology consists of eight main stages, as shown in Figure 16.13. The *first stage* is the characterization of urban network attributes for deriving a base network for route design. The *second stage* deals with the insertion of average travel times into the links of the road network, while the *third stage* defines the service area. The *fourth stage* focuses on a method to estimate potential demand for trips on a designed shuttle route. This method includes the development of a potential demand measure for each link on the road network, based on urban and spatial criteria, and using density and walking-distance parameters. This measure is used as an input to an optimization model developed in the *fifth stage*. The model enables the automated design of an optimal circular route complying with a given total travel-time constraint. Test runs using the optimization model are performed in the *sixth stage*, showing that its complexity is high and cannot be efficient for medium and large-size networks. Accordingly in the *seventh stage*, an alternative heuristic algorithm is proposed and developed. The algorithm, which enables the automatic design of circular routes, ensures good, mostly but not always optimal, results. The *eighth stage* examines the heuristic algorithm on different networks.

16.7.1 Base road network

Theoretically, there could be a situation in which all city streets are used for the passage of transit vehicles; in actuality, this is not the case. For various reasons, some of the network links will never be used for the passage of transit vehicles; therefore, these links can *a priori* be excluded from the planning process. Such an action will define a more reduced transit-road network (the base network). Working with a base network allows the planner to reduce the complexity of the problem and enables the transit agency and the passengers to rely on a simpler network that is easier to understand and operate. Determination of a base road network consists of the following elements: street characteristics (width, slope, parking arrangements); spacing between parallel streets; and safety considerations or any other criterion set by the transit agency's constraints on a case-by-case basis.

According to the stages in Figure 16.13, the travel time in each network link is determined after creating the base network. This value can be either measured directly or calculated according to the length of the link and the average travel speed. The next stage in Figure 16.13 is to define the 'service area' that can be served reasonably by a single circular

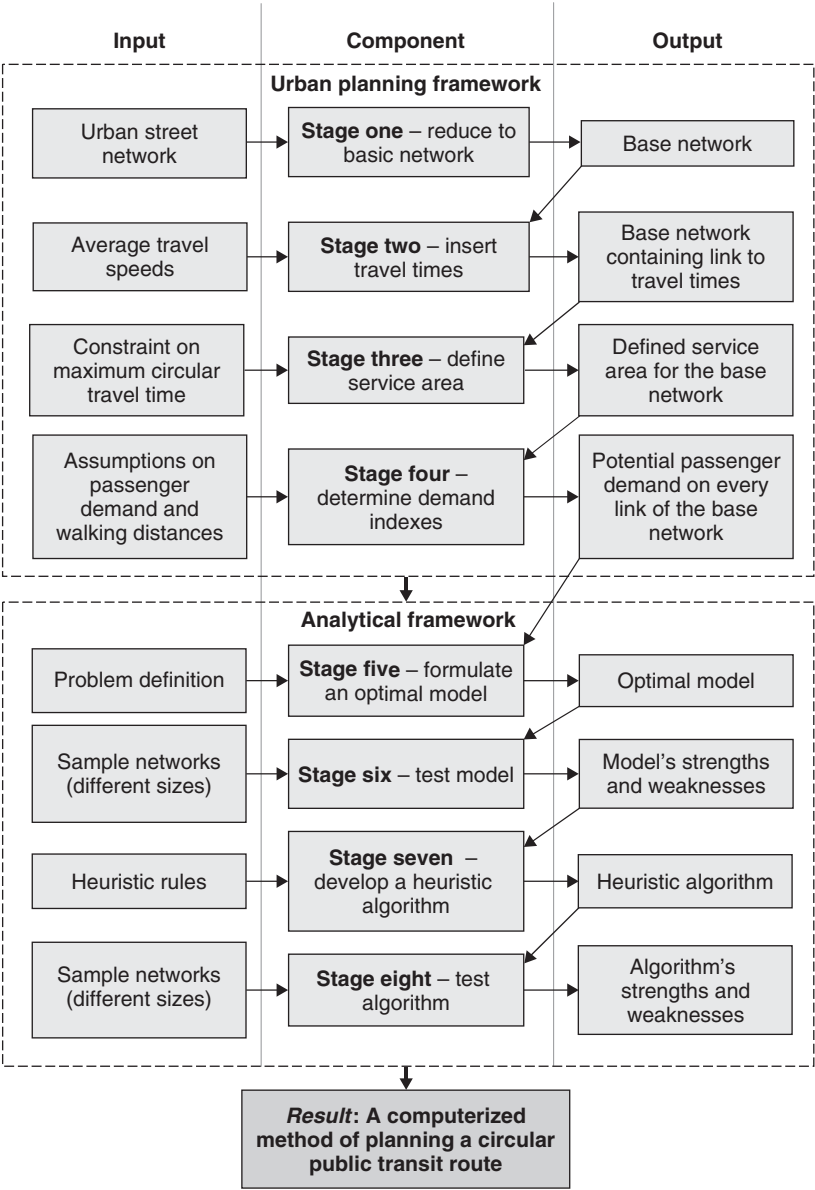


Figure 16.13 Urban planning and analytical frameworks for a circular route design

route. The size of the service area depends on the route’s travel-time constraint; i.e. links that cannot be covered in the framework of the travel time allocated should be eliminated from the base network. The proposed process includes testing of each of the base network links in order to determine whether a circular route can be created from the link to the main

node and back while complying with the time constraint. For convenience, the main node will continue to be referred to as the train station. Should the duration of the shortest route (in travel-time values) from the link to the train station and back exceed the time constraint, such a link should be eliminated from the network. Applying this procedure on all links of the base network will result in a further reduction of the network and the elimination of distant links that should not have been taken into consideration in the first place.

16.7.2 Potential passenger demand

The demand potential is assessed on the basis of spatial data from the urban network, and is subject to a number of assumptions. Methods for assessing transit demand, which were presented in Chapter 11, include the counting of passengers, conducting surveys, using known databases of residential addresses and workplaces, and weighting of socio-economic and demographic data of the region. In addition to these methods, following is a two-phase alternative method that uses spatial and demographic data.

Phase A: Calculation of average walking distance per each link

This phase considers the maximum acceptable walking distance to the shuttle route for potential passengers. This variable is known to be part of a 'catchment area'. It is customary to assume that it encompasses an area of up to 400 metres from every side of the road link, although some passengers will be willing to walk a longer distance under certain conditions. Figure 16.14 illustrates the catchment areas for pedestrians in a case-study area of downtown Haifa (Israel) as polygons on the base network. In this case, the catchment area amounts to 300 metres from every side of the route. In Figure 16.14, one can identify areas where overlapping is created between two or more polygons; the assumptions discussed below are used for such overlapping cases.

The first assumption is that passengers choose the road link nearest to the point of origin as the waiting point for the shuttle vehicle. The second assumption is that passengers will walk to the nearest possible waiting point by taking the shortest 'bee line' distance (a reasonable assumption, since passengers in many cases can walk to the stop in an almost straight line through parks or alleys). Therefore, in cases in which catchment areas overlap, every point in the overlapping area can be attributed to the nearest bee-line link. A value assessing the average walking distance in the area will now be defined.

The average walking distance $\overline{wd}_{i,j}$ on link (i, j) is influenced by land uses within the catchment area and is calculated as follows:

$$\overline{wd}_{i,j} = \frac{\sum_{b=1}^m \overline{pop}_b \cdot wd_b}{\overline{pop}_{i,j}} \quad (16.2)$$

where m represents the total number of buildings in the catchment area; b is a certain building in the catchment area; $\overline{pop}_b$ represents the amount of population (residents, employees) in building b; wd_b is the walking distance from building b to the nearest road link; i and j represent network nodes; and $\overline{pop}_{i,j}$ is the amount of population in the catchment area between nodes i and j.

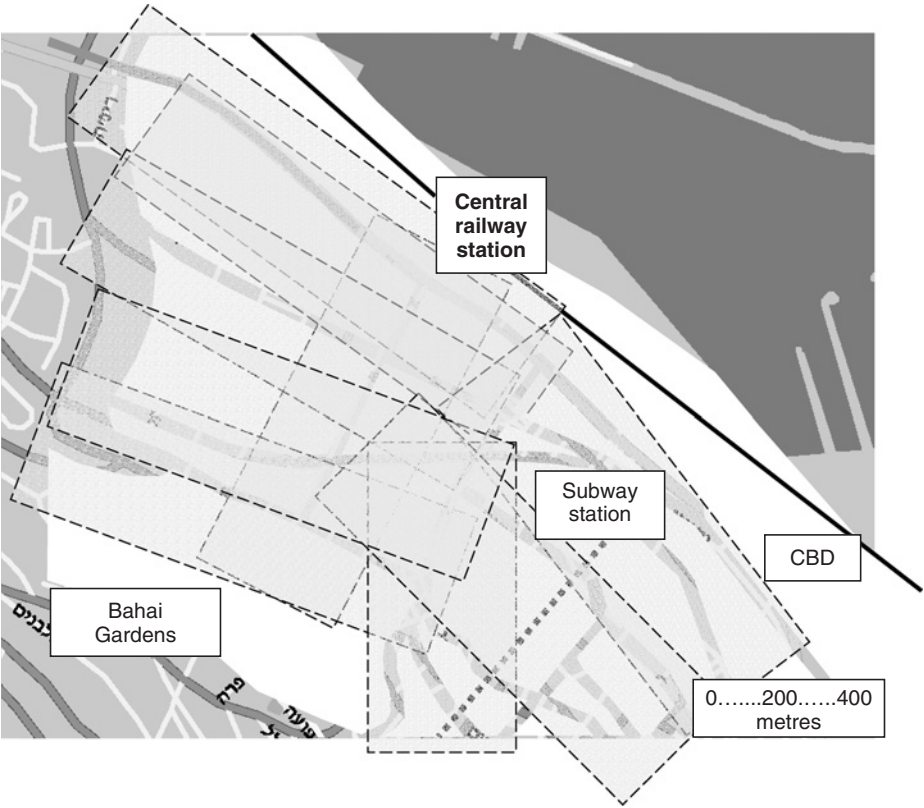


Figure 16.14 Catchment areas for the base network of the case study (city of Haifa)

Phase B: Assessment of the demand potential as dependent on walking distances

At this phase, the demand potential from each of the catchment areas can be assessed, based on two more assumptions. The first of these assumptions is that the entire population in the pedestrian catchment area can be related to as potential users, regardless of their motorization and socio-economic levels; this means that there is a correlation between building density and demand potential. The second assumption is that the demand for trips also depends on the walking distances from the land-uses around the road to the shuttle stop; that is, the further the land-use from the road link, the less the demand potential.

Based on these assumptions, a demand potential index is proffered for every network link (i, j) as follows:

$$pd_{i,j} = \frac{\overline{pop}_{i,j}}{\overline{wd}_{i,j}} \tag{16.3}$$

where $pd_{i,j}$ expresses the ratio between the amount of population at each link and the average walking distance per link. In other words, the demand potential measure (index) in

Equation (16.3) is in direct proportion to the population density and in inverse proportion to the average walking distance.

The two-phase method described for estimating demand requires quality geographic information. This method was implemented in the case study shown in Figure 16.14, with bands around each network link. When no exact data existed on the number of residents in each block, assessments were carried out based on a multiplication of the number of building floors (using the number of apartments per floor) and the approximate size of households. The outcome for the case study in Figure 16.14 is presented in the schematic network of Figure 16.15.

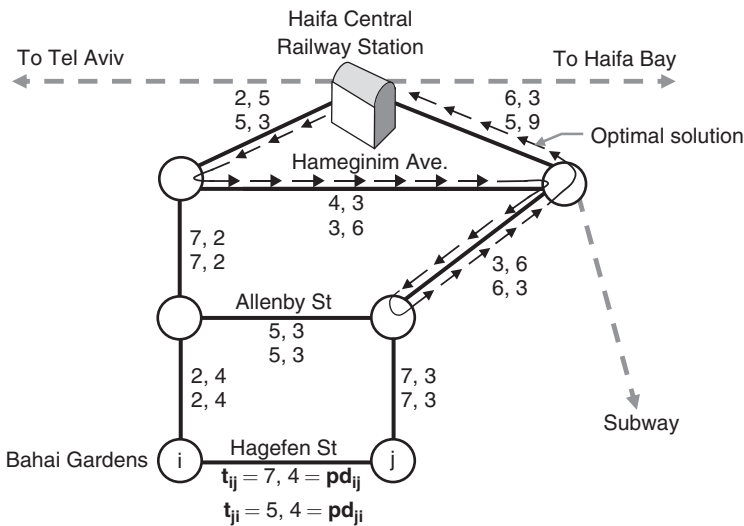


Figure 16.15 Schematic description of the case-study network and its optimal circular route solution for $T = 25$ minutes

16.8 Optimal routing design: algorithm

This section contains two main parts. First, an optimization model formulation is described, using operations research (OR) terms (see Section 5.3 in Chapter 5). Second, a heuristic algorithm, following Jerby and Ceder (2007), is presented for handling large-size networks.

16.8.1 Optimization model

Let $G = \{N, A\}$ be a network of streets in an urban area, with sets of nodes and links (arcs) designated N and A , respectively. Each (i, j) link is associated with an average travel time of t_{ij} minutes by time-of-day, and demand potential index pd_{ij} . The potential demand value is assumed to be homogeneously distributed at each link.

The model described produces a circular route r , starting and ending at node n_1 . The model maximizes the demand potential of passengers while complying with the constraint of maximum circulating travel time T . Its formulation is as follows:

$$\max \sum_{i,j \in r} pd_{i,j} \cdot y_{i,j} \quad (16.4)$$

Subject to:

$$\sum_{i,j \in r} t_{ij} \cdot y_{ij} \leq T \quad (16.5)$$

$$\sum_j y_{ij} = \sum_k y_{ki} \quad \text{for } (i,j) \in r, (k,i) \in r \quad (16.6)$$

$$S = \{s_1, s_2, \dots, s_M\} \quad \text{in which } s_m \in N, n_1 \notin s_m, |s_m| \geq 2 \quad (16.7)$$

$$\sum_{\substack{i \in s_m, j \in s_m \\ i \notin L_u, j \notin L_u}} y_{ij} \leq |s_m| - 1 \quad \forall s_m \in S, 1_u \in L \quad (16.8)$$

$$pd_{ji} = 0.2 * pd_{ij} \quad \forall (i,j) \in r \quad (16.9)$$

$$y_{ij} = \begin{cases} 0, & \text{if } (i, j) \in r, (i, j) \in A \\ 1, & \text{if } (i, j) \notin r, (i, j) \in A \end{cases} \quad (16.10)$$

Equation (16.4) contains a $\{0, 1\}$ variable y_{ij} in which $y_{ij} = 1$ means that the link (i, j) is part of r , or zero otherwise. Equation (16.5) represents the maximum travel-time constraint. Equation (16.6) implements a condition for creating a closed circular route in the network; it follows the known OR Euler theorem (e.g. see Eiselt *et al.* 1995), in which the number of links incoming and outgoing every node is equal. Equation (16.7) defines S as a set of all the network-node combinations, which include at least two nodes without node n_1 . It also ensures that the number of route links connecting the nodes of any subset s_m of S will be smaller than the number of nodes in the set (resulting in a tree or a forest).

Equation (16.8) ensures that there will be no group of nodes along r in which the number of links connecting the nodes is equal to or larger than the number of nodes. In other words, Equation (16.8) requires that all chosen links must be connected to each other, thus avoiding the formation of more than one continuous circular route. Equation (16.8) also ensures the existence of a situation in which two (or more) circular routes create overlapping in some of the nodes (e.g. figure 8-shaped routes). For this purpose, let $L = \{\ell_1, \ell_2, \dots, \ell_u\}$ be a set of all the possible one-circle routes passing through node n_1 . The constraint should be operated for each of the group L routes. The condition established in Equation (16.8) eliminates a situation in which additional circular routes are created that are detached from a tested ℓ_k route; instead, it enables the creation of sub-routes that may have overlapping nodes with r .

Equation (16.9) postulates that the demand potential in the opposite direction for a checked link is significantly smaller: 20% less. In practice, this means that the majority of

passengers will board the shuttle upon its first arrival. This modelling limits the model's tendency to repeat the same link in both directions. Finally, Equation (16.10) defines y_{ij} .

Note that Equations (16.5) and (16.9) increase in geometrical progression and that Equation (16.8) increases in a manner that cannot be described by a polynomial equation, hence causing a high level of complexity. The conclusion, therefore, is that solving the problem of real-life road networks through the formulation described is impractical. This complexity in the number of calculations warrants the examination of heuristic solutions that is given in the next section.

16.8.2 Heuristic algorithm

The algorithm described is based on an efficiency criterion. For that purpose, an impedance ratio, z_{ij} , is defined as follows.

$$z_{ij} = \frac{\omega \cdot t_{ij}}{\lambda \cdot pd_{ij}} \quad (16.11)$$

where ω , λ are coefficients.

As mentioned, the potential demand is a value that is in direct proportion to inhabitant density and in inverse proportion to walking distances. The efficiency criterion utilized in the algorithm is that of the shortest path between each two links of a selected group of links in accordance with Equation (16.11).

The heuristic algorithm is shown in a flow diagram in Figure 16.16. The algorithm constructs an efficient circular route from an initial group p of links ($p \leq P$) characterized by relatively high demand potential; it utilizes Equation (16.11) and the shortest path between the links in the selected group p . A small sub-group of links is chosen in every iteration of the algorithm. The algorithm examines the possibility of a circular route that will pass through the sub-group links and meet the travel-time constraint. Then the algorithm compares the potential demand of the candidate route to each of the routes examined so far. Finally, the best circular route found is chosen.

Several test runs were made enabling a comparison of the heuristic algorithm with the optimal model. The results are shown in Table 16.3. The first network is that of Figure 16.15; the seven networks that follow are small, randomly selected networks; last is a larger-size network. One important feature of the heuristic algorithm is group p , which is the number of initial mandatory links chosen for route r . Figure 16.16 defines P as the maximum number of such mandatory links. Table 16.3 shows the ratio between P and the number links in the network, A . $P = 4$ was found to provide reasonable computer running time, including the 42-link last tested network. Finally, the results support the heuristic algorithm, as they are the same as the optimal across all small networks.

16.9 Implementation stages

In order to secure the potential success of a new shuttle service, certain steps should be undertaken gradually and carefully. Essentially, these steps involve an initial quantitative analysis and then a pilot plan.

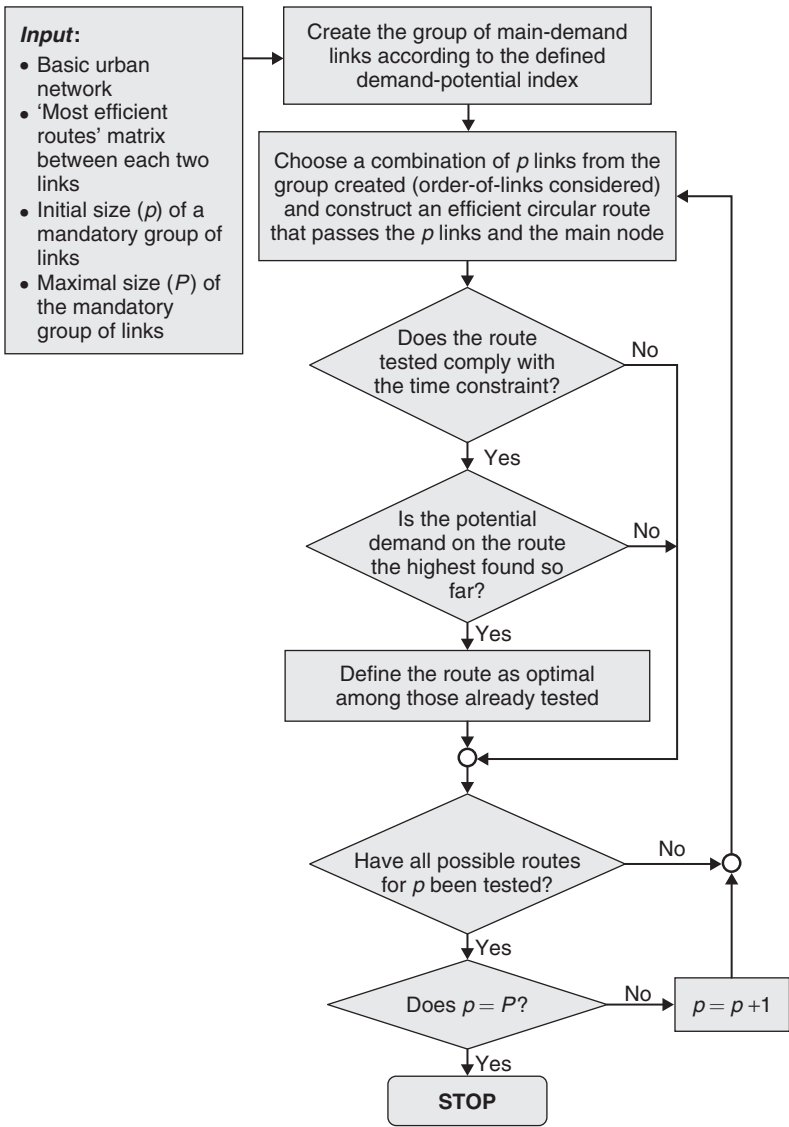


Figure 16.16 Flow diagram of the heuristic procedure

16.9.1 Initial analysis

Figure 16.17 schematically outlines five components that are necessary to complete the initial analysis: (1) constructing a base street network; (2) creating groups of fixed routes; (3) constructing short-turn, short-cut and bi-directional strategies; (4) creating a DRT-type of service; and (5) comparing different strategies with a given passenger demand.

Table 16.3 Summary of the heuristic algorithm runs, given different networks

Network tested	No. of nodes	P/A ratio	Objective function value* (optimal)	Objective function value* (heuristic)
from Figure 16.15	7	4/9	26.6	26.6
Random 1	7	4/9	191	191
Random 2	7	4/7	208.8	208.8
Random 3	7	4/7	206.4	206.4
Random 4	7	4/8	79.2	79.2
Random 5	7	4/9	338	338
Random 6	7	4/9	243.6	243.6
Random 7	8	4/9	204	204
Random 8 (large)	25	4/42	Too long running time	142.2

*In potential units of demand, following Equations (16.2) and (16.3)

Component 1 of Figure 16.17 uses site observations and measurements for a base road-network configuration, including traffic-light locations and an 85 percentile of time to and from the train-station platform. This 85 percentile of the different times observed ensures adequate walking time for the majority of people. The determination of the base road network considers the following elements: approximate (low, average, high) density of residential area, street characteristics (width, slope, parking arrangements), spacing between parallel streets and the road-network shape of each zone in the area.

Component 2 of the initial analysis creates the fixed routes to be considered. These circular routes can be constructed by the heuristic algorithm described in Section 16.8, including a measure of its demand potential. The maximum length of the routes to be selected is influenced by the number of shuttle vehicles available and, if given, the minimum frequency required.

Component 3 constructs the operational strategies that can continuously ensure an adequate level of service. These strategies, outlined and explained in Section 16.3, basically cover short-turn, short-cut and bi-directional possibilities. They can be analysed by a simulation tool similar to that explicated in Sections 16.4 and 16.5.

Component 4 in Figure 16.17 creates DRT strategies, given certain input elements. That is, the input is based on the minimum number of passengers for whom a vehicle may be sent for a pickup; the train schedule, in order to match the expected arrival time of the DRT vehicle with the train's arrival; the 85 percentile of walking time to and from the train-station platform; and average times required to pickup and drop-off passengers. The tool to create this type of dynamic, on-line routing is based on an algorithm (e.g. shortest path from point to point), using the simulation tool discussed in Sections 16.4 and 16.5.

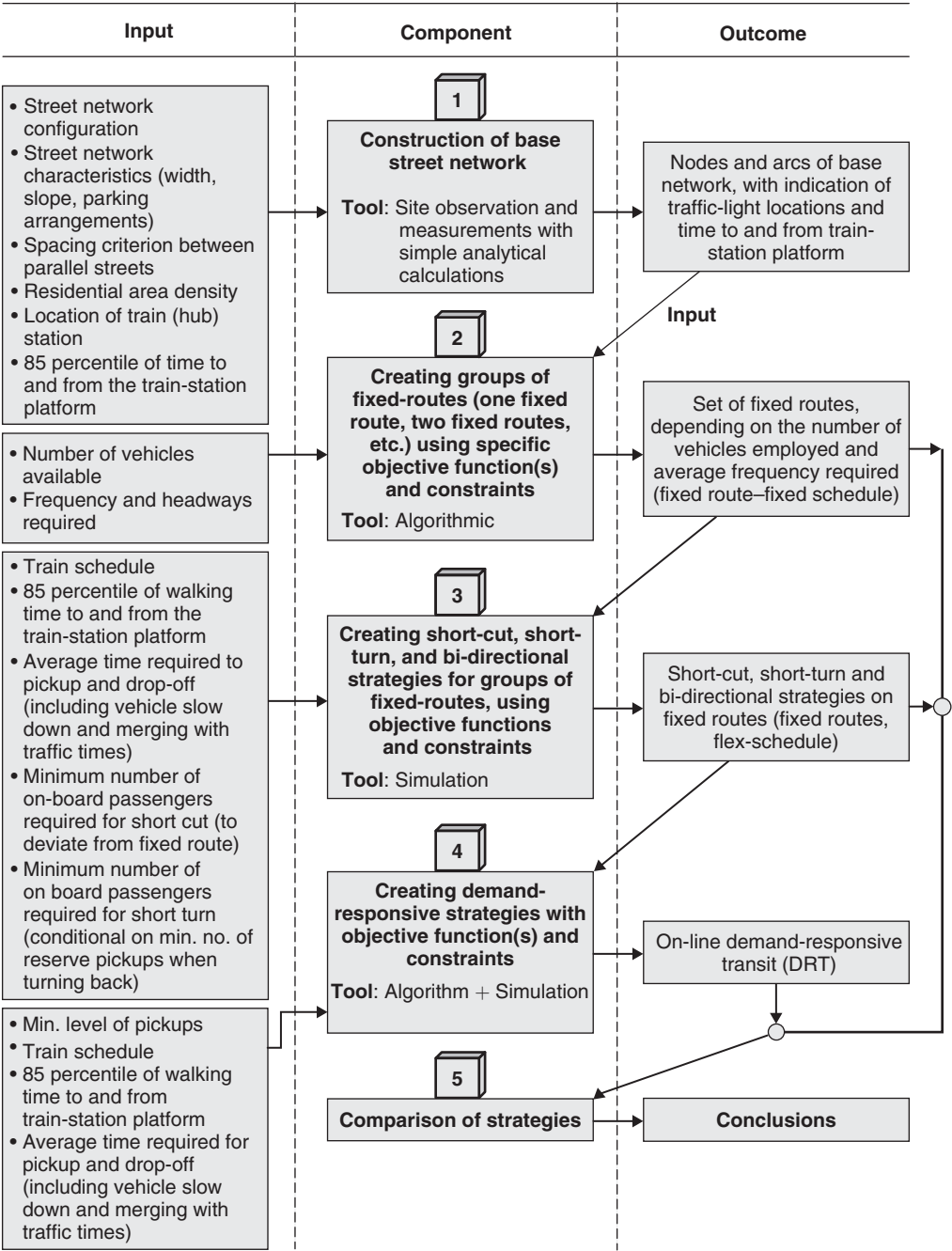


Figure 16.17 Practical methodology for constructing a feeder/shuttle service

Finally, component 5, which consists of the outcome of components 2, 3 and 4, compares the different strategies. This comparison can cover different demand levels, different numbers of available vehicles, and different input parameters (travel times, threshold values for dispatching a trip, short turn, short cut, etc.). The comparison will lead to a strategy that can better fit a given situation (by time-of-day and demand level).

16.9.2 Pilot plan

Once the analysis of the shuttle service is completed, the application of a pilot study is recommended. The implementation of such a pilot in the area being considered can follow, for example, the 12 steps shown in Figure 16.18. These 12 steps can serve as a framework for a master plan, with each outcome of a step becoming an additional input for the next step, except for Step 6.

The pilot master plan starts with a demand analysis by time-of-day and day-of-week in order to find the origin-destination pattern and consumer-oriented features.

The second step is to establish (if done previously in Sections 16.7 and 16.8) or design the fixed-routing and stop system. The third step is to determine the base frequencies and timetables for each route or, alternatively, to set the operational strategies in accordance with Section 16.3. The fourth step is to determine the number and size of the shuttle vehicles and to create the chain of trips (vehicle schedules) that will serve the fifth step, constructing the crew schedules.

The pilot plan continues in Step 6 with the establishment of effective information channels and instruments (e.g. call centre, Internet, newspapers, radio, TV, mail leaflets) that will lead to the development of user-friendly communication procedures between the users and a selected operating agency in the next step (Step 7). Step 8 constructs the DRT operational strategies without the use of the fixed routing/stop/schedule system. Step 9 determines the testing scenarios of the pilot while Step 10 presents the process for selecting an adequate operator. Step 11 uses proper advertising tools to approach an operable pilot study. Finally, the last step of the plan (Step 12) is aimed at improving the instruments, procedures, and strategies with the use of innovative intelligent transportation systems (ITS).

16.10 Literature review and further reading

This section describes models for the design of shuttle or feeder routes. Only models corresponding with a fixed route fall within the scope of this review.

Wirasinghe (1980) formulates an analytical model for a system in which bus routes feed a railway line during peak hours. It is assumed that the street network is a perfect rectangular grid and that there is no need for synchronization between the feeders and the rail line. Two cases are presented. In the first case, the locations of railway stations are given, and the optimal density and headways of the buses determined. In the second case, the spacing of feed points is not given, and therefore all three parameters (station spacing, route spacing and headways) are optimized.

Kuah and Perl (1987) present an optimization methodology for the design of an integrated bus–rail system. A mathematical programming problem is formulated in which the itinerary

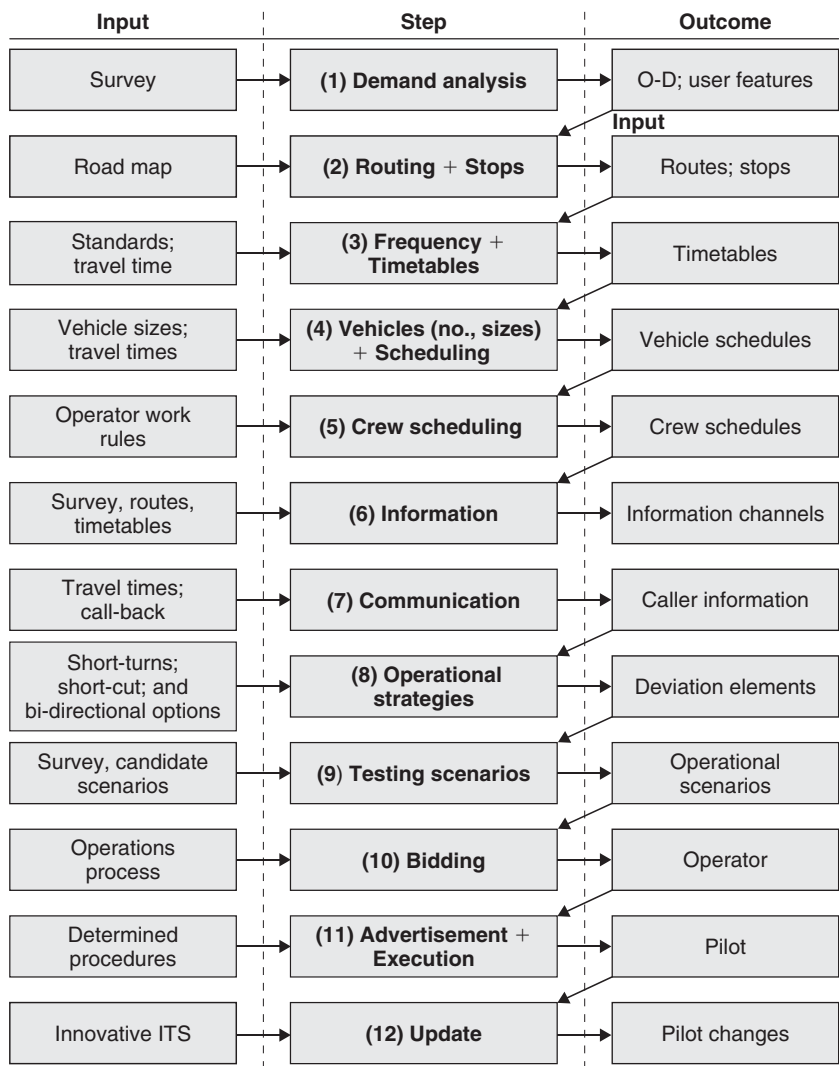


Figure 16.18 Overview of a possible feeder/shuttle pilot master plan

and frequency of each route are determined under a fleet-size constraint. In addition, a heuristic solving method is developed. For the basic problem, a many-to-one demand pattern is assumed. This demand pattern, which is assumed in most feeder-route models, means that passengers from various origins travel to either one destination or various destinations, but they must pass one transfer hub that is viewed as their destination when the feeder system is designed. The authors show, however, that the model can be generalized to represent many-to-many demands. It enables simultaneous network-wide optimization, meaning that the bus system designed may serve more than one rail line. The model is not based on binding

assumptions regarding the street network; streets may form any structure, and zones do not have to be rectangular. In addition, the feeder routes are not necessarily perpendicular to the railway line.

The same researchers (Kuah and Perl, 1988) describe an analytical model for the design of buses feeding an existing rail line. The feeder routes in this case are perpendicular to the rail line in a grid-shaped network. The variables optimized are headways, route spacing and stop-spacing. The paper examines three different stop-spacing policies: uniform network-wide stop-spacing, uniform line-specific stop-spacing that may change from one route to another, and stop-spacing that can vary along the route. They conclude that headways and route spacing are not sensitive to system characteristics.

Chang and Schonfeld (1991) present a model for designing feeder routes to one transfer station in a simple rectangular service area. The many-to-one demand is elastic to service quality and fare. The model allows the designed service to change, along periods of the day, according to demand characteristic, operating costs and vehicle speeds, which are given separately for each period. The suggested model makes it possible, therefore, to determine the optimal route spacing that does not change during the day, with varying route headways within each period. Models are formulated and compared for four types of conditions: steady fixed demand, cyclical fixed demand, steady equilibrium demand, and cyclical equilibrium demand. The results for the four cases are compared. All demand patterns (with minor exceptions) are found to have an optimal constant ratio between the headway and route spacing.

Chang and Lee (1993) and Chang and Yu (1996) develop mathematical models for optimizing fixed-route and flexible-route feeder services in a simple rectangular area. Separate models are formulated for each of the two alternatives with the objective of maximizing welfare under a subsidy constraint; the optimal results of both types of service are compared. Decision variables are route spacing, headways, and fares; their optimal values are examined under several possible demand functions.

Chang and Schonfeld (1993) propose an analytical model for the design of parallel routes that feed a central terminal. Demand is assumed to vary during the day, and therefore different headways are assigned to different day periods. Other decision variables – zone size, route length, and route spacing – are determined without differences between periods.

The model developed by Chien and Schonfeld (1998) strives for a joint optimization of a rail route and its feeder buses. The variables optimized are rail-line length, feeding-station spacing, bus headways, bus stop-spacing and bus routespacing. Street network is assumed to be a simple grid, with the feeders perpendicular to a single rail line. On the other hand, the model is also based on the realistic assumption of many-to-many demand.

Chien (2000) presents a model for defining bus-feeder routes and their headways. One central route is fed into one major station, and the demand is many-to-one. The service area is shaped as a partial grid; that is, a grid with some of the links missing. The form of each zone must be rectangle-like in a simple grid, but zone size may be heterogeneous. The model enables the user to decide the number of feeder routes to be defined.

Chien, *et al.* (2001a) introduce a methodology for the design of a feeder-bus route feeding a major inter-modal transfer station in a suburban area. The model is formulated as a programming problem with geographic, capacity and budget constraints. Decision variables are route locations and headways. The street network has a partial grid pattern, and the demand pattern is many-to-one. Two solution algorithms are suggested. The first is an

exhaustive search algorithm that promises, in principle, an optimal solution; however, it is very time-consuming in most realistic networks. The second suggested algorithm is a genetic algorithm; i.e. an algorithm in which the search for optimal solutions is based on the natural evolution process. In each iteration of a genetic algorithm, a set of solutions is selected based on the previous set; solutions that show good performance in each iteration have a higher probability of being selected for the next iteration. Genetic algorithms are efficient in terms of calculation time, but may result in only a near-optimal solution.

Chien *et al.* (2001b) compare fixed and flexible route systems serving as feeders for a transportation centre close to a rectangular service area. Optimal vehicle size, route spacing, and headways are determined for both types of service. The models assume a probabilistic demand, which varies over periods of the day. Passengers' value of time is assumed to be non-additive; i.e. the value that a passenger ascribes to a two-minute period is higher than twice the value ascribed to a one-minute period. The researchers find it best to operate conventional buses at peak hours and flexible routes in off-peak periods. The threshold value of demand at which one system is more cost effective is calculated.

Shrivastava and Dhingra (2001) present a heuristic algorithm for the development of bus routes feeding railway stations. The algorithm uses different node-selection and insertion strategies to determine route itineraries under many-to-many demand. Routes are developed for two types of time criteria. The first is a maximum-demand-deviated, shorter, time-path criterion; this means that when nodes are inserted into the middle of a route, the length of the deviated route should not exceed a given acceptable limit. The second criterion is a path-extension time criterion, meaning that if nodes are inserted at the end of a route, then the maximum route length in terms of travel time should not exceed a certain maximum value. Use of the suggested method is possible in any network, no matter the street-network structure. A case study is presented, in which the use of this method enabled serving the same demand as previously with a smaller number of routes, a smaller fleet, and better schedule adherence because of the shorter routes.

Another paper by Shrivastava and Dhingra (2002) presents a methodology for designing a coordinated schedule for existing bus routes feeding a railway station. The model presented aims at minimizing total operating costs and total transfer times under the constraint that each transfer time should be executed within given boundaries. The model is formulated as a nonlinear, non-convex programming problem with many variables, and therefore it is difficult to solve. A genetic algorithm is suggested to provide near-optimal results. The results show that the designed schedule is most sensitive to the parameters (such as the transfer penalty). The conclusion is that the value of this penalty should be carefully considered.

Characteristics of the models reviewed are summarized in Table 16.4.

Exercise

Given the following street network connected to a train station. Each link direction (there are two directions of travel per link, marked by arrows) is characterized by a pair of numbers: average travel time (in minutes, on the left) and demand potential ratio (in units of Equation (16.3), on the right). The maximum circulating travel time is given as $T = 30$ minutes; in addition, $\omega = 1.0$, $\lambda = 1.0$, $p = 2$ and $P = 2$.

Table 16.4 Summary of models reviewed, with emphasis on fixed-route models

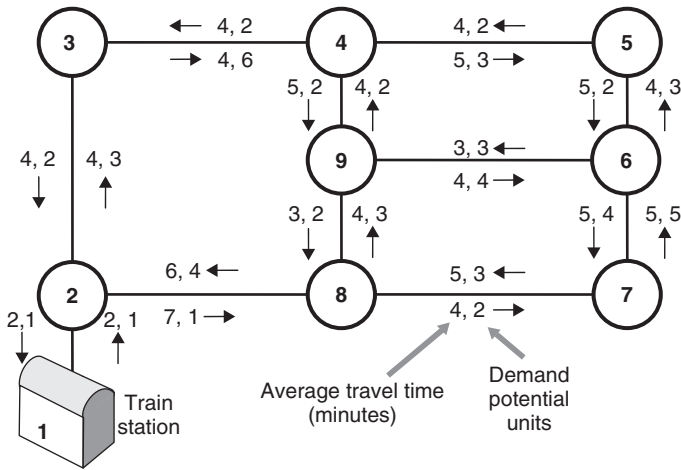
Source	Model formulation	Decision variables	Street network structure	Feeding system	System being fed	Special demand features
Wirasinghe (1980)	Analytical model	Feeder spacing and headways. Locations of feeding points – optional	Simplistic	Parallel routes	One rail route	
Kuah and Perl (1987)	Programming problem	Feeder itineraries and headways	Realistic	Any routes	Several rail routes	Many-to-one and many-to-many
Kuah and Perl (1988)	Analytical model	Feeder spacing, headways, and stop-spacing	Simplistic	Parallel routes	One rail route	
Chang and Schonfeld (1991)	Analytical model	Feeder spacing and day-period-specific headways	Simplistic	Parallel routes	One transfer station	Multiple-period elastic demand
Chang and Lee (1993); Chang and Yu (1996)	Analytical model	Feeder spacing, headways and fares	Simplistic	Parallel routes	One rail route	Elastic demand
Chang and Schonfeld (1993)	Analytical model	Feeder spacing, length and day-period-specific headways	Simplistic	Parallel routes	One transfer station	Multiple-period

(Continued)

Table 16.4 Summary of models reviewed, with emphasis on fixed-route models (continued)

Source	Model formulation	Decision variables	Street network structure	Feeding system	System being fed	Special demand features
Chien and Schonfeld (1998)	Iterative algorithm	Rail line length, location of feeding points, feeder spacing and headways, feeder stop-spacing	Simplistic	Parallel routes	One rail route	Many-to-many
Chien (2000)	Iterative algorithm	Feeder itineraries and headways (the user determines number of routes)	Semi-realistic (partial grid)	Any routes	One transfer station	
Chien <i>et al.</i> , (2001a)	Programming problem	Feeder itinerary and headway	Semi-realistic (partial grid)	One feeder route	One transfer station	
Chien <i>et al.</i> , (2001b)	Analytical model	Feeder spacing, headways and vehicle size	Simplistic	Parallel routes	One transfer station	Multiple-period probabilistic demand
Shrivastava and Dhingra (2001)	Iterative algorithm	Feeder itineraries	Realistic	Any routes	Several transfer stations	Many-to-many
Shrivastava and Dhingra (2002)	Programming problem	Railway and feeder schedule	Realistic	Any routes	One/several transfer stations	

- (a) Create the matrix of the impedance ratio z_{ij} .
- (b) Assume that the initial group of the highest potential demand units contains 5 links (out of 22 links, both directions) of the given network. Find the best single route that will maximize potential demand while complying with the constraint T.



References

- Borndorfer, R., Grottschel, M., Klostemeier, F. and Kuttner, C. (1999). Telebus Berlin: Vehicle scheduling in a dial-a-ride system. In *Computer-aided Transit Scheduling. Lectures Notes in Economics and Mathematical Systems*, **471** (N. H. W. Wilson, ed), pp. 391–422, Springer-Verlag.
- Ceder, A. and Yim, Y. B. (2002). *Integrated Smart Feeder/shuttle Bus Service*. (California PATH Report). Institute of Transportation Studies, University of California at Berkeley.
- Cervero, R. (1998). *Paratransit in America: Redefining Mass Transportation*. Praeger.
- Chang, S. K. and Schonfeld, P. M. (1991). Multiple period optimization of bus transit systems. *Transportation Research*, **25B**, 453–478.
- Chang, S. K. J. and Lee, C. J. (1993). Welfare comparison of fixed- and flexible-route bus systems. *Transportation Research Record*, **1390**, 16–22.
- Chang, S. K. and Schonfeld, P. M. (1993). Optimal dimensions of bus service zones. *Journal of Transportation Engineering*, **119**, 567–585.
- Chang, S. K. and Yu, W. J. (1996). Comparison of subsidized fixed- and flexible-route bus system. *Transportation Research Record*, **1557**, 15–20.
- Chien, S. and Schonfeld, P. M. (1998). Joint optimization of a rail transit line and its feeder bus system. *Journal of Advanced Transportation*, **32**, 253–282.
- Chien, S. (2000). Optimal feeder bus routes on irregular street networks. *Journal of Advanced Transportation*, **34**, 213–248.
- Chien, S., Yang, Z. and Hou, E. (2001a). Genetic algorithm approach for transit route planning and design. *Journal of Transportation Engineering*, **127**, 200–207.

- Chien, S. I., Spasovic, L. N., Elefsiniotis, S. S. and Chhonkar, R. S. (2001b). Evaluation of feeder bus systems with probabilistic time-varying demands and nonadditive time costs. *Transportation Research Record*, **1760**, 47–55.
- Eiselt, H.A., Gendreau, M. and Laporte, G. (1995). Link routing problems, part I: The Chinese postman problem. *Operations Research*, **43**, 231–242.
- Fu, L. (1999). Improving paratransit scheduling by accounting for dynamic and stochastic variations in travel time. *Transportation Research Record*, **1666**, 74–81.
- Fu, L. (2001). Simulation model for evaluating intelligent paratransit systems. *Transportation Research Record*, **1760**, 93–99.
- Fu, L. and Xu, Y. (2001). Potential effects of automatic vehicle location and computer-aided dispatch technology on paratransit performance. *Transportation Research Record*, **1760**, 107–113.
- Ioachim, I., Derosiers, J., Dumas, Y., Solomon, M. and Villeneuve, D. (1995). A request clustering algorithm for door-to-door handicapped transportation. *Transportation Science*, **29**, 35–139.
- Jerby, S. and Ceder, A. (2007). Optimal routing design for shuttle bus service. *Transportation Research Record* (forthcoming).
- Kuah, G. K. and Perl, J. (1987). A methodology for feeder-bus network design. *Transportation Research Record*, **1120**, 40–51.
- Kuah, G. K. and Perl, J. (1988). Optimization of feeder bus routes and bus-stop spacing. *Journal of Transportation Engineering*, **114**, 341–354.
- Melucelli, F., Nonato, M., Crainic, T. G. and Guertin, F. (2001). Adaptive memory programming for a class of demand responsive transit systems. In *Computer-aided Scheduling of Public Transport*. Lectures Notes in Economics and Mathematical Systems, **505** (S. Voss and J. R. Danuna, eds), pp. 253–273, Springer-Verlag.
- Salzborn, F. J. M. (1972). Optimum bus scheduling. *Transportation Science*, **6**, 137–148.
- Shrivastava, P. and Dhingra, S. L. (2001). Development of feeder routes for suburban railway stations using a heuristic approach. *Journal of Transportation Engineering*, **127**, 334–341.
- Shrivastava, P. and Dhingra, S. L. (2002). Development of coordinated schedules using genetic algorithms. *Journal of Transportation Engineering*, **128**, 89–96.
- Wilson, N. H. M., Sussman, J. M., Higonnet, B. T. and Goodman, L. A. (1970). Simulation of a computer-aided routing system (CARS). *Highway Research Record*, **318**, 66–76.
- Wirasinghe, S. (1980). Nearly-optimal parameters for a rail-feeder bus system on a rectangular grid. *Transportation Research*, **14A**, 33–40.
- Yim, Y. B. and Ceder, A. (2006). Smart feeder/shuttle bus service: Consumer research and design. *Journal of Public Transportation*, **9**, 97–121.

17 Service Reliability and Control



Chapter 17 Service Reliability and Control

Chapter outline

- 17.1 Introduction
 - 17.2 Measures of reliability and sources of unreliable service
 - 17.3 Modelling of reliability variables
 - 17.4 Passenger waiting time at a stop
 - 17.5 Advanced reliability-based data and control
 - 17.6 Techniques to resolve reliability problems
 - 17.7 Literature review and further reading
 - Exercises
 - References
-

Practitioner's Corner

Transit-service reliability problems are very real and have had significant impact on transit efficiency and productivity. Transit reliability can be defined as dependability in terms of time (waiting and riding), passenger load, vehicle quality, safety, amenities and information. There is never a good uncertainty or a bad certainty in transit operations. The question arises as to whether remedies exist to service reliability problems and, if so, whether they are implementable and exhaustive. This chapter will attempt to answer this question through a disaggregate perspective of individual elements causing unreliable service.

This chapter contains six main parts, following an introductory section with examples of passenger complaints as reported in the media. Section 17.2 discusses measures of reliability, sources and indicators of unreliable service, and the bus-bunching phenomenon. Section 17.3 provides the variables affecting service reliability and models for estimating travel time and dwell time. Section 17.4 describes and analyses statistical models for passenger waiting time at transit stops. Section 17.5 focuses on introducing innovations into an advanced public transit system – e.g. automatic vehicle locators and automatic passenger counters – their objective being to collect and control reliability-based data. Section 17.6 furnishes an overview of techniques to improve reliability problems; it covers the subjects of better planning, on-line control strategies, and vehicle-priority schemes. Section 17.7 reviews work done on three themes: measures of reliability, passenger waiting time, and control strategies. The chapter ends with exercises.

Practitioners may skip Section 17.4 as they read through the chapter. We argue that transit agencies usually assess measures of system performance that do not reflect what individual passengers perceive, thus warranting measurements at the disaggregate level in order to capture passengers' perceptions. We make the following suggestion to passengers; the story is illustrative:

Never let a bus driver know you're in a hurry; it will work against you. This brings us to the story. A priest and a bus driver arrive in heaven. The angel Gabriel approaches them and asks them to follow him so he can show them their new houses. They come to

a nice neighbourhood with villas. Gabriel pulls out a key and gives it to the bus driver: “This is your home”, the angel says, pointing to one of the villas. Gabriel continues to walk with the priest and, after a while, they arrive at a poor neighbourhood, with small houses that are almost tottering. The angel pulls out a key and gives it to the priest: “This is your home”, he says, pointing to one of the small houses. The priest, looking completely puzzled, asks: “How come that is for me? After serving God all my life, you give me such a house and you give the bus driver a villa?” Gabriel replies: “Because when you’re praying, everyone falls asleep, when he is driving – everyone prays”.

17.1 Introduction

Service reliability in transit operations has been receiving increasing attention as agencies become faced more and more with the immediate problem of providing credible service while attempting to reduce operating costs. Unreliable service has been cited as the major deterrent to existing and potential passengers. For example, Balcombe *et al.* (2004), in a UK practical-transit guide, report that passengers’ perception of local bus services is interpreted in the following ranking of importance in weights (given here in parenthesis) that add up to 100: Reliability (34), Frequency (17), Vehicles (14), Driver behaviour (12), Routes (11), Fares (7) and Information (5). In other words, for instance, it is twice as important from a passenger’s perspective to improve reliability as to increase the frequency of service.

There are many humoristic sayings about transit reliability that represent part of the passengers’ accumulated frustration on this subject. Some of these sayings: (a) Passengers waiting long give a look you could have poured on a waffle; (b) Bus and men are alike – both aren’t there when you need them; (c) To go nowhere, follow the crowd or stay on the bus (stuck in a traffic jam); (4) Ad on a bus: “Love is like a bus – if you missed one, another will come along”, to which one can add: like in love, you don’t know when it will come and whether it will have room for you; (d) When the bus you are on is late, the bus you want to transfer to is on time (one of Murphy’s laws).

Along the line of comprehending the magnitude of service-reliability problems in regard to transit demand, it is also interesting to read passengers’ complaints. The following are typical and have appeared in the media; they remind us what Alfred North Whitehead said: “We think in generalities, we live in details”.

Misleading frequency claim

From the *New Zealand Herald*, 16 May 2006, in an article by Claire Trevett (‘Link bus adverts “misleading”’).

Stagecoach’s claim that its Link bus service runs every 10 minutes has been ruled misleading after a passenger complained of sometimes waiting for up to 40 minutes . . .

Of the claim that the buses ran ‘every 10 minutes,’ the board said: “There was a passenger expectation that a bus would be available every 10 minutes on the route and, as that had not been the case in the consumer’s experience, the Complaints Board ruled that the website advertisement was misleading. . .”.

Jean McGeorge, a law student in her 30s: “They are really erratic. You’ll get two buses at the same time and then nothing for 20 minutes and that’s incredibly frustrating”.

Missed connections

From the *San Diego Union-Tribune*, 2 March 2006, in a Letter to the Editor (“Transit schedules are a nightmare”).

In addition to the common complaints about public transit not being on time and going where people want to go, the biggest problem I have had is in making connections.

Example 1: I am in class at Cajamarca College until 8:50 p.m., but the No. 858 bus departs at 8:44 p.m., leaving me to wait until 10:02 p.m. As a result, I don’t get home until 11:30 p.m., then have to leave again at 6:30 a.m. to be at my office in time for work.

Example 2: I often have meetings in University Heights and transfer from the Green Line trolley to the No. 11 bus at San Diego State University. The No. 11 leaves at about the same minute that the trolley arrives, with no time to get upstairs to catch it.

Example 3: I can’t begin to count how many times I’ve seen the No. 855 bus leaving Grossmont Transit Center right as the trolley arrives. The point is that the schedules don’t seem to take each other into consideration. And just for fun, try to get from Santee Town Center to Scripps La Jolla.

SCOTT WESELIS, *Santee*

Lack of amenities

From *BBC.co.uk* on 28 October 2005, by a participant Gerry Chatham.

Train announcements are rare. Customer services are a joke. Trains always packed – lucky to get a seat on some of the newer trains and if you do – you’re at risk of catching DVT because of lack of leg room.

Three out of my five morning trains this week were either delayed or cancelled. Some (not all) staff are rude/unhelpful; lack of staff at train stations. Inability to purchase a ticket on a Monday morning as ticket offices are now like ghost towns. No visible security presence at stations late at night. Dirty stations; expensive fares. . . .

It sounds like I am being very negative – I have good reason. Kent is a lovely place to live. It would be excellent if *we, the customers*, were offered value for money, clean trains that run on time, and overall a train company that puts the customers first on their list of priorities. Just look at the comments left by everyone else here, then make up your own mind.

Bus bunching

From the book by Larson and Odoni (1981), taken from a Sunday edition of the *Washington Post*.

I have long been trying to discover why Metro buses on Wisconsin Avenue are regularly bunched during rush hours instead of being evenly spaced. It is a common experience to wait for 10 or 15 minutes in rain or snow, and then find three or four buses coming along nose to tail.

As Metro refuses to reply to letters on this subject, I can only assume that it has something to hide.

From the passenger’s point of view, the advantages of even spacing are obvious – shorter waits and buses that are evenly filled, instead of being packed up at the front and half-empty behind. Traffic congestion is also eased.

Evidently, either the company or its drivers must prefer bunching. I continue to wonder why.

E. PETER WRIGHT

The above citations cover four service-reliability problems from a broader list: misleading timetables, missed connections, lack of amenities and bus bunching. In a general sense, a service-reliability problem can be defined as degradation to system performance because of: (1) uncertainties in the operating environment, (2) lack of suitable data for efficient operations planning, (3) improper service design, (4) improper service monitoring and control, and (5) failures in executing designed schedules. This chapter attempts to highlight the main ingredients of transit-service reliability and provide some insight into possible remedies to alleviate problems.

17.2 Measures of reliability and sources of unreliable service

Prior to analysing the elements of transit reliability, we ought to view acceptable measures of both reliability and perceived reliability and the sources of unreliable service. Abkowitz *et al.* (1978) defined service reliability as the invariability of service attributes that influence the decisions of travellers and transit providers. This section, too, focuses its examination on the variability of attributes that are of concern to passengers and agencies.

17.2.1 Attributes and measures of reliability

Transit-related attributes that vary by time or space may be distributed. Therefore, the (statistical) characteristics of the distributions form the base for constructing measures of reliability. Measures such as mean (average) value and variance, coefficient of variation (ratio of standard deviation to mean), and percentage of observations for a value greater than the mean can represent the compactness and skewness measures of an attribute's distribution. Following are three lists of attributes, from the passenger's and agency's perspectives and exogenous attributes, all of which tend to vary by time of day, day of week, week of season and space.

Reliability attributes of concern to passengers

- Waiting time
- Boarding time
- Seat availability
- In-vehicle time
- Alighting time
- Total travel time
- Transfer time
- Missed connections
- Pre-trip information time
- Pre-trip time required for changes in access path

Reliability attributes of concern to the agency

- Dispatching according to schedule adherence
- On-route schedule adherence

- Headway distribution
- Individual-vehicle headway
- Load-counts distribution
- Individual-vehicle load count
- On-time pullout
- Missed trips
- Breakdowns
- Late (crew) report (arrival)
- Driver proficiency
- Dispatcher and street-inspector proficiency

Reliability exogenous attributes

- Traffic delays
- Road and other accidents
- Weather

Figures 17.1 and 17.2 illustrate the main aforementioned attributes, the exogenous attributes being associated with the agency. Figure 17.1 shows the process in which the attributes of concern to passengers take place; if each attribute's probability of complying with passengers' expectation is high, then the service is perceived as reliable. Similarly, if the probability that a given run will comply with the agency's expectation is high for each attribute in Figure 17.2, then the agency may perceive the service as reliable. High probability coincides with small variance for each attribute.

Conceptually, the importance of attributes being negatively or positively perceived by individual passengers depends on their preferences. For instance, frequent users will place smaller value on reliable pre-trip information time than will infrequent users. In addition, there are quality-based attributes that are difficult to quantify. These attributes especially concern passengers' satisfaction. Figure 17.1 presents ride comfort as one of these attributes; there are also expectations in regard to the existence and proper functioning of on- and off-vehicle amenities.

Concerning reliability-related standard measures, Figure 1.4 in Chapter 1 contains five such standard items, their common criteria, and corresponding remarks as follows. These five standards are only examples, and more can be established from the lists of attributes; their criteria may vary by time, space and agency.

Another theme affecting the reliability of system performance is maintenance reliability. Examples of maintenance-based reliability attributes:

- Missed trips
- Kilometres and hours of service lost because of road calls
- Number of vehicles not available for service
- Number of late starts
- Number of vehicles overdue for inspection
- Number of operator-trouble reports
- Absenteeism

Standard item	Typical criterion range *	Remarks
Schedule adherence**	Min 80% on time (0–5 minutes behind schedule) at peak period, 90% otherwise	This guideline is usually relaxed for short headways
Timed transfer	Max of 3–8 minutes vehicle wait at transfer point	Used more in smaller agencies
Missed trips**	Min 90%–95% of scheduled trips are OK	Missed trips may also not comply with trip-reliability criterion
Passenger safety**	Max 6–10 passenger accidents per 10^6 passengers; Max 4–8 accidents per $1.6 \cdot 10^5$ vehicle-km	Depends on updated safety data
Public complaints	Limits on no. of complaints per driver/pass/time period	Public comments and complaints always received

* Reflects data mainly from the US

** Standards commonly in use

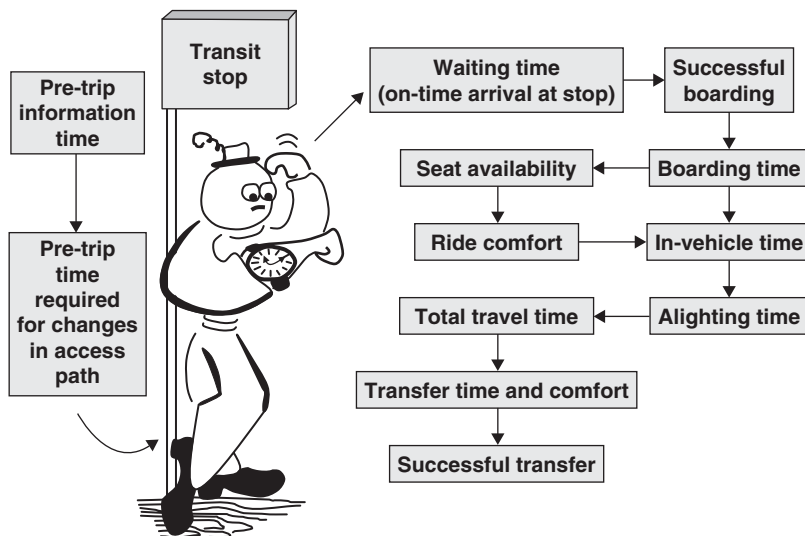


Figure 17.1 Flow diagram of reliability attributes of concern to the passenger; sources of unreliable service are characterized by high variance

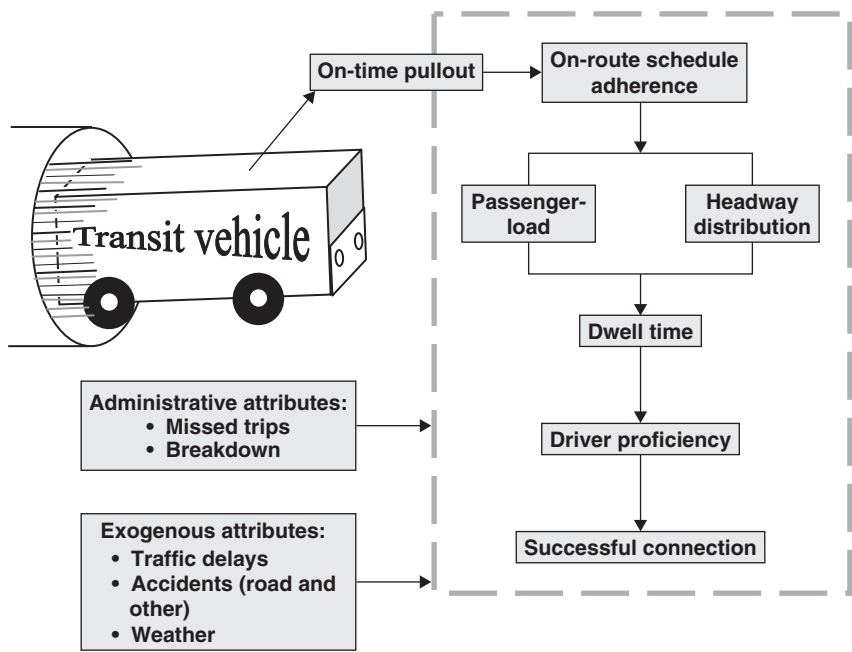


Figure 17.2 Flow diagram of reliability attributes of concern to the agency; sources of unreliable service are characterized by high variance

It is important to analyse these attributes historically and to disaggregate information by type of service and equipment.

17.2.2 Sources of unreliable service

Transit reliability problems can stem from a number of service factors. Abkowitz *et al.* (1978) divided these factors into two groups: ‘environmental’ and ‘inherent’. The first group includes such factors as traffic signals, changes in traffic conditions, variation in demand, and availability of crew and vehicles on any given route and day. The second group includes setting and distribution of waiting times, headways, travel times, boarding and alighting times, and transfer times on any given route and day.

Some causes of unreliable service are chronic and known in advance; suitable planning and adjustments can address these causes. Other causes are unpredictable in nature and require real-time responses, preferably taken from a library of options. The sources of unreliable service may be found in the following lists of indicators of developing reliability problems.

Planning indicators of developing reliability problems

The items in this list, unlike those in the other lists, do not always/necessarily create reliability problems but can serve as diagnostic indicators.

- Long urban route
- Network of routes requiring many individual (O-D) transfers
- Lack of feeder services
- Short spacing between stops
- Problematic stop location
- Single daily average running (travel) time
- Sticking, in principle, to even headways
- Different forms of fare payment
- Poor in- and off-vehicle amenities
- Poor security

Operational indicators of developing reliability problems

- Missed trips
- Late pullouts
- Poor on-time dispatching
- Significantly late/early trips
- Bunching
- Uneven loads
- Overloaded vehicles
- Unpredictable passenger demand
- Missed transfers
- High variance of scheduled headway and/or running (travel) time
- Insufficient/extended layover time
- Absenteeism
- Road calls
- Breakdowns
- Passenger complaints
- Dispatcher complaints
- Driver complaints
- Bad press
- Road and other accidents

Maintenance indicators of developing reliability problems

- Lack of vehicles available for service
- Long hours/kilometres between preventive maintenance activities
- Large number of vehicles overdue for inspection
- Large percentage of old vehicles
- Inadequate replacement policies and contingency plans
- Poor quality level of spare parts
- Intensive vehicle use
- Poor vehicle design
- Inadequacy of maintenance facilities
- Lack of data on vehicle histories and maintenance effectiveness
- Improper maintenance-engineering staff

Trouble-shooting in all the above three themes in terms of suitable recurrent analysis or even in the form of a checklist can eliminate the need for diagnoses and improve service reliability.

Finally, we will describe one of the most irritating phenomena in urban bus operations, called ‘bunching’ or ‘pairing’ (see passenger complaints in Section 17.1). Undoubtedly bus bunching is a major cause of unreliable service. Figure 17.3 describes three possible sources of the bunching phenomenon: (1) delayed vehicle, (2) early-dispatched and/or speeding vehicle, and (3) unexpected passenger overflow. Each of these sources or some combination of them can lead to the creation of two or more buses arriving nose to tail.

The base of Figure 17.3 consists of an even-headway timetable and random passenger arrivals, from which three vehicle trajectories are shown on a time–space diagram; therefore, the slope of each trajectory between two adjacent stops constitutes the average vehicle’s speed. In all three cases described, bunching occurs between the second and third vehicles at Stop 2. Part (a) of the figure illustrates the case in which traffic conditions cause the second vehicle to slow down before Stop 1. This slowing down becomes more pronounced at each successive stop, as more and more passengers accumulate after the departure of the first vehicle. Meanwhile, the third vehicle finds fewer and fewer passengers waiting (shortened headways between the second and third vehicles), and eventually it comes together with the second vehicle, in our case at Stop 2. Part (b) in Figure 17.3 presents a scenario in which the driver of the third vehicle is in a rush, departing early and speeding up. In part (c), the second vehicle confronts passenger overflow at Stop 1, which forces it to extend its dwell time at that stop. Thus, this vehicle departs Stop 1 late, as is the case in part (a). These three cases, in addition to creating bunching, result in imbalanced loading on the second and third vehicles.

17.3 Modelling of reliability variables

Following the diagnosis of the sources and causes of the development of reliability problems, this and the next sections will present models and methods covering key aspects of service reliability. These key aspects are time-related issues that result from the operating environment. This section provides the basic notation of the variables, the common formula for calculating dwell time, and an analysis of travel time. The next section presents an analysis of passenger waiting time at the stops. Table 17.1 lists the main variables and parameters of the analysis to follow.

The prediction of travel time, thus the arrival time at stops, consists of this information: the running time R_{ik} and the dwell time D_{ik} . The variation measure of the running time can be defined as

$$r_{ik} = \frac{1}{2} \text{Var}[R_{ik} - R_{(i-1)k}] = \frac{1}{2} \{ \text{Var}[R_{ik}] + \text{Var}[R_{(i-1)k}] - \text{Cov}[R_{ik}, R_{(i-1)k}] \} \quad (17.1)$$

where:

$\text{Var}[\dots]$ and $\text{Cov}[\dots]$ designate the variance and covariance of the random variable in brackets. If R_{ik} and $R_{(i-1)k}$ are uncorrelated, r_{ik} is just the variance of the running time; r_{ik} is also the covariance between $(R_{(i+1)k} - R_{ik})$ and $(R_{ik} - R_{(i-1)k})$.

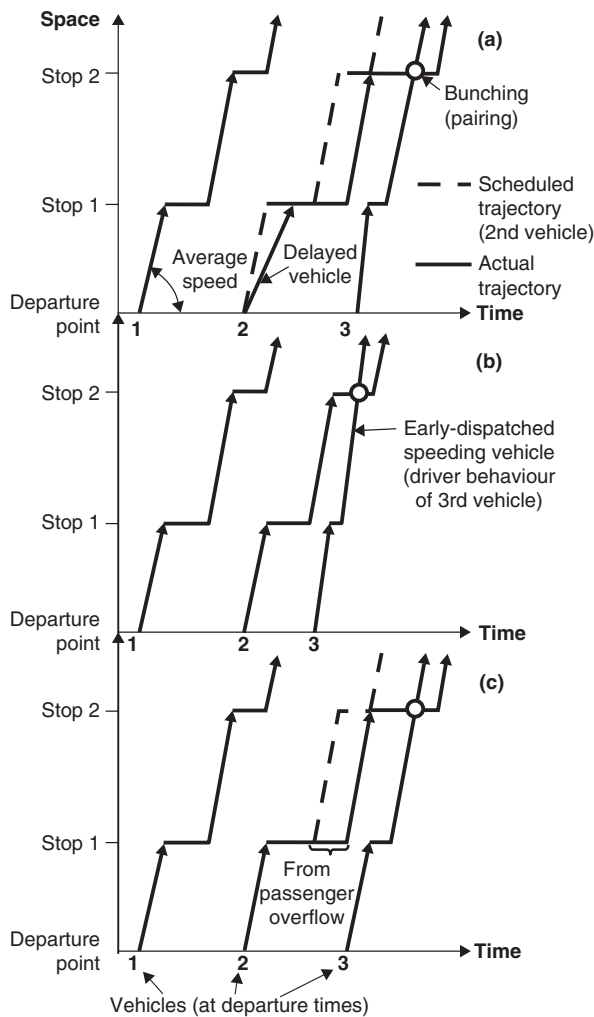


Figure 17.3 Three typical processes creating a pair of vehicles (at Stop 2), notwithstanding scheduled even headways and random passenger arrivals: part (a), with second vehicle delayed because of traffic; part (b), with third vehicle dispatched early and speeding up because of driver behaviour; and part (c), with extra demand at Stop 1 for the second vehicle

The correlation expressed in Equation (17.1), which exists in real-life, between the running times of two consecutive vehicles can be considered in analytical or simulation models.

17.3.1 Dwell time

The dwell time depends on the number of boarding and alighting passengers. Usually it can be expressed by the following linear model.

Table 17.1 Notation and definition of reliability-based variables and parameters

Notation	Definition
P_{ik}	Passenger load on-board trip i , upon departure from stop k
λ_{ik}	Average passenger arrival rate for trip i , at stop k
H_{ik}	Headway between trips i and $i-1$, upon departure from stop k
R_{ik}	Running (travel) time of the vehicle serving trip i between departure time from stop $k-1$ and arrival time at stop k
r_{ik}	Variance-based variation measure of R_{ik}
D_{ik}	Dwell time of the vehicle serving trip i , at stop k , including the time required for acceleration and deceleration ($D_{ik} = 0$ if it does not stop at k)
b	Dead time portion of the dwell time, including the time required for acceleration and deceleration ($b = 0$ if no stop at k)
b_{ik}	Number of passengers boarding the vehicle serving trip i , at stop k
A_{ik}	Number of passengers alighting from the vehicle serving trip i , at stop k
Δ_B	Marginal dwell time per boarding passenger
Δ_A	Marginal dwell time per alighting passenger

$$D_{ik} = \begin{cases} b + \Delta_B \cdot B_{ik} + \Delta_A \cdot A_{ik}, & \text{single-door vehicle, if } B_{ik} > 0 \text{ or } A_{ik} > 0 \\ 0, & \text{single-door vehicle, if } B_{ik} = A_{ik} = 0 \\ b + \max(\Delta_B \cdot B_{ik}, \Delta_A \cdot A_{ik}), & \text{double-door vehicle, if } B_{ik} > 0 \text{ or } A_{ik} > 0 \\ 0, & \text{double-door vehicle, if } B_{ik} = A_{ik} = 0 \end{cases} \quad (17.2)$$

where:

the dead time b is made up of the time to open and close the door(s), the time taken between sequential alighting and boarding events on single-door vehicles, the time taken by the driver to check the traffic: it includes, by convention, the penalty for stopping because of the consequent deceleration and acceleration.

Equation (17.2) assumes that boarding and alighting passenger flows are distinct for double-door vehicles. For a given number of boarding and alighting passengers, we can assume that there is no intrinsic randomness. If there is (e.g. different groups of passengers carrying different sizes of luggage), its variance can be assumed to be fixed and thus added to the running-time variation in Equation (17.1).

The following factors influence dwell time:

- Form of payment
- Entrance characteristics

- Conflicts between boarding and alighting passengers
- Available space near drivers
- Traffic-flow characteristics
- Passengers characteristics

Following are practical (rounded) values of the marginal dwell time per boarding and alighting passenger; these and the dead-time component of Equation (17.2) were extracted from a number of studies, especially in the UK (Balcombe *et al.*, 2004).

Boarding time per passenger

- $\Delta_B = 1.5$ sec – pay conductor
- $\Delta_B = 2.5$ sec – flat fare – with change
- $\Delta_B = 3.0$ sec – flat fare – with change
- $\Delta_B = 5.0$ sec – with change
- $\Delta_B = 6.5$ sec – automated fare box

Alighting times per passenger

- $\Delta_A = 1.5$ sec – without hand baggage and parcel
- $\Delta_A = 3.0$ sec – moderate hand baggage
- $\Delta_A = 5.0$ sec – considerable baggage

Dead time per stop

- $b = 2.0$ sec – pay conductor
- $b = 5.5$ sec – otherwise, but with extra space near driver
- $b = 7.0$ sec – without extra space

17.3.2 Analysis of travel time

From a reliability standpoint, one of the more crucial input elements in the planning process is vehicle travel (running) time. Often the term ‘travel time’ is broken down into the following string: walk from origin – wait for vehicle – ride in vehicle – walk to destination. Here we will apply this term only to the riding or running time, and exclude layover (recovery) time. The majority of the agencies do not build slack time into the scheduled travel time to compensate for the variability of vehicle travel time; average travel time is then utilized for all scheduling tasks. For convenience let R_i be the travel time of the vehicle serving trip i across the entire transit route. This travel time depends on the trip time (hour, day, week, season), number of passengers, and the habits of each individual driver.

Jeong and Rilett (2005) reviewed prediction models for travel time, including historical data-based models, regression models, time series models and neural network models. They selected artificial neural network models to predict R_i , given real-time information on traffic congestion and transit-vehicle dwell times at stops. The researchers found that, given the availability of real-time data, their model outperformed both the historical data-based model and the regression model in terms of prediction accuracy. The increasing use of advanced public transit systems (see Section 17.5 below) by transit agencies allows for better predictions of both R_i and R_{ik} in real time; this undoubtedly can be used for reducing

passenger uncertainties at stops and on the vehicles, and eventually for improving service reliability.

While Sections 17.5 and 17.6 focus on new technologies and techniques to improve real-time operation, the present section provides a practical, statistical-based analysis for deploying R_i at the planning level. It is based on Ceder (1981), which utilized four major procedures: (1) exclusion of outliers, (2) division of the day into intervals, (3) union of intervals, and (4) union of days for weekly and seasonal cross-sections.

Description of general method

The main objective of data processing is to create a database on vehicle travel times for different cross-sections of the day, week, and season (d-w-s). From a statistical viewpoint, the existence of a large data bank that can be systematically updated means that an *analysis of variance* (concerned with differences between means of groups) can be carried out (based on the normality distribution assumption) in order to estimate the effects of d-w-s (the independent variables) on the dependent variable, R_i . Afterwards, the independent variables that were found to affect the dependent variable (at a desired level of significance) can be analysed with the use of *contrasts* (comparison of the differences between pairs or combinations of means). Finally, it is possible to apply a multiple variable regression analysis.

In order to effectively carry out an analysis of variance on d-w-s, a data bank must be accumulated over a yearly period, on the assumption that there will be no physical changes along the transit route or in the location or in quantity of stops. Because it is well known that transit routes are liable to physical changes in a dynamic fashion, it is questionable whether the data bank for R_i values can rely on a yearly base. Moreover, it is desirable that the process that determines the R_i relationship allow for the intermediate involvement of those responsible for data collection. This involvement can be expressed by the identification of outliers, whose cause is known, and by the practical decisions relating to changes in statistical criteria, which are combined in the process, for different transit routes.

As a result, a number of criteria can be utilized for characterizing outliers; dividing the day into time used; dividing the week into homogeneous days; and dividing the year into seasons. If no physical changes occur for a given transit route over a yearly period, and a data bank for R_i values exists, there is a possibility of carrying out an analysis of variance and contrasts for d-w-s. Once again, it should be noted that the objective is not to build a statistical model for simulation or control, but to construct a value system for R_i that can be of greater aid in realistic planning than can a single mean value over a specified period of time (e.g. year, as some transit agencies do).

A possible method of analysing R_i consists of four main components: (1) exclusion of outliers (OUTLIER procedure), (2) division of day into intervals (INTERVAL procedure), (3) union of intervals (UNIT procedure), and (4) union of days for weekly and seasonal cross-sections (UWEKDAY-USEAS procedure).

The first three components relate to the data on a daily basis, and the fourth component serves as a tool for determining the division of weeks and seasons. The procedure starts with assembling data into two sets: (a) general data, which include the number of vehicles in the set, route number, origin, destination, the day (or days) of the week, and the season; and (b) specific data, which include vehicle departure time, travel time and type of vehicle.

The outliers are deleted from these data, and then a number of methods are tested for dividing the day into intervals (division for each hour, each two hours, each three hours, and one daily average). Then statistical tests are conducted to determine the possibility of unifying the intervals. The procedure continues for the series of days, and then the possibility of unifying the days is examined. The next step checks the possibility of unifying a number of days from different seasons; this is possible if the division of days in each season (or at least those chosen for unification) is similar. More details are shown in Ceder (1981).

Example of bus data set

For a given bus route (in Haifa, Israel) departing at a frequency of either every 15 or 30 minutes, data were collected for two days (on the same day of the week for two weeks) at a total of 126 data points. The data points are shown in Figure 17.4, as well as the results obtained after the INTERVAL and UNIT components. In the OUTLIER procedure, five outliers were found and are marked on the upper portion of Figure 17.4; they also appear at the start of Table 17.2. The latter exhibits a computer record.

Following a run of the INTERVAL procedure, the method designated C in Table 17.2 was chosen; that is, a division of the day into three-hour intervals, when the first hour is considered separately, as shown in the upper portion of Figure 17.4 and in Table 17.2. The record in Table 17.2 is accompanied by F values, where

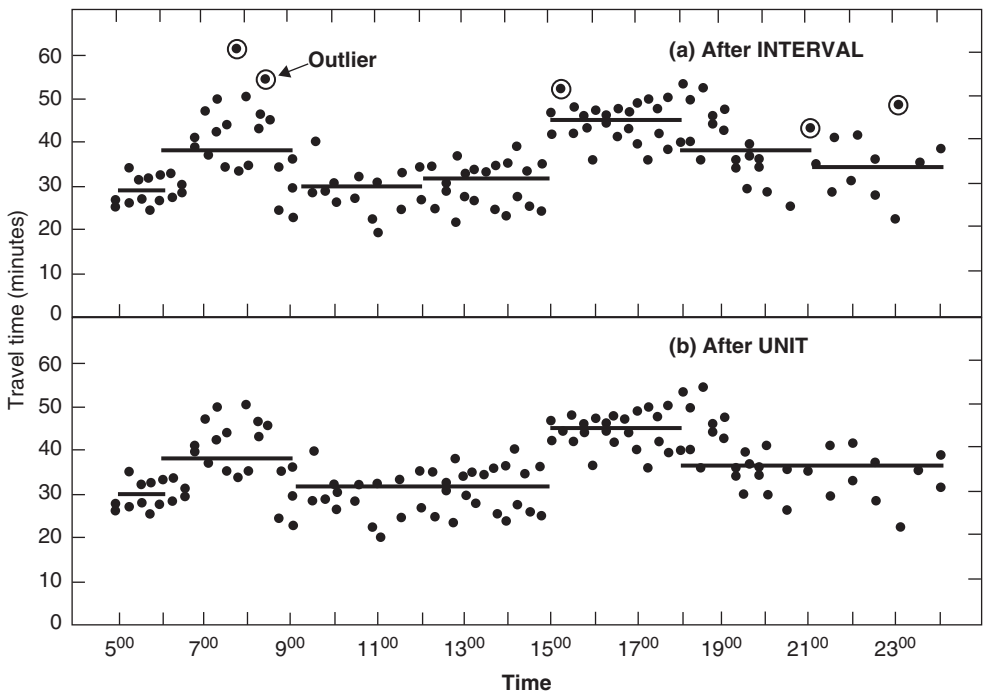


Figure 17.4 An example of travel time data points for one bus line with mean values determined by the INTERVAL procedure in part (a) and by the UNIT procedure in part (b)

Table 17.2 Example of data analysis using *OUTLIER*, *INTERVAL* and *UNIT* for a given bus route; summary of computer record

Stage	Description																								
Input	126 data points (see Figure 17.4)																								
OUTLIER	(1) departure 7:45, travel time = 62 minutes (2) departure 8:30, travel time = 55 minutes (3) departure 15:15, travel time = 53 minutes (4) departure 21:00, travel time = 44 minutes (5) departure 23:00, travel time = 49 minutes																								
INTERVAL	Method A: division by one hour Method B: division by two-hour periods Method C: division by three-hour periods Method D: single average for the whole day Steps of comparisons: (A vs. D) → (A vs. C) → (A vs. B), until significant <i>F</i> difference found at the 0.05 level (1) A vs. D: <i>F</i> -calculated = 1.92, <i>F</i> -table = 1.35; conclusion: continue (2) A vs. D: <i>F</i> -calculated = 1.15, <i>F</i> -table = 1.35; conclusion: accept																								
UNIT	Same-means hypothesis is based on <i>t</i> -test at the 0.05 level																								
	<table><tr><th>Integer interval</th><th>Actual interval</th><th>Mean value (minutes)</th><th>Standard deviation (min.)</th></tr><tr><td>5:00 < R_i ≤ 6:00</td><td>5:00 to 6:00</td><td>29.4</td><td>3.5</td></tr><tr><td>6:00 < R_i ≤ 9:00</td><td>6:15 to 9:00</td><td>38.2</td><td>8.1</td></tr><tr><td>9:00 < R_i ≤ 15:00</td><td>9:30 to 15:00</td><td>30.1</td><td>6.1</td></tr><tr><td>15:00 < R_i ≤ 18:00</td><td>15:15 to 18:00</td><td>44.8</td><td>4.5</td></tr><tr><td>18:00 < R_i ≤ 24:00</td><td>5:00 to 6:00</td><td>36.9</td><td>7.0</td></tr></table>	Integer interval	Actual interval	Mean value (minutes)	Standard deviation (min.)	5:00 < R _i ≤ 6:00	5:00 to 6:00	29.4	3.5	6:00 < R _i ≤ 9:00	6:15 to 9:00	38.2	8.1	9:00 < R _i ≤ 15:00	9:30 to 15:00	30.1	6.1	15:00 < R _i ≤ 18:00	15:15 to 18:00	44.8	4.5	18:00 < R _i ≤ 24:00	5:00 to 6:00	36.9	7.0
	Integer interval	Actual interval	Mean value (minutes)	Standard deviation (min.)																					
	5:00 < R _i ≤ 6:00	5:00 to 6:00	29.4	3.5																					
	6:00 < R _i ≤ 9:00	6:15 to 9:00	38.2	8.1																					
	9:00 < R _i ≤ 15:00	9:30 to 15:00	30.1	6.1																					
	15:00 < R _i ≤ 18:00	15:15 to 18:00	44.8	4.5																					
18:00 < R _i ≤ 24:00	5:00 to 6:00	36.9	7.0																						

$F = (\text{found variation of the group averages})/(\text{expected variation of the group averages}).$

The data then proceed to the UNIT procedure. Of the seven intervals in method C, two are united and only five intervals remain, as shown in the lower portion of Figure 17.4 and in Table 17.2. This example does not include the UWEKDAY-USEAS procedure.

The last five intervals (each with a mean and standard deviation) are transmitted to the bus company’s scheduling department. This information can be transmitted either automatically or manually. Determination of the *R_i* value for planning purposes will depend on the degree of certainty that the bus will arrive at its destination on time. Table 17.2 contains both the mean value of *R_i* and its standard deviation for two possible objectives: (1) to utilize the standard deviation for introducing slack time to account for the variability of *R_i*; and (2) to allow the scheduler more flexibility in creating/adjusting blocks, which in certain cases are already performed by a specific driver.

17.4 Passenger waiting time at a stop

This section attempts to clarify some of the probabilistic issues of passenger behaviour at transit stops. The section consists of two parts. First, the distribution of waiting times is analysed for the case of random passenger arrivals. This distribution is described by the two families of headway distributions, the deterministic and the exponential, discussed in Chapter 12, Section 12.2.2, which can approach the two extremes. Second, the mean waiting-time formulation is interpreted using observed data from two studies; in addition, two explicit expressions for the mean waiting time are derived for the case of a suburban rail station. The majority of this section follows Ceder and Marguier (1985).

17.4.1 Waiting-time distributions

In Chapter 12, Section 12.2.2, Equation (12.1) presented the basic relationships between the expected (mean) waiting time $E(w)$ and the mean $E(H)$ and variance $\text{Var}H$ of the time headway. This known equation for $E(w)$ is based on two assumptions: (a) passengers can always board the first vehicle to depart (no overloading situations), and (b) the passenger random-arrival rate at the stop is independent of the vehicle-departure process and constant over the period. Following is the derivation of Equation (12.1) on a per-route basis, in which $f(H)$ is the probability density function for headway H (from a passenger's perspective, compared to $f_H(t)$, which is from the system's perspective); $\bar{w}(\bar{H})$ = mean waiting time per passenger for $\bar{H}$, which is $\frac{H}{2}$; λ = mean arrival rate of passengers, from which λH is the number of passengers arriving in the course of H minutes.

$$\left(\begin{array}{l} \text{total number of} \\ \text{passengers arriving} \\ \text{at the transit stop} \end{array} \right) = \int_{H=0}^{\infty} \lambda \cdot H \cdot f(H) dH$$

$$\left(\begin{array}{l} \text{mean waiting time} \\ \text{for passengers} \\ \text{arriving randomly} \end{array} \right) = E(w) = \frac{\text{Total waiting time}}{\text{Total number of passengers}}$$

$$= \frac{\int_{H=0}^{\infty} \lambda \cdot H \cdot \left(\frac{H}{2} \right) \cdot f(H) dH}{\int_{H=0}^{\infty} \lambda \cdot H \cdot f(H) dH} = \frac{\frac{\lambda}{2} E(H^2)}{\lambda \cdot E(H)}$$

The use of $\text{Var}(H) = E(H^2) - E^2(H)$ results in the known mean waiting-time formula shown in Equation (12.1):

$$E(w) = \frac{E(H)}{2} \left[1 + \frac{\text{Var}(H)}{E^2(H)} \right]$$

Thus the mean passenger waiting time is minimized at half the headway when the variance of the headway is zero.

In practice, assumptions (a) and (b) above may not always hold, which would then argue in favour of even-load headways, as was analysed and proposed in Chapter 4, rather than even headways. For instance, there are surges in the passenger-arrival rate at the start and end of a factory shift and a school day.

Under the assumption of random passenger arrivals at the stop, the following quality relates the waiting-time distribution to the headway distribution: Let $f_w(t)$ be the waiting-time probability density function.

$$f_w(t) = \frac{\int_t^{\infty} f_H(u) du}{E(H)} = \frac{\bar{F}_H(t)}{E(H)} = F \cdot \bar{F}_H(t) \quad (17.2)$$

where:

F is the frequency of service (vehicle-arrival rate), $F_H(t)$ is the cumulative distribution function of the headway H (system's perspective as in Chapter 12), and $\bar{F}_H(t) = 1 - F_H(t)$. The derivation of this relationship, which is shown in Larson and Odoni (1981), is related to the phenomenon commonly referred to as *random incidence*.

Particular examples of the application of Equation (17.2) are provided by the deterministic-headway and exponential-headway cases. The following is obtained for the deterministic headway in which $H = \frac{1}{F}$:

$$\bar{F}_H(t) = \begin{cases} 1, & t \leq \frac{1}{F} \\ 0, & t \geq \frac{1}{F} \end{cases}$$

$$\text{hence, } f_w(t) = \begin{cases} \frac{1}{F}, & t \leq \frac{1}{F} \\ 0, & t \geq \frac{1}{F} \end{cases} \quad \text{and } E(w) = \frac{1}{2F} = \frac{E(H)}{2}.$$

In this case, the waiting time is uniformly distributed between 0 and the fixed headway $\frac{1}{F}$.

For the exponential-headway case, the following is obtained:

$$f_H(t) = F \cdot e^{-F \cdot t}, \quad \bar{F}_H(t) = e^{-F \cdot t}$$

hence

$$f_w(t) = F \cdot e^{-F \cdot t}$$

$$\text{and } E(w) = \frac{1}{F} = E(H).$$

In this case, the waiting time is exponentially distributed with the same parameter F on the headway.

The application of Equation (17.1) to the ‘power’ family of headway distributions can be obtained by integrating $F_H(t)$ in Equation (12.5) into Equation (17.2):

$$f_w(t) = \begin{cases} F \left(1 - \frac{1 - C^2}{1 + C^2} \cdot F \cdot t \right)^{\frac{2C^2}{1 - C^2}}, & 0 \leq t \leq \frac{1 + C^2}{1 - C^2} \cdot \frac{1}{F} \\ 0, & t \geq \frac{1 + C^2}{1 - C^2} \cdot \frac{1}{F} \end{cases} \quad (17.3)$$

For the second family (gamma distributed) of headway distributions, Equation (12.6) can be integrated in a close form for values of C^2 , such that $\frac{1}{C^2}$ is an integer. In these cases, the gamma distribution is an Erlang distribution.

Nonetheless, the results for these cases also give a good sense of what the results will be for other intermediate values of C^2 . Integrating $f_H(t)$ into Equation (12.6) and applying Equation (17.2) yields:

$$\begin{aligned} f_w(t) &= F \cdot \frac{\left(\frac{F}{C^2} \right)^{\frac{1}{C^2}}}{\Gamma\left(\frac{1}{C^2}\right)} \sum_{k=0}^{\frac{1-C^2}{C^2}} \frac{\left(\frac{F \cdot t}{C^2} \right)^k}{k!} \cdot e^{-\frac{F \cdot t}{C^2}} \cdot \frac{1 - C^2}{C^2} \cdot \frac{1}{\left(\frac{F}{C^2} \right)^{\frac{1}{C^2}}} = \\ &= F \cdot \sum_{k=0}^{\frac{1-C^2}{C^2}} \frac{\left(\frac{F \cdot t}{C^2} \right)^k}{k!} \cdot e^{-\frac{F \cdot t}{C^2}}, \quad t \geq 0 \end{aligned} \quad (17.4)$$

The two families of waiting-time distributions obtained are illustrated in Figure 17.5 in two parts. Part (a) shows the time distributions derived from the headway ‘power’ distributions for a mean headway of ten minutes and for service-reliability measures C^2 of 0, $\frac{1}{4}$, $\frac{1}{3}$, $\frac{1}{2}$, and 1. Part (b) shows the waiting-time distributions derived from the headway gamma distributions for the same values of $E(H)$ and C^2 as in part (a). For the extreme situations, $C^2 = 0$ and $C^2 = 1$, both families give identical curves because these values must correspond, respectively, to the deterministic headway and exponential headway cases. For $C^2 = 0$, the uniformity of the waiting-time distribution (between 0 and 10) can be noticed. For $C^2 = 1$, the curve is exponential. For both families, the waiting-time probability densities are always decreasing. This is a general property, implied by Equation (17.2).

17.4.2 Average waiting time

Consider a transit route in which the vehicle-departure times follow the timetable and the traffic flow is not congested. In such situations, which are characterized by relatively low

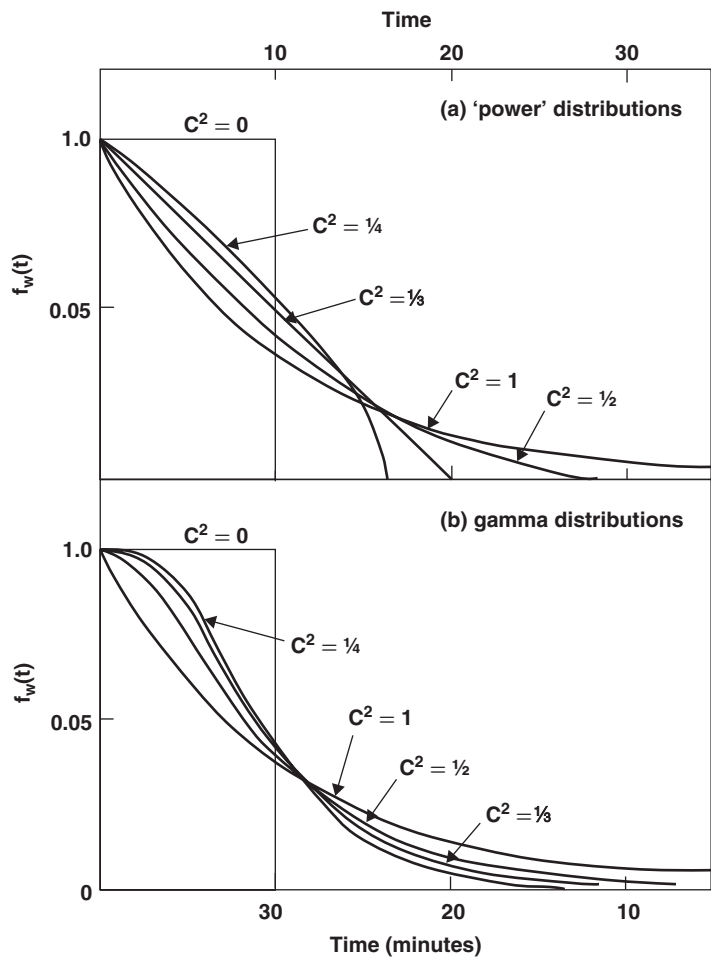


Figure 17.5 Various waiting-time distributions showing (a) the ‘power’-distributed headways and (b) the gamma-distributed headways for a mean headway of ten minutes and different values of C^2

(between days) variability, some passengers will attempt to follow the arrival pattern of vehicles in order to reduce their waiting time.

Jolliffe and Hutchinson (1975) observed buses in suburban London with published timetables and examined between-day variations. They found no correlations between the between-day variability of bus-departure times and headways. When they considered the (between-day) expected (mean, average) waiting time for passengers arriving at the bus stop at time t , they found that the waiting-time function had a minimum wait, $w_{\min}$ before the average departure time of each bus, the corresponding t (to $w_{\min}$) being the optimal arrival time. The authors showed how, based on departure-time variability, it is possible to compute $w_{\min}$ theoretically. The observed average waiting time $\bar{w}$ was found to be higher than $w_{\min}$ and lower than the expected value $E(w)$ given by Equation (12.1), indicating that

some passengers do time their arrival. It was also concluded that some passengers never wait – if they see a bus coming, they rush to board it. Jolliffe and Hutchinson (1975) proposed to represent the passenger-arrival pattern by considering three proportions: (1) passengers who arrive coincidentally with a bus (see and rush); (2) a proportion p for passengers who, based on the timetable and experience, arrive at the optimal time and wait only for $w_{\min}$; and (3) passengers who arrive at random and wait on average for $E(w)$.

Jolliffe and Hutchinson found the proportion of coincidental arrivals to be relatively low (16%), similar to what other studies have observed (Ceder and Marguier, 1985). The mean waiting time for non-coincidental arrivals could then be expressed as follows:

$$\bar{w} = p \cdot w_{\min} + (1 - p) \cdot E(w) \quad (17.5)$$

Jolliffe and Hutchinson used $w_{\min}$ to determine the proportion p , and proposed this relationship:

$$p = 1 - e^{-\alpha g} \quad (17.6)$$

where $g = E(w) - w_{\min}$, and α is a constant (found by them to be 0.131 for the peak and 0.015 for the off-peak period data).

Let us consider a suburban rail station. Assume that the service is very reliable, say $C^2 = 0$, with posted timetables, and that passengers cannot see coming trains in advance. Using the above approach, which should be appropriate in this case, and inserting Equation (17.5) into Equation (17.6), one obtains:

$$\bar{w} = E(w) - g(1 - e^{-\alpha g}) \quad (17.7)$$

For this deterministic headway, the expected waiting time (of randomly arriving passengers) is $E(w) = E(H)/2$; and Equation (17.6) becomes:

$$\bar{w} = \frac{E(H)}{2} - g(1 - e^{-\alpha g}) \quad (17.8)$$

Huddart (1973) in London observed, as expected, that for transit services with small headways, passengers arrive at random. However, if the service is sufficiently infrequent, passengers may take advantage of this information to arrive at the boarding point just before the vehicle is due to arrive. Huddart considers this situation to be rare for urban bus services, but fairly common at a suburban railway station.

If we consider the passenger-arrival rate $\lambda(t)$ as monotonically and continuously varying with time, as shown by the typical passenger-arrival pattern given by Huddart (1973), the mean waiting time can be expressed as:

$$\bar{w} = \frac{\int_{-E(H)}^0 -t \cdot \lambda(t) dt}{\int_{-E(H)}^0 \lambda(t) dt} \quad (17.9)$$

For example, the data in Huddart (1973) can be fitted into a regression analysis for an exponential function of the type $\lambda(t) = a \cdot e^{bt}$, where a, b are fitted parameters. For different passenger-arrival rates between three and 15 minutes before a train departure, the following model was calibrated: $\lambda(t) = 52.5e^{0.3}$ with $r^2 = 0.95$ (measure of regression analysis, indicating a good fit). Thus,

$$\begin{aligned}\bar{w} &= \frac{1}{b} - \frac{E(H) \cdot e^{-bE(H)}}{1 - b \cdot e^{-bE(H)}} \\ &= \frac{E(H)}{2} - \frac{b \cdot E(H)^2}{12} + \frac{b^3 \cdot E(H)^4}{720} + \dots\end{aligned}\quad (17.10)$$

The results of the two approaches are exhibited by Equations (17.9) and (17.10). These two expressions do not have quite the same form, but both verify $\bar{w} = E(H)/2$ when passengers do not attempt to time their arrivals to reduce their waiting time; i.e. for $g = b = 0$.

17.5 Advanced reliability-based data and control

One of the known potential remedies for passenger complaints concerning reliability problems is the use of advanced public transit/transportation systems (APTS). APTS offer new technologies to improve the mobility, convenience and safety of transit passengers, and to increase passenger demand. While it is still a challenge to use APTS efficiently in real time, an analysis of their data reveals weaknesses in regard to reliability. These weaknesses can be repaired by the transit personnel (management, planners, schedulers, dispatchers, mobile supervisors and inspectors). This section describes the main APTS features and their required data.

The US Department of Transportation (2003) issued a report on APTS based on a deployment tracking survey that was conducted over the Internet for 2002. This survey covered 593 transit agencies in the US, representing about half of the existing agencies; they were asked about existing and planned APTS. In addition, the US Federal Transit Administration issued a report by Hwang *et al.* (2006) on the state of the art of APTS, which emphasized the need for integrating APTS with emerging intelligent transportation systems (ITS) and ITS trends. Hwang *et al.* (2006), from lessons learned, stressed the need for the provision of better data, standards and voice communication in order to: (1) improve transit planning, maintenance, operations and incident management, and (2) facilitate coordination, integration, and interoperability with transportation providers and public safety organizations.

The main features contained in the US survey and reports, some of which appear in Khattak and Hickman (1998), are assembled according to system category in Table 17.3. Common APTS terms used in this table are AVL (automatic vehicle location), APC (automatic passenger counter), CAD (computer-aided dispatch), AFP (automatic fare payment), and ATIS (advanced traveller information system). Other known terms appearing in the literature, but not used here, are AVM (automatic vehicle monitoring), AVL/C (automatic vehicle location and communication), and a different CAD (computer-assisted design).

Table 17.3 *Advanced public transit systems (APTS) applications*

APTS applications	System name	Features
Monitoring systems	AVL	Automatic position determination by means of dead-reckoning (using the vehicle's odometer and compass), GPS, signposts (transmitted signals picked up by vehicle), ground-based radio, and real-time reporting
	APC	Automatic counts of boarding and alighting passengers (e.g. use of treadle mats or infrared beams placed by the door)
	Advanced communication	Digital radio (binary information) and/or trunked radio (computer selection of frequency)
	CAD	Dispatch software as a support tool
	Automated operations of software	Software that displays vehicle positions, vehicle and agency data, and communications information
	Silent alarm	Emergency signal triggered by the driver; possible hidden microphone for dispatcher/other listeners, on-board surveillance camera
Fare payment systems	Vehicle component monitoring	Automatic remote measurement of engine oil pressure, engine temperature, electrical system, tire pressure, etc.
	AFP	Payment by smart card, magnetic stripe card, credit/debit card, etc.
Traveller information systems	Multi-carrier fare integration	Fare scheme that covers more than one transit service provider
	ATIS	Pre-trip, in-terminal, and in-vehicle real-time passenger information
Multi-purpose information systems	Multi-modal traveller information	Available information covering multiple modes (i.e. transit and traffic or different transit modes)
	Vehicle probe	Automatic data from transit vehicles for estimating traffic travel times and speeds, and flow conditions
	Mobile data terminal	Wireless device that can send and receive information
	Information sharing	Sharing of information on traffic and incidents among agencies

(Continued)

Table 17.3 *Advanced public transit systems (APTS) applications (Continued)*

APTS applications	System name	Features
Multi-purpose information systems <i>(continued)</i>	Mobility manager	Coordination of travel requests and vehicle dispatching (multiple agencies)
	Manual priority	Extended green – activated by driver
Traffic signal control	Automatic priority	Automatic extension of green phase

Current practices in transit agencies show that sufficient data seldom exist for both service operations planning and improving reliability problems. Manual data collection is costly and, consequently, must be used sparingly. As a result, detailed information on passenger demand and service characteristics is generally not available at the route level. Without this information, the efficient deployment of transit service commensurate with demand is impossible. Thus, a major reason for transit agencies to be interested in the use of APTS data is the hope of gaining badly needed information at a greatly reduced unit cost. The resulting data is expected to improve utilization of vehicles and crew, as well as to resolve some reliability problems.

Useful, automatically collected data are extracted from AVL, APC, CAD and AFP systems. Their features appear in Table 17.3. These data sources have the potential to enhance schedule efficiency and to improve service reliability.

A general program for developing operations planning applications is depicted in Figure 17.6. The program as a whole intends to serve transit routes with or without AVL/APC/CAD/AFP data. Its major objectives are: (1) to improve management and operations by developing, improving, validating, and testing models and procedures for transit operations planning; (2) to improve levels of service through increased reliability resulting from better control and response; (3) to improve productivity and efficiency by better matching supply and demand; (4) to reduce data gathering, processing and reporting costs; and (5) to develop vital components for a management information system pertaining to operations and passenger behaviour.

The applications for AVL/APC/CAD/AFP data in Figure 17.6 are rooted in two streams of data flows: passenger-load data and vehicle running-time data. Some of the methods of handling the running-time data are discussed in Section 17.3. The use of load data and its appraisal appeared in Chapters 3 and 4. Moreover, AVL/APC/CAD/AFP data can improve route design, schedule adjustments, and vehicle scheduling. The outcome of best exploitation of the data is better operational strategies, which are required to reduce reliability problems.

In an implementation study utilizing AVL and APC systems, Kimpel *et al.* (2006) utilized operations data to improve schedules and reduce service reliability problems. The data used were recovered by the TriMet transit agency (Portland, Oregon, US), using an

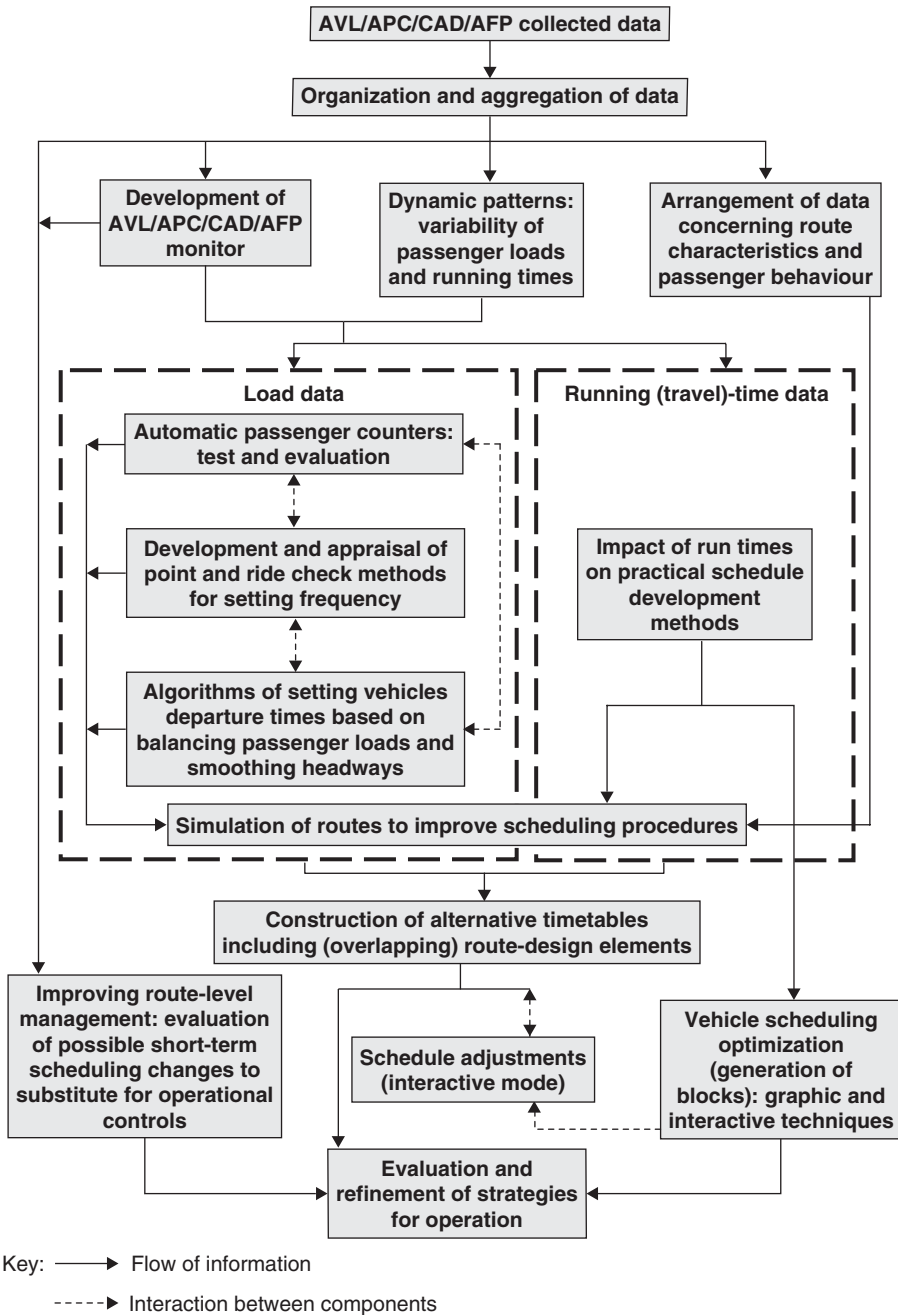


Figure 17.6 Program to develop applications for AVL/APC/CAD/AFP data

automated CAD system. At TriMet, 100% of the bus fleet is equipped with AVL technology, while approximately 72% of the vehicles are equipped with APC. TriMet obtains bus-location information at regular time intervals and transmits it to two dispatch centres in real time. The data-collection component of the AVL and APC systems include arrival and departure times, dwell times, door openings, lift operations and maximum speed since the previous stop, as well as the number of boarding and alighting passengers on APC-equipped vehicles.

The findings of Kimpel *et al.* (2006) show that the on-time performance measure (percentage arriving on time at stops) for fifteen frequent-service bus routes rose from 73%–74% to 75.5%, which is slightly above the agency standard of 75%. In addition, they claim that TriMet's scheduling and operations management has benefited from the analysis of the AVL and APC data, especially in reports generating capabilities. The study concluded by recommending that efforts to improve schedule efficiency should centre on reducing excess runs and layover times, as well as employing supervisory actions to reduce operator variability.

17.6 Techniques to resolve reliability problems

Essentially there are three basic mechanisms for improving transit reliability: (1) improve planning and scheduling, (2) improve priority techniques for transit vehicles, and (3) improve operations control. Abkowitz *et al.* (1978) classified reliability-improvement strategies as preventive and corrective/restorative. Preventive strategies are fundamentally associated with mechanisms (1) and (2); corrective/restorative strategies, with mechanism (3). TranSystems *et al.* (2006), in a study on attracting and retaining transit passengers, suggested that these mechanisms should address one or more of the following parameters: travel time, convenience, comfort, perceived personal security/safety and perceived 'image' of the transit system. What follows is a description of tactics, strategies and possible actions in conjunction with each of the three mechanisms, and an extended example of a holding strategy.

17.6.1 Improved planning and scheduling

Basically all the methods presented in this book are intended to improve transit planning and scheduling; inherently better reliability measures are incorporated into the resultant improvement. Examples of specific actions, some of which can be found in TranSystems *et al.* (2006), are presented in the following lists, by category.

Area coverage actions

- Increased route coverage; new routes
- Old set of routes replaced by a new set
- Service expansion; new local circulators/shuttles
- New/improved feeder services
- New/improved timed transfers; improved route coordination
- New/improved transit centres

Route restructuring actions

- Interlinings introduced
- Route shortening, extension, realignment, and removal
- Revised operating strategies
- Route splitting
- Joining routes to form a single route
- New/improved zonal, express, and local services
- Reduced number of stops

Scheduling amendment actions

- Increased span of service; longer late night and weekend service hours
- Increased service frequency
- Average even-load headways introduced
- Adjusted departure times to suit new connections
- Modified running (travel) times by time-of-day
- Increased layover time
- Deadheading trips introduced for necessary reinforcement
- Shifts in departure times introduced as a corrective strategy

General system's organization actions

- New/improved geographic, O-D, route, stop and transfer-point database
- New/improved frequency, timetable and transfer-time database
- New/improved stop, station, transit centre and park-and-ride amenities
- New/improved passenger amenities
- New/improved real-time service information
- New/improved vehicles
- Increased security and safety

These types of actions are typically implemented in an effort to seek a better, more reliable transit service.

TranSystems *et al.* (2006) described examples of successful actions taken in the US. Here are two typical examples: (1) Washington Metropolitan Area Transit Authority extended late-night weekend hours on Metrorail from 02:00 to 03:00 (a.m.) on Saturday and Sunday mornings. The trial programme, which ran for 18 months, was projected to attract an additional 3,000 passengers; actually performance exceeded these expectations by 20%, and it was decided to make the extension permanent. (2) Bangor (Maine) Area Comprehensive Transportation System introduced a new fleet of low-floor buses for quicker, easier boarding and alighting; this action resulted in increased passenger demand, which rose by 8% between 2003 and 2004. It is customary in the US to have an approximately 200-metre distance between bus stops, compared with about 320 metres in Europe. In Europe, moreover, one can usually find the name of the stop and updated timetable information at each bus stop.

17.6.2 Improved priority techniques for transit vehicles

Transit-vehicle priority techniques are extensive, and many bus-priority strategies have been demonstrated worldwide. Traditionally, priority is granted for transit-vehicle operation at stops, at intersections and by preferential/exclusive lanes. Usually there is a trade-off between granting priority to buses and improving traffic flow for the other vehicles. Local authorities (especially those in elected positions) are often reluctant to provide this priority. In other words, with a hint of a humour, in order to convert a street lane into strictly a bus lane (thereby increasing bus reliability and reducing traffic congestion, by helping to switch from automobile to buses), one must first prove that there is no traffic congestion on that street; this is analogous to the common experience that in order to secure a loan, one must first show that he/she doesn't need it.

Priority at stops

Transit-vehicle stop locations at intersections should normally be designed to prefer the far-side location (after crossing the intersection). In this way, the bus will not block traffic intending to turn; it will have easier pull-in and pull-out manoeuvres, and will experience fewer conflicts with pedestrians. However, a near-side stop is more appropriate at locations where transit vehicles make a turn and where the crossing street is one-way. Other useful strategies are to prohibit parking near stops, to grant pull-out priority (reducing the merging time with the traffic), and to extend the pavement at the stop location (eliminating a pull-out manoeuvre and easing the boarding process).

Priority at intersections

At intersections, transit priority can be divided into *passive* and *active* schemes. Traffic engineers usually employ passive priority at intersections in order to utilize four measures: (1) exempt transit vehicles from turning prohibitions so as to facilitate transit routes, (2) extend the green interval at signalled intersections for non-stopping transit vehicles, (3) divide the green interval into two parts within the same cycle, and (4) provide preference to streets carrying transit routes through YIELD and STOP signs. Active priority permits transit vehicles to pre-empt traffic signals, using in general one or a combination of the three following procedures: (1) immediate priority upon the arrival of the transit vehicle, (2) priority dependent on the crossing-street traffic queue, and (3) priority granted only to transit vehicles with late arrivals.

Priority through preferential/exclusive lanes

Preferential treatment to transit vehicles on street lanes can be categorized according to these three lists:

Type of preferential lane

- Exclusive curb lane
- Semi-exclusive curb lane (shared only with cars about to turn)
- Exclusive median lane (with stop island)

- Exclusive lane in the centre of a street
- Transit vehicle malls (known as bus malls; limited to pedestrians and buses)
- Exclusive freeway/highway lanes
- Ramp bypass (for entering a freeway/highway during traffic congestion)
- Congestion bypass (exclusive lanes to bypass traffic bottlenecks)

Integration with traffic flow

- With-flow lane (by pavement markings and signs; enforcement problems)
- Contra-flow lane (easy enforcement)
- Exclusive lanes

Period of operation

- Single-peak operation
- Two-peak operation
- Permanent operation

Some exclusive lanes for transit vehicles are shared with high-occupancy vehicles (taxis, or a certain minimum number of people in a car, for encouraging carpools).

Another known type of priority involves BRT (bus rapid transit): a flexible, rubber-tire form of rapid transit that combines stations, vehicles, services, priority lanes and intelligent-system elements into an integrated system with a unique identity. TranSystems *et al.* (2006) described BRT as a system of buses with such features as signal priority, dedicated right-of-way, automated and off-vehicle fare collection, automated information systems, level boarding, modern vehicles, bus shelters with enhanced amenities and unique graphics identity (painted on the bus). Two known BRT systems in South America are found in Brazil and Ecuador:

- Curitiba, Brazil:** Curitiba runs one of the early BRT systems; the system features multi-application smart cards (used for transit as well as other applications).
- Quito, Ecuador:** This BRT system uses controlled-access stations adjacent to the exclusive lanes.

An example of a successful BRT implementation is that run by the Los Angeles County Metropolitan Transportation Authority, which in 2000 launched the service in several of Los Angeles' heavy corridors. TranSystem reported that bus ridership in the Wilshire-Whittier and Ventura Blvd. corridors has increased by 20% and 50%, respectively, since the implementation of BRT; furthermore, up to one-third of BRT passengers had previously not been transit users.

In Europe, numerous transit-priority projects have been executed; for example, in Athens, Dublin, Munich, Turin, Vienna and Zurich. Based on Ceder (2004), lessons can be learned from these six case studies: Table 17.4 summarizes the benefits gained from the implementation of transit-priority schemes in these cities.

17.6.3 Improved operations control

Fundamentally there are two distinctive real-time transit-performance disruptions: (1) deviations from the schedule (timetable), but not necessarily creating an imbalance between

Table 17.4 *Examples of transit-priority results in six European cities*

City (population)	Benefits gained
Athens (4 million)	• Reduction of travel time and its variance on bus lanes • 10% increase in patronage on some bus routes
Vienna (1.6 million)	• Reduction in travel time on bus lanes • Possible (not clear) increase in patronage
Munich (1.2 million)	• Reduction of 19% in travel time for tram priority at one traffic signal
Dublin (1 million)	• No reduction in travel time, but a reduction in its variance on bus lanes • Reduction in boarding times • Increased patronage with new fare scale • Increase in revenue • Reduction in the variance of travel time and headways using AVL system
Turin (1 million)	• Reduction in tram travel time, and more reliable tram reliability using AVL (called SIS) system
Zurich (0.36 million)	• Increase of 19 km/hr in average vehicle speed because of priority at traffic signals, AVL and passenger-information systems

supply and demand; (2) creation of an imbalance between supply and demand (overloaded and almost empty vehicles), but not necessarily deviating from the schedule. Given that these disruptions are known in real-time (e.g. by an APTS), corrective and restorative control strategies can take place.

The main real-time control strategies are shown in the following list.

- Holding the vehicle (at terminal or at mid-route point)
- Skip-stop operation
- Adding a reserve vehicle
- Changes in speed (not above the lawful speed limit)
- Interlining operation
- Deadheading operation
- Short-turn operation
- Short-cut operation
- Leapfrogging operation with the vehicle ahead.

The first strategy on the list can be used for improving on-time performance (scheduled-based), eliminating bunching (headway-based) and responding to unexpected demand; however, it has an adverse effect on on-board passengers. This holding strategy is further discussed and analysed in the next section. The skip-stop operation is meant to pass stops at which no passenger wishes to alight; however, it has an adverse effect on waiting passengers at those stops. Adding a reserve vehicle as a reinforcement action is suitable in locations where unexpected demand may appear. Changes in speeds (slowing down or speeding) can

improve on-time performance and eliminate bunching; in the slowing-down cases, it may be better valued by the on-board passengers than is the holding strategy. Interlining (if commonly not in the schedule), deadheading and short-turning operations can be used as corrective actions for possible no-show/lateness situations. A real-time short-cut decision to convert a local service into an express (or semi-express) service between two points is feasible if it fits the destinations of all on-board passengers. Finally, the leapfrogging operation will serve to correct a load-imbalance scenario between two following vehicles, as well as ease the bunching phenomenon.

The main drawback of possible real-time control actions is the lack of prudent modelling and software that can activate these actions by automatic/semi-automatic/manual mode. Figure 17.7 illustrates schematically a computer-aided, real-time control system; such a system can be employed in a transit-control centre to allow for best exploitation of the real-time information.

Another means of improving transit-operations control can be approached from the passenger side. That is, the match between supply and demand can be improved by utilizing intelligent real-time passenger-information systems. Pre-trip information has the potential

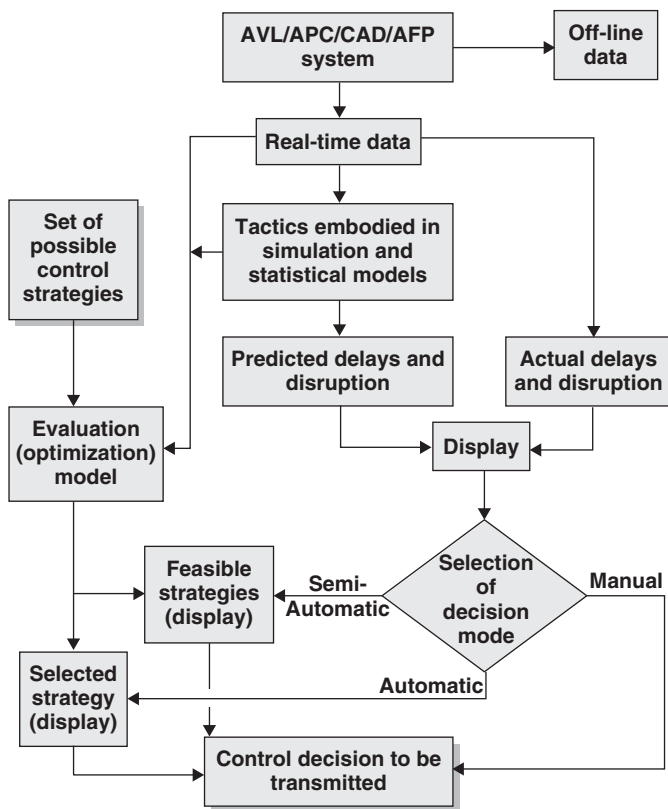


Figure 17.7 Schematic diagram of a computer-aided real-time control system

to change demand by time-of-day and to direct the passengers to the best service. En-route information can reduce passenger uncertainty, thus improving service reliability.

Current communication technology offers the opportunity to convey traveller information via a variety of media: cell phones; in-terminal systems; variable message signs at bus stops, train stations and platforms; and in-vehicle systems, by signs or voice. Lessons may also be learned from the following four major European cities:

- (a) **London:** London Transport offers several types of traveller-information service; e.g. the Journey Planner, bus stop-specific timetables and maps, and the Countdown bus-arrival information system; Oyster, a regional smart card system.
- (b) **Paris:** RATP has deployed Modouls, a regional smart card payment system, with multiple-application agreements for retail, telephone calls and more; ALTAIR, a real-time information system that provides information on-board and at bus stops; and SIEL, which provides information on regional rail service.
- (c) **Turin, Italy:** A public-private partnership provides transit information, including itinerary planning, variable message signs and at-stop displays.
- (d) **Helsinki, Finland:** Advanced traveller-information services include the Journey Planner; the transit agency's real-time system (HELMi); personal mobile traveller and traffic-information service; use of smart cards; and real-time information via the Internet.

17.6.4 Holding strategy: approximate method

Vehicle holding strategy is considered a possible remedy for reliability problems for three main reasons: (1) prevents bunches from forming, (2) ensures that scheduled connections (transfers) are made, and (3) minimizes total passenger waiting time. Holding strategies, therefore, take place at dispatching points, connection points and major stops (e.g. max load point).

Traditionally, there are two classes of holding strategies: scheduled-based and headway-based. The first strategy aims at holding the vehicle until its scheduled departure time; it is especially useful for schedules with large headways. The second strategy, which usually fits schedules with short headways, has the objective of minimizing the total passenger delay at stops. This optimization framework is the result of reducing headway variability (saving delay) and creating extra delay (because of the holding strategy).

Undoubtedly the use of a holding strategy is complicated in practice because of the uncertainty involved (in reaching its objective), its adverse effect on the on-board passengers, and its impact on real-time operations at the network level. This situation may justify the use of an approximation (isolated) method to describe the development of basic holding rules. In general, Wirasinghe (2003) describes acceptable approximate methods in transit services, including dispatching policies, headway setting and scheduling travel times.

An early foundation for approximately analysing dispatching and holding strategies in transit operation was set down by Newell (1971, 1977, 1982). Section 17.7 reviews chronologically the essence of control strategies to improve transit reliability. Newell (1971) and Wirasinghe (2003), in which Newell's work is discussed, present the best transit-vehicle dispatching policy for minimizing total passenger waiting time, subject to a fixed number of departures (dispatches). Newell found that the best dispatching rate was proportional to (i) the square root of the random passenger-arrival rate for large (not constrained) vehicle

capacities, and to (ii) the random passenger arrival rate, otherwise. Osuna and Newell (1972) presented control strategies for dispatching immediately or holding a vehicle for a hypothetical bus-route loop with only one service and control point. They used dynamic programming and queuing techniques for distributed travel times and uniform passenger-arrival rates. Following is an approximate analysis for a holding strategy inspired by Newell's work.

Figure 17.8 illustrates the process of passengers arriving at a max load stop and then departing. Part (a) of the figure refers to an even-load timetable and part (b) to an even-headway timetable. In each part, the upper curve describes the accumulated number of passengers according to the scheduled departure times; the lower curve exhibits a real-life situation in which one early and one late arrival are introduced. The notations used in Figure 17.8, some of which appear in Chapters 3, 4 and 7, are these: P_m is passenger load at the max load point; d_0 is the desired occupancy on a single vehicle; t_i and t'_i are the i -th scheduled and real departure times at the max load point, respectively; λ_{im} is the average uniform random passenger-arrival rate between the $(i-1)$ th and the i -th departures at the max load stop; H , H_i and H'_i are the scheduled even headway, scheduled even-load headway, and real headway, respectively; and Δ_1 , Δ_2 are the possible holding times for an early arrival and the extra-late arrival time, respectively.

As expected, the accumulated number of passengers in the upper curve in part (a) of Figure 17.8 reaches d_0 when a departure takes place. This is not the case in part (b), assuming a different λ_{im} for the even-headway case. Improving reliability usually coincides with reducing the total waiting time at transit stops. In the lower curves of parts (a) and (b) in Figure 17.8, the shaded area represents additional passenger-minutes delay compared to the scheduled (expected) situation. Although it is impossible to reduce the additional delay incurred by the late vehicle arrival (by Δ_2 minutes), it is conceivable that the total passenger delay can be reduced by holding the early arrival vehicle (by Δ_1 minutes). It should be noted that in real time, the situation of late arrivals may also be ameliorated or avoided by the use of skip-stop or short-cut strategies.

In an average and approximated sense, the dashed area of Δ_1 in part (a) of Figure 17.8 is $H'_1 \cdot \lambda_{1m} \cdot \Delta_1$, being the additional passenger-minutes incurred if the t'_1 departure is held until t_1 . Otherwise, the additional approximated delay for passengers expected to board the t_2 departure is $\lambda_{1m} \cdot \Delta_1 \cdot H_2$. Consequently, a simple strategy is this: hold the vehicle only if $H'_1 \cdot \lambda_{1m} \cdot \Delta_1 < \lambda_{1m} \Delta_1 \cdot H_2$. However, because of the uncertainty about H_2 , its expected value is considered; i.e. $E(H_2)$. In general, the strategy for even-load timetables is to hold the vehicle if

$$H'_i < E(H_{i+1}) \quad (17.11)$$

and to dispatch immediately, otherwise. The expected (mean) value $E(H_{i+1})$ requires data for constructing the probability density function (distribution) of H_{i+1} for all departures i , $i \neq 1$.

The simple rule of Equation (17.11) also applies to part (b) of Figure 17.8, the case of even headways. That is, the vehicle is held from t'_1 departure until t_1 if

$$H'_i < E(H) \quad (17.12)$$

and dispatched immediately, otherwise.

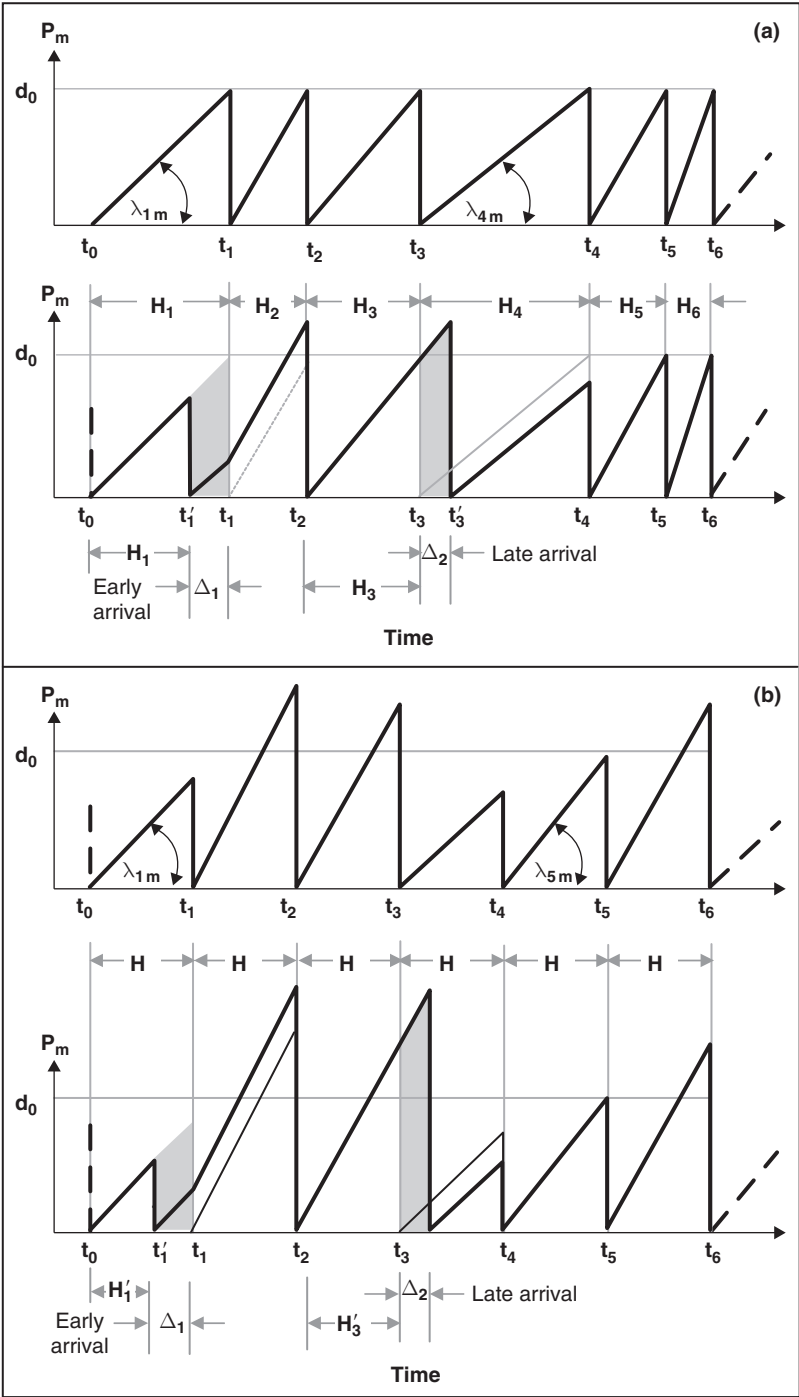


Figure 17.8 Cumulative number of passengers at the max load stop for an even-load timetable in part (a) and an even-headway timetable in part (b), in which the anticipated, scheduled situation in each part is in the upper curve, and the real-life scenario is in the lower curve

17.7 Literature review and further reading

The reliability of a public transit service is considered one of the main factors influencing the level of passengers' satisfaction. There is a very abundant literature on various aspects of transit reliability. This review describes, in chronological order, selected papers that focus only on the following subjects:

- Development of quantitative measures of service reliability.
- Models of passengers' waiting time at stops as a function of service reliability.
- Suggested control strategies for improving reliability.

17.7.1 Measures of service reliability

Service reliability is an abstract term. In order to include reliability considerations in a detailed engineering design and to evaluate the differences between existing and suggested service alternatives, it is necessary to describe reliability in mathematical terms. This section includes a review of quantitative measures of service reliability.

One of the earliest discussions of transit-service reliability measures was made by Polus (1978), who focused on bus services. Reliability was defined as the amount of consistency from day to day associated with any operational performance measure; the proposed reliability indicator being the inverse of the standard deviation of travel times.

Silcock (1981) mentions some traditional reliability measures:

- Number of buses taking x minutes longer than scheduled.
- Percentage of buses that depart from one minute early to four minutes late.
- Average waiting time of passengers.
- Excess waiting time of passengers.
- The difference between the actual average waiting time and the calculated waiting time.

Turner and White (1990) formulated reliability measures in order to compare services provided by different bus sizes. They suggest the following formula for excess waiting time (EWT) as a measure of reliability:

$$\text{EWT} = \frac{T}{2} \left(\frac{1}{N} - \frac{1}{S} \right) + \frac{N \cdot \text{VarH}}{2T}$$

where T = time duration of the reliability survey, N = number of observed buses during the reliability survey, S = number of scheduled buses during the reliability survey and VarH = variance of observed headways.

Henderson *et al.* (1991) analysed the advantages of using odd ratios of on-time performance as a measure of service reliability. An odd ratio is the percentage of trains or buses that arrive on time, divided by the percentage of those that were not on time. It should be emphasized that an odd ratio is not the ratio of on-time arrivals to total arrivals, since the denominator is not that of total arrivals. Because of this formulation of the ratio, the manner in which its value changes as the number of deviations from the schedule increases is not

linear; the authors claim that this formulation gives a better representation of passengers' perception of reliability than do most other measures. For example, the reliability of a transit service with a 70% probability of arriving on time ($70/30 = 2.33$) is eight times worse than the reliability of a service with 95% on-time arrivals ($95/5 = 19$). Hence, odd ratios as reliability measures show passengers' high sensitivity to the consistency of service punctuality.

Rudnicki (1997) defined several unreliable performance measures. A suggested unreliability measure is the chance that a passenger will have to wait for a later departure because of to a too-early departure of the expected vehicle. An inconvenience measure is the average excess waiting time. A degree-of-punctuality factor is the average value of a predetermined function, which receives a value between 0 (worse) and 1 (best) according to the deviation from the scheduled time. Another unreliability factor is the difference between the 95th percentile of actual waiting time and the planned waiting time. The paper explains which of the measures represent the interests of the passengers and which are more suitable to represent the operator's point of view.

Hallowell and Harker (1998) present a method of predicting the schedule reliability of a partially double rail line on which delays are caused by the need for one train to wait for the passing train. Using a simulation model for the design of a reliable timetable, the researchers describe measures of departure-time inaccuracy: at origin, arrival at destination, delay time from origin to destination, and slack time from origin to destination (slack time is defined as planned ride time minus free running time).

Carey (1999) developed measures of estimating service reliability in advance. The first proposed measure depends on the probability of delay for a specific train, resulting either from schedule failure of this specific train or from a delay caused by previous trains. Another measure depends not only on the chance of such delay but also on the magnitude of the delay. Both of these measures require a probability density function (PDF) of route delay as the input. Additionally, the author suggests two measures whose calculation does not require a given PDF. One is based on the standard deviation and the mean of headways. The other derives from the idea that a small initial delay grows larger if the delayed train, on the way to its intended platform, has to cross paths used by another train; the fewer paths crossed, the more reliable the service. The measures developed by Cary are normalized, such that they all receive values between 0 and 1, with 1 representing the highest reliability level.

Yin *et al.* (2002) develop a simulation-based approach to assess transit reliability, taking into account the interaction between network performance and passengers' route-choice behaviour. Measures are developed for three types of reliability: travel-time reliability (TTR), schedule reliability (SR), and waiting-time reliability (WTR). TTR is the probability that the average actual travel time is less than a given threshold. SR is the probability that an actual line is frequency larger than a given threshold. WTR is the probability that passengers' average waiting time is less than a given threshold, taking into account the possibility of passengers' being unable to board the first vehicle that arrives owing to insufficient vehicle capacity. A Monte Carlo simulation approach is used, incorporating a stochastic user-equilibrium transit-assignment model with a capacity constraint and elastic frequencies. Analysing simulation results using the measures mentioned is very effective in evaluating the trade-off among alternatives for service improvement.

Characteristics of the reliability measures reviewed are summarized in Table 17.5.

Table 17.5 *Characteristics of works dealing with passenger waiting time as a function of reliability*

Source	Proposed measures	Source of input data
Polus (1978)	Inverse of standard deviation of travel times	Travel-time measurements for different days
Silcock (1981)	Number of late buses, percentage of buses that depart on time, average waiting time, excess waiting time, difference between actual and calculated wait	Vehicle performance survey/simulation
Turner and White (1990)	Excess waiting time	Vehicle performance survey
Henderson <i>et al.</i> (1991)	Odd ratio = percentage of on-time arrivals divided by percentage of not-on-time arrivals	Vehicle performance survey/simulation
Rudnicki (1997)	Factor of unreliability on account of unpunctuality, factor of inconvenience owing to unpunctuality, degree of punctuality as a factor of operation regularity, factor of unreliability on account of irregularity	Vehicle performance survey
Hallowell and Harker (1998)	Departure-time error at origin, arrival-time error at destination, delay time from origin to destination, slack time from origin to destination	Simulation
Carey (1999)	A measure depending on the chance of delay and a measure depending on the magnitude of delay	PDF of route delay
	A measure based on standard deviation and mean of headways	Vehicle performance survey/simulation
	A measure of the chance of delay owing to the need to wait for a crossing train	Information about number of rail paths crossed
Yin <i>et al.</i> (2002)	Probability that the average actual travel time is less than a given threshold, probability that an actual line frequency is larger than a given threshold, probability that the passengers' average waiting time is less than a given threshold	Simulation

17.7.2 Passenger waiting time as a function of reliability

It is often claimed that passengers are more sensitive to changes in their waiting time than to the other periods comprising the trip. Naturally, waiting time depends on the motion of the anticipated vehicle, and service is considered more reliable from this point of view if that motion is always the same. Prediction of vehicle-time variability is a complex task in

itself, since it is influenced by numerous factors, such as traffic conditions and passenger flow while boarding and alighting. Waiting time also depends on other factors, some of which have to do with the passengers' behaviour. Several models for estimating waiting time will now be reviewed.

The traditional formula for average passenger time, depending on headway variability, is shown in Equation (12.1) in Chapter 12, and discussed in Section 17.4.1 in this chapter. This formula was used by Osuna and Newell (1972), as well as by other authors. For the sake of clarity, it is presented again:

$$E(w) = \frac{\bar{H}}{2} (1 + C_H^2)$$

where $E(w)$ = average waiting time, $\bar{H}$ = mean headway, and C_H = coefficient of variation in headways (standard deviation / mean).

This formula is based on the assumption that passengers arrive at a stop randomly. The average waiting time is greater than half the headway because more passengers arrive during long headway intervals (before the arrival of a delayed vehicle) than during short intervals (after the arrival of a delayed vehicle). The average wait, therefore, is longer when headways are less uniform.

Turnquist (1978) developed a model of bus and passenger arrivals at a bus stop. He investigated the impact of reliability on the average waiting time under the assumptions that the bus arrival time behaves according to a log-normal distribution and that its mean and variance are known. The bus-arrival-variance variable plays the role of the reliability indicator. Calculation of waiting time takes into account both passengers who arrive at the station randomly and passengers who plan their arrival time. If the proportion of random arrivals is known, then this model makes it easy to calculate waiting time for planning purposes.

Bowman and Turnquist (1981) developed another model for evaluating average waiting time as a function of service reliability and frequency. This model is based on more complex assumptions regarding passenger behaviour: it takes into account the decision-making process that passengers go through when they choose a time to arrive at the bus stop. The model predicts the distribution of passenger arrivals when the mean and average of bus headways are given. It enables operators to make a realistic prediction of how a change in service reliability will alter the overall waiting time. This prediction is sensitive to the differences in passenger behaviour among lines with different frequencies. The authors conclude that waiting time is much more sensitive to schedule reliability than to service frequency.

Guenther and Sinha (1983) described another model for estimating waiting times. Their model is unique in that it assumes that schedule reliability is influenced by the level of vehicle maintenance: when there are not enough vehicles available to meet service demand, reliability is severely affected. The tool presented consists of three models: a maintenance model, a reliability model and a performance model. The input for the maintenance model is the number of spare buses and the number of mechanics: its output is a dependability factor that expresses the point at which maintenance problems will lead to service failure. This factor is calculated by using simulation. The dependability value is an input for the reliability model, which determines the average waiting times both of passengers who

arrive randomly and of those who come at a pre-planned time. Waiting-time values are transferred to the performance model, which yields predictions of ridership and an evaluation of general system performance. Implementing the proposed set of models requires some computer tools as well as calibrated values of demand elasticities, which may not always be available.

Table 17.6 summarizes the models reviewed.

Table 17.6 *Characteristics of works on passenger waiting time as a function of reliability*

Source	Required input	Assumptions regarding passenger arrival at station	Assumptions regarding vehicle arrival at station
Osuna and Newell (1972)	Standard deviation and mean of headway	Random	None
Turnquist (1978)	Standard deviation and mean of headway	Some passengers arrive randomly, others plan their arrival time	Log-normal distribution
Bowman and Turnquist (1981)	Standard deviation and mean of headway, probability distribution function of vehicle arrival	A utility function is ascribed to each possible arrival time	None
Guenthner and Sinha (1983)	Number of spare buses and mechanics	Some passengers arrive randomly, others plan their arrival time	Log-normal distribution

17.7.3 Control strategies for improving reliability

Transit services are considered unreliable if they suffer from significant travel-time variability. Improvement in service reliability is, therefore, likely to occur if actions are taken to reduce time variability. Such actions are usually aimed either at having control over vehicle adherence to its schedule or at introducing route patterns with low time variability. This section presents several strategies for improving reliability. It should be noted that short-turn strategies, although mentioned in some papers in this context, are referred to in another chapter of this book.

Osuna and Newell (1972) presented the conflict between the desire for minimal route time and the need to slow down system performance in order to increase its regularity. They formulated a control problem that aims at increasing punctuality, thereby achieving minimal waiting time. Dispatching and holding strategies are introduced: dispatching strategies help in deciding what the actual headways would be between successive departures from the starting terminal, while holding strategies mean that departure time is also controlled at other timepoints along the route. The performance of controlled and uncontrolled systems is analysed, based on a PDF of travel time, which is required as input.

Other basic assumptions, declaratively simplistic, refer to an idealized network. For example, only single-stop routes with one or two buses are discussed. Still, this paper represented a significant discussion of schedule control, since it led to many other research works in this area.

Lesley (1975) suggested a procedure for designing a realistic schedule on a route along which bus progress is monitored. A reliability index is calculated for each stop on the basis of the level of headway variance. Stops at which the index is at least twice the average are chosen as timepoints for holding control. The required slack time at each point is calculated.

Newell (1977) developed a method for determining the slack time allocated to each timepoint when holding control is applied. With the aim of reducing vehicle time while keeping schedule regularity, the lower boundary of slack time is calculated. Some of the basic assumptions are simplistic, such as ignoring alighting times (which is still a realistic assumption) and supposing no correlation between adjacent buses. Passengers are assumed to arrive at stops according to a Poisson distribution, and bus delay is assumed to behave according to Fokker-Planck. Slack time is calculated for each stop, not just selected ones.

Koffman (1978) compared several real-time strategies for improving headway reliability on a single-direction bus route with a single control point. Employing a simulation model, the author examined the following strategies: holding, skipping stops, introducing bus-priority signalling and reducing dispatch uncertainty. All control strategies are analysed under different headway thresholds. The resulting conclusion is that the last two strategies seem the most promising.

Jordan and Turnquist (1979) discussed reliability improvement through the introduction of a special pattern for urban routes. According to this pattern, the whole path is divided into several zones, and each bus makes stops only in a specific zone. A dynamic programming problem is formulated in order to determine the optimal number of zones, points where zones switch, and the number of buses serving each zone. The objective is to find a zone structure that minimizes trip-time average and variance. The model developed includes sub-models for predicting dwell time at stops, distribution of headways, mean and variance of waiting time, and mean and variance of travel time; intensive calibration is therefore required. The dynamic programming model is applied in a case study, which yields the conclusion that even a very simple zone structure can lead to a substantial improvement in reliability and a decrease in bus-fleet size.

Turnquist and Bowman (1980), using a set of simulation experiments, investigated the effect of network structure on service reliability. Transit networks on which route paths form a radial pattern are compared to grid-shaped networks. The researchers found that even though passengers were obliged to transfer more in grid networks, the uncertainty in regard to the length of the transfer delay was smaller than in radial networks. The combined travel-time uncertainty is also smaller in grid networks. The authors conclude that a concentration of transfers at the centre node of radial networks has a more disruptive effect on reliability than does the dispersed distribution of transfers that occurs in grid networks.

Furth and Nash (1985) suggested a method for improving adherence to a preplanned timetable, for which all routes started at a specific terminal. According to this method, the pool of vehicles that belongs to the terminal serves all the round trips leaving that terminal in a 'first-in-first-out' sequence, instead of being assigned to specific trips in advance.

Features of this strategy are presented, and the reliability of the resulting service is estimated. A PDF of bus travel times is needed as input.

Abkowitz *et al.* (1986) determined which single point was optimal for holding control. They found that the location of the control point was sensitive to the passenger-boarding profile. The optimal control point is usually located just prior to a group of stops having a large number of boarding passengers; that is, toward the peak segment of the route (max load point). Many other papers published in the 1970s and 1980s describe methods of designing optimal holding strategies. Most of these methods use simulation techniques to optimize the value of a waiting-time threshold, above which a stop is declared a timepoint for controlling schedule adherence.

Seneviratne (1990) calibrated an expression for headway standard deviation as a function of the number of timepoints. This function is a second-degree polynomial, meaning that there is an optimal number of points and that increasing the points beyond that number will lead only to diminished reliability.

Li *et al.* (1993) presented various real-time dispatching strategies to be used at the terminal of a single route: instructing bus drivers to skip stops, directing them to cut out parts of their route, or simply adjusting layover times. The design of these strategies is expressed as a nonlinear program and as a stochastic binary integer program. Simulation is used for testing this approach, but it can also be used to evaluate pre-established schedules or decisions regarding stop-skipping strategies. Special attention is given to assessing the impact of bus-location information on the dispatching decision. The researchers find no need for an automatic vehicle-location system that covers the entire network for the purpose of real-time dispatching. Location information from a limited number of points seems to be sufficient.

Wirasinghe and Liu (1995) formed a dynamic programming model for the design of holding strategies for an isolated bus with fixed demand. The number and location of timepoints are determined, as is the amount of slack time at each point.

Carey (1998) formulated a problem of optimizing slack time between successive activities in a vehicle schedule. Increasing slack time reduces the risk that a delay in one activity will affect the next activity; on the other hand, allocating more spare time may cause the activity to take longer because of behavioural reasons. The author claims that some percentage of scheduled rides will always depart a few minutes late, independently of the length of the slack time before the scheduled departure. A PDF of activity time is a required input. Slack times that give minimum total operating costs are calculated. It should be noted that many other papers by this author deal with delay control and various other aspects of transit reliability.

Hallowell and Harker (1998) proposed a simulation method mentioned earlier in this section. Their method can be used for determining slack times allocated to stations along a railway line in order to increase schedule reliability. The authors compare features of real-time scheduling and master scheduling. They also discuss the difficulty in using simulation methods for improving performance; it is due to the complexity of the computation tasks involved.

Adamski and Turnau (1998) developed a tool for real-time schedule-control decision-making. The input for the model is real-time information about deviations from scheduled times or from regular headways along a transit route; the output is a set of recommended decisions regarding vehicle dispatching. The tool developed includes a sub-model that

calculates the time spent at stops. Many other papers by Adamski deal with dispatching control.

Eberlein *et al.* (1998) discuss the real-time strategy of deadheading; i.e. running empty from the terminal to the first station at which passengers are allowed to board, while skipping several stations en route. A programming model is formulated in order to determine which vehicles to deadhead and how many stations to skip.

O'Dell and Wilson (1999) developed a real-time decision-support system for rail-transit operations. The system includes a deterministic model of a rail system and mixed-integer programming formulations for the choice of an optimal holding strategy. Several holding strategies are discussed: hold at all stations, hold only at the first station after the disruption occurs, and hold at a fixed chosen station. These are investigated with and without train-capacity constraints. The paper also includes a dwell-time sub-model. Attention is given to assumptions on the order in which trains from different branches enter junctions; this is claimed to have a major effect on system performance. The results show that holding strategies at one station (whether the first after the disruption or any station) are almost as effective as holding at all stations. They recommend holding control at the first station after the disruption, since it is the easiest to implement.

Eberlein *et al.* (1999) considered several control strategies: two station-skipping strategies, holding at a given station, and combinations of the two. They formulated mathematical models for all control problems under the assumption that real-time vehicle-location information was available, and investigated the advantages and disadvantages of each control type. The models are deterministic, but are also examined under various stochastic conditions. The results show that a combination of station-skipping and holding strategies can have very effective results on system performance. It is also shown that combined strategies are more efficient than any single strategy, although a holding strategy is slightly preferable. The authors find that the optimal holding strategy depends mainly on the route-headway pattern at the holding station and is independent of passenger demand along the route.

Shen and Wilson (2001) developed a disruption control model for rail-transit systems. The model, a deterministic, mixed-integer program, includes a dwell-time sub-model. Several control strategies are compared. The main conclusion is that best system performance is achieved when holding and short-turn strategies are combined.

Liu and Wirasinghe (2001) presented another simulation model for the design of a holding control. The simulation model consists of several sub-models that predict performance during various journey stages: linked travel time, passenger arrival, passenger-alighting demand, dwell time, and dispatching time at the starting terminal. The output of the overall model is the choice of timepoints and the slack times allocated to each of them.

Fu and Yang (2002) developed two simulation models for choosing timepoints for cases in which holding control should be implemented. In both models, timepoints are located at every stop where headways exceed some threshold. In the first method, only the headway between a bus and the preceding bus is checked; the second method checks, in addition, the headway between each bus and the following bus. The researchers find that control is most useful when the control stops chosen are those with the highest demand and close to the middle of the route. The number of control points is also investigated – scenarios with one or two or all stops with control are examined. Control at all stops reduces waiting time, but it also results in a significant increase in vehicle time. Control at two stops, one of which is

the starting terminal, is found to be the optimal solution, since it enables the operator to achieve a reduction in waiting time with no increase in in-vehicle time and only a slight increase in bus-travel time.

Dessouky *et al.* (2003) examined simulated systems in which holding and dispatching strategies are used; they also investigated the dependence of system performance on available technologies. Systems in which communication, vehicle locating and counting technologies are available or unavailable are compared. The analysis addresses seven different holding-strategy combinations, each of which differs in its definition of holding conditions and in its approach to the use of holding for connecting buses. The results show that advanced technology is most advantageous when there are many connecting buses; the schedule slack is then close to zero, and the headway is large.

The major characteristics of the models reviewed are summarized in Table 17.7. For practical use, the main difference between analytic methods and methods based on simulation should be mentioned. Analytical methods require less detailed information about all network components; however, they usually rely on many assumptions. The problem is then to simplify the representation of the system performance. Methods that use simulation provide much more realistic results, but the input that they require is much more detailed.

Exercises

- 17.1 Given a bus route on which passengers' mean waiting time, $E(w)$, can be estimated by $E(w) = E(H^2)/2E(H)$, where H is the headway variable:
 - (a) Explain and show (mathematically) how $E(w)$ can be reduced, using the waiting-time formula. Consider headway distributions ranging from regular (deterministic) to exponential (random).
 - (b) Outline feasible (practical) ways to reduce $E(w)$, given fixed passenger demand.
- 17.2 With high volumes of bus services on major arterial roads, it becomes very expensive and often ineffective for bus agencies to use curbside inspectors and mobile supervisors. The inspector's or supervisor's knowledge of situations is limited to his/her immediate vicinity, with no indication of conditions elsewhere. Accordingly, bus drivers may be instructed in some cases to act in a fashion that is counter-productive from the system point of view. This represents only one group of problems that can be resolved by introducing an Automatic Vehicle Location (AVL) system. In general, the real-time control provided by an AVL system enables the operator to increase both the utilization of each bus and the service level as seen by each passenger. Furthermore, the substantial amount of off-line AVL data can be used to improve the entire bus-planning process.
 - (a) List all possible strategies that supervisors can accommodate in the control room while using an advanced AVL system. Note that these strategies are carried out by instructing a driver or a group of drivers to follow a certain action.
 - (b) Outline the major potential benefits of (adequately used) AVL off-line data.

Table 17.7 *Characteristics of works on control strategies for improving reliability*

Source	Control type	Method	Place of control	Suggested solution
Osuna and Newell (1972)	Dispatching, holding	Analytical	Single control point	
Lesley (1975)	Holding	Simulation	Number and location of points are optimized	
Newell (1977)	Holding	Analytical	All stops are controlled	
Koffman (1978)	Real-time holding, stop-skipping, bus-priority signalling, dispatching	Simulation	Single control point (terminal)	Bus-priority signalling and reducing dispatching uncertainty
Jordan and Turnquist (1979)	Zone scheduling	Dynamic programming		
Turnquist and Bowman (1980)	Network structure	Simulation		Grid network
Furth and Nash (1985)	Real-time dispatching	Analytical	Terminal	
Abkowitz <i>et al.</i> (1986)	Holding	Combination of simulation and analytical	Location of control point is optimized	Optimal control point is toward the peak segment of the route
Seneviratne (1990)	Holding	Simulation	Number of control points is optimized	
Li <i>et al.</i> (1993)	Real-time dispatching/stop-skipping	Nonlinear program/ binary integer program (simulation is used for testing the method)	Terminal of a single route	

Wirasinghe and Liu (1995)	Holding	Dynamic programming	Number and location of points are optimized	
Carey (1998)	Dispatching	Analytical	Terminal	
Hallowell and Harker (1998)	Holding (real time/master)	Simulation	All stops are controlled	
Adamski and Turnau (1998)	Real-time dispatching	Simulation	All stops are controlled	
Eberlein <i>et al.</i> (1998)	Real-time deadheading (stop-skipping)	Programming model	Terminal	
O'Dell and Wilson (1999)	Real-time holding	Analytical/mixed-integer programming	All stations/first station after disruption/fixed strategies are investigated	Holding at first station after disruption
Eberlein <i>et al.</i> (1999)	Real-time stop-skipping, holding, combinations	Analytical		Combination of strategies
Shen and Wilson (2001)	Holding, short turn	Mixed-integer program		Combination of strategies
Liu and Wirasinghe (2001)	Holding	Simulation	Number and location of points are optimized	
Fu and Yang (2002)	Holding	Simulation	1, 2 and all strategies are investigated	Holding at 2 points
Dessouky <i>et al.</i> (2003)	Holding, dispatching	Simulation	Seven alternatives are compared	

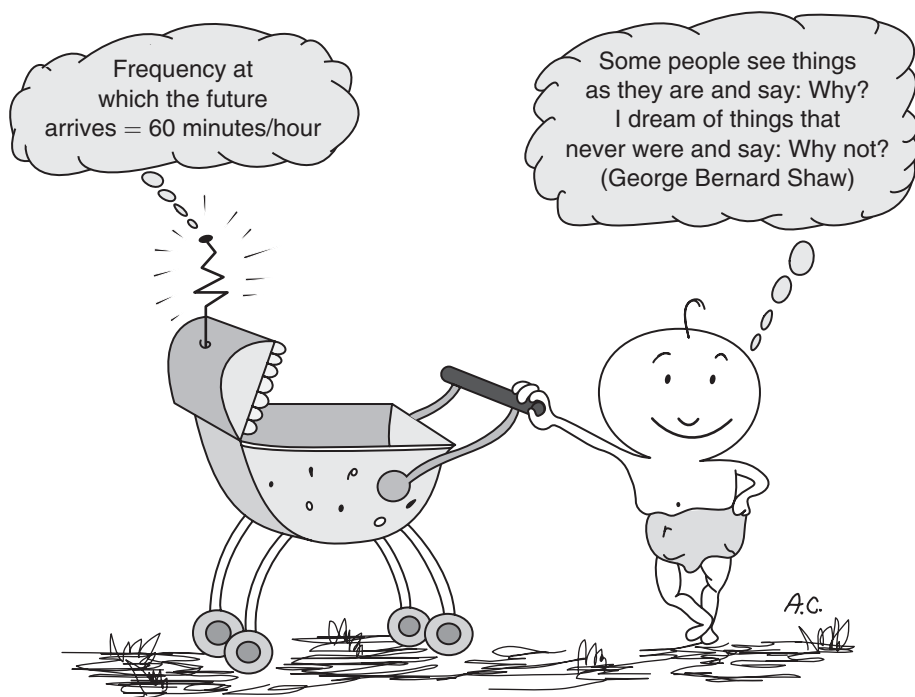
References

- Abkowitz, M., Eiger, A. and Engelstein, I. (1986). Optimal control of headway variation on transit routes. *Journal of Advanced Transportation*, **20**, 73–88.
- Abkowitz, M., Slavin, H., Waksman, R., Engelstein, I. and Wilson, N. H. M. (1978). *Transit Service Reliability*. US Department of Transportation, Final Report UMTA MA-06-0049-78-1.
- Adamski, A. and Turnau, A. (1998). Simulation support tool for real-time dispatching control in public transport. *Transportation Research*, **32A**, 73–87.
- Balcombe, R., Mackett, R., Paulley, N., Preston, J., Shires, J., Titheridge, H., Wardman, M. and White, P. (2004). *The Demand for Public Transport: A Practical Guide*. TRL Report, TRL593, TRL Limited.
- Bowman, L. A. and Turnquist, M. A. (1981). Service frequency, schedule reliability, and passenger wait times at transit stops. *Transportation Research*, **15A**, 465–471.
- Carey, M. (1998). Optimizing scheduled times, allowing for behavioral response. *Transportation Research*, **32B**, 329–342.
- Carey, M. (1999). Ex-ante heuristic measures of schedule reliability. *Transportation Research*, **33B**, 473–494.
- Ceder, A. and Marguier, P. H. J. (1985). Passenger waiting at transit stops. *Traffic Engineering and Control*, **26**, 327–329.
- Ceder, A. (2004). New urban public transportation systems: Initiatives, effectiveness, and challenges. *ASCE Journal of Urban Planning and Development*, **130**, 56–65.
- Dessouky, M., Hall, R., Zhang, L. and Singh, A. (2003). Real-time control of buses for schedule coordination at a terminal. *Transportation Research*, **37A**, 145–164.
- Eberlein, X. J., Wilson, N. H. M. and Barnhart, C. (1998). The real-time deadheading problem in transit operations control. *Transportation Research*, **32B**, 77–100.
- Eberlein, X., Wilson, N. H. M. and Bernstein, D. H. (1999). Modeling real-time control strategies in public transport operations. In *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **471** (N. H. M. Wilson, ed.), pp. 325–346, Springer-Verlag.
- Fu, L. and Yang, X. (2002). Design and implementation of bus-holding control strategies with real-time information. *Transportation Research Record*, **1791**, 6–12.
- Furth, P. G. and Nash, A. B. (1985). Vehicle pooling in transit operations. *Journal of Transportation Engineering*, **111**(3), 268–279.
- Guenther, R. P. and Sinha, K. C. (1983). Maintenance, schedule reliability, and transit system performance. *Transportation Research*, **17A**, 355–362.
- Hallowell, S. F. and Harker, P. T. (1998). Predicting on-time performance in scheduled railroad operations: Methodology and application to train scheduling. *Transportation Research*, **32A**, 279–295.
- Henderson, G., Adkins, H. and Kwong, P. (1991). Subway reliability and the odds of getting there on time. *Transportation Research Record*, **1297**, 10–13.
- Huddart, K. W. (1973). Bus priority in greater London: Bus bunching and regularity of service. *Traffic Engineering and Control*, **14**, 592–594.
- Hwang, M., Kemp, J., Lerner-Lam, E., Neuerburg, N. and Okunieff, P. (2006). *Advanced Public Transportation: State of the Art Update 2006*. Report FTA-NJ-26-7062-06.1, Federal Transit Administration, US Department of Transportation.

- Jeong, R. and Rilett, L. R. (2005). Prediction model of bus arrival time for real-time application. *Transportation Research Record*, **1927**, 195–204.
- Jolliffe, J. K. and Hutchinson, T. P. (1975). A behavioral explanation of the association between bus and passenger arrivals at a bus stop. *Transportation Science*, **9**, 248–282.
- Jordan, W. C. and Turnquist, M. A. (1979). Zone scheduling of bus routes to improve service reliability. *Transportation Science*, **13**, 242–268.
- Khattak, J. A. and Hickman, M. (1998). Automatic vehicle location and computer aided dispatch systems: Commercial availability and development in transit agencies. *Journal of Public Transportation*, **2**(1), 1–26.
- Kimpel, T. J., Strathman, J. G., and Callas, S. (2006). Improving scheduling through performance monitoring using AVL and APC data. In *Computer-aided Scheduling of Public Transport*. Lecture Notes in Economics and Mathematical Systems (M. Hickman, S. Voss, and P. Mirchandani, eds), Springer-Verlag (to appear).
- Koffman, D. (1978). A simulation study of alternative real-time bus headway control strategies. *Transportation Research Record*, **663**, 41–46.
- Larson, R. C. and Odoni, A. R. (1981). *Urban Operations Research*. Prentice Hall.
- Lesley, L. J. S. (1975). The role of the timetable in maintaining bus service reliability. In *Proceedings of Operating Public Transport Symposium*, University of Newcastle upon Tyne.
- Li, Y., Rousseau, J. M. and Gendreau, M. (1993). Real-time dispatching public transit operations with and without bus location information. *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **430**, 296–308.
- Liu, G. and Wirasinghe, S. C. (2001). A simulation model of reliable schedule design for a fixed transit route. *Journal of Advanced Transportation*, **35**, 145–174.
- Newell, G. F. (1971). Dispatching policies for a transportation route. *Transportation Science*, **5**, 91–105.
- Newell, G. F. (1977). Unstable Brownian motion of a bus trip. In *Statistical Mechanics and Statistical Methods in Theory and Applications* (U. Landman, ed.), Plenum Press.
- Newell, G. F. (1982). *Applications of Queueing Theory*. 2nd edition. Chapman & Hall.
- O'Dell, S. W. and Wilson, N. H. M. (1999). Optimal real-time control strategies for rail transit operations during disruptions. In *Computer-aided Transit Scheduling*. Lecture Notes in Economics and Mathematical Systems, **471** (N. H. M. Wilson, ed.), Springer-Verlag.
- Osuna, E. E. and Newell, G. F. (1972). Control strategies for an idealized public transportation system. *Transportation Science*, **6**, 57–72.
- Polus, A. (1978). Modeling and measurement of bus service reliability. *Transportation Research*, **12**, 253–256.
- Rudnicki, A. (1997). Measures of regularity and punctuality in public transport Operation. *Preprints of the 8th IFAC/IFIP/IFORS Symposium*, **2**, 678–683.
- Seneviratne, P. N. (1990). Analysis of on-time performance of bus services using simulation. *Journal of Transportation Engineering*, **116**, 517–531.
- Shen, S. and Wilson, N. H. M. (2001). An optimal integrated real-time disruption control model for rail transit systems. In *Computer-aided Scheduling of Public Transport*. Lecture Notes in Economics and Mathematical Systems, **505** (S. Voss, and J. R. Daduna, eds), Springer-Verlag.
- Silcock, D. T. (1981). Measures of operational performance for urban bus services. *Traffic Engineering and Control*, **22**, 645–648.

- TranSystems Corp., Planner Coll., Inc. and Crikelair, T. Assoc. (2006). Elements needed to create high ridership transit systems: Interim guidebook. *TCRP Report 32*, Transportation Research Board, Washington, DC.
- Turner, R. P. and White, P. R. (1990). Operational aspects of minibus services. *Transport and Road Research Laboratory Contractor Report*, **185**, 1–5.
- Turnquist, M. A. (1978). A model for investigating the effects of service frequency and reliability on bus passenger waiting times. *Transportation Research Record*, **663**, 70–73.
- Turnquist, M. A. and Bowman, L. A. (1980). The effects of network structure on reliability of transit service. *Transportation Research*, **14B**, 79–86.
- US Department of Transportation (2003). *Advanced Public Transportation Systems Deployment in the United States*. 2002 Update.
- Wirasinghe, S. C. (2003). Initial planning for urban transit systems. In *Advanced Modeling for Transit Operations and Service Planning* (H. K. Lam and M. G. H. Bell, eds), Elsevier Ltd.
- Wirasinghe, S. C. and Liu, G. (1995). Determination of the number and locations of time points in transit schedule design. In *Passenger Transportation* (M. Gendreau and G. Laporte, eds), Baltzer Science.
- Yin, Y., Lam, W. H. K. and Miller, M. A. (2002). A simulation-based reliability assessment approach for congested transit network. *Advanced Modeling for Transit Operations and Service Planning*, Croucher Advanced Study Institute.

18 Future Developments in Transit Operations



Chapter 18 Future Developments in Transit Operations

Chapter outline

- 18.1 Introduction
 - 18.2 Multi-Agent Transit System (MATS)
 - 18.3 Vehicle encounters on road segments
 - 18.4 Developments in transit automation
 - 18.5 Literature review and further reading
 - 18.6 Concluding remark
-
- References
-

Practitioner's Corner

Albert Einstein once said: “I never think of the future. It comes soon enough”. Indeed, the rate at which the future seems to be arriving provides a good reason to plan and design ahead; this should be an integral part of any routine transit-operations planning undertaking. In this last chapter, I will attempt to touch the near future while being fully aware that choice, and not chance, determines destiny.

Our times, especially the past decade, have known extreme revolutions, especially in communications, and these have brought about a whole new culture, including changes in language and behaviour. The Internet and the cellular-phone have altered the way we live. These two revolutions have far-reaching consequences, especially for the rapid development of technologies supporting them: fast switching, self-healing distributed wireless and mobile networks, proximity and location technologies, nano-technologies, motion control, mobile and miniscule power supplies, new user interfaces and user experience. All these developments are gradually trickling into the field of transit service.

This chapter starts with an introductory section describing a new (future) concept of automatic passenger transfers along road segments rather than at transit stops and then discusses technological issues in transit-system automation. It continues with three main parts. Section 18.2 details the new concept outlined in the introductory section, that of a multi-agent transit system composed of the following agents: transit vehicles, passengers, road segments, transit agency and local authority (or government). This section defines each agent and their interrelationship. Section 18.3 analyses the probability of the simultaneous arrival of two or more transit vehicles at a given road segment; it calculates the road-segment encounter, with any point along the segment constituting a possible encounter point, and discusses possible on-line tactical deployment measures. Section 18.4 moves to the technology perspective, describing existing automated transit systems and proposed future concepts, such as the dual-mode personal rapid transit system. The chapter ends with a literature review.

Fundamentally, practitioners are encouraged to visit all sections of this chapter. The chapter offers challenges for devising alternative transit solutions that have yet (if at all) to be introduced or that may need to negotiate the barrier of implementation.

Until the future actually reveals itself, practitioners can, and should, improve their readiness for future transit operations under new conditions.

Finally, two pieces of advice for practitioners with regard to future developments: (1) future solutions to existing transit problems are, in fact, existing solutions (some of which may be found in this book) to future problems; (2) whatever you do, try to act like an optimistic scientist, who will never cry over spilled milk, because 88% of it is just water in any case.

18.1 Introduction

Niels Bohr facetiously cautioned: “Prediction is very difficult, especially of the future”. Nonetheless, future developments in transit operations are already reflected in a vast amount of articles. Although we cannot change the direction of the wind (evolution of lifestyles, land-use patterns), we can adjust the sails (create attractive transit systems that will naturally shift people from the car to public transit vehicles). This sail-adjustment realization relies heavily on the ongoing development of new technologies.

Innovative technologies in transit operations have three major objectives:

- to increase the productivity of transit operations, particularly through the introduction of automation;
- to improve safety, performance, and service capabilities and to achieve this in a cost-effective manner;
- to support transit-related priorities at the national level, such as energy conservation, safety, central city revitalization and environmental protection.

Essentially the new technologies should pursue the goal of a prudent, well-connected transit system as defined in Section 13.5 in Chapter 13: *An advanced, attractive transit system that operates reliably and relatively rapidly, with smooth (ease of) synchronized transfers, part of the door-to-door passenger chain*. Interpretation of each component of this definition appears in Section 13.5. What follows are introductory remarks, a literature review of one possible future transit-operation design concept, and an overview of transit-system automation.

18.1.1 New concept using multi-agent systems

Advanced transit systems, such as dial-a-ride (e.g. Dial, 1995) or other demand-responsive systems, are not widely used and have never replaced existing conventional transit systems. The flexible routing and scheduling approach can be combined into a new concept consisting of elements suitable for a mass transit system. This new concept has the following main characteristics: (a) the use of a multi-agent system; (b) real-time, multi-legged trip planning, based on each passenger’s attributes; (c) transit stops independent of synchronized transfers; and (d) deployment of operational tactics (hold, skip stop, etc.) to overcome the stochastic nature of the transit system.

Models that treat the stochastic nature of the transit system apparently were not developed specifically for public transit, but for vehicle-routing models (e.g. Gendreau *et al.*, 1996;

Ghiani *et al.*, 2003). A public transit model that uses variable demand was developed by Chien *et al.* (2001), but it does not contain the combination of variable travel time, passenger demand and transfer probability that will be presented in this chapter.

Following the description of synchronized transfers in Chapter 6 and the reliability problems in Chapter 17, the reader will recognize the uncertainty of a planned simultaneous arrival of two vehicles at an existing stop. In order to alleviate this uncertainty of simultaneous arrivals, Sections 18.2 and 18.3 introduce a new concept; the extension of the commonly used single-point encounter (at a single transit stop) to a road-segment encounter (any point along the road constitutes a possible encounter point). Such a change in concept will reduce the uncertainty of meeting at a certain point and will enable more flexibility in deploying the operational tactics of Section 17.6.3 in Chapter 17.

The new concept, which relies on on-line information, allows two vehicles to meet in time and space so as to eliminate the need for waiting time at transfer points. Moreover, if we do not permit conventional transit systems to obstruct our imagination, the possibility exists of physically arranging for two vehicles that arrive at the meeting point to align together either longitudinally or side-by-side in order to allow passengers an easy transfer. This is like the meeting of satellites in space. Such a perfect transfer arrangement is illustrated schematically in Figure 18.1.

Realization of the new concept can be based on multi-agent systems (MAS) as described in Section 18.2. Van Dyke Parunak (1997) defined MAS as collections of autonomous agents within an environment that interact with each other for achieving internal and/or global objectives. Minsky (1986) argued that an intelligent system could emerge from non-intelligent parts. His definition of the 'Society of Mind' makes use of small, simple processes, called agents, each of which performs some simple action, and the combination of all these agents forms an intelligent society (system). Bradshaw (1997) classified agents by three attributes: autonomy, cooperation and learning; he defined four agent types from these attributes: collaborative, collaborative learning, interface and smart. The most interesting type so far as the public transit system is concerned is the collaborative agent. Such an agent is simple and can perform tasks independently, but can collaborate with other agents if necessary in order to achieve a better solution. Jiamin *et al.* (2003) developed an MAS for a bus-holding algorithm. The authors treat each bus-stop as an agent; the agents negotiate with each other, based on marginal cost calculations, to devise minimal passenger waiting-time costs.

Lastly, as we noticed in Chapter 14, designing a public transit system produces conflicts among the parties involved. The agency will usually prefer a small number of routes with transfer centres in order to reduce fleet size. The passenger will prefer more routes and fewer transfers because transfers naturally decrease comfort and reliability. This conflict may intensify in areas with variable and sparse demand. The solution is to combine the advantages of a transfer-based system (lower costs) with models that increase the comfort of transfers and reduce waiting and travel time. Table 18.1 summarizes these options.

18.1.2 Transit-system automation

Emerging technologies, both those being implemented now and those to be implemented in the near future, are surrounding us at an increased pace. The Internet and mobile phone revolutions are creating a technological breakaway from tradition that is affecting every aspect

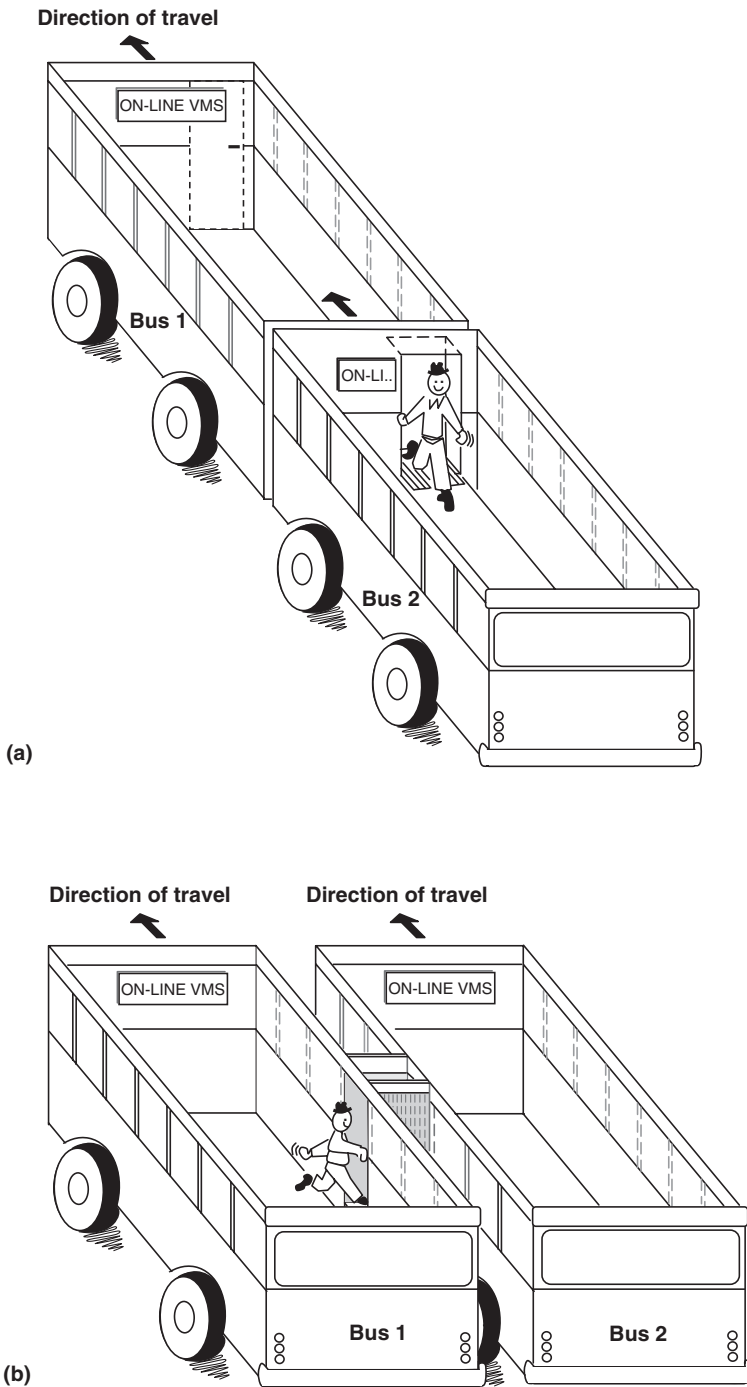


Figure 18.1 Schematic illustration of two buses aligning together longitudinally in Part (a) and side-by-side in Part (b), each allowing an easy transfer for passengers

Table 18.1 *Perception of transit options, by party*

Option \ Party			
	Passenger	Agency	Authority
More routes	Better service	More expensive	Better coverage
More transfer	Reduced comfort and reliability	Less expensive	More user complaints
Smart transfers	Reduced waiting and travel time	Less expensive and increased ridership	Increased use of public transit

of life; it is a revolution that cannot be stopped, and so affects transit operations, as well. One of the basic claims that is frequently made to delay or reject the introduction of a new technology is this: “Show us where these new systems are running in the world”. The answer to this question will be given shortly, once these systems are introduced. The revolution is happening, and it is essential in all fields of conveyance: baggage and cargo handling, elevator and walkway systems, industrial conveyors, and of course mass transit. The quest for what new technologies are searching is echoed in the advertising of many large logistics companies, “We take Anything, Anywhere, Anytime”. Even the European Union Voyager project (EU Report, 2005), the main research project on the future of transportation, concluded that what is needed and what the public demands is a high quality, door-to-door ‘mobility service’.

The automation revolution has been realized in the past few years, especially in the fields of computers, telephones and mobile networking. The automation revolution has brought technology to the point where creating an automated transit system is a workable option. Rosenbloom and Fielding (1998) in TCRP (Transit Cooperative Research Program) Report 28 state that single-occupant drivership can only be reduced with reverse-transit solutions, services to employers or universities, pooling incentives, and restructuring routing and feeder services. These changes are actually being implemented with the aid of the new technologies.

Another important aspect of transit-system automation is the development of special information technologies. Two technologies deserve attention: RFID (radio frequency identification) and Proximity Information. RFID utilizes a radio frequency transmitter and receiver to identify items and their location. Usually, these are extremely small, low-cost devices that use very little or no power and draw their needed power from the air; such devices can be used to identify transit-vehicle locations and driving paths, prevent the theft of goods, and enable cashier-less supermarkets.

Proximity Information, termed the ‘Location Revolution’, is a technology that started around 2001 for finding locations; for instance, giving correct route and map directions, creating personalized advertisements according to the types of vehicles or mobile phones passing by, and even locating lost children. Together, these two technologies are paving the way to produce automated transit vehicles and driverless systems, as well as on-line tactics to improve transit reliability.

18.2 Multi-Agent Transit System (MATS)

The arguments of Chapters 6, 13 and 14 demonstrate that the use of transfers in public transit has the advantages of reducing operations costs and introducing more flexible, efficient route planning. In contrast, the main drawback, from the passengers' point of view, is the inconvenience of travelling multi-legged trips. In the attempt to diminish waiting time caused by transfers, the following sections will elaborate the new Multi-Agent Transit System (MATS) concept.

The MATS proposed here will be composed of the following agents: public transit vehicles, passengers, road segments, transit agencies and transit authorities. In order to construct the system as a whole, it is necessary to define each agent and the interrelationship of agents. Since the system is complex and cumbersome, each agent will be explored separately; this will also make it easier to define the system's elements. Figure 18.2 presents the main activities and interrelationships of the proposed system.

The following is a short description of the main activities of each agent shown in Figure 18.2.

Passenger agent: The passenger agent plans the trip according to passenger preferences, and based on the real-time information available from road-segment agents (travel time) and vehicle agents (routes and dwell times). Passengers will input to their cell phones or personal digital assistant (PDA) every trip desired from point A to point B. The agent, referring to each passenger preference, will search for the best trip based on the public transit data available at that time. The passenger will choose among the possible options and book a trip in the system. Then the passenger will be notified by SMS or a different means of communication about path changes owing to, for example, traffic congestion or a system-wide optimization deploying tactics that change the planned schedule of route legs. In response to the route changes, the passenger agent will try to find a better route, if possible changing the existing planned path. The agent itself can be a small software program running on a cell phone or PDA.

Road-segment agent: The road-segment agent can reside physically, as part of the road infrastructure (e.g. at a traffic-light control), or virtually, as part of a multi-agent software system. The agent is responsible for the following activities: travel-time estimation, encounter probability estimation, improving the system's objective function, and instructing vehicles on tactical deployment. Each road-segment agent continuously collects local traffic-flow information and estimates the travel time. The agent evaluates the encounter probability (for vehicles transferring passengers), based on the adjacent road-segment travel-time estimations and vehicle locations. This probability is described in the next section. Using dynamic programming, each agent or group of agents calculates the optimal tactical deployment that will optimize the system's objective function, which is the total expected travel time.

Vehicle agent: The vehicle agent can be part of the onboard AVL system or, virtually, part of a multi-agent software system. The agent estimates the vehicle's dwell time according to booked trips and demand forecasts, using travel-time estimations (from the road-segment agents) and estimated deviations from the planned timetable schedule. The vehicle agent receives instructions for the deployment of tactics from the road-segment agent in order to improve the system's objective function.

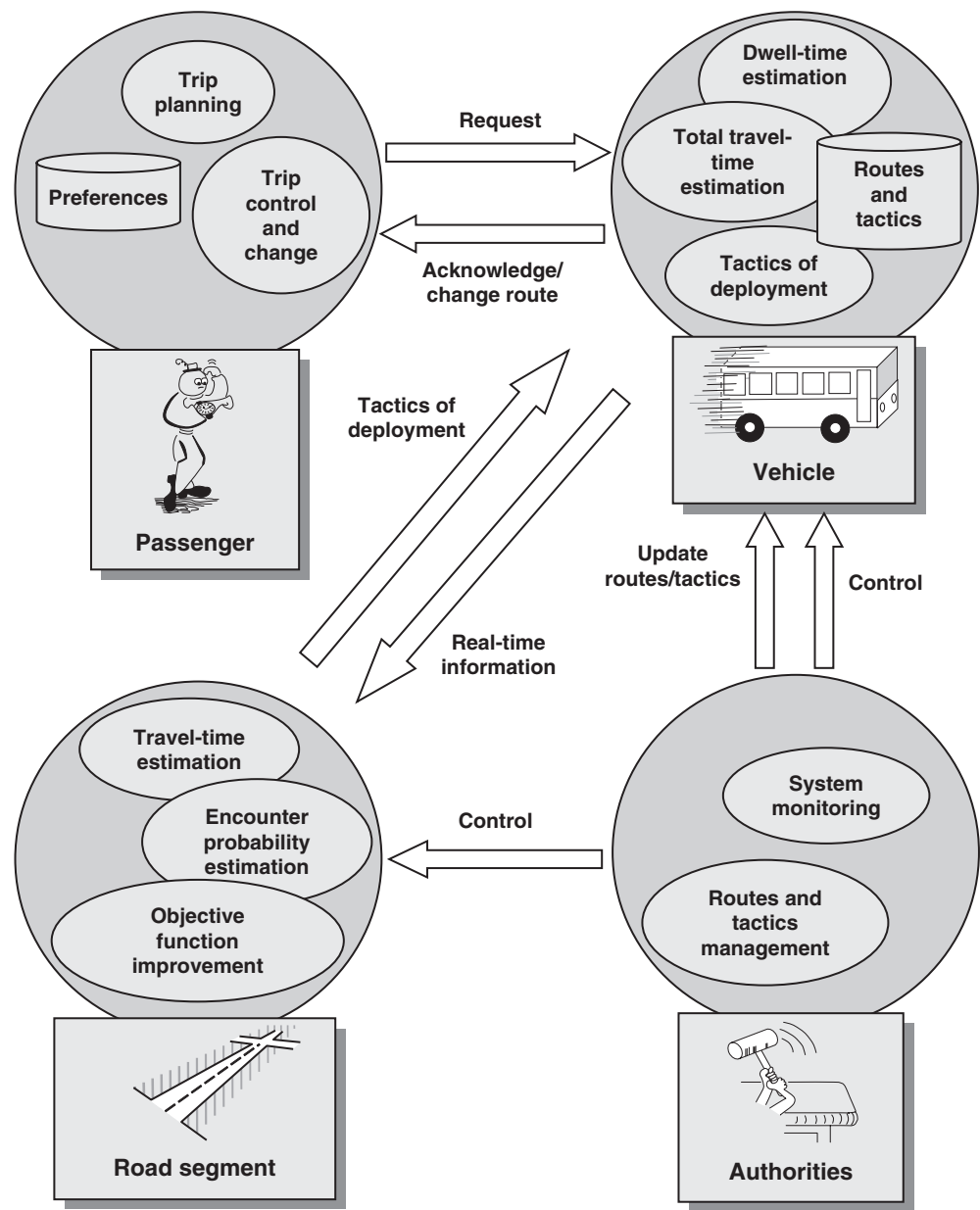


Figure 18.2 Agents' activities and relationships

Agency agent: This agent is responsible for designing and managing the transit network, for updating timetables, and for configuring the possible operational tactics available on each road-segment for each route.

Authority agent: The authority (local authority or federal government) is responsible for monitoring system performance according to the determined/decided indicators.

The MATS offers the following benefits, which are inherent in the multi-agent approach.

Extensibility: Easily allows the system's growth and adding of new resources. Each vehicle has computation power to contribute to the entire network. Adding a new vehicle is similar to plugging in a new computer to a local area network (LAN).

Fault tolerance: The proposed system will handle failures. Critical operations that are heavily dependent on computation and that are built on a standard central computing architecture must have a redundancy system in order to maintain a certain level of service. Redundant systems are expensive and cumbersome; however, MAS agents are distributed. Consequently, if some agents are down, the others will continue to perform because of their autonomous capability. Table 18.2 presents the mode of operation and outcome for different communication scenarios in case of communications disturbances.

Table 18.2 Communication scenarios

Scenario	Mode of operation	Outcome
Full communication	On-line collaboration	Synchronized transfers; total travel-time reduction
Partial communication		
No communication	Autonomous; according to timetables	Ordinary transit system with reliability problems

Scalability: Distributed systems can theoretically grow without limit (e.g. the Internet).

Adaptability: Changing rules and transmitting data to the agent is quick and simple, similar to the spreading of a virus.

Efficiency: Negotiations between agents can reach an optimal or a near-optimal solution efficiently (e.g. see Raiffa, 1982).

Distributed problem solving: Cooperating agents can distribute sub-tasks to other agents that are idle and can contribute their computer time to solve these sub-tasks (e.g. see Smith, 1980, and Davis and Smith, 1983).

Stability: The use of a closed set of operations tactics for each road-segment in order to eliminate solution sets that are not stable.

The following is a simple example of the benefits of MATS. On-board a shuttle heading for the train station are 30 passengers, of whom 20 plan to take the 12:20 train. The train headway is 30 minutes, and trains are assumed to depart on-time. Along the shuttle route, 4 more passengers are waiting (say, for the system's call-back; see Chapter 16) to take this shuttle service. According to the estimated arrival time, there is a probability of 0.60 of reaching the train station at 12:15 and a probability of 0.40 of reaching the station at 12:25. In this case, the expected total waiting time for the train will be $(0.6 \cdot 5 + 0.4 \cdot 25) \cdot 24 = 312$ minutes. If the shuttle deploys a short-cut tactic, the estimated arrival-time probability will be increased to 0.95 for 12:15, in which case the 4 waiting passengers will miss the 12:20 train and will have to board the next shuttle (the average headway being 20 minutes).

The total new waiting time for the 20 passengers will be $(0.95 \cdot 5 + 0.05 \cdot 25) \cdot 20 = 120$ minutes. For the existing 4 waiting passengers, and assuming that the next shuttle for the 12:50 train does not execute any operational tactic, the total waiting time will be $(0.6 \cdot 5 + 0.4 \cdot 25 + 20) \cdot 4 = 132$ minutes. In terms of system optimization, the short-cut tactic results in $120 + 132 = 252$ minutes total waiting time, compared with 312 minutes otherwise; thus, the tactic will be preferred.

18.3 Vehicle encounters on road segments

MATS encapsulates activities and processes that the agents continuously perform in parallel with one another. Some of these activities have been described in this book and can easily be integrated into MATS: estimating travel time (see Section 17.3.2 in Chapter 17), estimating dwell time (see Section 17.3.1 in Chapter 17), and finding the best route (see Chapters 14 and 16). However, this section presents a step-further method, in which MATS incorporates new modelling concepts rooted in the determination of encounter probabilities.

18.3.1 Defining an encounter probability

Transit-vehicle behaviour can be illustrated by a time-space trajectory, an example of which is shown in Figure 18.3. The notations S_0 and S_3 represent the entrance and exit of the road-segment, respectively; t_1 is the entry time of the vehicle to the road segment; t_2 and t_4 are the arrival times at Stops S_1 and S_2 , respectively; and t_3 and t_5 are the departure times from these stops. The slopes from t_1 to t_2 and from t_4 to t_5 show the average travel time (based on average speed) between two consecutive stops. The average travel speed depends on traffic-flow characteristics, speed limits, geometric design, and driver behaviour. The

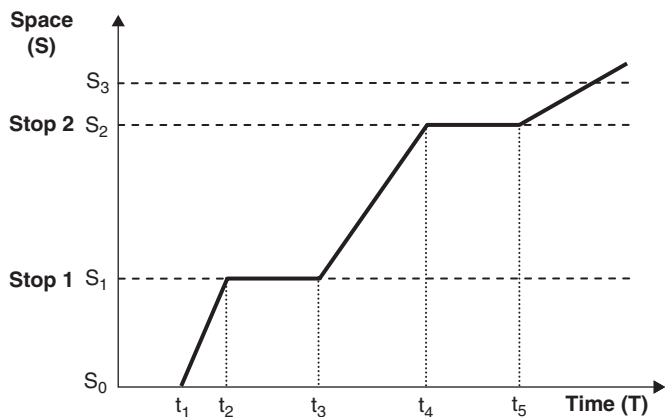


Figure 18.3 Time-space (T-S) diagram illustrating transit-vehicle trajectory behaviour on a road-segment

dwelling time depends on the number of passengers boarding and alighting, vehicle configuration (number of doors, entry and exit doors and payment methods), and driver behaviour.

The probability of an encounter between two transit vehicles, especially buses, is a function of road-segment characteristics and vehicle travel characteristics. Road-segment characteristics include segment length, number of stops, distance between adjacent stops and travel-time estimation. Vehicle characteristics include road-segment, entry time and the number of passengers boarding and alighting at each stop. The following assumptions are made: (a) a planned encounter (based on a given timetable) exists for two vehicles on a road segment; and (b) segment travel time for an arrival vehicle is dependent on the travel time of the preceding vehicle. The first assumption is essential because transfers are planned on pre-defined road segments (where two transit vehicle lines intersect) rather than on ad-hoc transfers. The second assumption, more realistic in urban bus systems, assumes the dependency of travel times between sequential vehicles (experiencing similar traffic characteristics) as long as their headway is not too large.

Transit-vehicle behaviour is taken into account when constructing an estimation model for the probability of an encounter. Such a model considers that: (1) randomness exists for travel time and dwelling time; and (2) dependency exists between the travel-time distributions of the vehicles examined. Figure 18.4 illustrates, through a time–space diagram, an example

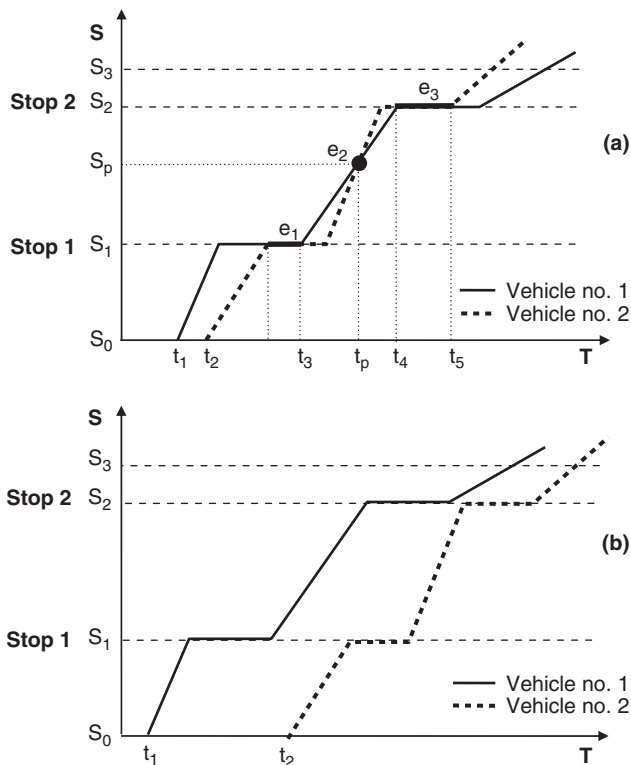


Figure 18.4 (a) Public transit vehicles making an encounter; (b) public transit vehicles missing an encounter

of possible encounter situations between two vehicles and a possible miss of encountering. Part (a) in Figure 18.4 shows two vehicles that encounter each other three times (designated e_1 , e_2 and e_3) along the road-segment: while at Stop 1, along the segment between stops, and while at Stop 2. In contrast, part (b) of Figure 18.4 shows the case in which, because of a large headway, the two vehicles missed their planned encounter; as a result, the passengers who wanted to transfer missed their connection and will experience a longer wait.

18.3.2 Estimating encounter probability

A JAVA-based simulation was developed for estimating encounter probabilities. The simulation emulates the behaviour of two vehicles along a road-segment, given road-entry time distributions, stop locations on the road-segment, the number of passengers (boarding and alighting at each stop), and travel-time distribution parameters between stops. Each vehicle acts as an independent entity, using the threading capabilities of JAVA (the ability to emulate the parallel processing of the simulation entities).

Time-space trajectories are constructed for each simulation run. With the use of line-intersection techniques, the number of encounters (point and trajectory intersections as in part (a) in Figure 18.4) is accumulated as is the number of missed encounters. Using this information, the following formula may be established:

$$EP = \frac{SR - ME}{SR} \quad (18.1)$$

where EP is the encounter probability estimator, SR is the number of simulation runs and ME is the number of missed encounters.

This simulation model can be integrated into the road-segment agent. The calculation of EP will take place continuously for each group of transit vehicles planned for transferring passengers.

18.3.3 Deployment of tactics

The stochastic nature of transportation networks enforces a reduction in encounter probabilities. Hence, in order to improve encounter probabilities, a change of tactics must be imposed on the vehicles. A list and description of possible on-line tactics in transit operations appears in Section 17.6.3 in Chapter 17. The main real-time control tactics for the encountering-coordination problem are as follows:

- Holding the vehicle (at terminal, mid-route point).
- Skip-stop operation.
- Changes in speed (but not above the lawful speed limit).
- Short-turn operation.
- Short-cut operation.

A set of tactics will be available for deployment for each road-segment and vehicle. The particular tactic, or tactics deployed, will be decided by the transit agency. For example, the agency may decide to employ a short-cut tactic if, according to demand forecasts, passengers seldom wait to embark along the section(s) skipped.

18.3.4 Improving encounter probabilities

Chapter 17 indicates that an important aspect of operating a public transit system is the use of on-line information to change vehicle routes and/or timetables dynamically. The changes are made in order to increase the system's objective functions (total travel time and encounter probability for transferring passengers). Estimating an encounter probability is crucial to deciding whether or not an encounter is plausible. Moreover, information of a maximum encounter probability is important for deciding whether it is worthwhile to change tactics (which might cause increased travel time for waiting passengers).

In order to achieve maximum overlap of time-space trajectories, a time shift must be placed on a vehicle's road-segment entry time. Figure 18.5 illustrates an overlapping situation between two vehicles that depart at the same time from Stop 1. Note that their dwell times at Stops 1 and 2 are not the same. This S-shaped overlap has a time length of $(T_b - T_a)$ and a space length of $(S_2 - S_1)$. Two areas of Figure 18.5 are defined by A_1 and A_2 in units of space multiplied by time; A_1 is between S_1 and S_0 , and A_2 is between S_3 and S_2 . The simulation program calculates these areas for an average speed between each two adjacent stops for each vehicle. These areas are used to calculate an average weighted headway, H_w , in time units between the two vehicles along the road segment as follows.

$$H_w = \frac{A_1 + A_2}{S_3 - S_0} \quad (18.2)$$

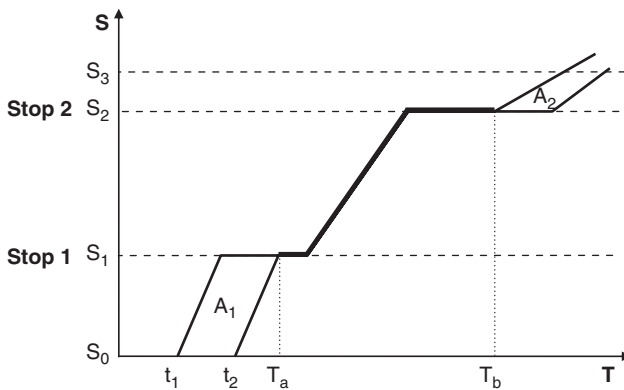


Figure 18.5 Time-space diagram illustrating an overlapping situation

The rationale for Equation (18.2) is that the smaller the H_w , the larger is the overlapping area. For two parallel trajectories, H_w will be the same as the constant headway between the two trajectories. An algorithm was constructed, first, to calculate H_w for each S-shaped overlap that exists between two adjacent stops; second, to determine the overlapping case in which H_w is the minimum.

An example of calculating encounter-probability was constructed with four scenarios of different headways between two vehicles once they entered a road segment. For each scenario, H_w was calculated, and the estimate of the encounter probability was extracted from 100 simulation runs. Table 18.3 summarizes the results. It can be seen that Scenario (c) has the smallest H_w and the maximum encounter probability.

Table 18.3 Simulation results

Scenario	Headway at entry point (sec)	Weighted average headway (sec)	Encounters	Missed encounters	Encounter probability
(a)	220	153.6	90	58	0.42
(b)	130	51.4	263	7	0.93
(c)	105	45.7	400	0	1.00
(d)	10	59.3	144	25	0.75

Another part of presenting the results, involves the observation of a three-dimensional (3D) figure of the simulation outcome; that is, 3D histograms of time–space encounters. These histograms show where the encounters are most likely to occur. Such a histogram, aside from being a tool for visualizing the simulation, can help the transit planner compare the most probable encounter locations. This information, combined with the geometric features of the road segment and the existing location of the transit stops, may help in relocating the stops so as to expedite the transfer process. Figures 18.6 and 18.7 present two different 3D histograms of the simulation runs. Figure 18.6 depicts an urban-scenario example (a 3.5 km stretch of road-segment), in which most encounters occur at the transit stops rather than along the road-segment.

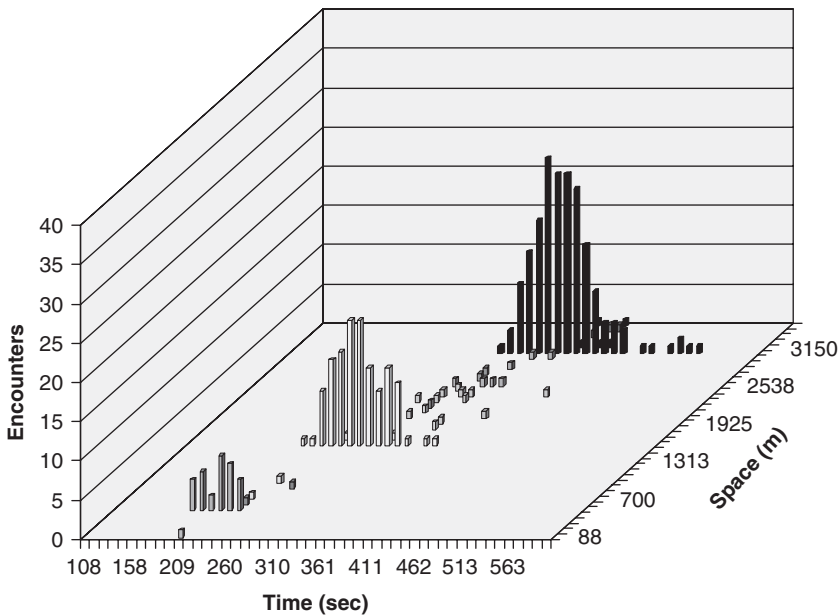


Figure 18.6 A time–space encounter histogram for an urban scenario

Figure 18.7 illustrates the example of an intercity scenario. In this example, the road-segment is ten times longer (35 km) and most of the encounters occur along the segment rather

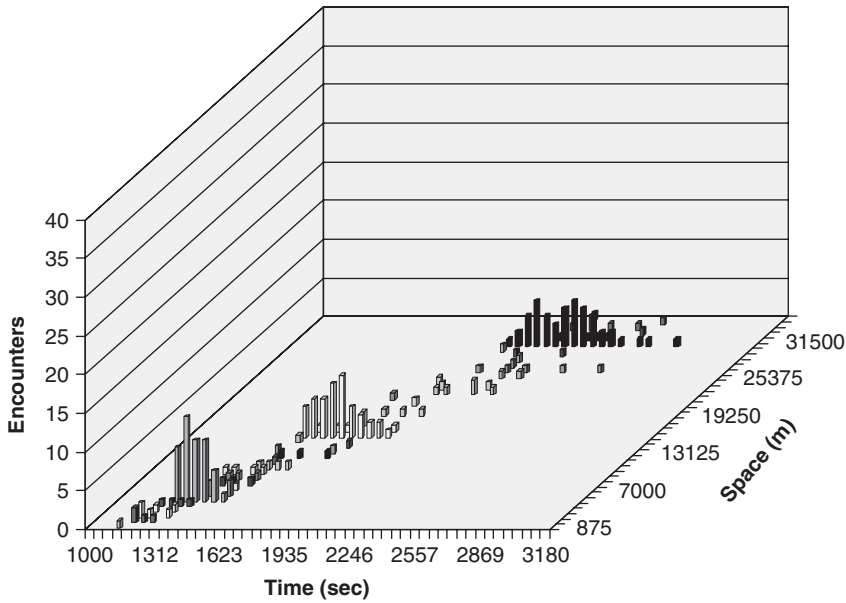


Figure 18.7 Time-space encounter histogram for an *intercity scenario*

than at transit stops. The importance of allowing transfers other than at stops is emphasized by the encounter probability of 0.78 when encounters are permitted everywhere, compared with 0.54 when encounters are permitted only at the stops. Prohibiting not-at-stop transfers does not directly imply a reduction in the encounter probability, as vehicles can match speeds and perform a transfer at the next stop; however, it inflicts a time loss because of schedule changes (being not on time), not to mention a reduction in reliability performance.

18.3.5 Objective function improvement

Estimating encounter probability adds a new operational objective function; namely, the total transit-system travel time (waiting, on-board and transfer times). The establishment of this function will now be described. $V(i, j)$ is designated as the vehicle number on the i -th route leg for the j -th passenger in constructing the following equation:

$$PTT_{i,n} = \sum_{j=1}^{n-1} (TT_{V(i,j)} + ETR_{V(i,j),V(i,j+1)}) + TT_{V(i,n)} \quad (18.3)$$

where $PTT_{i,n}$ is the total travel time for passenger i , the trip having n legs; $TT_{V(i,j)}$ is the estimated travel time on leg j ; and $ETR_{V(i,j),V(i,j+1)}$ is the expected transfer time between leg j and leg $j + 1$. The expected transfer time is

$$\begin{aligned} ETR_{V(i,j),V(i,j+1)} = & EP_{V(i,j),V(i,j+1)} \cdot TR_{V(i,j),V(i,j+1)} \\ & + (1 - EP_{V(i,j),V(i,j+1)}) \cdot [Pr_{V(i,j),V(i,j+1)} \cdot W_{V(i,j),V(i,j+1)} \\ & + (1 - Pr_{V(i,j),V(i,j+1)}) \cdot W_{V(i,j)+1,V(i,j+1)}] \end{aligned} \quad (18.4)$$

where $EP_{V(i,j),V(i,j+1)}$ is the encounter probability, $TR_{V(i,j),V(i,j+1)}$ is the direct transfer time; $Pr_{V(i,j),V(i,j+1)}$ is the conditional probability that vehicle V_1 will arrive ahead of vehicle V_2 , given that they will not meet along the road-segment; and $W_{V(i,j),V(i,j+1)}$ is the headway between two vehicles.

The new objective function is the total travel time obtained by the sum of all passengers' travel times as per Equation (18.3). Such an objective function can be optimized by a dynamic programming (DP) method (see Section 5.3 in Chapter 5 on operations research, and Appendix 10.A in Chapter 10 for an example of DP formulation). Following the calculation of both current and maximum encounter probability estimators, operational tactics (discrete decision variables) are examined for minimizing the new objective function. Combining the distributed computing capabilities of MAS with the DP approach can produce an efficient algorithm for on-line use in optimizing real-world public transit systems.

18.4 Developments in transit automation

Several automatic transit systems have already been implemented, using such technologies as magnetic location, distributed networking, and linear induction motors (LIM). The most common, well-known automated transit systems are automatic train operations (ATO), automatic people movers (APM), elevated rail systems for automatic guided vehicle (AGV), and personal rapid transit (PRT).

Examples of known ATO systems are the Paris Metro, BART (California), Metrorail (Washington, DC) and MTR (Hong Kong). Known APM systems are implemented especially in airports, such as JFK, Newark, Düsseldorf, Frankfurt and Schiphol in Amsterdam. APM systems are also available in light-rail metro systems in Toronto, Vancouver, Detroit and New York.

18.4.1 Elevated rail systems

Elevated monorails have been in use for more than 110 years, or since 1890, with not one known fatality. Outside the city, elevating systems benefit nature by not creating an ecological barrier. An elevated system cannot crash with cars; it consumes less land and it is not susceptible to traffic congestion. In some noted cases, elevated rail was the only system able to work after heavy earthquakes that brought down bridges and freeways; e.g. Red Line California 1994, Osaka 1999, and Seattle 2001.

In urban areas, the problem with elevated rail systems usually concerns the right-of-way for residents who live on floors at the level of travel. Certainly, this is not a problem at airports or business districts. Heavy monorails as a solution are still extremely expensive, and once installed have further difficulties.

The advanced systems – APM, AGV, and PRT – constantly failed to be realized over the past 30 years. However, some progress has recently been made in that direction. In London, Heathrow Airport has ordered a limited system that employs the ULTRA concept, a PRT-prototype system developed in Wales for the European Union, to take passengers from the long-term parking lot to the main airport terminal. In South Korea, a system that evolved from Raytheon's PRT is being installed. At Uppsala University in Stockholm, Sweden, the European Union is finally building a large test track for an automatic system, code-named 'Vectra'. And

in Dubai City, UAE, there is now a tender for PRT to work a large-scale city and hotels loop. Furthermore, a research project is in progress in Germany for an automatic urban railcar system (to operate on the NeoVAL system). Figure 18.8 illustrates several PRT systems and one AGV system related to these projects.

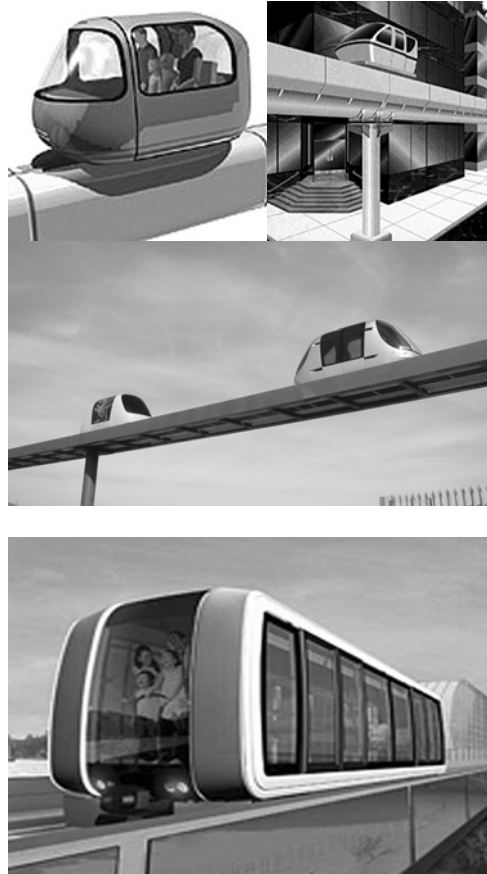


Figure 18.8 PRT (upper illustrations) and AGV systems

18.4.2 Elevator-PRT and dual-mode concepts

Often, a PRT system is seen as a limited solution because it does not take into consideration heavy cargo or large groups of people, such as big families or school classes. In some city transportation committee meetings that are debating light-rail solutions, PRT is usually brought up to show how ‘futuristic’ a solution it is, and then discarded (PRT, 2004). One possible, and faster, way to introduce PRT systems may be via the elevator industry, which has been searching for a true ‘door-to-door’ elevator. In other words, an elevator that can go between buildings and reach parking lots. The joint elevator-PRT concept has been

exhibited in several futuristic movies. Figure 18.9 illustrates such an example as visualized for the movie *Minority Report*.



Figure 18.9 Joint elevator-PRT concept envisioned for the movie, “*Minority Report*”

Finally, there is a dual-mode concept suggested by Reynolds (2006). This consists of environmentally clean transit vehicles and automobiles that use guideways at different speeds (location-based) as one (main) mode and ordinary driving as a second mode. The ‘driving’ on the guideways is done by an automatic computer-controlled system. Reynolds justified his concept with the following argument: “A hundred years ago most rural people got water into their houses by carrying it in buckets from a well in the yard. Those old buckets delivered very little water compared to modern plumbing, since a pipe can provide continuous flow. The same observation applies to transportation: Buses and trains are like one-at-a-time buckets. Cars are supposed to run or flow continuously on the highways, and they will flow continuously on the guideways. The big spaces we leave between cars on the highways are necessary because of the unsynchronized traffic, the limitations of human drivers, and the limitations of automobile braking and tire traction”.

18.5 Literature review and further reading

Recent years have seen intensive discussion in the literature about various issues relating to transit systems of the future. Naturally, much of this discussion concerns technological

development; however, some conceptual issues, mode-specific issues, and others have been analysed, as well. Some of these works are reviewed in this section.

18.5.1 Overview

Several publications present an overview of new technologies in transit. In a very comprehensive review of advanced transit systems, the US Department of Transportation (2000a) divides these systems into five groups. The first group consists of fleet-management systems, which include automatic vehicle location, transit-operations software, communication systems, geographical information systems, automatic passenger counters and systems for traffic priority. The other groups are traveller-information systems, electronic payment systems, systems for demand management and intelligent transit vehicles. The discussion of each group contains a description of the technology, its benefits, a review of the state of the art and existing systems and examples of application.

In a separate publication, the US Department of Transportation (2000b) identifies and quantifies the benefits derived from current applications of the advanced systems described in the previous publication. A very detailed discussion of the various technologies addresses their benefits and costs, including monetary values based on actual experience. Applications of fixed bus routes, demand-responsive transit, and various rail types are presented.

Blythe *et al.* (2000) present an overview of intelligent transportation system (ITS) applications on public transit. The authors identify specific advantages that ITS can offer transit in terms of travel-demand management, infrastructure management, vehicle management, information provision and multimodal ticketing. The paper then describes a range of applicable technologies, including smart cards, bus priority systems, automatic vehicle location, trip planning and on-board information systems.

18.5.2 Advanced systems categorized by technology

A considerable number of papers focus on new technologies that are presently used in various transit systems. Several more recent examples follow. Further literature review of related subjects appears in Section 17.7 of Chapter 17.

The World Wide Web

Vorvick and Dueker (1998) describe an application that delivers real-time bus-schedule information to Internet users. The system enables a bus rider to receive information on the schedule deviation of a specific bus at a specific time. Fingerle and Lock (1999) present considerations in choosing the architecture of a journey-planner system on the Internet, as well as practical issues for its design. The discussion refers to various levels of geographical coverage and to a case in which the data contains information about numerous transit companies and modes. The authors define the content and structure of the database and the actions performed by the system in response to complex queries. Stone *et al.* (2001) describe an Internet-based system for the support of a transit managers' decision. The system is intended to help transit managers identify their transit needs and choose potential technological solutions; it combines various decision-making factors, such as transit-service area, service type, daily ridership, and fleet size in order to assist in matching appropriate

technologies to various needs. Sun and Bernstein (2002) demonstrate the advantages of using dynamic XML (rather than HTML) to provide transit timetable information on the Internet. Maclean and Dailey (2002) discuss the provision of real-time bus-departure information to users of mobile devices with wireless Internet access. The system makes use of real-time vehicle reports to predict travel times to future locations. Trepanier *et al.* (2002) present a hybrid algorithm that has been implemented in transit-information websites. The algorithm uses heuristics for the calculation of urban transit itineraries, including information on pedestrian access and egress paths, route sequences, schedules and stops. A 'totally-disaggregate approach' and object-oriented modelling are used for specifications of the pedestrian network, the route network and passenger demand.

Geographical Information Systems (GIS)

Huang and Peng (2002) design an object-oriented GIS data model for transit-trip planning systems. To enable efficient data retrieval, despite the complex, dynamic nature of transit networks, the authors model the components of the transit network as space-time entities that have start times, end times and lifetime spans. A time-map object controls the creation and destruction of these objects; in the process of a network search, only the active components of the system are built into the topology.

Web-based GIS

Both Smith (2000) and Peng and Huang (2000) describe systems that combine the advanced processing abilities of GIS with the availability of the Internet. The systems enable users to interactively plan their transit-trip itineraries, often using friendly graphical features. The papers discuss such issues as the design of a suitable path-finding algorithm and the creation of the database.

Automatic Vehicle Location (AVL)

Greenfield (1999) develops a methodology for evaluating the accuracy of an AVL system in which bus location is presented on a digital map and for testing whether bus location is consistent with map coordinates. Horbury (1999) illustrates how historical AVL data can be used to identify segments of a bus route that would benefit most from bus-priority measures and to improve scheduling by highlighting locations with the greatest deviation from schedule. The paper also presents a methodology that uses historical AVL data and on-bus passenger counts to calculate the passenger-arrival rate at stops and to estimate annual patronage, bus speeds, and dwell times. Dailey *et al.* (2001) offer an algorithm to predict the arrival time of a transit vehicle up to one hour in advance; the method is based on time series created from AVL data using various historical statistics. Lee *et al.* (2001) study the effect of an AVL system on schedule adherence and operator behaviour. Gillen *et al.* (2001) develop measures of performance productivity for transit systems of varying sizes and locations; they use these measures as a basis for examining the potential contribution of alternative AVL applications. Hounsell and Wall (2002) identify eight alternative architectures for AVL systems combined with urban traffic-control systems and review their implementation.

Global Positioning System (GPS)

It should be noted that GPS is often used as a method of AVL; since many papers discuss GPS applications separately, we will do the same here. Lin and Zeng (1999) discuss real-time predictions of bus-arrival times using GPS data. They develop arrival-time-estimation algorithms and deal with the issues of route representation and GPS data screening for identifying data quality and delay patterns. Lin and Padmanabhan (2002) describe the development of procedures for creating bus-route information as part of an urban transit-inventory database; GPS is used to automate the generation of the inventory. A method is presented for capturing route information on the basis of a sequence of vehicle-location data.

18.5.3 Advanced systems categorized by purpose

Some papers place the main emphasis, not on the type of technology used, but on its purpose. Following is a review of important publications of this sort.

Passenger information

Gildea and Sheikh (1996) compare various applications of technology in systems that provide information to transit users. They then focus on an Internet-based information service that provides route, schedule and fare information covering a dozen transit operators. Peng and Jan (1999) assess different means of transit-information delivery. They define several types of information systems (pre-trip, in-terminal, in-vehicle) and describe various available technologies (audio media, video media, multimedia). An evaluation framework is developed to assess different media on the basis of accessibility, versatility, interactivity, capacity, friendliness, cost and ease of implementation. It should be noted that many of these papers (as do studies categorized by technology) describe passenger-information systems (e.g. Lin and Zeng, 1999; Dailey *et al.*, 2001; Maclean and Dailey, 2002; Lin and Padmanabhan, 2002).

Data collection

Turner (1996) analyses advanced techniques for travel-time data collection: electronic distance-measuring instruments, computerized and video number-plate matching, cellular phone tracking, automatic vehicle identification, automatic vehicle location and video imaging. These are not specific to transit, but have many related applications. For each technique, the author discusses the necessary equipment and procedures, applications, advantages and disadvantages. Several papers develop methods for bus arrival-time prediction based on data collected by AVL systems or passenger-counting systems. The model proposed by Jeong and Rilett (2005) uses an artificial neural network concept.

Bus priority

Balke *et al.* (2000) describe the concept of intelligent bus priority at signalled intersections. The concept uses real-time information about the bus location to predict when, within the signal cycle, the bus should be given priority; this is done by using phase extension, phase insertion, or early return strategies without losing overall coordination. Arrival time at bus stops along the way is taken into account.

18.5.4 Advanced systems categorized by mode

Some publications focus on advanced systems that suit a certain transit mode. These are illustrated in the following paragraphs.

Bus

Levinson (2001) presents an overview of the role of bus transit in the twenty-first century; suggested opportunities for further development and progress include policy issues, management issues, service design, choice of vehicles, and technology. Hounsell and Wall (2002) summarize applications of intelligent systems to improve bus services, among them AVL, bus priority, automatic ticketing systems, automatic camera-enforcement systems, and variable message signs. Siuru (2003) summarizes a broader review of bus systems that attempt to meet sustainability criteria. Bus Rapid Transit systems are highlighted, but various other technologies are also discussed in the context of achieving energy savings, improved air quality, reduced emissions, improved performance, increased capacities, greater durability and longer service life.

Light rail

Campion *et al.* (2000) discuss future opportunities and technological changes with light-rail transit. The opportunities discussed relate to the influence on urban form, concentrated development around stations, multimodal corridors, partnerships, new passenger markets, facilitation of multimodal connections and phase implementation of lines. Technological changes include barrier-free fare collections, track and roadbed, and traction electrification.

Personal Rapid Transit (PRT)

The idea of intelligent public vehicles for private use has been discussed in the literature for a decade, but some small-scale cases of implementation have induced a new debate over practical issues. Anderson (2000) reviews the rationale for PRT and the process needed to develop it. Glazebrook and Subramanian (1997) offer a detailed discussion of Personal Public Transport, which is based on a concept similar to PRT. Several other issues were brought up earlier in this chapter.

Demand-responsive transit (DRT) and paratransit

Many authors anticipate that future transit will be more heavily based on demand-responsive services; the design and assessment of such services are discussed in many papers. Stone *et al.* (1993) evaluate software for a computerized paratransit operation. Dial (1995) presents a dial-a-ride service, based on autonomous vehicles that are able to negotiate with one another in order to assign passengers to the vehicle representing the least-cost approach. Farwell (1998) discusses the operation of routes with flexible itineraries. Spring *et al.* (1998) analyse aspects of employing AVL technologies for paratransit services. Bakker (1999) describes the experience gained while introducing large-scale DRT systems; the focus is on costs, benefits, level of use, and other practical issues. Benjamin and Sakano (2000), who present a geographical information system applied to a dial-and-ride service, analyse

the quality of the service and consumer response. Horn (2004) develops an algorithm for planning multi-legged trips that can minimize travel time on both DRT and fixed routes.

18.5.5 The future of public transit

Several publications offer conceptual prophecies in regard to the transit services of the future. Rosenbloom and Fielding (1998), who study transit markets of the future, make more practical recommendations. They discuss the effect of various trends (demographic, economic, social, land-use and transport policy) on the demand for transit, and identify opportunities for maintaining current transit markets and creating new markets. Kane (2000), who conducted a survey among urban planners, historians and architects, formulated several likely characteristics of future transportation (not just transit) systems. Some of them are very relevant to the implementation of future transit policies, such as privatization, an increased dependence of revenues on actual usage, greater integration of transportation modes, personalization of transit services, and increased environmental friendliness of vehicles.

18.6 Concluding remark

This is the last section of the book, and thus it deserves a final word, which will be given by way of symbolism. There is a Swedish proverb: “Worry often gives a small thing a big shadow”, about which Marie Curie said: “Nothing in life is to be feared. It is only to be understood”. Understanding the nature of transit operations problems is already half way towards reaching adequate solutions. Finally, an African proverb cautions: “Smooth seas do not make skilful sailors”. Indeed, experiencing problems in transit operations provides an important lesson that, together with the solutions and thoughts proposed in this book, can hopefully end with a significantly improved transit service.

References

- Anderson, J. E. (2000). A review of the state of the art of personal rapid transit. *Journal of Advanced Transportation*, **34**, 3–29.
- Bakker, P. (1999). Large-scale demand responsive transit system – A local suburban transport solution for the next millennium. *Proceedings of Seminar E* held at the AET European Transport Conference, Homerton College, Cambridge, UK, **P433**, 109–125.
- Balke, K. N., Dudek, C. L. and Urbanik, T. (2000). Development and evaluation of intelligent bus priority concept. *Transportation Research Record*, **1727**, 12–19.
- Benjamin, J. M. and Sakano, R. (2000). Issues on the application of an advanced public transit system to dial-a-ride service. *Journal of Public Transportation*, **3**, 49–65.
- Blythe, P. T., Rackliff, T., Holland, R. and Madeean, J. (2000). ITS applications in public transport: Improving the service to the transport system user. *Journal of Advanced Transportation*, **34**, 325–345.
- Bradshaw, J. M. (1997). *Software Agents*. MIT Press.

- Campion, D. R., Larwin, T. F., Schumann, J. W. and Wolsfeld, R. P. Jr. (2000). Light rail transit: Future opportunities and changes. *Transportation in the New Millennium*. <<http://gulliver.trb.org/publications/millennium/00066.pdf>>.
- Chien, S. I., Spasovic, L. N., Elefsiniotis, S. S. and Chhonkar, R. S. (2001). Evaluation of feeder bus systems with probabilistic time-varying demands and non additive time costs. *Transportation Research Record*, **1760**, 47–55.
- Dailey, D. J., Maclean, S. D., Cathey, F. W. and Wall, Z. R. (2001). Transit vehicle arrival prediction: Algorithm and large-scale implementation. *Transportation Research Record*, **1771**, 46–51.
- Davis, R. and Smith, R. G. (1983). Negotiation as a metaphor for distributed problem solving. *Artificial Intelligence*, **20**, 63–109.
- Dial, R. B. (1995). Autonomous dial-a-ride transit-introductory overview. *Transportation Research*, **3C**, 261–275.
- EU Report (2005) *Voyager: Transportation in 2020. A Project Approach*. European Union Research.
- Farewell, R. G. (1998). Demand-driven transit operations: Flex-route services. *Transportation Quarterly*, **52**, 31–43.
- Fingerle, G. P. and Lock, A. C. (1999). Practical issues in prototyping national public transport planning system using journeyweb protocol. *Transportation Research Record*, **1669**, 46–52.
- Gendreau, M., Laporte, G. and Seguin, R. (1996). Stochastic vehicle routing. *European Journal of Operational Research*, **88**, 3–12.
- Ghiani, G., Guerriero, F., Laporte, G. and Musmanno, R. (2003). Real-time vehicle routing: Solution concepts, algorithms and parallel computing strategies. *European Journal of Operational Research*, **151**, 1–11.
- Gildea, D. and Sheikh, M. (1996). Applications of technology in providing transit information. *Transportation Research Record*, **1521**, 71–76.
- Gillen, D., Chang, E. and Johnson, D. (2001). Productivity benefits and cost efficiencies from intelligent transportation system applications to public transit: Evaluation of advanced vehicle location. *Transportation Research Record*, **1747**, 89–96.
- Glazebrook, G. and Subramanian, S. (1997). Personal public transport in Australia: Developments and prospects. *Journal of Public Transportation*, **3**, 45–69.
- Greenfeld, J. (1999). Performance evaluation of automated vehicle locator and digital map accuracy in transit application. *Transportation Research Record*, **1666**, 45–51.
- Horbury, A. X. (1999). Using non-real-time automatic vehicle location data to improve bus services. *Transportation Research*, **33B**, 559–579.
- Horn, M. E. T. (2004). Procedures for planning multi-leg journeys with fixed-route and demand-responsive passenger transport services. *Transportation Research*, **12C**, 33–55.
- Hounsell, N. and Wall, G. (2002). New intelligent transport systems applications in Europe to improve bus service. *Transportation Research Record*, **1791**, 85–91.
- Huang, R. and Peng, Z. R. (2002). Object-oriented geographic information system data model for transit-trip planning systems. *Transportation Research Record*, **1804**, 205–211.
- Jeong, R. and Rilett, L. R. (2005). Prediction model of bus arrival time for real-time application. *Transportation Research Record*, **1927**, 195–204.
- Jiamin, Z., Bukkapatnam, S. and Dessouky, M. (2003). Distributed architecture for real-time coordination of bus holding in transit networks. *Intelligent Transportation Systems, IEEE Transactions on Computers*, **4**, 43–51.

- Kane, A. R. (2000). Transportation in the new millennium. *Transportation Quarterly*, **54**, 5–9.
- Lee, Y. J., Chos, K. S., Hill, D. L. and Desai, N. (2001). Effects of automatic vehicle location on schedule adherence for a mass transit administration bus system. *Transportation Research Record*, **1760**, 81–90.
- Levinson, H. S. (2001). Bus transit in the 21st century: Some perspectives and prospects. *Transportation Research Record*, **1760**, 42–46.
- Lin, W. H. and Zeng, J. (1999). Experimental study of real-time bus arrival time prediction with GPS data. *Transportation Research Record*, **1666**, 101–109.
- Lin, W. H. and Padmanabhan, V. (2002). Simple procedure for creating digitized bus route information for intelligent transportation system applications. *Transportation Research Record*, **1791**, 79–84.
- Maclean, S. D. and Dailey, D. J. (2002). Wireless internet access to real-time transit information. *Transportation Research Record*, **1791**, 92–98.
- Minsky, M. L. (1986). *The Society of Mind*. Simon and Schuster.
- Peng, Z. R. and Jan, O. (1999). Assessing means of transit information delivery for advanced public transportation systems. *Transportation Research Record*, **1666**, 92–100.
- Peng, Z. R. and Huang, R. (2000). Design and development of interactive trip planning for a web-based transit information system. *Transportation Research*, **8C**, 409–425.
- PRT Cyberspace Dream Keeps Colliding with Reality*. (2004). Light Rail Now Publication. <http://www.lightrailnow.org/facts/fa_prt001.htm>.
- Raiffa, H. (1982). *The Art and Science of Negotiation*. Belknap Press of Harvard University.
- Reynolds, F. D. (2006). *The Revolutionary Dualmode Transportation System*. E-book <<http://faculty.washington.edu/jbs/rev/revcontents.htm>>.
- Rosenbloom, S. and Fielding, G. J. (1998). *Transit Markets of the Future – The Challenge of Change*. TCRP Report 28, Transportation Research Board, Washington, DC.
- Siuru, W. D. (2003). New comprehensive report on future bus transit systems. *Mass Transit*, **28**, 44–48.
- Smith, R. G. (1980). The contract net protocol: High-level communication and control in a distributed problem solver. *IEEE Transactions on Computers*, **29**, 1104–1113.
- Smith, B. L. (2000). Using geographic information systems and the world wide web for interactive transit-trip itinerary planning. *Journal of Public Transportation*, **3**, 37–50.
- Spring, G. S., Collura, J. and Black, K. P. (1998). Evaluation of automatic vehicle location technologies for paratransit in small and medium-sized urban areas. *Journal of Public Transportation*, **1**, 43–60.
- Stone, J. R., Nalevanko, A. and Tsai, J. (1993). Assessment of software for computerized paratransit operation. *Transportation Research Record*, **1378**, 1–9.
- Stone, J. R., Ahmed, T. and Valevanko, A. (2001). Internet-based decision support for advanced public transportation systems technology. *Transportation Research Record*, **1731**, 63–67.
- Sun, W. and Bernstein, D. (2002). XML-based transit timetable system. *Transportation Research Record*, **1804**, 151–161.
- Trepanier, M., Chapleau, R. and Allard, B. (2002). Transit itinerary calculation on the web based on a transit user information system. *Journal of Public Transportation*, **5**, 13–32.
- Turner, S. M. (1996). Advanced techniques for travel time data collection. *Transportation Research Record*, **1551**, 51–58.

- US Department of Transportation (2000a). *Advanced Public Transportation Systems: The State of the Art*. Transit Admin. Report. Federal Transit Administration, Washington, DC.
- US Department of Transportation (2000b). *Benefits Assessment of Advanced Public Transportation System Technologies*. Transit Admin. Report. Federal Transit Administration, Washington, DC.
- Van Dyke Parunak, H. (1997). Go to the ant: Engineering principles from natural multi-agent systems. *Annals of Operations Research*, **75**, 69–101.
- Vorvick, J. and Dueker, K. J. (1998). Transit time internet access. *Transportation Research Record*, **1618**, 180–185.

Answers to Exercises

Chapter 3

3.1 (a)

Parameter	Time period				
	6:00–7:00	7:00–8:00	8:00–9:00	9:00–10:00	10:00–11:00
A (pass/km)	444.2	1796	1186	1349	952.4
ρ	0.82	0.64	0.51	0.71	0.72
Method proposed	1*	1	1	1	1
χ^2 test between Methods 1 and 2 data	Calculated $\chi^2 = 2.66$ $\chi^2_{\alpha=0.05} = 9.49$		* Fox Max load methods, use Method 1		

(b)

Method	Time period									
	6:00–7:00		7:00–8:00		8:00–9:00		9:00–10:00		10:00–11:00	
	F*	H*	F	H	F	H	F	H	F	H
Method 1	2	30	7.80	8	6.08	10	6	10	4.44	14
Method 2	2	30	7.80	8	6.47	9	6.36	9	4.44	14
Method 3	2	30	5.85	10	4.85	12	4.50	13	3.17	19
Method 4 (20%)	2	30	7.80	8	6.08	10	6	10	4.44	14

* F in (veh/hr), H in minutes.

Solutions for (d) and (e)

Time period	Method	(d)		(e)
		Excess load (passenger-km)	Empty space-km	Range of x = possible reduced single tariff (€/km)
6:00–7:00	3	0 (Minimum frequency provided)	0	–
	4(20%)	0	0	–
7:00–8:00	3	191.4	510.6	$x < 8.00$
	4(20%)	0	0	–
8:00–9:00	3	161.8	420.2	$x < 7.79$
	4(20%)	13.8	124.2	$x < 27.00$
9:00–10:00	3	179.4	378.6	$x < 6.33$
	4(20%)	14.4	93.6	$x < 19.50$
10:00–11:00	3	119.6	258.4	$x < 6.48$
	4(20%)	0	0	–

- 3.2 (a) 6:00–7:00, Max load = 148 passengers at Stop 3,
7:00–8:00, Max load = 99 passengers at Stop 2
(b) $F_2(6:00-7:00) = 3.7$, $F_2(7:00-8:00) = 3.0$
(c) $F_{4(40\%)}(6:00-7:00) = 3.2$, $F_{4(40\%)}(7:00-8:00) = 3.0$

Chapter 4

- 4.1 (a) Timetable at point A using Method 3: 07:00, 07:26, 07:59.
Using Method 4 (30%): 07:00, 07:20, 07:45, 08:21.
(b) With $c = 60$, only Method 3 results are changed: 07:00, 07:25, 07:55.
Method 4 results same as in (a).
- 4.2 (a) Even-headway timetable based on Method 2: 6:00, 6:30, 7:00, 7:08,
7:15, 7:23, 7:31, 7:38, 7:46, 7:54, 8:00, 8:11, 8:21, 8:30, 8:39, 8:48,
8:58, 9:07, 9:16, 9:26, 9:35, 9:45, 9:54, 10:05, 10:19, 10:32, 10:46,
11:00.
Even-headway timetable based on Method 4 (20% criterion): 6:00, 6:30,
7:00, 7:08, 7:15, 7:23, 7:31, 7:38, 7:46, 7:54, 8:02, 8:12, 8:22, 8:32,
8:42, 8:51, 9:01, 9:11, 9:21, 9:31, 9:41, 9:51, 10:01, 10:15, 10:28, 10:42,
10:55.

- (b) Even-headway timetable based on Method 2: 6:00, 6:30, 7:00, 7:08, 7:15, 7:23, 7:31, 7:38, 7:46, 7:54, 8:00, 8:07, 8:15, 8:22, 8:30, 8:37, 8:45, 8:52, 9:00, 9:07, 9:15, 9:22, 9:30, 9:37, 9:45, 9:52, 10:00, 10:12, 10:24, 10:36, 10:48, 11:00.

Even-headway timetable based on Method 4 (20% criterion): 6:00, 6:30, 7:00, 7:08, 7:15, 7:23, 7:31, 7:38, 7:46, 7:54, 8:00, 8:07, 8:15, 8:22, 8:30, 8:37, 8:45, 8:52, 9:00, 9:10, 9:20, 9:30, 9:40, 9:50, 10:00, 10:12, 10:24, 10:36, 10:48, 11:00.

- (c) Total number of departures for the four timetables above: 28, 27, 32 and 30, respectively. Single-route fleet size is 6, and is the same for all timetables.

- 4.3** (a) With Method 2, the use of cumulative loads at Stop 3 (6:00–7:00) and stop 2 (7:00–8:00) leads to the following timetable at Stop 1: 6:17, 6:36, 6:53, 7:14, 7:34, and by extrapolation 7:54.
- (b) (i) Certainly not all the loads at each stop on each vehicle in a given hour are the same.
- (ii) On departure 6:17, at Stop 2, the average load will be $52 \cdot 17/15 \approx 59$ (above 40); and on departure 7:14, at Stop 3, the average load will be 80 (above 60).
- 4.4** (a) Passenger arrival rates (passengers/minute) are $240/60 = 4$ and $180/60 = 3$; and the even headways are $64/4 = 16$ and $24/3 = 8$ minutes, respectively, for the two hours. Hence, the fourth departure, if at 7:00, will have 48 passengers; in order to comply with 44 passengers, it will depart at 6:59. The first departure after 7:00 will, therefore, be at 7:13. The complete timetable: 6:16, 6:32, 6:48, 6:59, 7:13, 7:21, 7:29, 7:37, 7:45, 7:53, and 8:01.
- (b) Using Method 2, we check for another possible departure, with 64 passengers as the desired load. Since this time will move beyond 7:00, we change the desired load to 24, and so end with a departure before 7:00. This contradiction suggests the use of the procedure at (a).

Chapter 5

The adjusted load between 6:00 and 7:00 for Stop A is $25 + 65 + (25/35) \cdot 67 = 138$; and for Stops B and C, 148 and 85 passengers, respectively. Between 7:00 and 8:00, the adjusted loads are 178, 97, and 114 for Stops A, B, and C, respectively. Hence, Method 2 is used for frequency determinations:

$$F_2(6:00-7:00) = \max(148/40, 2) = 3.70 \text{ veh/hr, } H_2(6:00-7:00) \approx 16 \text{ minutes}$$

$$F_2(7:00-8:00) = \max(178/60, 2) = 2.97 \text{ veh/hr, } H_2(7:00-8:00) \approx 20 \text{ minutes.}$$

The results, which appear in the table that follows, are based on applying the cumulative-frequency-curve approach for the first procedure and the cumulative-load-curve approach for the two other procedures.

Observed data	Departure	6:00	6:35	7:10	7:45	8:00			
	Max load point	A	B	B	A	A			
	no. of passengers	25	72	82	84	75			
Evenly spaced headways	Departure	6:00	6:16	6:32	6:48	7:06	7:26	7:46	8:06
	Max load point	A	A	B	B	B	A	A	A
	no. of passengers	25	55	33	36	43	46	51	97
Even average load at hourly Max load point	Departure	6:11	6:30	6:51	7:20	7:45	8:00		
	Max load point	A	B	B	A	A	A		
	no. of passengers	45	39	47	60	60	75		
Even average load at individual Max load point	Departure	6:08	6:27	6:45	7:12	7:39	7:54	8:05	
	Max load point	A	B	B	B	A	A	A	
	no. of passengers	40	40	40	40	60	60	50	

Chapter 6

- 6.1 The upper bound $Z^* = 8$. Both Synchro-1 and Synchro-2 give the optimal results of 4 meetings, with these departure times: Route I at 0, 7; Route II at 2, 9; Route III at 4.
- 6.2 The upper bound $Z^* = 15$. Synchro-1 results in 6 meetings. Synchro-2 results in the optimal solution, with 8 meetings. The optimal solution is based on these departure times: Route I at 0, 8, 16; Route II at 7, 13, 21; Route III at 6, 14, 22; and Route IV at 8, 20, 30.
- 6.3 *Hint:* The objective function given by Equation (6.1) is changed to

$$\text{Max} \sum_{k=1}^{M-1} \sum_{i=1}^{F_k} \sum_{q=k+1}^M \sum_{j=1}^{F_q} \sum W_{lkjqn} Z_{ikjqn}$$

where w_{lkjqn} represents the number of passengers at node n who need to be transferred between the vehicle of the i -th departure on route k and the vehicle of the j -th departure on route q (all passenger-transfers). A heuristic procedure can combine Synchro-1 with this change.

Chapter 7

- 7.1 The solution can use Salzborn’s idea (1972) for single routes by combining the findings: construct three curves representing the frequencies for each route (y-axis) by time-of-day (x-axis) and the sum of the two routes. For route 1, the largest number of buses that departs in any time interval of 3 hours is 26 (4–7 p.m.). For route 2, there are 28 independent departures (buses) (6–9 a.m.). However, the two time intervals do not overlap; hence, the minimum fleet size required is **not** the sum of the two. This minimum is found from the curve representing the sum of frequencies, which from 6–9 a.m. is $18 + 18 + 16 = 52$ buses.

- 7.2** The augmenting-path algorithm starts by constructing a network-flow similar to that in Figure 7.5. Four s - t paths can be found using the last four departures. The minimum s - t cut separates the first four departures and t from the rest of the network. Hence, $\text{Min } N = n - \text{Max } Z = 8 - 4 = 4$, based on Theorem 7.2. The four blocks are (by trip no.): [1 – 5 – 8], [2 – 6 – 7], [3], and [4].
- 7.3** (i) The DF at terminal a starts with a maximum of $D(a) = 3$ between 7:30 and 8:00; and at b , $D(b) = 0$ between 5:00 and 6:00. The URDHC procedure is then applied with the latest DH insertion possible. Based on the NT rule, two DH trips, DH_1 and DH_2 , are inserted from b to arrive at a , s_1^a at the start of the maximum interval (7:30); one DH trip, DH_3 , a compensatory act, is inserted from a to arrive at b at 8:30. These three DH trips result in $D(a) = 1$, $D(b) = 0$.
- (ii) The sum function (of a , b , and depot), $g(t, S)$, reveals that $G(S) = 4$. In order to find the minimum fleet size required of S , $d(\text{depot}, t)$ is constructed, and it results in $D(\text{depot}) = 3$ between 6:40 and 9:00. Hence, $G(S)$ indeed is the minimum fleet size required (i.e. the sum of all determined $D(k)$, $k = a, b, \text{depot}$).
- (iii) The four chains are [1-DH1-6], [2-5], [3-DH2-7], [4-DH3-8].

Chapter 8

- 8.1** Minimum fleet size is 22 vehicles for the case with only shifting, having the following FIFO blocks: [1–10–20–34], [2–11], [3–27], [4–23–43], [5–15–35], [6–17–44], [7–18–40], [8–38], [9–31], [12–32–45], [13–33], [14–29], [16–41], [19–36], [21–46], [22–47], [24], [25], [26–37], [28–42], [30], [39].
Minimum fleet size is also 22 vehicles for the case with only DH insertion trips, having the following FIFO blocks: [1–10–20–34], [2–17–44], [3– DH_3 –39], [4– DH_4 –30], [5–15–35], [6–23–42], [7–18–40], [8–38], [9–31], [11], [12–32], [13–33], [14–29], [16– DH_1 –46], [19–36], [21–45], [22– DH_2 –43], [24–47], [25], [26–37], [27], [28–41].
Minimum fleet size is 20 vehicles for the case with both shifting and DH trip insertion, having the following FIFO blocks: [1–10–20–33], [2–14–29], [3–27], [4– DH_1 –30], [5–15–38], [6–17–44], [7–18–40], [8–32–45], [9–23–43], [11–34], [12–31], [13– DH_2 –39], [16–41], [19–36], [21–46], [22–35], [24–47], [25], [26–37], [28–42].
- 8.2** (1) Minimum fleet size is six vehicles (indicated at 7:20 on the single DF);
(2) There are 4 shiftings required to arrive at a minimum fleet size of four vehicles: Trip 3 backward by 3 minutes; Trip 9 forward by 5 minute; Trip 8 forward by 5 minutes; but because of the shifts of Trips 4 and 8, another shift is required, that of Trip 10 forward by 5 minutes.
(3) The original shuttle-bus schedule is coordinated with the train schedule; hence, shifting the shuttle schedule indeed will have an adverse effect on this coordination.
(4) In a single terminal (terminal k) operation, there is no need for the improved lower bound; because the reason is that $g(t, S) = d(t, k)$.
(5) The four FIFO blocks are [1–8–6], [2–9], [7–5], [3–4–10].
- 8.3** $G(S) = 3$, $G'(S') = 4$, and $G''(S'') = 5$.

- 8.4 (i) $G'(\bar{S}_{sf}) = 5$, and $G''(\bar{S}_{sf}) = 5$.
 (ii) $G(S_{sf}) = 3$, $G'(S'_{sf}) = 3$, and $G''(S''_{sf}) = 4$.
- 8.5 Follow Equations (8.6) to (8.8), established for the $\Delta^{j\pi(-)}$ criterion, and derive similarly the criterion for $\Delta^{j\pi(+)}$.

Chapter 9

- 9.1 (a) Three main steps are performed. First, $N_1 = 2 + 1 + 1 = 4$, $C_1 = 4 \cdot 10 = 40$, and $G''(S)$ is calculated to be 4. Second, $N_2 = 2 + 1 + 1 + 2 = 6$, $C_2 = 2 \cdot 10 + 1 \cdot 8.5 + 1 \cdot 4 + 2 \cdot 5 = 42.5$. Third and the best solution, is $N = 2 + 1 + 1 + 1 = 5 \neq G''(S)$, $C = 2 \cdot 10 + 1 \cdot 8.5 + 1 \cdot 4 + 1 \cdot 5 = 37.5$. Following are the five blocks, using the FIFO rule, by trip number and type of vehicle: type I–[5], [1-8-DH₄-12], type II–[2-6], type III–[4-DH₃-7-DH₁-10], and type IV–[3-DH₂-9-11].
- (b) The first main step in (a) is the same. Second, $N_2 = 3 + 3 = 6$, $C_2 = 3 \cdot 10 + 3 \cdot 4 = 42$. Third and best is $N = 3 + 1 = 4 = G''(S)$, $C = 3 \cdot 10 + 1 \cdot 4 = 34$. Following are the four (FIFO) blocks: type I–[1-8-12], [2-6-DH₃-9], [3-DH₁-5-DH₄-10], type II–[4-DH₂-7-11].
- 9.2 (a) The new unit costs do not satisfy condition (b) of Step 6 in algorithm VTSP; therefore, DH₁ and DH₂ cannot be inserted. The new unit costs concerning condition (b) result in $10 - (5 + 3) = 2 > 0$. With DH₃, but without DH₄, we obtain: $n_1 = 2$ (type I: $c_1 = 10$), $n_2 = 4$ (type II: $c_2 = 5$), $n_3 = 1$ (type III: $c_3 = 3$), and $C = 2 \cdot 10 + 4 \cdot 5 + 1 \cdot 3 = 43$. Following are the seven (FIFO) blocks: type I–[1-9], [3-12], type II–[2-DH₃-7-11], [8-13], [4-10], [6], type III–[5-14].
- (b) The change starts at Step 6 of algorithm VTSP; four new DH trips are DH₁, DH₂, DH₃, DH₄, from which we obtain: $n_1 = 2$, $n_2 = 4$, $n_3 = 0$, and $C_3 = 2 \cdot 10 + 4 \cdot 11 + 0 \cdot 6 = 64$. Following are the six (FIFO) blocks: type I–[1-9], [5-DH₂-12], type II–[2-DH₃-7-11-DH₄-14], [3-DH₁-8], [4-10], [6-13].
- (c) Similar situation to (b) to obtain: $n_1 = 3$, $n_2 = 2$, $n_3 = 1$, and $C_3 = 3 \cdot 12 + 2 \cdot 8 + 1 \cdot 9 = 61$. Following are the blocks: type I–[1-9], [3-DH₁-8], [5-DH₂-12], type II–[4-10], [6-13], type III–[2-DH₃-7-11-DH₄-14].
- 9.3 Three main steps are performed. First, $N_1 = 3 + 3 + 0 = 6$, $C_1 = 6 \cdot 5 = 30$, and $G''(S)$ is calculated to be 6. Second, $N_2 = 3 + 3 + 2 = 8$, $C_2 = 3 \cdot 5 + 3 \cdot 4 + 2 \cdot 3 = 33$. Third, and the best solution: $N = 3 + 2 + 1 = 6 = G''(S)$, $C = 3 \cdot 5 + 2 \cdot 4 + 1 \cdot 3 = 26$. Following are the six (FIFO) blocks: type I–[1-11], [6-DH₁-9], [5-DH₃-12], type II–[2-8-14], [4-7-DH₂-13], type III–[3-10].
- 9.4 (a) Let AB travel speed = $x \cdot$ MB travel speed in which $0 < x < 1$. For $P > 30$, we obtain $x = 2/3$. For 10% less MB travel speed, $x = 11/15$. By calculating the \$ saving to be greater than the additional MB operational cost, we arrive at $P > 60$ passengers.
- (b) Using the above definition of x and with P as a variable, we obtain the inequality $P > (10(9 - 5))/(1 - x)$, where $x = 5/9$ for $P = 0$. Therefore, the lower

bound of AB travel speed is $(5/9) \cdot$ MB travel speed; and the upper bound is simply the MB travel speed.

- (c) Turning the inequality in (b) into an equation and examining its first derivative reveals that there is no optimum P for the range considered for x.

Chapter 10

10.1 $D(k) = 9$.

- (i) Algorithm T_mF results in thirteen joinings, from which only two (see below in bold) comply with the $T_{\max}$ constraint. The [arrival-departure] joinings are by trip numbers: **[10-22]**, [11-14], [12-25], [13-16]; **[17-28]**, [18-23], [19-24], [20-27], [21-29], [25-30], [26-31], [32-39], [33-40].
- (ii) The use of the FIFO rule results in the following thirteen joinings by trip numbers, which none of which complies with the $T_{\max}$ constraint: [10-14], [11-15], [12-16], [13-22], [17-23], [18-24], [19-27], [20-28], [21-29], [25-30], [26-31], [32-39], [33-40].

10.2 The algorithm Dijkstra is applied for the two blocks. Block 1 is divided into two sets of pieces: $[a-b-c-d]$ and $[d-c-b]$, with a set's cost being 9 and 4, respectively (total cost of block is 13). Block 2 is divided into three sets of pieces: $[a-b]$, $[b-c]$, and $[c-a-d]$ with a set's cost being 3, 3, and 12, respectively (total cost of block is 18).

10.3 The Roster procedure ends with d_4^3 and d_3^6 left untreated; thus, a new roster is manually created: R_4 covering Wed. and Sat. The minimum number of drivers derived is five: three with R_1 , one with R_3 , and one with R_4 . That is, $[d_1^1-d_1^2-d_1^3-d_4^4-d_3^5]$, $[d_2^1-d_2^2-d_2^3-d_2^4-d_1^5]$, $[d_4^1-d_3^2-d_3^3-d_3^4-d_2^5]$, $[d_1^6-d_4^7]$, $[d_3^6-d_4^3]$.

Chapter 11

11.1 Revenue = $20q - 0.04q^2$ with a maximum of \$2500 per hr at $q = 250$ tickets/hr and $p = \$10$. The function $p - q$ is inelastic for p between 0 and 10, and elastic for p between 10 and 20; for $p = 0$, it is perfectly inelastic; for $p = 20$, it is perfectly elastic; and for $p = 10$, we obtain a unit elastic point.

11.2 (a) Price elasticity of demand for transit trips is -0.3 (1% reduction in fare would lead to a 0.3% increase in transit patronage). A reduction of 33.33% (from \$1.20 to \$0.80) results in 22,000 hourly passengers; the company will then lose $24,000 - 17,600 = \$6400$.

(b) Car price cross-elasticity of demand is 0.2; a \$0.60 rise is 15% of \$4, leading to an increase in patronage, from 20,000 to 20,600 (+3%).

11.3 (a) Bus, light rail, and car utilities are -4.3 , -5.35 , and -5.35 , respectively. Utilizing Equation (11.10) results in a modal split of 58.8%, 20.6%, and 20.6% for the bus, light rail, and car, respectively.

(b) For an imposed parking fee of \$1.20, the utility of cars becomes -10.15 ; the modal split is then 73.92%, 25.87%, and 0.21% for the bus, light rail, and cars, respectively.

Chapter 12

- (a) There are 5 possible ‘single’ travel options: [Route 1(A-B)]; [Route 2(A-X) and Route 3(X-B)]; [Route 2(A-Y) and Route 3(Y-B)]; [Route 2(A-X), Route 3(X-Y), and Route 4(Y-B)]; [Route 2(A-Y) and Route 4(Y-B)]. In addition, these options may be combined; for example, the passenger may decide to take the first route that arrives at A (either Route 1 or 2). In the combined option, we assume that the passenger will take the route according to its frequency proportion (inverse of the headway).
The table below summarizes the possible sets of combinations for each node.

Node	Attractive routes (route-exit-node)	Waiting time (minutes)	Route probabilities			
			1	2	3	4
A	1-B	6.0	1	–	–	–
A	2-X	6.0	–	1	–	–
A	2-Y	6.0	–	1	–	–
A	1-B,2-X	3.0	0.5	0.5	–	–
A	1-B,2-Y	3.0	0.5	0.5	–	–
X	2-Y	6.0	–	1	–	–
X	3-Y	15.0	–	–	1	–
X	3-B	15.0	–	–	1	–
X	2-Y,3-Y	4.3	–	0.71	0.29	–
X	2-Y,3-B	4.3	–	0.71	0.29	–
Y	3-B	15	–	–	1	–
Y	4-B	3	–	–	–	1
Y	3-B,4-B	2.5	–	–	0.17	0.83

- (b) The expected travel times for each option can be calculated using this table, assuming that travel times are composed of waiting and in-vehicle times only. For example, the first row in the table, corresponding to the option of taking Route 1, gives an expected travel time of 31 minutes (6 + 25). The expected travel time for each combination is found by averaging the route probabilities. For example, the option indicated in the 4th row (either Route 1 or 2), combined with boarding Route 3 at X, gives an expected travel time of $(3 + 0.5 \cdot 25) + (3 + 0.5 \cdot (7 + 15 + 8)) = 33.5$ minutes.

- (c) The minimum expected travel time is obtained by the following combination: take either Route 1 to B or 2 to Y; at Y, take either Route 3 or 4. The expected travel time is $(3 + 0.5 \cdot 25) + (3 + 0.5 \cdot (13 + 2.5 + 0.17 \cdot 4 + 0.83 \cdot 4)) = 15.5 + 12.75 = 27.5$ minutes.

Chapter 13

- 13.1** In the beginning, $D(a) = 2$, $D(b) = 2$, $D(c) = 1$, and $P(a) = 2$, $P(b) = 2$, $P(c) = 3$. Shifting trips 5 and 2 reduces $D(c)$ by one. Shifting trip 1 to the right reduces $P(b)$ by one, but $P(a)$ increases; hence, requiring the shifting of trip 4. **Solution:** (a) $D(a) = 1$, $D(b) = 2$, $D(c) = 1$; (b) $P(a) = 1$, $P(b) = 1$, $P(c) = 3$.
- 13.2** In the beginning, $D(a) = 2$, $D(b) = 1$, $D(c) = 0$, $D(d) = 2$, and $P(a) = 2$, $P(b) = 1$, $P(c) = 2$, $P(d) = 2$. It is impossible to reduce $D(k)$, $k = a, b, c, d$. Shifting trips 6 and 7 reduces $P(a)$ by one, but $P(c)$ and $P(b)$ increase; hence, requiring the shifting of trip 3 to the left and trip 9 to the right. **Solution:** (a) same as in the beginning; (b) $P(a) = 1$, $P(b) = 1$, $P(c) = 2$, $P(d) = 2$.
- 13.3** In the beginning, $D(a) = 2$, $D(b) = 2$, $D(c) = 0$, and $P(a) = 2$, $P(b) = 2$, $P(c) = 2$. It is impossible to reduce $D(k)$, $k = a, b, c$. Shifting trips 6 and 7 reduces $P(a)$ by one, but $P(c)$ increases; hence, requiring the shifting of trip 3 to the left and also inserting a DH trip from c (at 8:30) to reduces $P(c)$. **Solution:** (a) same as in the beginning; (b) $P(a) = 1$, $P(b) = 2$, $P(c) = 1$.
- 13.4** (a) Minimum of two stops is required; they can be located at nodes 5 (a stop for nodes 1, 3, 4, 6) and 2 (a stop for node 2), as well as around these stops in a variety of possibilities.
- (b) The minimax distance for four stops is $\ell = 250$ metres, which is determined by the stop located at the midway point on arc (4,6); the locations of the three other stops are: on arc (1,3), between 50 metres from node 1 to 50 metres before node 3; on arc (3,4), between 150 metres from node 3 to 150 metres before node 5; and around node 2, by 250 metres from node 2 on arcs (1,2), (2,3), and (2,4).
- (c) The single stop will be located on arc (2,3) at a distance of 1050 metres from node 2 and 50 metres away from node 3; thus, the minimax is $\ell = 1050$ metres, which will be fully used by the demand at nodes 2, 4 and 6.

Chapter 14

- 14.1** (a) Two alternative settings: Routes A and B, with 14 and 11 departures, respectively; and Routes A and B, with 15 and 10 departures, respectively; the second alternative is selected with headways of 8 (Route A) and 12 (Route B) minutes because the headways are integers; for Route B, they are clock headways.
- (b) Minimum fleet size required for Routes A and B are 13 and 7 vehicles, respectively.

- (c-1) There are four combinations of single routes: 1-2-4-3; 1-2-3-4; 1-4-2-3; and 1-4-3-2.
- (c-2) Adequate criterion: minimum passenger-hours.
- (c-3) Load profiles for Routes 1-2-4-3, 1-2-3-4, 1-4-2-3, and 1-4-3-2 are [1500,960,840], [1500,960,600], [1500,1780,840], and [1500,1780, 1380], respectively.
- (c-4) Minimum fleet size: 30, 30, 25 and 30 vehicles for Routes 1-2-4-3; 1-2-3-4; 1-4-2-3; and 1-4-3-2, respectively.
- (c-5) Route 1-4-2-3 is selected, with 64,600 passenger-minutes compared with 67,800 for the two existing routes.

14.2 (a) The calculated measures are given in this table:

Route	PH _r (pass-hr)	EH _r (pass-hr)	DPH _r (pass-hr)	w _r (pass-hr)
1-2-4-6	670.8	236.7	75	53.6
4-5-3	383	72	1.7	27.2
1-8-7	117	21.7	7	14.4

- (b) Total DPH, including transfers, is 338.7 passenger-hours.
- 14.3
- (a) Calculated frequency (from load profile) in veh/hr and headway in minutes for fast train and fast ferry are, respectively: [F = 3, H = 20] and [F = 4, H = 15]. Thus, the average waiting time is 10 minutes for the train, and 7.5 for the ferry; the hourly cost, therefore, is \$3120 for the train, and \$2340 for the ferry.
 - (b) Lost cost for empty-seat hours (utilizing the load profile) is \$14,700 for the train, and \$12,600 for the ferry; the lost cost for travel time is \$3660 and \$6630 for the train and ferry, respectively.
 - (c) Hourly income is \$67,700 and \$53,200 for the train and ferry, respectively; the profit (using the loss in (a) and (b) above), therefore, is \$46,220 and \$31,630 for the train and ferry, respectively.
 - (d) Thus from (c), the preferred mode will be the fast train.
 - (e) Some of the neglected cost elements in the actual analysis are these: vehicle cost, operational cost (equipment, labour, maintenance, fuel, etc.), real-estate and land costs, terminal/berth construction cost, and road/sea accident cost.

Chapter 15

- 15.1
- (a) Constructing the deficit functions $d(a,t)$ and $d(c,t)$ results in $D(a) = 6$ and $D(c) = 5$ vehicles, with a single DH trip from c (at 7:00) to a , $D(a) = 5$; thus, $N_{min} = 10$ vehicles. The Minimax H procedure is then applied, and this results in Minimax H = 14 minutes ($a \rightarrow c$, 6:00–7:00), = 18 ($a \rightarrow c$, 7:00–8:00), = 15 ($c \rightarrow a$, 6:00–7:00 and 7:00–8:00). Hence, in the $a \rightarrow c$ direction, four trips will end at b (c is the Max load point); and in the $c \rightarrow a$ direction, three trips will start at b (Max load point) while creating the timetable with maximum short-turn

trips. The arrivals at **b** for the $a \rightarrow c$ direction are as follows, with a single asterisk for arrivals ending at **b**: [6:20; 6:27*; 6:32; 6:40; 6:47; 6:55*; 7:01; 7:08*; 7:15; 7:25*; 7:33; 7:40; 7:50; 7:55; 8:03]. The departures at **b** for the $c \rightarrow a$ direction are as follows, with a single asterisk for departures starting at **b**: [6:25; 6:35; 6:43*; 6:50; 6:57*; 7:05*; 7:15; 7:20; 7:35; 7:43; 7:50; 7:57; 8:05; 8:13*; 8:18]. Creating $d(a,t)$, $d(b,t)$, and $d(c,t)$ for the schedule, with maximum short-turn trips, results in $D(a) = 6$, $D(b) = 0$, and $D(c) = 5$, and with a single DH trip from **b** (at 7:08) to **c**, $D(c) = 4$; thus, $N'_{\min} = 10$ vehicles $= N_{\min}$. In this particular exercise, therefore, it doesn't make sense to introduce short-turn trips.

- (b) Listing the trips in increasing order of departure times, with departures at **a** coming before **c** for the same departure times, results in a list of 31 trips. Note that trip number 18 on the list is a DH trip between 7:00 and 7:28 from **c** to **a**. Applying the FIFO procedure results in the following ten blocks, by trip number determined in the list of trips: [1,11,22], [2,12,21,31], [3,13,24], [4,14,23], [5,16,26], [6,15,25], [7,18,27], [8,17,28], [9,20,29], and [10,19,30].

- 15.2 (a) Using component FIRST from Chapter 6 at the single synchronization point **b** results in $\text{Max}(5,8,6,7) \leq d \leq \text{minimum}(10,13,15,20)$; thus, the minimum (even) headway is $d = 8$ minutes. Setting the first departure from terminal **c** at 6:00 will determine the remaining departures with 8-minute headways. That is, the first departure from **a** will be at 6:05, from **e** at 6:15, and from **k** at 6:05, and all will meet at **b** at 6:25. This is for maximizing the number of meetings at **b**. The 12 departures from terminal **a** are [6:05; 6:13; 6:21; 6:29; 6:37; 6:45; 6:53; 7:01; 7:09; 7:17; 7:25; 7:33]; and from **c**: [6:00; 6:08; 6:16; 6:24; 6:32; 6:40; 6:48; 6:56; 7:04; 7:12; 7:20; 7:28].
- (b) Constructing $d(a,t)$ and $d(c,t)$ results in $D(a) = 5$ and $D(c) = 7$ vehicles, with no possibility of inserting DH trips; thus, $N_{\min} = 12$ vehicles. The Minimax H procedure is then applied, and this results for all hours and directions of travel in Minimax H = 16 minutes. Hence, four trips in the $a \rightarrow c$ direction will start at **b** (Max load point), and four trips in the $c \rightarrow a$ direction will end at **b** (**c** is the Max load point) while creating the timetable with maximum short-turn trips. Because **b** is a meeting point, the departure and arrival times at **b** are the same for both directions; these times appear in the following list, with a single asterisk for departures starting at **b** and a double-asterisk for arrivals ending at **b**: [6:25; 6:33; 6:41*; 6:49**; 6:57*; 7:05**; 7:13; 7:21(*)**; 7:29; 7:37; 7:45(*)**; 7:53]; note that 7:21 and 7:45 are both starting and ending times for short-turn trips. Creating $d(a,t)$, $d(b,t)$, and $d(c,t)$ for the schedule with maximum short-turn trips results in $D(a) = 3$, $D(b) = 1$, and $D(c) = 7$; thus, $N'_{\min} = 11$ vehicles $< N_{\min}$. The last step is an attempt to extend i.e. minimize, the short-turn trips, which results in extending the 6:41 departure at **b** to start at 6:21 from **a**, and extending the 7:05 arrival at **b** to arrive to **c** at 7:20.

Chapter 16

- (a) The following are the 22 impedance ratios, z_{ij} , with their corresponding links in parenthesis: 2(1,2); 2(2,1); 1.33(2,3); 2(3,2); 7(2,8); 1.5(8,2); 0.66(3,4); 2(4,3);

1.66(4,5); 2(5,4); 2.5(4,9); 2(9,4); 2.5(5,6); 1.33(6,5); 1.25(6,7); 1(7,6); 1(6,9); 1(9,6); 1.66(7,8); 2(8,7); 1.33(8,9); 1.5(9,8).

- (b) The 22 links in the network are arranged in parenthesis in decreasing order according to the following pd_{ij} units:

6 (3,4); **5** (7,6); **4** [(8,2),(6,7),(9,6)]; **3** [(2,3),(7,8),(8,9),(6,9),(4,5),(6,5)];
2 [(9,8),(3,2),(4,3),(5,4),(4,9),(9,4),(8,7),(5,6)]; **1** [(2,1),(1,2),(2,8)].

The highest pd_{ij} group selected, with 5 links, contains units with $pd_{ij} = 6, 5, 4$. Given the initial group of 5 links and $p = 2$, there are 20 arranged (the order-of-links considered) combinations (sub-groups) of links that may be considered for inclusion in the circular route: [(3,4),(7,6)]; [(3,4),(8,2)]; [(3,4),(6,7)]; [(3,4),(9,6)]; [(8,2),(6,7)]; [(8,2),(9,6)]; [(9,6),(6,7)]; [(8,2),(7,6)]; [(6,7),(7,6)]; [(9,6),(7,6)]; [(7,6),(3,4)]; [(8,2),(3,4)]; [(6,7),(3,4)]; [(6,7),(8,2)]; [(9,6),(8,2)]; [(6,7),(9,6)]; [(7,6),(8,2)]; [(7,6),(6,7)]; [(7,6),(9,6)]; [(9,6),(3,4)].

For each pair, a circular route is constructed using the shortest-path criterion of z_{ij} between the links and the train. From the efficient routes found, each complying with the constraint of $T = 30$ minutes, the best route is determined (by node number): **1-2-3-4-9-8-2-1**; the one selected has a 26-minute circular travel time and a pd_{ij} sum of 19.

Chapter 17

17.1 Answers to (a) and (b) can be found within the chapter itself.

17.2 Same as 17.1; answers to (a) and (b) can be found within the chapter itself.

Author Index

A

Abkowitz, M., 155, 158, 525, 528, 546, 561, 564
Adamski, A., 114, 156, 157, 561, 562, 565
Adkins, H. (*see* Henderson, G. *et al.*, 555, 557)
Ahmed, T. (*see* Stone, J. R. *et al.*, 587)
Ahuja, R. K., 130, 174, 193, 200, 205, 293, 317, 383
Albom, R., 39, 40
Allard, B. (*see* Trepanier, M. *et al.*, 588)
Alter, C. H., 399, 401
Amedeo, R. O. (*see* Koutsopoulos, H. N. *et al.*, 76, 78)
Anantharamaiah, K. M., 477
Anderson, J. E., 590
Assad, A. (*see* Bodin, L. *et al.*, 291)
Ausman, J. (*see* Ryus, P. *et al.*, 400, 401)
Axhausen, K. W., 445

B

Baaj, M. H., 446, 448 (*see also* Shih, M. C. *et al.*, 156, 159)
Bakker, P., 590
Balas, E., 385
Balcombe, R., 322–329, 523, 533
Balke, K. N., 589
Ball, M., 294 (*see also* Bodin, L. *et al.*, 291)
Bamford, C. G., 38, 40
Banihashemi, M., 194, 195, 306, 308 (*see also* Haghani, A. *et al.*, 194, 195)
Banks, J. H., 76, 78
Barnes, K. E., 39, 40
Barnhart, C., 410 (*see also* Eberlein, X. J. *et al.*, 114, 562, 565)
Bartlett, T. E., 131, 179
Barua, B., 39, 40
Barzily, Z. (*see* Silman, L. A. *et al.*, 410)
Beasley, J. E., 306, 308
Becker, A. J., 157, 158

Bel, G. (*see* Dubois, D. *et al.*, 410)
Bell, M. G. H., 345, 358, 359 (*see also* Kurauchi, F. *et al.*, 360, 362)
Ben Akiva, M., 338, 339, 340
Benjamin, J. M., 590
Ben-Shabat, E., 339, 340
Bernstein, D. H., 359, 361, 588 (*see also* Eberlein, X. J. *et al.*, 562, 565)
Bertossi, A., 252
Bianco, L., 300 (*see also* Mingozi, A. *et al.*, 306)
Bielli, M., 447, 449 (*see also* Bianco, L. *et al.*, 300)
Black, K. P. (*see* Spring, G. S. *et al.*, 590)
Bladikas, A. K. (*see* Spasovic, L. N. *et al.*, 446, 448)
Bly, P. H., 268, 273, 274
Blythe, P. T., 587
Boberg, J. (*see* Barua, B. *et al.*, 39, 40)
Bodin, L., 291
Bodin, R. (*see* Ball, M. *et al.*, 294)
Boile, M. P. (*see* Spasovic, L. N. *et al.*, 446, 448)
Booz Allen, 38, 40
Borndorfer, R., 487
Boschetti, M. A. (*see* Mingozi, A. *et al.*, 306, 308)
Bouzaïene-Ayari, B., 359, 361
Bovy, P. H. L., 447, 449
Bowman, L. A., 588–560, 564
Boyce, D. E., 322
Bradshaw, J. M., 572
Branco, I. (*see* Costa, A. *et al.*, 251; Daduna, J. R. *et al.*, 6, 290)
Brand, D. (*see* Nelson, M. *et al.*, 155, 159)
Bruggman, J. (*see* Heathington, K. W. *et al.*, 409)
Bukkapatnam, S. (*see* Jiamin, Z. *et al.*, 572)
Bunt, P. D., 476, 477
Bunyan, R. E., 409

- C
 Callas, S. (see Strathman, J. G. et al., 478;
 Kimpel, T. J. et al., 544, 546)
 Champion, D. R., 590
 Cao, E. B., 306, 308
 Caprara, A., 300
 Caramia, M. (see Bielli, M. et al., 447, 449)
 Carey, M., 556, 557, 561, 565
 Carotenuto, P. (see Bielli, M. et al., 447, 449)
 Carraresi, P., 300, 306, 308 (see also Bertossi,
 A. et al., 252)
 Carrick, R. J. (see Bamford, C. G. et al., 38,
 40)
 Cassidy M. J. (see Madanat, S. M. et al., 399,
 401)
 Catanas, F., 300
 Cathey, F. W. (see Dailey, D. J. et al., 589)
 Ceder, A., 4, 13, 52, 55, 84, 88, 87, 91, 94,
 114–115, 121, 126, 128, 132, 141, 142, 148,
 154, 158, 171, 173, 176, 179, 184, 188, 200,
 210, 211, 221, 251, 315, 325, 330, 331, 345,
 346, 348, 349, 350, 352, 353, 381, 410, 419,
 438, 442, 444, 458, 462, 465, 483, 498, 503,
 507, 534, 535, 537, 541, 549 (see also
 Tykulsker, R. J. et al., 294; Yin, Y. et al., 410,
 417, 418, 419, 556, 557)
 Cepeda, M., 360, 362
 Cervero, R., 487
 Chan, K. S. (see Lam, W. H. K. et al., 359)
 Chang, E. (see Gillen, D. et al., 588)
 Chang, S. K., 272, 274, 515, 517
 Chapleau, R. (see Trepanier, M. et al.,
 588)
 Chen, H. L., 447, 449
 Chernicoff, W. P., 9
 Chhonkar, R. S. (see Chien, S. et al., 272, 516,
 572)
 Chiang, K. H. (see Haghani, A. et al., 195)
 Chien, S., 272, 274, 515, 516, 518, 572
 Chisholm, R. (see Pine, R. et al., 3)
 Chmiel, W., 156, 157
 Cho, C., 445, 448
 Chos, K. S. (see Lee, Y. J. et al., 588)
 Chriqui, C., 358, 360
 Christofides, N., 381, 382
 Clement, R., 306, 308
 Clever, R., 156, 158
 Coello Coello, C. A., 428, 439
 Cohon, J. L., 439
 Collura, J. (see Spring, G. S. et al., 590)
 Cominetti, R., 359, 360, 361 (see also Cepeda,
 M. et al., 360, 362)
 Conely, W., 430, 437
 Cooper, M. (see Ryus, P. et al., 400, 401)
 Correa, J., 359, 360, 361
 Costa, A., 251
 Crainic, T. G. (see Melucelli, F. et al., 487)
 Crider, L. (see Guttenplan, M. et al., 400, 401)
 Crisalli, U., 340, 345, 358 (see also Nuzzolo,
 A. et al., 359)
 Cushman, K. (see Schneider, J. B. et al., 155, 159)
- D
 Daduna, J. R., 6, 156, 158, 170, 290
 Dailey, D. J., 588, 589
 Darby-Dowman, K., 305, 308
 Daskin, M. S., 322
 Davis, R., 577
 De Cea, J., 358, 361
 De Palma, A., 113, 114
 Deffebach, C. (see Schneider, J. B. et al., 155,
 159)
 Dell Amico, M., 193, 195
 Dell Site, P., 477
 Deo, N. (see Syslo, M. M. et al., 430)
 Derosiers, J. (see Ioachim, I. et al., 487)
 Desai, N. (see Lee, Y. J. et al., 588)
 Desauliniers, G. (see Haase, K. et al., 194, 195,
 306)
 Desilets, A., 156, 158
 Desrochers, M., 6, 290, 306, 308
 Desrosiers, J., 170 (see also Haase, K. et al.,
 194, 195, 306)
 Dessouky, M., 114, 563, 565 (see also Jiamin,
 Z. et al., 572)
 Dhingra, S. L., 516, 518
 Dial, R. B., 358, 360, 409, 571, 590 (see also
 Ball, M. et al., 294)
 Dijkstra, E., 292, 315
 Di-Miele, F., 114
 Dressler, O., 55
 Driscoll, M. K. (see Abkowitz, M. et al., 155,
 158)
 Dubois, D., 410
 Ducker, K. J. (see Strathman, J. G. et al., 478)
 Duckstein, L. (see Goicoechea, A. et al., 441)
 Dudek, C. L. (see Balke, K. N. et al., 589)
 Dueker, K. J., 587
 Dumas, Y. (see Desrosiers, J. et al., 170;
 Ioachim, I. et al., 487)

E

Eberlein, X. J., 114, 562, 565
 Edvardson, B. (*see* Friman, B. *et al.*, 400)
 Eiger, A. (*see* Abkowitz, M. *et al.*, 561, 564)
 Eiselt, H. A., 508
 Elefsiniotis, S. S. (*see* Chien, S. *et al.*, 272, 516, 572)
 Engelstein, I. (*see* Abkowitz, M. *et al.*, 561, 564)
 Ernst, A., 300
 Evans, J., 409
 Even, S., 174

F

Farewell, R. G., 381
 Farvolden, J. M., 410
 Fernández, J. E., 358, 361
 Fielding, G. J., 574, 591
 Filippi, F., 477
 Fingerle, G. P., 587
 Fischett, M. (*see* Dell Amico, M. *et al.*, 193, 195)
 Fischetti, M., 307, 309
 Florian, M., 356, 358, 359, 360 (*see also* Cepeda, M. *et al.*, 360; Wu, J. H. *et al.*, 359)
 Ford, L. R. Jr., 172, 174, 187, 193
 Fores, S., 306, 307, 309
 Freling, R., 194, 195, 290, 307, 308, 309 (*see also* Huisman, D. *et al.*, 194, 195, 307, 308)
 Friedrich, M., 339, 340
 Friman, M., 400, 401
 Fu, L., 487, 562, 565
 Fulkerson, D. R., 172, 174, 187, 193
 Furth, P. G., 38, 40, 51, 52, 75, 78, 114, 338, 339, 340, 476, 477, 560, 564

G

Gallo, G., 114, 300 (*see also* Bertossi, A. *et al.*, 252)
 Gao, Z. Y. (*see* Lam, W. H. K. *et al.*, 359)
 Garling, T. (*see* Edvardson, B. *et al.*, 400)
 Gavish, B., 171
 Gehner, C. D., 154, 159
 Gendreau, M., 571 (*see also* Bouzaïene-Ayari, B. *et al.*, 359, 361; Eiselt, H. A. *et al.*, 508; Li, Y. *et al.*, 561, 564)
 Gerhart, R. L. (*see* Strathman, J. G. *et al.*, 478)
 Gertsbach, I., 131, 179, 188
 Ghiani, G., 572
 Gildea, D., 589
 Gillen, D., 588
 Girard, L. (*see* Carraresi, P. *et al.*, 306)

Glazebrook, G., 590
 Gleason, E. (*see* Kyte, M. *et al.*, 155, 158)
 Goicoechea, A., 441
 Golany, B. (*see* Ceder, A. *et al.*, 142, 148, 154)
 Golden, B. (*see* Bodin, L. *et al.*, 291)
 Gonen, D., 171
 Gonzalez, H. (*see* Ceder, A. *et al.*, 410)
 Gonzalez, O. (*see* Ceder, A. *et al.*, 410)
 Goodman, L. A. (*see* Wilson, N. H. M. *et al.*, 487)
 Greenfeld, J., 588
 Griffin, D. (*see* Strathman, J. G. *et al.*, 478)
 Gronau, R., 271, 273, 274
 Grotschel, M. (*see* Borndorfer, R. G. *et al.*, 487)
 Guan, J. F., 355
 Guenther, R. P., 558, 559
 Guerriero, F. (*see* Ghiani, G. *et al.*, 572)
 Guertin, F. (*see* Melucelli, F. *et al.*, 487)
 Gur, Y. J., 339, 340
 Gurevich, Y., 131, 179, 188
 Guttenplan, M., 400, 401
 Gwilliam, K. M., 268, 274

H

Haase, K., 194, 195, 306
 Haghani, A., 194, 195, 306, 308
 Hakimi, S. L., 383
 Hall, R. W., 155, 158 (*see also* Dessouky, M. *et al.*, 114, 563, 565)
 Hallowell, S. F., 556, 557, 561, 565
 Hamerslag, R. (*see* Van Nes, R. *et al.*, 445)
 Hamilton, 38
 Handler, G. Y., 381, 382, 383, 386
 Hansen, D. R. (*see* Goicoechea, A. *et al.*, 441)
 Harker, P. T., 556, 557, 561, 565
 Heathington, K. W., 409
 Henderson, G., 555, 557
 Hendrickson, C., 409
 Henk, R. H., 399, 401
 Hensher, D., 328, 329, 400, 401
 Heydecker, B. G. (*see* Zhou, J. *et al.*, 330)
 Hickman, M., 6, 290, 359, 361, 542
 Higonnet, B. T. (*see* Wilson, N. H. M. *et al.*, 487)
 Hill, D. L. (*see* Lee, Y. J. *et al.*, 588)
 Hobeika, A. G., 445, 448
 Hoff, G. C. (*see* Heathington, K. W. *et al.*, 409)
 Holland, R. (*see* Blythe, P. T. *et al.*, 587)
 Horbury, A. X., 588
 Horn, M. E. T., 591
 Horowitz, A. J., 399, 401
 Hou, E. (*see* Chien, S., 515, 518)

Hounsell, N., 588, 590
 Hsia J. S. (see Barua, B. *et al.*, 39, 40)
 Hsu, P. S. (see Ben Akiva, M. *et al.*, 338, 339)
 Huang, R., 588
 Hubbard, S. M., 399, 401
 Huddart, K. W., 478, 541, 542
 Huisman, D., 194, 195, 307, 308 (see also
 Freling, R. *et al.*, 194, 195, 307)
 Hurdle, V. F., 113
 Hutchinson, T. P., 349, 540, 541
 Hwang, M., 542

I

Ibrahim, W. H. W. (see Madanat, S. M. *et al.*,
 399, 401)
 Immers, B. H. (see Van Nes, R. *et al.*, 445)
 Ioachim, I., 487
 Israeli, Y., 345, 352, 353, 410, 419, 438, 442,
 444

J

Jan, O., 589
 Jansson, J. O., 268, 274
 Jansson, K., 269, 273, 274
 Jeong, R., 533, 589
 Jerby, S., 483, 503, 507
 Jiamin, Z., 572
 Jiang, H. (see Ernst, A. *et al.*, 300)
 Johnson, D. (see Gillen, D. *et al.*, 588)
 Jolliffe, J. K., 349, 540, 541
 Jordan, W. C., 560, 564
 Josef, R. (see Abkowitz, M. *et al.*, 155)

K

Kane, A. R., 591
 Karlaftis, M. G. (see Washington, S. P. *et al.*,
 33, 34)
 Kemp, J. (see Hwang, M. *et al.*, 542)
 Keudel, W., 410
 Khasnabis, S., 77, 78
 Khattak, J. A., 542
 Kim, D., 410
 Kimpel, T. J., 544, 546 (see also Strathman,
 J. G. *et al.*, 478)
 Klemmt, W. D., 155, 156, 158
 Klostemeier, F. (see Borndorfer, R. G. *et al.*, 487)
 Knoblauch, M. (see Ryus, P. *et al.*, 400, 401)
 Knox, R. R. (see Heathington, K. W. *et al.*, 409)
 Kocur, G., 409
 Koffman, D., 560, 564

Koshi, M., 447, 449
 Kottenhoff, K., 325, 326
 Koutsopoulos, H. N., 76, 78, 114
 Kowalit, J. S. (see Syslo, M. M. *et al.*, 430)
 Krishnamoorthy, M. (see Ernst, A. *et al.*, 300)
 Kroon, L., 114, 307, 309
 Kuah, G. K., 409, 513, 515, 517
 Kuo, S. H. F. (see Lee, K. K. T. *et al.*, 270)
 Kurauchi, F., 360, 362
 Kuttner, C. (see Borndorfer, R. G. *et al.*, 487)
 Kwan, A. S. K., 193, 194, 195, 306, 308
 Kwan, R. S. K., 193, 194, 195, 307 (see also
 Kwan, A. S. K. *et al.*, 306)
 Kwong, P. (see Henderson, G. *et al.*, 555, 557)
 Kyte, M., 155, 158

L

Lam, W. H. K., 359, 361 (see also Yin, Y. *et al.*,
 556, 557; Zhou, J. *et al.*, 330)
 Lamont, G. B. (see Coello Coello, C. A. *et al.*,
 428, 439)
 Lampkin, W., 410
 Landis, B. W. (see Guttenplan, M. *et al.*, 400, 401)
 Laporte, G. (see Eiselt, H. A. *et al.*, 508;
 Gendreau, M. *et al.*, 571; Ghiani, G.
 et al., 572)
 Larson, R. C., 524, 538
 Larwin, T. F. (see Campion, D. R. *et al.*, 590)
 Le Clercq, F., 358, 360
 LeBlanc, L. J., 76, 78
 Lee, C. J., 515, 517
 Lee, K. K. T., 156, 158, 270, 274
 Lee, Y. J., 588
 Lentink, R. M. (see Freling, R. *et al.*, 194, 195,
 307)
 Lerner-Lam, E. (see Hwang, M. *et al.*, 542)
 Lesley, L. J. S., 560, 564
 Levinson, H. S., 590
 Li, Y., 561, 564
 Libre, M. (see Dubois, D. *et al.*, 410)
 Lin, W. H., 589
 Lindh, C., 326
 Lindsey, R., 113, 114
 List, G. F., 445, 448
 Litman, T., 327, 328
 Liu, G., 561, 562, 565
 Lo, H. K., 447, 449
 Löbel, A., 193, 195
 Lock, A. C., 587
 Lourenco, H. R., 300, 306, 308

M

- Macbriar, I. D., 39, 40
 MacDonald, R. (see Bamford, C. G. *et al.*, 38, 40)
 Macke, P. P. (see Ben Akiva, M. *et al.*, 338, 339, 340)
 Mackett, R. (see Balcombe, R. *et al.*, 322–329, 523, 533)
 Mackie, P. J. (see Gwilliam, K. M. *et al.*, 268, 274)
 Maclean, S. D., 588, 589 (see also Dailey, D. J. *et al.*, 588, 589)
 Madanat, S. M., 399, 401
 Madeean, J. (see Blythe, P. T. *et al.*, 587)
 Magnanti, T. L. (see Ahuja, R. K. *et al.*, 130, 174, 193, 200, 205, 293, 317, 383)
 Mahmassani, H. S., 446, 448 (see also Shih, M. C. *et al.*, 156, 159, 274)
 Mandel, M. (see Nelson, M. *et al.*, 155, 159)
 Mandl, C. E., 410
 Mannering, F. L. (see Washington, S. P. *et al.*, 33, 34)
 Marcotte, P. (see Wu, J. H. *et al.*, 359, 361)
 Marguier, P. H. J., 345, 346, 348, 349, 350, 352, 537, 541
 Marlin, P. G., 114
 Marwah, B. R., 445, 448
 Marx, E., 381
 Maxwell, R. R., 157, 159
 McCollom, B., 38, 40
 McLeod, D. S. (see Guttenplan, M. *et al.*, 400, 401)
 Melucelli, F., 487
 Mesquita, M., 194, 195
 Miller, E. J., 476, 477
 Miller, J. (see Heathington, K. W. *et al.*, 409)
 Miller, M. (see Yin, Y. *et al.*, 410, 417, 418, 419)
 Mingozi, A., 306, 308 (see also Bianco, L. *et al.*, 300)
 Minieka, E., 381, 382, 383
 Minsky, M. L., 572
 Mirchandani, P. (see Hickman, M. *et al.*, 6, 290)
 Mitra, G., 305, 308
 Mohan, S., 448, 449
 Mora, J. G., 9
 Morello, E. (see Nguyen, S. *et al.*, 339, 340)
 Mott, P. (see Friedrich, M. *et al.*, 339, 340)
 Mourikas, K. (see Dessouky, M. *et al.*, 114)
 Murugesan, R., 400, 401
 Musmanno, R. (see Ghiani, G. *et al.*, 572)

N

- Nalevanko, A. (see Stone, J. R. *et al.*, 590)
 Nash, A. B., 560, 564
 Nash, C. A., 278 (see Gwilliam, K. M. *et al.*, 268, 274)
 Nauss, R. M. (see Marlin, P. G. *et al.*, 114)
 Naverrete, G., 39, 40
 Navick, D. S., 339, 340
 Nelson, M., 155, 159
 Neuerburg, N. (see Hwang, M. *et al.*, 542)
 Newell, G. F., 113, 552, 553, 558, 559, 560, 564
 Nguyen, S., 339, 340, 358, 359, 360 (see also Bouzaïene-Ayari, B. *et al.*, 359, 361)
 Niemeyer, J. (see Pine, R. *et al.*, 3)
 Noekel, K. (see Friedrich, M. *et al.*, 339, 340)
 Nonato, M. (see Carraresi, P. *et al.*, 306; Melucelli, F. *et al.*, 487)
 Norris, S., 300
 Nowroozi, A. (see Dessouky, M. *et al.*, 114)
 Nuzzolo, A., 340, 345, 358, 359, 362

O

- O'Dell, S. W., 562, 565
 Odijk, M. A. (see Freling, R. *et al.*, 194, 195, 307)
 Odoni, A. (see Koutsopoulos, H. N. *et al.*, 76, 78, 114)
 Odoni, A. R., 524, 538
 Okunieff, P. (see Hwang, M. *et al.*, 542)
 Oldfield, R. H., 268, 273, 274
 O'Neill, K. K. (see Tykulska, R. J. *et al.*, 294)
 Orlin, J. B. (see Ahuja, R. K. *et al.*, 130, 174, 193, 200, 205, 293, 383)
 Ortuzar, J. de D., 322, 331, 336
 Osuna, E. E., 113, 553, 558, 559, 564

P

- Padberg, M. W., 385
 Padmanabhan, V., 589
 Païas, A., 306, 308
 Paixao, J. M. P., 170, 194, 195, 300, 306, 308 (see also Costa, A. *et al.*, 251; Daduna, J. R. *et al.*, 6, 290; Freling, R. *et al.*, 194, 195, 290, 307; Lourenco, H. R. *et al.*, 300, 306, 308)
 Pallotino, S., 358, 359, 360 (see also Nguyen, S. *et al.*, 339, 340)
 Passy, U. (see Silman, L. A. *et al.*, 410)
 Patnaik, S. B. (see Marwah, B. R. *et al.*, 445, 448)
 Paulley, N. (see Balcombe, R. *et al.*, 322–329, 523, 533)

- Peeters, L., 114
 Peng, Z. R., 588, 589
 Perl, J., 409, 513, 515, 517
 Pine, R., 3
 Polus, A., 399, 401, 555, 557
 Portugal, R. (*see* Lourenco, H. R. *et al.*, 300, 306, 308)
 Powell, W. B., 410
 Prashker, J. (*see* Ceder, A. *et al.*, 381)
 Pratt, R., 409
 Preston, J. (*see* Balcombe, R. *et al.*, 322–329, 523, 533)
 Prioni, P., 400, 401
 Proll, L. (*see* Fores, S. *et al.*, 306, 307, 309)
- R**
 Rackliff, T. (*see* Blythe, P. T. *et al.*, 587)
 Rahin, M. A., 193, 194, 195
 Raiffa, H., 577
 Rainville, W. S., 3
 Rama Moorthy, N. A., 400, 401
 Ramirez, A. I., 446, 448
 Rapp, M. H., 154, 159
 Reynolds, F. D., 586
 Rhoades, M. (*see* Marlin, P. G. *et al.*, 114)
 Ricciardelli, S. (*see* Bianco, L. *et al.*, 300; Mingozzi, A. *et al.*, 308)
 Richardson, A. J., 409
 Richardson, T., 39, 40
 Rilett, L. R., 533, 589
 Robillard, P., 358, 360
 Rosenbloom, S., 574, 591
 Rossetti, M. D., 39, 41
 Rousseau, J. M., 6, 156, 158, 290 (*see also* Li, Y. *et al.*, 564)
 Rubin, J., 385
 Rudnicki, A., 556, 557
 Rudraraju, R. K., 77, 78
 Russo, F. (*see* Nuzzolo, A. *et al.*, 340, 359, 362)
 Ryus, P., 400, 401
- S**
 Saalmans, P. D., 410
 Sakano, R., 590
 Salkin, H., 437
 Salzborn, F. J. M., 87, 131, 155, 159, 168, 179, 485
 Schmöcker, J.-D., 345, 358, 359 (*see also* Kurauchi, F. *et al.*, 360, 362)
 Schneider, J. B., 155, 159
 Schonfeld, P. M., 156, 158, 272, 274, 409, 446, 448, 515, 517, 518 (*see also* Lee, K. K. T. *et al.*, 270, 274)
 Schumann, J. W. (*see* Champion, D. R. *et al.*, 590)
 Schweitzer, P. (*see* Gavish, B. *et al.*, 171)
 Seguin, R. (*see* Gendreau, M. *et al.*, 571)
 Seneviratne, P. N., 446, 448, 561, 564
 Shefer, D., 399, 401
 Sheffi, Y. (*see* Tykulsker, R. J. *et al.*, 294)
 Sheikh, M., 589
 Shen, S., 478, 562, 565
 Shen, Y., 307, 308
 Shih, M. C., 156, 159, 270, 274
 Shires, J. (*see* Balcombe, R. *et al.*, 322–329, 523, 533)
 Shlifer, E. (*see* Gavish, B. *et al.*, 171)
 Shrivastava, P., 516, 518
 Sier, D. (*see* Ernst, A. *et al.*, 300)
 Silcock, D. T., 555, 557
 Silman, L. A., 410
 Simon, J., 338, 340
 Singh, A. (*see* Dessouky, M. *et al.*, 563)
 Sinha, K. C., 558, 559
 Siuru, W. D., 590
 Slavin, H. (*see* Abkowitz, M. *et al.*, 525, 528, 546)
 Smith, B. L., 588
 Smith, B. M., 193, 306, 308
 Smith, L. D. (*see* Marlin, P. G. *et al.*, 114)
 Smith, R. L., 445
 Smith, R. G., 577
 Sodhi, M., 300
 Soehodo, S., 447, 449
 Solomon, M. (*see* Ioachim, I. *et al.*, 487)
 Solomon, M. M. (*see* Desrosiers, J. *et al.*, 170)
 Soumis, F., 306, 308 (*see also* Desrosiers, J. *et al.*, 170)
 Spadoni, M. (*see* Bianco, L. *et al.*, 300)
 Spasovic, L. N., 446, 448 (*see also* Chien, S. *et al.*, 272, 516)
 Spielberg, F., 157, 158
 Spiess, H., 356, 358, 360
 Spring, G. S., 590
 Stanley, K. (*see* Kyte, M. *et al.*, 155, 158)
 Stemme, W., 155, 156, 158
 Stern, H. I., 114, 121, 132, 173, 176, 179, 184, 188, 210, 211 (*see also* Ceder, A. *et al.*, 381)
 Stone, J. R., 587, 590
 Strathman, J. G., 478 (*see also* Kimpel, T. J. *et al.*, 544, 546)
 Subramanian, S., 590

Sun, W., 588
 Sussman, J. M. (*see* Wilson, N. H. M. *et al.*, 487)
 Syslo, M. M., 430

T

Tal, O., 142, 148, 158
 Taplin, M., 329
 Tarjan, R. E., 174
 Teaf, D. (*see* Ryus, P. *et al.*, 400, 401)
 Tisato, P., 271, 274
 Titheridge, H. (*see* Balcombe, R. *et al.*, 322–329, 523, 533)
 Tom, V. M., 448, 449
 Tong, C. O., 339, 340, 359, 361
 Toth, P. (*see* Dell Amico, M. *et al.*, 193, 195)
 Tozzi, J. (*see* Abkowitz, M. *et al.*, 155, 158)
 Trepanier, M., 588
 Tsai, J. (*see* Stone, J. R. *et al.*, 590)
 Tsao, S., 409
 Tsygalnitsky, 338, 339, 340
 Turnau, A., 561, 565
 Turner, K. (*see* Strathman, J. G. *et al.*, 478)
 Turner, R. P., 555, 557
 Turner, S. M., 589
 Turnquist, M. A., 558, 559, 560, 564
 Tykulsker, R. J., 294

U

Umrigar, F. S. (*see* Marwah, B. R. *et al.*, 445)
 Urbanik, T., 39, 40 (*see also* Balke, K. N. *et al.*, 589)

V

Valevanko, A. (*see* Stone, J. R. *et al.*, 587)
 Van Dyke Parunak, H., 572
 Van Nes, R., 445, 447, 448, 449
 Van Veldhuizen, D. A. (*see* Coello Coello, C. A. *et al.*, 428, 439)
 Vandebona, U., 409
 Vijayaraghavan, T. A. S., 477
 Villeneuve, D. (*see* Ioachim, I. *et al.*, 487)
 Viola, P., 381
 Vorvick, J., 587
 Voss, S., 6, 155, 156, 158, 159, 290 (*see also* Hickman, M. *et al.*, 290)

W

Wagelman, A. P. M. (*see* Freling, R. *et al.*, 194, 195, 307, 307, 309; Huisman, D. *et al.*, 194, 307, 308)

Waksman, R. (*see* Abkowitz, M. *et al.*, 525, 528, 546)
 Wall, G., 588, 590
 Wall, Z. R. (*see* Dailey, D. J. *et al.*, 588, 589)
 Wan, Q. K., 447, 449
 Wang, W., 400, 401
 Wardman, M. (*see* Balcombe, R. *et al.*, 322–329, 523, 533)
 Washington, S. P., 33, 34
 Weinstein, A., 39, 40
 White, P., 555, 557 (*see also* Balcombe, R. *et al.*, 322–329, 523, 533)
 Willumsen, L. G., 322, 331, 336
 Wilson, N. H. M., 4, 6, 51, 52, 75, 78, 114, 141, 290, 478, 487, 562, 565 (*see also* Abkowitz, M. *et al.*, 525, 528, 546; Eberlein, X. J. *et al.*, 114, 562, 565; Koutsopoulos, H. N. *et al.*, 76)
 Wirasinghe, S. C., 77, 78, 113, 513, 552, 561, 562, 565 (*see also* Guan, J. F. *et al.*, 355)
 Wolsfeld, R. P. Jr. (*see* Campion, D. R. *et al.*, 590)
 Wong, S. C., 339, 340, 359, 361
 Wren, A., 6, 193, 290, 306, 308 (*see also* Fores, S. *et al.*, 306, 307, 309; Kwan, A. S. K. *et al.*, 306)
 Wu, J. H., 359, 361

X

Xu, Y., 487

Y

Yan, S., 447, 449
 Yang, H. (*see* Guan, J. F. *et al.*, 355; Lam, W. H. K. *et al.*, 359)
 Yang, X., 400, 401, 562, 565
 Yang, Z. (*see* Chien *et al.*, 515)
 Yim, Y. B., 483, 498
 Yin, Y., 410, 417, 418, 419, 556, 557
 Yu, W. J., 515, 517

Z

Zeleny, M., 442
 Zeng, J., 589
 Zhang, L. (*see* Dessouky, M. *et al.*, 563, 565)
 Zhang, X. (*see* Barua, B. *et al.*, 39, 40)
 Zhou, J., 330

This page intentionally left blank

Subject Index

A

Advanced public transit/transportation systems (APTS), 542–546
fare payment, 543
monitoring, 543
multi-purpose information, 543–544
traffic-signal control, 544
traveller information, 543
Advanced technologies, 587–589
AVL, 588
GIS, 39, 588
GPS, 589
web-based GIS, 588
world wide web, 587–588
Advanced traveller information system (ATIS), 542–546
AFP – *see* Automatic fare payment (AFP)
Agency, 15–16, 413–416, 525–528, 576
network route design perspective, 413–416
reliability attributes, 525–528
service viability, 15–16
AGV – *see* Automatic guided vehicle (AGV)
Alighting time, 525, 533
APC – *see* Automatic passenger counter (APC)
APM – *see* Automatic people movers (APM)
APTS – *see* Advanced public transit/transportation systems (APTS)
Arithmetic mean, 31
Assignment, 354–358
on chart, 16, 29
combined frequency, 356–357
network synthesis, 356–357
O-D path, 357–358
passenger-choice strategy, 356
practical characteristics, 346
route-choice based, 354–358
transit network, 346, 355–356
ATIS – *see* Advanced traveller information system (ATIS)

ATO – *see* Automatic train operation (ATO)
Authority constraints, 5
Automatic fare payment (AFP), 542–545
Automatic guided vehicle (AGV), 584–585
Automatic passenger counter (APC), 24, 28, 72, 94, 542–545
Automatic people movers (APM), 584
Automatic train operation (ATO), 584
Automatic vehicle location (AVL), 24, 39, 542–543, 588
Automatic vehicle location and communication (AVLC), 542–545
Automatic vehicle monitoring (AVM), 24, 542
Automation, 584–586
APM, 584
ATO, 584
dual-mode concept, 586
elevated rail, 584–585
elevated-PRT, 585–586
LIM, 584
Average, 32
AVL – *see* Automatic vehicle location (AVL)
AVLC – *see* Automatic vehicle location and communication (AVLC)
AVM – *see* Automatic vehicle monitoring (AVM)

B

Baggage size, 24
Balanced scheduling, 191–195
formulation, 193–195
BART – *see* Bay Area Rapid Transit
Bay Area Rapid Transit (BART), 483–484, 494–496
Block, 5, 7, 169–170, 190
shuttle service, 486
Blocking – *see* Vehicle scheduling

Boarding time, 533
 Branching route, 24
 BRT – *see* Bus rapid transit (BRT)
 Bus rapid transit (BRT), 418–419, 549, 590

C

CATI – *see* Computer-aided telephone-interview (CATI)
 CBD – *see* Central business district (CBD)
 Central business district (CBD), 9, 52, 168, 369, 446
 Chains of trips – *see* Block
 Chi-square statistic, 55–56
 Clock headway, 54, 87, 100
 Combined frequency, 356–357
 Comfort, 14, 15, 325
 Communication, 14, 389, 571
 Company – *see* Agency
 Comparison measures, 5, 87–88
 Complexity, 128–130
 efficient algorithm, 130
 exponential-time algorithm, 130
 NP-complete, 130, 170, 286, 290, 430, 439
 worst-case scenario, 130, 144
 Computer assisted design, 542
 Computer-aided dispatch, 487, 543
 Computer-aided telephone-interview (CATI), 498–499
 Concessions, 15
 Confidence interval, 33
 Confidence level/coefficient, 33
 Connection point – *see* Transfer point
 Connectivity, 367–368, 389–398
 attractiveness, 389
 attributes, 389
 qualitative, 389–390
 quantitative, 390
 detecting weakness of, 394–395
 base data, 398
 finding bottleneck, 395–398
 inter-modal chain/path, 394–395
 inter-route chain/path, 394–398
 measures, 389–394
 MOCP, 371–373
 network-flow model, 397–398
 rapidly, 389
 reliability, 389
 smoothness (ease), 389
 standards, 371–374

 synchronized, 389
 transit path, 388–389
 Control, 549–552
 on charts, 16, 29
 strategies, 550–551
 Cost, 5
 elements, 5
 operating, 9, 11, 13, 15, 267
 per passenger, 11, 13
 recovery ratio, 11, 13
 Cost per covered cell (CPCC), 431, 434–438
 CPCC – *see* Cost per covered cell (CPCC)
 Crew, 5
 assignment, 5, 281, 295–299
 list, 5
 rosters, 5, 8, 300–305
 wages, 15
 work rules, 5, 295–299
 Crew rostering, 300–305
 heuristic approach, 301–305
 problem (CRP), 300–305
 RNN algorithm, 303–305, 317–318
 Roster, 302–305
 Crew scheduling, 4–5, 8–9, 280–312
 arrival-departure joinings, 282–283
 case study, 294–299
 on charts, 16, 29
 DH trips, 297–298
 hollows, 282–283
 maximum unpaid times, 286–290
 optimization, 283–286
 problem (CSP), 290–294, 305–307
 SCP, 290–294, 295–297
 SPM, 290–294
 SPP, 290–293, 299–300
 sub-functions, 281
 vehicle-chain construction, 282–290
 FIFO rule, 285–290
 Crowding level, 10, 12, 52–56, 358, 399
 desired occupancy, 52–56
 maximum allowable standees, 52–56
 timetable with minimum, 130–137
 fleet-size formula, 131
 lower bound, 130, 136
 upper bound, 130–131, 136
 variable scheduling, 135–137
 CRP – *see* Crew-rostering problem (CRP)
 CSP – *see* Crew-scheduling problem (CSP)
 Customer survey, 498–502

D

- Data, 23
 - adequate, 23
 - coordination, 23
- Data collection, 23, 67, 589
 - automated, 23
 - manual, 23
 - techniques, 24–26
- Data requirement, 27–30
 - group-of-routes level, 27–28
 - regional level, 27–28
 - route level, 27–28
- Dead time (of dwell time), 533
- Deadhead check, 26, 28
 - interlining route, 26
 - shortest path, 26, 29
- Deadheading (DH) time, 5, 26, 29
 - matrix, 8
- Deadheading (DH) trip, 7, 176–193
 - crew scheduling, 297–298
 - heuristic algorithm, 183–188
 - insertion, 176–190
 - VTSP, 251–266
- Decision, 23, 51
 - intelligent, 23, 51
- Deficit function (DF), 7, 176–188, 209–236
 - crew scheduling, 282–290
 - hollows, 282–283
 - vehicle-chain construction, 282–290
 - description of, 241–247
 - fixed schedule, 176–188
 - hollow, 177, 188
 - network route design, 417
 - short-turn trips, 457–479
 - surplus-function related, 374–378
 - URDHC, 181–188, 225–229, 254–263, 378–379, 472
 - URSC, 220, 228–229, 254
 - variable schedule, 209–236
 - VTSP, 251–266
 - website of, 183, 228
- Demand (of passengers), 321–340
 - on charts, 16, 29
 - elasticity, 326–330
 - arc elasticity, 326
 - cross-elasticity, 327, 329
 - direct elasticity, 326–327, 329
 - point elasticity, 326
 - external factors, 323
 - alternative modes, 323
 - cost, 323
 - economic conditions, 323
 - funding initiatives, 323
 - land use, 323
 - policies, 323
 - population characteristics, 323
 - travel conditions, 323
 - factors affecting, 322–326
 - coordination, 323
 - fare collection and structure, 323
 - marketing, 323
 - partnership, 323
 - service adjustments, 323
 - function, 326–327
 - mode choice elements, 324
 - price, 324
 - quality of service, 324
 - socio and demographic, 324
 - trip characteristics, 324
 - potential, 505, 507
 - willingness to pay, 325
 - comfort, 325
 - on-board service, 325
 - quality satisfaction, 325
 - timetable factors, 325
- Demand forecasting, 330–340
 - attribute weighting, 334–336
 - attributes, 332
 - input–output, 332
 - MNL model, 336–338
 - modal split, 332
 - O-D estimation, 338–340
 - see also* Origin-destination
 - survey, 331–334
- Demand-responsive transit (DRT), 487–490, 590–591
- Departure time, 5, 214–220
 - left-shift limit, 220
 - modified shortest-path related, 464–467
 - right-shift limit, 220
 - shifting of, 5, 214–220, 224–229
 - tolerance, 215–218
- Depot, 188–189
- Depot-constrained – *see* Balanced scheduling
- DF – *see* Deficit function (DF)
- DH – *see* Deadheading (DH)
- Distribution, 30–32
 - bell-shaped, 32
 - normal, 33
- DP – *see* Dynamic Programming (DP)

- Driver, 7, 8
 - equality rules, 8
 - list of, 8
 - priority rules, 8
- DRT – *see* Demand-responsive transit (DRT)
- DT – *see* Departure Time (DT)
- Duty, 5, 8, 165, 281, 298
 - composition, 8
 - length of, 8
 - type of, 8
- Duty rosters, 5, 300–305
- Dwell time, 28, 531–533, 562, 575
 - boarding, alighting and dead times, 531, 532, 533
 - influencing factors, 532–533
- Dynamic programming (DP), 113, 306, 339–340, 560, 561
- E**
- Environmental Protection Agency (EPA), 371–372
 - MOBILE5, 371
 - MOBILE6, 372
 - PART5, 371–372
- EPA – *see* Environmental Protection Agency (EPA)
- Equity, 9, 399
- Estimation, 32–33
 - confidence interval, 33
 - confidence level/coefficient, 33
 - error of, 32, 33
 - goodness of, 32
 - precision, 32, 33
 - tolerance, 32, 33
- Evaluation standards, 11–13
- Even headway, 7, 54, 86, 90–94, 100–113
- Even load, 7, 86, 94–113
 - on individual vehicles, 121–127
 - load-tolerance criterion, 233–236
 - minimum frequency, 125
- F**
- Fare, 24, 322, 323, 325, 331, 333, 370
 - category, 24, 28, 38
 - type, 26
- Farebox, 24, 28, 38, 39
- Feeder, 16, 29, 155, 482, 515
- Feeder service – *see* Shuttle service
- FIFO – *see* First-in, first-out (FIFO)
- First-in, first-out (FIFO), 169–170, 188, 224, 259, 486
 - vehicle-chain construction, 285–286, 486
- Fleet size, 5, 179–184
 - formula, 131, 178
 - lower bound, 5, 179–183, 210–214, 220–223
 - minimum, 5, 179, 181–183, 188
 - NT rule, 184–185, 210, 257, 262
 - optimization with fixed, 130–137
 - reduced, 15
 - shuttle (circular) route, 485–486
 - single-route, 87–88, 100–108, 168–170, 215–218
 - URDHC, 181–188, 225–229
 - URSC, 228–229
- Frequency, 6, 49–80
 - combined for assignment, 360–361
 - cumulative curve, 91–92
 - determination of, 27–28, 49–80
 - real-life example, 64–73
 - histogram, 30–32
 - load profile (ride check) methods, 56–60
 - max load (point check) methods, 52–56
 - Method 1, 53–56, 72–77, 86, 92–93, 96–102, 108
 - daily max load point, 53–56
 - Method 2, 53–56, 58–60, 86, 91–94
 - hourly max load point, 54–56
 - Method 3, 58–60, 75–77, 86, 100–113
 - considered load level, 58
 - empty space (seat)-km, 56–60
 - Method 4, 58–60, 61–65, 75–77, 86, 93, 95–113
 - minimum even-load standard, 125
- Future development, 569–594
 - advanced technologies, 587–589
 - AVL, 588
 - GIS, 39, 588
 - GPS, 589
 - web-based GIS, 588
 - world wide web, 587–588
 - multi-agent systems (MAS), 571–572
 - multi-agent transit system (MATS), 575–578
 - agents, 575–576
 - benefits, 577–578
 - system automation, 572–574
 - RFID, 574

- vehicle encounters, 578–584
 - improvements, 581–584
 - probability of, 578–580
 - tactics, 580
 - time-space diagram, 578–583

G

- Gamma distributions, 348–352
 - proportion boarding each route, 349–352
- Gantt chart, 229–232
- Garages, 8
- Geographical information systems (GIS), 588
- GIS – *see* Geographical information systems (GIS)
- Global Positioning Systems (GIS), 589
- GPS – *see* Global Positioning Systems (GIS)

H

- Headway, 5, 49–80
 - clock, 54, 86, 87, 99, 100
 - coefficient of variation, 347
 - determination of, 27–28, 49–80
 - distributions, 348–352, 538–540
 - exponential, 350, 538
 - gamma distributions, 348–352, 539–540
 - power distributions, 348–352, 539–540
 - even, 7, 54, 86, 90–95, 100, 530
 - lower limit, 10, 12, 372
 - policy, 8, 51–53, 75
 - upper limit, 10, 12, 372
 - variance, 249, 537–538, 560
- Heuristic procedure, 128
 - crew scheduling, 281–290
 - network route design, 430–438
 - shuttle-service routing, 509–510
 - synchronization, 144–154
 - vehicle scheduling – *see* Deficit function
- Histogram, 30–32
 - mound-shaped, 31–32
- Hollow, 177, 188
 - crew scheduling, 282–283
 - short-turn trip, 470–474

I

- Information system, 14
 - off-line, 14
 - on-line, 14, 15

- Integer programming, 130, 133, 136, 430, 433
- Interchanges, 5
- Interlining, 5, 87
 - route, 26
 - scheduling, 167, 170
- Inter-modal chain/path, 394–398
- Inter-route chain/path, 394–398
- IP – *see* Integer programming (IP)

L

- Land-use characteristics, 5
- Layover time, 5, 28
- Line – *see* Route
- Linear programming (LP), 204–205, 445
- Load, 7, 52–56
 - considered load level, 58, 59
 - cumulative curve, 95
 - desired occupancy, 53, 56, 64, 68–72
 - level, 10, 12
 - maximum, 52–56
 - vehicle, 29
- Load factor, 8
- Load profile, 25–26, 56–60, 77, 86, 97, 122
 - density, 61–64
- Log-normal model, 61–64
- Lower bound, 5, 179–183
 - fleet-size, 179–183, 210–214, 210–214, 220–223
 - fixed schedule, 131, 172–176, 210–214
 - variable schedule, 220–223
 - max-flow, 205
- LP – *see* Linear programming (LP)

M

- MA matrix – *see* Matrix of binary parameters
- Market, 26
 - opportunity, 26
 - segmentation, 26, 28, 29
- MAS – *see* Multi-agent systems (MAS)
- Matrix of binary parameters, 428–430, 431, 434–435, 442
- MATS – *see* Multi-agent transit system (MATS)
- Max load, 52–56, 74, 86, 88
 - daily point, 53–56
 - hourly point, 53–56
 - on individual vehicles, 121–127
 - point, 24, 25, 26, 52–56
 - short-turn trips, 460–461

- Max-flow, 172–176, 200–205
 - algorithm, 200–205
 - augmenting-path, 174–176, 200–205
 - bipartite network, 1747
 - definitions, 200
 - labelling algorithm, 202
 - lower bound, 205
 - minimum-arc path, 176, 202
 - node with capacity, 205
 - sink node, 174
 - source node, 174
 - technique, 172–176
 - undirected arcs, 205
 - unlimited capacity, 205
- Meal break, 8
- Measures of connectivity performance (MOCP), 371–374
- Measures of system performance (MOSP), 371–374
- Merriam-Webster's Collegiate Dictionary, 30
- METRO, 9, 11, 18
- MIP – *see* Mixed integer programming (MIP)
- Mixed integer programming (MIP), 144, 478, 562
- MNL model – *see* Multinomial logit (MNL) model
- MOCP – *see* Measures of connectivity performance (MOCP)
- Monitoring, 10
 - level, 10
- MOSP – *see* Measures of system performance (MOSP)
- Motivation, 2–4
- Multi-agent systems (MAS), 571–572
- Multi-agent transit system (MATS), 575–578
 - agents, 575–576
 - benefits, 577–578
- Multinomial logit (MNL) model, 331, 336–338
- Multi-objective, 438–445
- Multiple point checks, 24
- N**
- Natural logarithm (e), 32
- NCTRD, 38, 40
- Network, 4, 5, 6–7, 407–449
 - coordinated, 15
 - design – *see* Network route design
 - level, 10
 - nodes and arcs, 8, 9, 200–205, 315–318
 - of transit routes, 355
- Network route design, 4, 5, 6–7, 16, 29, 407–449
 - complete set of routes, 428–438
 - formulation, 428–430
 - MA matrix, 429–430
 - CPCC, 431–438
 - algorithm, 434–438
 - heuristic approach, 430–434
 - numerical example, 434–437
 - methodology, 419–428
 - six elements, 419–421
 - multi-objective technique, 438–445
 - evaluation of, 442–444
 - iterative process, 439–440
 - master problem and SDP, 440–442
 - numerical experience, 444–445
 - O-D, 419, 420
 - objective functions, 410–419
 - applications, 417–419
 - BRT, 424
 - calculation of, 414–417
 - formulation, 412–417
 - four criteria, 410–411
 - three perspectives, 410
 - SCP, 420, 421–422, 425, 426, 428, 429–430
- Network synthesis, 356–357
- Next terminal (NT) rule, 184, 185, 210–211, 257
- NLP – *see* Nonlinear programming (NLP)
- Nonlinear programming, 430, 438
- Normal curve, 32
- NP-complete, 130, 170, 286, 290
- NT rule – *see* Next terminal (NT) rule
- O**
- O-D – *see* Origin-destination (O-D)
- On-board, 14, 26
 - features, 14
 - load, 24
- Operation, 4, 23
 - astutely, 23
 - strategies, 370
 - tactics, 580, 584
- Operational parking conflicts – *see* Surplus function (SF)
- Operational planning, 4–8
 - decomposition process, 4–6
- Operations research (OR), 127–130, 171
 - cost-flow network problem, 251–252
 - DP, 306, 307, 584

integer programming, 171
 IP, 430
 LP, 204–205, 430
 MIP, 144, 252
 multi-commodity network flow, 252
 network-flow model, 395–397
 NLP, 430
 optimum stop location, 381–388
 SCP, 290, 295, 297, 299, 306, 307
 shortest-path algorithms, 315–318
 modified, 464–466
 SPM, 291–294
 SPP, 290–291, 299, 300, 306
 transportation problem, 171
 vehicle-chain construction, 282–290
 Operator – *see* Agency
 Optimization, 127–130
 constraints, 128
 heuristic procedure, 128
 objective function, 128
 symbols, 128
 variables, 128
 Optional timetables, 85–87
 OR – *see* Operations research (OR)
 Origin-destination (O-D), 14, 34–35, 338–340
 chain, 14
 demand, 414, 419–420, 421–422
 estimation, 338–340
 matrix, 35–36
 per passenger, 24
 shortest path, 27, 28

P

Passenger, 7
 alighting time, 29
 boarding time, 29
 count of, 7, 24
 gender of, 24
 load – *see* Load
 measures, 11
 safety, 10, 12, 372, 527
 shelters, 10, 12, 372
 survey, 24, 26
 per trip, 11, 13
 per vehicle-hour, 11, 13, 373
 per vehicle-km, 11, 13, 373
 willingness to pay, 325–326
 comfort, 331
 on-board service, 331
 quality satisfaction, 331

 timetable factors, 331
 Passenger assignment – *see* Assignment
 Passenger demand – *see* Demand
 Patronage, 5, 13, 15, 332
 Peak load factor – *see* Max load
 Peak load point – *see* Max load
 Performance, 11
 on-time, 14, 28, 29
 relative, 11, 13
 Personal rapid transit (PRT), 9, 584–585, 590
 Pi (π), 35
 Planning, 4–8
 intelligent, 23
 level, 10
 process, 4–9
 Point check, 24–25, 28, 52–56, 86
 branching route, 24
 criterion, 60–64
 max load point, 25, 52–56
 multiple point checks, 24
 route segment, 24
 strategic point check, 24
 transfer point, 24
 Policy headway, 8, 52, 53, 75
 Population, 24, 26, 30, 323
 survey, 24, 26, 28
 Power distributions, 348–352, 539–540
 proportion boarding each route, 352–354
 Practice, 23
 Precision, 32, 37
 Priority treatments, 548–549
 BRT, 549
 exclusive lanes, 548
 at intersection, 548
 results, 550
 at stops, 548
 Probability, 31
 Productivity, 28
 assessment, 28
 Proportion, 33
 of observations, 33
 PRT – *see* Personal rapid transit (PRT)
 PRT (report), 584–586, 590
 Public complaints, 10, 12, 372, 527
 Public transportation – *see* Transit
 Pull-ins, 8
 Pull-outs, 8

Q

QUATTRO. 9, 11

R

Radio frequency identification (RFID), 574

Real-life example

frequency determination, 51, 56, 77

synchronization, 154

vehicle-type scheduling, 251–253

Rearrange-node-number (RNN) algorithm,
317–318

Recovery time, 5, 28

Reliability, 522–565

APTS, 542–546

fare payment, 543

monitoring, 543

multi-purpose information, 543–544

traffic-signal control, 544

traveller information, 543

attributes, 525–528

of agency, 525–528

exogenous, 526–528

maintenance-based, 526, 528

of passengers, 525–528

on charts, 16, 29

control of, 542–546, 559–563

holding, 552–554

strategies, 550–552, 559–563

data related, 542–546

dwell time, 531–533

boarding, alighting and dead times, 533

influencing factors, 532–533

holding strategy, 552–554

even-headway, 553–554

even-load, 553–554

headway-based, 552

scheduling-based, 552

improved, 15

measures of, 525–530, 555–557

modeling variables, 530–536

passenger waiting time, 537–542

distributions, 537–539

priority treatments, 548–550

BRT, 549

exclusive lanes, 548

at intersection, 548

results, 550

at stops, 548

real-time information, 552

solution techniques, 546–554

control, 549–552

holding, 552–554

planning and scheduling, 546–547

priority for vehicles, 548–549

sources of problems, 528–530

maintenance indicators, 529

operational indicators, 529

planning indicators, 528–529

travel time, 533–536

bus data set, 535–536

method, 534–535

waiting time, 348–352, 537–542, 557–559

average of, 539–542

gamma distributions, 348–352, 539–540

power distributions, 348, 352, 539–540

Relief point, 5

location, 8

Revenue, 11

counts, 28

fare category, 29

RFID – *see* Radio frequency identification
(RFID)

Ride check, 24–25, 28, 56–61, 86, 99

criterion, 60–64

load profile – *see* Load profile

passenger counts, 24, 77

Ridership, 9, 370

see also DemandRNN algorithm – *see* Rearrange- node-
number (RNN) algorithm

Rotation rules, 5

Route, 5, 408–452

complete set of, 428–438

connectivity, 6, 10, 12, 372

see also Connectivity

coverage, 10, 12

design – *see* Network route design

directness, 10, 12, 372

express, 27–28

feasible, 6

fixed, 5

length, 10, 12, 372

level, 10, 27–28

local, 27–28

new, 28

optimal (in network design), 408–452

optimal (for shuttle service), 503–509

overlapping, 6, 10, 12, 372

structure, 10, 12, 372

Route Choice, 344–354

difference of waiting times, 353–354

frequency proportion curve, 354

frequency-share curves, 351–352

- proportion of boarding each route, 349–352
 - gamma distributions, 348–350
 - power distribution, 348–350
- proportions for regular vehicle arrivals, 352–354
- waiting-time dilemma, 346–347
 - assumptions, 349
- waiting-time strategy, 345–349
- Routing strategies, 487–498
 - bi-directional, 488–498
 - fixed route, 488–498
 - flexible route, 488–498
 - flexible schedule, 488–498
 - optimal routing (for shuttle service), 503–509
 - short cut, 488–498
 - short turn, 488–498
- Running time, 5, 28, 29
 - estimation of, 32–33
 - measurement, 31
- S**
- Sample size, 29, 33–37
 - binomial experiment, 35
 - contingency table, 35
 - multinomial experiment, 35
 - O-D survey, 34–37
 - precision table, 36, 43–48
 - upper bound, 35
- Schedule – *see* Timetable; Scheduling
- Scheduling, 7, 50, 321
 - adherence, 10, 12, 372
 - crew – *see* Crew scheduling
 - fixed, 7, 163–205
 - improved, 15
 - software of, 6, 55
 - see also* Deficit function website
 - surplus function, 374–376
 - variable, 7, 218–221
 - minimum crowding, 130–137
 - vehicle – *see* Vehicle scheduling
- SCP – *see* Set covering problem (SCP)
- SDP – *see* Set deletion problem (SDP)
- SDT – *see* Shifting departure time (SDT)
- Seat, 14
 - availability, 14
- Service, 5, 9–11
 - acceptable, 14
 - adjustments, 323
 - delivery, 9
 - design – *see* Service design
 - new, 28
 - operational parking conflicts, 374–381
 - short-turn, 27–28
 - span of, 10, 12
 - standards, 5, 9–11
 - strategies – *see* Service design strategies
 - zonal, 27–28
- Service design, 10, 366–388
 - connectivity – *see* Connectivity
 - elements, 368–374
 - cost estimation, 370
 - crew scheduling, 370
 - data collection system, 370
 - fare policy, 370
 - frequency, 369–370
 - information systems, 370
 - measures of performance, 370
 - network size and coverage, 369
 - network structure, 369
 - passenger amenities, 370
 - potential market, 369
 - public timetable, 369–370
 - revenue estimation, 370
 - ridership estimation, 370
 - route classification, 369
 - route coordination, 369
 - route structure, 369
 - schedule coordination, 370
 - setting standards, 370
 - span of service, 371, 372
 - vehicle scheduling, 370
 - optimum stop location, 381–388
 - strategies, 370
 - adjusted routing, 371
 - adjusted scheduling, 371
 - area coverage, 371
 - improved amenities, 371
 - new forms, 371
 - new routing, 371
 - new scheduling, 371
- Service reliability – *see* Reliability
- Set covering problem (SCP), 290, 295, 297, 299, 300, 306, 307, 385
 - network route design, 419–421, 428, 445–449
- Set deletion problem, 446–449
- Set partitioning problem (SPP), 290, 291, 299, 300, 306
- SF – *see* Surplus function (SF)

- Shelter, 10, 12
- Shifting, 5, 210–224, 225–231
 - departure time (SDT), 5, 210–224, 225–231
 - surplus-function related, 378–380
 - early departure criterion, 233–236
 - left-shift limit, 220
 - lower bound related, 221–223
 - minimum crowding model, 130–132
 - right-shift limit, 220
 - tolerance, 215–218
 - VTSP, 251–266
- Short turn trips, 10, 12, 456
 - candidate points, 458–460
 - on chart, 16, 28, 29
 - DH to service trips, 468–470
 - excluding departure times, 462–467
 - Minimax H algorithm, 462–467
 - modified shortest path, 464–467
 - max load point, 460–461, 469, 473
 - maximum extensions, 467–476
 - hollows, 472
 - at short-turn points, 470–476
 - URDHC, 472–476
 - methodology, 458–461
 - deficit function, 458–461
 - objectives, 457–458
 - shuttle service, 488–490
- Shortest path, 26
 - algorithms, 315–318
 - Dijkstra, 315–318
 - modified, for short turn, 464–467
 - O-D, 27–28
 - RNN, 317–318
- Shortest-path and matching (SPM), 291–294
- Shortlines – *see* Short turn trips
- Shuttle, 16, 29
- Shuttle service, 488–519
 - BART, 484, 494–498
 - case study, 494–498
 - DRT, 487–494
 - fleet-size for circular route, 485–486
 - blocks, 486
 - FIFO, 486
 - round-trip time, 485
 - implementation stages, 509–513
 - optimal routing, 503–509
 - algorithm, 507–509
 - base network, 503–505
 - heuristic approach, 509
 - model, 507–509
 - potential demand, 505–507
 - routing strategies, 487–498
 - bi-directional, 487–498
 - fixed route, 488–498
 - flexible route, 488–498
 - flexible schedule, 488–498
 - short cut, 488–498
 - short turn, 488–498
 - simulation, 490–494
 - input variables, 490–491
 - procedures, 491–494
 - survey, 498–502
 - attributes, 500–502
 - CATI, 498–499
 - waiting time, 501
- Simulation (of shuttle service), 490–494
 - input variables, 490–491
 - procedures, 491–494
- Single-route fleet size, 87–88, 94, 100–113, 215–218
- SPM – *see* Shortest-path and matching (SPM)
- SPP – *see* Set partitioning problem (SPP)
- Standard deviation, 31–32
- Standards, 4, 5, 9–11, 371
 - evaluation, 11–13
 - minimum even-load frequency, 125–127
- Standeers, 10, 12, 372
- Statistical tools, 30–37
- Statistics, 30–37
 - chi-Square, 55
 - coefficient of variation, 347, 348
 - log normal, 61–64
- Stop, 5, 24, 126, 548
 - optimum location, 381–388
 - realignment of, 28
 - spacing, 10, 12, 372, 517, 518
- Subsidy, 11, 15, 76, 78
 - per passenger, 11, 13
- Surplus function (SF), 366, 374–376
 - minimum parking spaces, 376
 - model, 374–376
 - reducing maximum of, 378–381
- Survey, 24
 - customer (for shuttle service), 500–502
 - interview-based, 26
 - passenger, 24, 26, 28
 - mailback, 26
 - O-D, 26, 34
 - on-board, 26, 28
 - population, 24, 26

- Synchronization, 139–160
 - on charts, 16, 29
 - maximum of, 140
 - OR model, 142–144
 - assumptions, 143–144
 - constraints, 143–144
 - MIP, 144
 - notation, 142
 - objective, 143
 - variables, 143
 - worst case, 144
 - real-life example, 154
 - Synchro-1 procedure, 145–146, 148, 150–151
 - CHOOSE, 145–146
 - FIRST, 145–146, 151
 - MIDDLE, 145–146, 151
 - Synchro-2 procedure, 146–148
 - MERGE, 147–151
 - MOVE, 147, 151
 - TRY-MOVE, 148–149, 151
 - trade-offs, 141
- T**
- Tactics, 570, 571, 575, 576, 580
- Tariff, 14
- TCRP (report 10), 3, 9, 11, 19, 409, 574
- TCRP, Digest, 405
- Terminal, 5
 - realignment of, 28
- Time-space diagram, 167, 530, 579, 581
- Timepoint, 8
 - single, 88–90
- Timetable, 4–8, 81–116
 - advanced, 119–137, 138–161
 - even individual loads, 121–127
 - minimum crowding, 130–137
 - synchronization, 139–160
 - automation, 98–113
 - comparison measures, 87–88
 - elements in practice, 85
 - even headway, 86, 90–94, 100–113
 - even load, 86, 94, 113
 - on individual vehicles, 121–127
 - load-tolerance criterion, 233–236
 - minimum frequency, 125
 - objectives, 84
 - optional, 85–87
 - public, 5, 86, 99
 - smooth transition, 90–94
 - special requests, 87
 - synchronization
 - test runs, 98–113
- Timetable development, 4–8, 82–116
 - on charts, 16, 29
- Tolerance, 7–8, 32
 - absolute, 33
 - absolute equivalent, 33
 - of estimation, 33
 - of even load, 233–236
 - recovery time, 5, 8
 - shifting departure time, 7–8, 215–218
 - VTSP, 253–266
- Transfer, 6, 10, 12, 376
 - coordinated, 14
 - counts, 29
 - feeder, 16
 - point, 141–142
 - shuttle, 16
 - synchronized, 14
 - timed, 10, 12, 376
- Transfer point, 24, 141–142, 156, 157
- Transit, 4
 - demand-responsive, 488–495, 590
 - network, 346, 395
 - viability, 13–15
 - agency perspective, 15
 - passenger perspective, 14–15
- Transit automation – *see* Automation
- Transit demand – *see* Demand
- Transit scheduling – *see* Scheduling
- Transit timetable – *see* Timetable
- TranSystems, 323–324, 369, 370, 546, 547, 549
- Travel mode, 29, 320, 336
- Travel time, 5, 8, 237, 337, 398, 501
 - acceptable, 14
 - analysis of, 533–536
 - bus data set, 535–536
 - method, 534–535
 - reduced, 15
- Trip, 5–8
 - end location, 8
 - end time, 8, 252
 - layover time, 8
 - missed, 10, 12, 372, 527
 - recovery time, 8, 242
 - short turn – *see* Short turn trips
 - start location, 8
 - start time, 8, 252
 - variable departure, 214–220
 - see also* Shifting

U

- Unit reduction DH chain (URDHC), 181–188, 225–229, 254–263, 378–379, 472
 - short-turn trips, 455–478
 - surplus-function related, 374–376
- Unit reduction shifting chain (URSC), 220, 228–229, 254, 378
- UMTA, 33, 36, 38
- URDHC – *see* Unit reduction DH chain (URDHC)
- URSC – *see* Unit reduction shifting chain (URSC)
- U.S. Department of Transportation., 542, 587

V

- Variable message sign (VMS), 14, 389, 552, 590
- Variance, 31–32
- Vehicle, 4–5, 7–8
 - capacity, 8
 - design, 14
 - hours, 11
 - km, 11
 - size – *see* Vehicle size
 - type – *see* Vehicle type
 - vehicle encounters, 578–584
 - improvements, 581–584
 - probability of, 578–580
 - tactics, 580
 - time-space diagram, 579–581
- Vehicle priority – *see* Priority treatments
- Vehicle schedule – *see* Block
- Vehicle scheduling, 4–8, 16, 29, 163–195
 - fixed, 163–188
 - max-flow technique, 172–176
 - fleet size – *see* Fleet size
 - Gantt chart, 229–232
 - multi-depot, 170–171
 - multi-route, 170–172

- task of, 173
- variable, 208–236
 - see also* Shifting
 - by vehicle size – *see* Vehicle size
 - by vehicle type – *see* Vehicle type
- website of DF, 183, 228
 - description of, 241–247

- Vehicle size, 16, 27–29, 266–275
 - determination of, 266–275
 - optimal size, 268–275
 - square-root formula, 267
 - standard bus vs minibuses, 266–267
- Vehicle type, 8, 16, 29, 249–264
 - optimal schedule, 251–253

- Vehicle-type scheduling problem (VTSP), 253–266
 - algorithm, 253–256
 - lower bound, 253–256
 - real-life example, 264–266
 - sensitivity analysis, 256–257
 - upper bound, 253
- VMS – *see* Variable message sign (VMS)
- VOYAGER, 574
- VTSP – *see* Vehicle-type scheduling problem (VTSP)

W

- Waiting time, 345, 367
 - average of, 539–542
 - on chart 14, 28, 29
 - difference of waiting times, 356–357
 - dilemma, 346–347
 - assumptions, 349
 - distributions, 348–358, 539–540
 - gamma distributions, 348–352, 539–540
 - power distributions, 348–352, 539–540
 - shuttle service, 501–502
 - strategy, 345–346
- Walking distance, 14, 381