Dynamic Capacity Constrained Transit Assignment

Jan-Dirk Schmöcker

Thesis submitted for the degree of Doctor of Philosophy

Centre for Transport Studies
Department for Civil and Environmental Engineering,
Imperial College London

September 2006
Abstract

Coping with an increasing demand for an existing transit network is one of the big transport challenges of the future. Already today the problem of full transit vehicles is experienced daily by commuters in many of the big conurbations around the world. The basis for finding any solution is a detailed analysis of where in the network passenger crowding occurs.

This thesis proposes an approach to transit assignment considering especially the following public transport properties: a) Transit vehicles have a finite capacity and there might be times of the day when the demand exceeds this capacity; b) in networks without published timetables passengers consider multiple routes and their actual route choice depends only on which vehicle arrives first (the so-called ‘common line problem’) and c) the analysis of capacity problems requires a dynamic approach as congestion builds up over the peak period.

Central to the approach is the introduction of a probability that passengers are not able to board the first vehicle arriving if this vehicle does not have sufficient available space. This “fail-to-board probability” is set in such a way that all the available space is expected to be used, with all demand exceeding the available capacity remaining on the platform.

Overcrowding and the resulting fail-to-board probability are further thought to deter passengers from attempting to board at overcrowded platforms if they have feasible
alternative routes and hence influence their route choice. Instead of focusing on the shortest route, passengers might consider taking longer but less congested routes. The Method of Successive Averages (MSA) is used to set the fail-to-board probability in an iterative process to find the risk-averse user equilibrium assignment.

To reflect changes in the fail-to-board-probability over time, time intervals are considered. Trips that can not be completed in one time interval are carried over to the next period. Trips are incomplete if a passenger fails to board at one stage during the journey, or if the journey is so long that the destination can not be reached within a single time period. It is assumed that those passengers who failed to board attempt to continue their journey from the same platform in the next time interval (including a chance that they might fail to board again at this platform). Those passengers with long trips are assumed to continue their journey from the last node they reached in the previous time interval.

The approach described in this thesis is illustrated with a case study of the London Underground network. It is shown that this approach is capable of finding the critical platforms in a large scale network.
Acknowledgements

First and foremost I want to thank my supervisor Prof. Michael Bell for his essential guidance and support during this PhD project. Some initial work on this topic started already during our time at Newcastle University in 2001 and only his effort to secure financial support enabled me to come to Imperial College London and start a PhD on this topic. His supervision throughout all these years has been outstanding, he was correcting mistakes but leaving me freedom to try own ideas.

The second person with an important influence on this work is Dr. Fumitaka Kurauchi from Kyoto University. The discussions with him about route choice in transit assignment have brought this work significantly forward. I am further grateful that he, Dr. Nobuhiro Uno and Prof. Yasounori Iida enabled me to spend one year at Kyoto University to finalise this thesis.

Further, I want to thank the Railway Technology Strategy Centre and the Department of Civil Engineering of Imperial College London for jointly funding this thesis. Thanks must further go to Transport for London for providing data and their comments on this work. I also want to acknowledge the support of many other members of the Centre for Transport Studies at Imperial College London and at Kyoto University during my stay there.

Finally, I want to thank Ayako, my family and many friends in London for their moral support during these years.
Declaration of Contribution

At various stages during my PhD collaboration has been taken place with colleagues working on the subject. My supervisor, Prof. Michael GH Bell, and Dr. Fumitaka Kurauchi of Kyoto University advised me at several stages. The network loading procedure used in this thesis and described in Chapter 4.5 was originally proposed by Prof. Michael G H Bell and published in Bell et al (2002). The common line formulation used in the thesis and described in 4.3 and 4.4 was developed by Dr. Kurauchi during his stay at Imperial College London in 2002/03 and published in Kurauchi et al (2003). I played a role in both developments and am a coauthor of both the above papers.

I hereby declare that besides the collaboration referred to above the work described in this thesis has been carried out by myself.

.................................

(Jan-Dirk Schmöcker)

.................................

(Prof. Michael G.H. Bell)
Table of Contents

Abstract ........................................................................................................................................ 2
Acknowledgements .................................................................................................................. 4
Declaration of Contribution .................................................................................................... 5
1. Introduction .......................................................................................................................... 10
   1.1. Background .................................................................................................................. 10
   1.2. Objectives of the thesis ............................................................................................. 13
   1.3. Thesis structure .......................................................................................................... 15
2. Approaches to Transit Assignment ...................................................................................... 17
   2.1. Introduction ................................................................................................................ 17
   2.2. Network characteristics determining the choice of model ...................................... 19
      2.2.1. High or Low Frequency? .................................................................................... 19
      2.2.2. Passenger information and service punctuality .................................................... 20
      2.2.3. Transfer behaviour of passengers ....................................................................... 21
      2.2.4. Demand or supply variations by time of day ....................................................... 22
      2.2.5. Crowding and congestion .................................................................................... 22
      2.2.6. Capacity problems ............................................................................................... 23
   2.3. Further issues and summary ...................................................................................... 24
3. Review of Frequency Based Assignment Approaches ...................................................... 27
   3.1. Introduction ................................................................................................................ 27
   3.2. Route choice in uncongested transit assignment models ........................................ 28
      3.2.1. Early approaches (before Spiess and Florian, 1989) .............................................. 28
      3.2.2. Route-section and strategy-based approaches ....................................................... 31
      3.2.3. Linear Programming formulation and solution algorithm to the common line problem
      37
   3.3. The effective frequency approach .............................................................................. 40
      3.3.1. Problem description .......................................................................................... 40
      3.3.2. Effective frequencies with practical capacities ...................................................... 41
      3.3.3. Effective frequencies with strict capacities .......................................................... 44
   3.4. Approach with fail-to-board probabilities ................................................................. 46
   3.5. Passenger distribution between different paths of one hyperpath (bus stop problem) ....................................................................................................................... 48
   3.6. Discussion .................................................................................................................... 51

4.1. Introduction ............................................................................................................. 54
4.2. Network description and its notation ........................................................................ 55
  4.2.1. Network representation ....................................................................................... 55
  4.2.2. Notation (Glossary) ........................................................................................... 59
4.3. The cost of a hyperpath ............................................................................................ 61
  4.3.1. Elementary paths and hyperpaths ................................................................. 61
  4.3.2. Transition probabilities .................................................................................... 61
  4.3.3. The cost elements of a hyperpath .................................................................... 63
  4.3.4. Bellman’s principle (The costs from a node) .................................................. 64
4.4. Finding the optimal hyperpath ................................................................................ 67
  4.4.1. Introduction ....................................................................................................... 67
  4.4.2. Hyperpath Search Algorithm .......................................................................... 68
4.5. Network Loading ................................................................................................... 70
  4.5.1. Arc and Node Volumes ...................................................................................... 70
  4.5.2. Ensuring strict capacity constraints .................................................................. 73
4.6. The network equilibrium ......................................................................................... 73
  4.6.1. Multiple hyperpaths to a destination ............................................................... 73
  4.6.2. Characterisation of the network equilibrium .................................................. 76
  4.6.3. Existence of a unique network equilibrium ................................................... 76
  4.6.4. A Note on Circular Lines ............................................................................... 77
4.7. Method of Successive Averages (MSA) ............................................................... 82
  4.7.1. Introduction ....................................................................................................... 82
  4.7.2. MSA algorithm ................................................................................................. 83
  4.7.3. Convergence of the MSA ............................................................................... 83
4.8. Assignment without consideration of common lines ............................................. 85
  4.8.1. Introduction ....................................................................................................... 85
  4.8.2. Calculation of Transition Probabilities in simplified network ......................... 86
  4.8.3. Calculation of fail-to-board probabilities (Correction Algorithm) ................... 87
  4.8.4. Risk averse assignment .................................................................................... 89
4.9. Case Study on a small example network .............................................................. 90
  4.9.1. Network description ......................................................................................... 90
  4.9.2. Effect of Risk-averseness ............................................................................... 94
  4.9.3. Effect of common lines .................................................................................... 95
4.10. Discussion ............................................................................................................ 96
  4.10.1. Summary ......................................................................................................... 96
  4.10.2. Distribution between Attractive Lines in Congested Situations ....................... 97
  4.10.3. Further Assumptions made in the CapCon approach ...................................... 101
5. Dynamic Network Loading ..................................................................................... 103
  5.1. Introduction ......................................................................................................... 103
  5.2. Network description and its notation .................................................................... 106
    5.2.1. Network representation .................................................................................. 106
    5.2.2. Additional notation ....................................................................................... 107
  5.3. Dynamic Network Loading ................................................................................. 108
    5.3.1. Dealing with excess boarding demand ......................................................... 108
    5.3.2. Dealing with trips unfinished within a time-interval ...................................... 110
  5.4. Finding the travel time of a hyperpath ............................................................... 112
    5.4.1. Introduction .................................................................................................... 112
    5.4.2. Algorithm to find Travel Time Matrix ......................................................... 113
5.4.3. Adjustments for inaccuracies in \( \mu_{ij} \)................................................................. 114

5.5. MSA for dynamic network loading................................................................. 116
  5.5.1. Introduction .................................................................................................. 116
  5.5.2. MSA with Time Loop ................................................................................ 117

5.6. Alternative formulation of the generalised cost function....................... 118
  5.6.1. Expected delays .......................................................................................... 118
  5.6.2. Considering variation of \( \bar{q} \) over time ..................................................... 120

5.7. Case Study on a small example network ................................................. 122
  5.7.1. Introduction ................................................................................................ 122
  5.7.2. Modelling different time interval durations ............................................ 122
  5.7.3. OD reliability ............................................................................................ 125
  5.7.4. Delay expectation and risk averseness .................................................... 126

5.8. Discussion ..................................................................................................... 128
  5.8.1. Summary .................................................................................................. 128
  5.8.2. Mingling in the dynamic case .................................................................... 129
  5.8.3. “Setting back” of unfinished trips ............................................................. 129
  5.8.4. Choice of Time Interval Duration ............................................................. 131
  5.8.5. Split between paths with equal costs ......................................................... 132
  5.8.6. Failure probability in future time intervals .............................................. 132

6. Further congestion effects – Model refinements ....................................... 134
  6.1. Introduction ................................................................................................. 134
  6.2. On-Board Congestion ................................................................................ 135
    6.2.1. Adjustment of the generalised cost function ........................................... 135
    6.2.2. Example: Comparison between crowding and fail-to-board ................. 136
    6.2.3. Perceived costs of sitting and standing in a crowded service ............... 139

6.3. Fail To Access ............................................................................................. 141
  6.3.1. Introduction .............................................................................................. 141
  6.3.2. Access to Fail Nodes ............................................................................... 142
  6.3.3. Ensuring strict capacity constraints (for platform access) ....................... 143
  6.3.4. Dealing with excess demand at platform access arcs ............................ 144
  6.3.5. Example: Restricting platform access .................................................... 145

6.4. Discussion .................................................................................................... 147
  6.4.1. Summary of Congestion Costs ................................................................. 147
  6.4.2. Further adjustments to the generalised cost function ......................... 149

7. Software implementation .............................................................................. 150
  7.1. Introduction ................................................................................................. 150
  7.2. Software structure ..................................................................................... 151
    7.2.1. Automatic creation of nodes and arcs .................................................... 151
    7.2.2. Flow Chart ............................................................................................ 153

7.3. Model Run Time Issues ............................................................................ 155
  7.3.1. Introduction .............................................................................................. 155
  7.3.2. Transition matrix structure ................................................................. 156
  7.3.3. Dividing the inversion process ................................................................. 157
  7.3.4. Matrix reorganisation ................................................................. 158
  7.3.5. Sparse matrix multiplication ................................................................. 159
  7.3.6. Optimised matrix inversion ................................................................. 160
  7.3.7. Enumeration of moves ......................................................................... 161

7.4. Summary and discussion .......................................................................... 162

8. Crowding and Capacity Problems in London ........................................... 164
  8.1. Introduction ................................................................................................. 164
8.2. Data description............................................................................................................. 165
  8.2.1. Network data............................................................................................................. 165
  8.2.2. Demand Data......................................................................................................... 168
  8.2.3. Reduction of network OD matrix to Inner Zone OD matrix ......................... 169
8.3. 3 hour model runs........................................................................................................... 173
  8.3.1. Comparison with flows estimated by London Underground......................... 173
  8.3.2. Comparison of arc flows with and without consideration of common lines .... 174
  8.3.3. Congestion during the 3 hour period................................................................. 176
8.4. Dynamic runs (15 minute model)............................................................................... 177
  8.4.1. Demand assumption.............................................................................................. 177
  8.4.2. Effect of risk-averseness...................................................................................... 178
  8.4.3. Comparison of capacity constrained assignment and assignment with crowding effects 181
8.5. Capacity reducing effects and an assignment example......................................... 183
  8.5.1. Consideration of service irregularities............................................................... 183
  8.5.2. London Underground’s research on crowding behaviour.............................. 184
  8.5.3. Assignment with lower capacity...................................................................... 186
8.6. Discussion.................................................................................................................. 188
9. Conclusions...................................................................................................................... 190
References.......................................................................................................................... 196
1. Introduction

1.1. **Background**

Jam-packed buses, metros and trains are experienced daily by passengers in big conurbations in many countries around the world. This means travelling to work in the morning is often a stressful experience. Crowded carriages cause discomfort to passengers and moreover are often the cause for delays. In London for example the costs related to crowding and delays of its public transportation network are estimated at £230m per year for the City of London alone. The stress caused from overcrowding is not only unpleasant but also reduces people’s productiveness at work (Oxford Economic Forecasting, 2003). The BBC and several London papers describe the problem more drastically and repeatedly report the “Daily Trauma” faced by passengers (e.g. BBC, 2003). The British government acknowledges this problem but also admits that it can not be solved in the coming years because every increase in capacity will be taken by slack demand (Department for Transport, 2004).

In Japan, and notably Tokyo, the problem of overcrowding is probably even more severe. The picture of “train pushers” whose job it is to make sure “slack” capacity is utilised has become famous. The investment in Tokyo’s rail network over the last decades has been substantial but nevertheless the increase in capacity is lower than the increase in population commuting to Tokyo (Ministry of Land, Infrastructure and Transport, 1992). The predicted growth in population for many major cities means that the problem of overcrowding will be further emphasised in the coming years.
Overcrowding is not only the cause of discomfort on the trains but also of safety risks on the platforms. Train operators are often less worried about safety and crowding on-board than about accident risks on the platforms, staircases and escalators. The UK’s Rail Safety and Standards Board (RSSB) for example pointed out that overcrowding at platforms and access links to platforms are one of the biggest hazards (RSSB, 2005). The RSSB further recommended in the same report that the throughput of ticket gates should be adjusted to control the crowding of platforms. In London therefore the operator sometimes close a whole station during the morning peak-hour for a period of around 10 minutes in order to allow the platforms to clear of passengers who failed to board overcrowded trains.

The management of crowds on their way to or from the platform is further of importance for large events, such as football match days. If the number of passengers accessing the platform is not artificially limited after the end of a match, the departure of services often becomes delayed because of passengers pressing onto the services and blocking the doors. These delayed departures then mean a de facto reduced service frequency and hence service capacity leading to even more overcrowding. This relationship between dwell time and crowding has been investigated by for example Lam et al (1999a) and Vucic (1981).

A further problem of overcrowding, besides discomfort, delays caused by dwell times and safety risks, is that in some situations the congestion is so severe that passengers are not able to board a service anymore. At some crowding level it becomes impossible for more passengers to board, or the train is at least so full that passengers prefer to wait for a subsequent service that might be less crowded.
London Underground for example found that at stations where passengers know that the subsequent service is emptier a substantial number of passengers did not board the first arriving train even though this train was not fully occupied (London Underground Limited, 1988).

Sound assignment methods are needed when one wants to assess the impact of adding links to existing networks as well as when one wants to make changes to an existing network or service pattern. This is the case for road-based transport as well as for public transport networks. Possible modifications to a public transport network include changes to service routes, their frequency or timetable, their service quality, changes in vehicle capacity as well as changes to the transit stops. All these modifications potentially influence the congestion in the network and with this the passengers’ route choice. Therefore the transit assignment tool needs to consider travel times and congestion effects carefully in the estimation of the generalised costs experienced by the passengers.

Compared to (road) traffic assignment, the following fundamental differences have to be addressed: Firstly, transit assignment has to consider the in-vehicle time as well as the waiting times at stations and the transfer between platforms. This introduces uncertainty in the travel time estimation and route choice. For example, at some stations passengers might have the choice between a local and an express service or between a number of services travelling different routes but all arriving at their destinations. If passengers only know the frequency of the services, they might be able to minimise their expected travel time by pre-selecting some services and choosing whichever service arrives first from this set of lines. This is frequently
referred to as the “common line problem” (CLP). Secondly, this chapter explains that in congested situations, the in-vehicle time for transit users often keeps relatively constant (except maybe for longer dwelling times), but the traveller might experience inconveniences like not getting a seat, crowding on the platform and longer interchanging times at the station. In some circumstances the traveller might not be able to board the first available vehicle because of insufficient space on board and in very severe circumstances access to the platforms might be restricted.

1.2. Objectives of the thesis

The primary objective of this thesis is to demonstrate a new method that solves the transit assignment problem in high frequency networks and considers capacity constraints as well as the common line problem. The review of existing literature shows that current assignment methods have some shortcomings, especially in the way capacity constraints are treated. Current software packages either do not consider these constraints or use the effective frequency approach, which will be reviewed with its advantages and shortcomings. In general, there is much literature on transit assignment and the common line problem, but transit assignment with capacity constraints has only received attention in recent years. It will be shown that the approach presented here can distinguish the impact of on-board crowding, of crowding on the platform possibly resulting in some passengers not getting on board and of crowding on the platform or at the station entrance leading to passengers not being able to get onto the platform. In order to show that the approach can be used
for larger applications, a case study will be conducted with the London Underground network.

A dynamic model will be developed that solves the problem of transit assignment with capacity constraints. A dynamic model is needed because congestion builds up over the peak-period. If one train is overcrowded, those passengers left behind will compete with newly arriving passengers for space on the next service meaning that the congestion gets even more severe. Further the assumption of constant demand as is the case with static models is difficult to justify. The London experience is that capacity problems so severe that passengers can not board a train only arise during the “peak of the peak”, if all services run according to schedule. This changes, however, if the service experiences any delays. In this case, large headways and bunching effects for subsequent services can lead to temporary overcrowding even during the off-peak period.

A secondary objective of this thesis is to analyse the impact of the common line problem on the assignment results. In recent years, all frequency-based assignment models are based on the assumptions made in the key paper by Spiess and Florian (1989). They explain that passengers develop route choice strategies under the assumption that passengers only know the frequency but not the exact arrival time of the next vehicles serving their transit stop. This leads to the assumption of the common line problem as explained above. Nowadays count-down information often at platforms and bus stops challenge the assumptions made in Spiess and Florian (1989). Full information provision is further assumed for schedule-based assignment meaning that schedule-based models do not consider common line issues. This will
be explained in more detail in this thesis and the sensitivity to assignment with and without common lines will be illustrated.

1.3. Thesis structure

Chapter 2 reviews the different approaches to transit assignment. All approaches can be categorised as either frequency- or schedule-based. The so-far existing frequency-based models are all static whereas there are schedule-based dynamic models. The model presented in this thesis is frequency-based and considers dynamic effects. It is therefore the objective of the following chapter to discuss strengths and weaknesses of both approaches, and to explain in which scenarios frequency-based models are more applicable.

Chapter 3 will then review previous work carried out in frequency-based transit assignment in more detail. Firstly, uncongested assignment methods are described. This is followed by a description of models that consider congestion and then by approaches that explicitly consider capacity constraints. Chapter 4 then describes a new approach to capacity constrained frequency-based transit assignment, which will be referred to as “CapCon”. The assumptions are explained and the results from small case studies are presented. The model considers common lines and it will be shown that the network description can be significantly simplified if the common line problem is ignored. Chapter 5 introduces the dynamic extension of the static model presented in Chapter 4. In this case, demand that exceeds the available capacity is kept in the network and assumed to attempt to board subsequent services.
Chapter 6 explains some further model extensions, in particular the consideration of platform access constraints. Chapter 7 summarises the model structure and discusses some possibilities to improve the calculation speed. The case study of London’s tube network is then presented in Chapter 8. Each of the following chapters concludes with a short summary and discussion of the main findings. Finally, Chapter 9 concludes the thesis and points out areas of further work.
2. Approaches to Transit Assignment

2.1. Introduction

Various transit assignment tools exist which differ significantly in complexity and spread of usage. There are a (small) number of widely accepted generic models and a (larger) number of bespoke models. The main distinction between models is whether they are frequency- or schedule-based. Further distinctions can be made, like whether stochastic or deterministic assignment is used and in which way the network is represented. Nuzzolo (2003), for example, distinguishes between diachronic and dual network representations and sets out the advantages of both. However, for practitioners in particular the choice between frequency-based and schedule-based models is an important one as both approaches have some inherent advantageous.

As the names already suggest frequency-based models consider only the (average) frequency of the services, whereas schedule-based approaches consider the exact timetable and model each single run of the services. Useful definitions for both approaches are given in Nuzzolo (2003):

‘The frequency-based approach considers services in terms of runs (lines). In this case run scheduled times are not considered explicitly, but we refer to the line headways, or to their inverse (the service frequencies), from which the name of the approach derives. Therefore we are not able to calculate explicitly attributes that users consider in relation to single lines, but we can refer only to ‘average values relative to lines’’
‘The schedule-based approach refers to services in terms of runs using the real vehicle arrival/departure time, and hence all the values of level of service attributes, evaluated at the time in which users make their choices, can be explicitly taken into account. This approach allows us to take into account the evolution in time of both supply and demand, as well as run loads and level of service attributes’

Frequency-based modelling constitutes the classical approach as it is simpler, requiring less input data and less computational power. Advanced frequency-based route choice models have been developed to reflect the choice passengers face in a public transport network where a number of lines would bring a passenger to his destination. Schedule-based approaches have been developed more recently and are becoming more widely used, partly because of increasing computational power. Since Tong and Richardson (1984) published their paper on scheduled-based assignment, the approach has been gaining in popularity (see for example, Wong and Tong, 2001; Nguyen et al., 2001 or Nielsen et al., 2001). Theoretical advances as well as case studies have been published, for example Wilson and Nuzzolo (2004).

Schedule-based approaches have been applied to very different scenarios. However, schedule-based models are not advantageous in every situation. The following literature review shows that in some circumstances also frequency-based models might be able to reflect passenger behaviour better and that the choice of model should depend on the questions discussed in the following section. This list of questions is not exhaustive as the focus is on the choice between schedule-based and
frequency-based models. A summary of the discussion in this chapter can also be found in Schmöcker and Bell (2006) and is also part of a “webtag” of the Department for Transport offering advice to local authorities and other transport planners (Department for Transport, 2005).

2.2. Network characteristics determining the choice of model

2.2.1. High or Low Frequency?
If services operate with a low frequency, the model should consider the difference between desired time of travel and actual vehicle departure and arrival times. It is generally accepted that the difference between desired arrival time (DAT) and vehicle arrival time (VAT) should be used for working trips and the difference between desired departure time (DDT) and actual vehicle departure time (VDT) for homebound trips (see for example Cascetta and Papola, 2003). Differences in desired and actual departure/arrival times are approached in Florian (1998) by considering the maximum lateness and earliness that is acceptable for passengers.

If services operate with a high frequency, it is generally sufficient to take DDT into account in the OD matrix and differences between DDT and VDT can be treated as a constant, for example half the headway. A commonly accepted threshold to distinguish high and low frequency services is 10-15 minute headways (e.g. Nuzzolo, 2003). If the service operates with a higher frequency the passenger arrivals can be assumed to be uniform, because passengers will often not check the timetable before
they start their journey (if one is available in the first place). However, if the service operates with a frequency less than the suggested threshold, travellers will turn up at the station for specific scheduled services.

From this it follows directly that frequency-based models are not suitable for services that operate with headways larger than some threshold. For low frequency services the modeller will need to consider further issues to decide whether to use a frequency- or schedule-based model.

2.2.2. Passenger information and service punctuality

If the traveller has reliable information on the arrival time of the vehicles he might choose the route ‘intelligently’ and not just take the next arrival from his choice set. The more timely information a traveller has, and the more reliable this information is, the more the choice will be run- rather than line-based. The common line problem described in the introduction will not apply in a full information environment. For example, passengers will know whether it is worth waiting for an express service and do not have to choose “whichever service arrives first”. Therefore, frequency-based models will be more suitable if services operate with low punctuality and/or a low level of user information.

Delays and irregularity have to be treated implicitly or explicitly in schedule-based models. An implicit treatment is possible by adding error terms to the path choice model. A Monte Carlo technique allows the explicit treatment of delays, as in Nuzzolo et al. (2001). It should be noted that nowadays with the increasing number of ATIS (Advanced Traveller Information Systems) passengers are often given very
precise information. In some cases, passengers are not only told the minutes left before the arrival of the next vehicle of each line but also how many seats are left. This information can only be modelled with schedule-based approaches.

2.2.3. Transfer behaviour of passengers

The previous section already illustrated that one needs to identify to what extent the common line problem exists. If passengers often change their pre-trip path choice en-route, or only make line choices en-route, a frequency-based approach considering a travellers ‘optimal strategy’ might handle this issue better. Schedule-based models will need to include some stochastic elements which might be difficult to calibrate in order to reflect the common line issue appropriately.

Whether or not the common line problem exists will depend also on the distance between stops. If bus stops of different lines or train platforms are adjacent, it can be assumed that passengers might consider lines from both stopping points when waiting for the next service. If the stops are far away from each other, passengers will have to decide for one stop and limit their choice set of potentially attractive lines. Another important point is the structure of the fare system. If a traveller’s ticket is only valid for one service and not transferable to another service, the common line problem will be reduced. For example, if the next vehicle arriving would bring the passenger closer to the destination but the service is from a different operator and would require the traveller to buy another ticket or incur a top-up fare, then the traveller will be less inclined to include this service in his/her set of attractive paths.
2.2.4. Demand or supply variations by time of day

Service regularity is a separate issue to punctuality. Regularity refers to the scheduled intervals between the arrivals of the vehicles and not to unplanned delays. Frequency-based models assume an equal share of passengers between the runs of this service. If a service is not scheduled to arrive with regular headways, say 00, 15, 30, 45 before the hour, but say 10, 15, 40, 45 after the hour, this might lead to line loading errors in frequency-based models. Further, a schedule-based approach is required if there is a major influx of passengers during a certain period (like an underground station connected to a train station that brings a large number of passengers to the underground network once every hour) in order to show overloading of certain services.

2.2.5. Crowding and congestion

Crowding often has an influence on network performance and raises several problems for the modeller. If the service deteriorates with increasing demand, this should be reflected in the link costs and the shortest path algorithm. Congestion might for example lead to longer dwell times at stops. The perceived cost will also increase through lower in-vehicle quality. The strategy approach of Spiess and Florian (1989) for frequency-based models has been reviewed by several authors in order to include congestion effects and derive a user equilibrium assignment (De Cea & Fernandez, 1993; Cominetti & Correa, 2001). However, if the crowding is severe and the service operates with high frequency, irregularity effects like “bus-bunching” might arise. One can only model this with a schedule-based model.
2.2.6. Capacity problems

Congestion and capacity problems in public transport assignment are not the same. This is for two reasons. Firstly, the cost function is not increasing continuously, but the finite capacity of public transport vehicles will lead to a step function; either a traveller can board the arriving vehicle or not, in which case the waiting time will increase by one headway. This might lead to impatience and frustration, so that the cost increase through additional waiting at the platform because of overcrowding should be differentiated from the expected waiting time. Secondly, capacity problems will only be experienced by boarders. Passengers on-board have priority and do not perceive an increase in cost (unlike in highway networks).

Schedule-based models can treat capacity problems explicitly and the modeller can see which runs suffer from capacity problems. However, changes in route choice behaviour should be checked carefully. Papola et al (2005) as well as Tong et al (2001) discuss approaches to capacity constrained schedule-based assignment and assume first in - first out (FIFO) rules for passengers waiting to board. Papola et al (2005) show a saw-shaped waiting time function which indicates from which arrival time onwards passengers will not be able to board the first (or even the first two, three, ...) service arriving because other passengers waiting in the queue before them have priority. FIFO might however not be true for long platforms where passengers mingle and hence it is rather those who ‘push more’ than those who wait longest who will get on the next vehicle (of course those who wait longest may eventually push more).
In the route choice algorithms of capacity constrained schedule-based assignment models it is further assumed that passengers know whether or not they can board the next service. This might also not be the case as it is difficult to judge whether one will join the queue before or after the cut-off point between those passengers able to board the first service and those who have to let one service go before being able to board.

In currently existing frequency based-models capacity problems are treated implicitly through effective frequency, as firstly suggested in De Cea and Fernandez (1993) and refined by several authors (discussed in Chapter 3). The idea is to increase the perceived costs of boarders through a local reduction in service frequency, reflecting the fact that the passenger may not be able to board a vehicle because of overcrowding. This approach can be criticised for two reasons: a) A cost increase based on the number of passengers wanting to board and spaces available is still a continuous cost function; b) an increase in cost does not prevent the network becoming overloaded. It might however be the case that a network cannot handle all the passengers wanting to travel in the time period modelled. Chapter 3 will explain the effective frequency approach in more detail and Chapter 4 will describe an alternative solution.

2.3. Further issues and summary

Besides these network characteristics, the scale of the network often determines the choice of model. Because of the more detailed network description and because of its dynamic nature, schedule-based approaches are computational more demanding. At the same time, the computation power available to model such systems is rapidly
increasing. Nevertheless, Wilson (2004) calls it an ‘open question’ whether schedule-based networks can handle large scale networks with reasonable speed, even in the future. Nielsen (2004), for example, talks about a run time of one week to assign demand with a schedule-based approach to the large-scale East Denmark model.

The above showed that schedule-based models are advantageous in many circumstances. Information for particular runs can only be obtained with such a model and frequency-based models are not applicable for systems with large or irregular headways. Further, with schedule-based models it is easier to show the impact of effects like irregular service arrivals, peaked demand distributions, or the provision of advanced traveller information.

It should however also be noted that the schedule-based approach is not the best in every situation. One might prefer to use a frequency-based approach for the following reasons:

- The model input is easier. Frequency based models require less detailed input data.
- Less detailed network representations also often lead to advantages in run time. Frequency-based models might therefore be preferred for the strategic modelling with large scale networks.
- If passenger arrivals and/or vehicle departures include some random element, the common line problem is easier to handle with frequency-based models.
- Most schedule-based models assume First-In-First-Out behaviour. Mingling among passengers who are already waiting for a long time and
those who have just arrived, which happens at least to some degree on long platforms, can however be more easily modelled by frequency-based models.

Finally, Sections 2.2.5 and 2.2.6 discussed some difficulties with frequency-based models, if congestion and overcrowding in the network exist. This thesis describes an approach to overcome some of these difficulties and still utilise the advantages of a frequency-based model.
3. Review of Frequency Based Assignment Approaches

3.1. Introduction

This chapter will review previous work carried out to solve the transit assignment problem with a frequency-based approach. The emphasis of the first part of this chapter is on the different network description and route choice models if the network contains common lines (3.2). The different approaches are roughly presented in chronological order. The approach presented by Spiess and Florian (1989) is discussed in more detail as it proved to be a key work for most publications since then. The second half of the chapter focuses on route choice and network loading in the presence of network crowding and capacity problems (3.3 to 3.4). In particular the effective frequency approach already mentioned in the previous chapter is reviewed (3.3). An alternative to this approach is to define the probability of passengers not being able to board the first service. A first attempt at this has been made in the TRANSEPT model which is discussed in Section 3.4.

Section 3.5 discusses work carried out to improve the estimate of the passenger split at a node between different arcs. In particular if there is congestion in the network a split according to nominal or effective frequencies might not be realistic. The chapter concludes by summarising the shortcomings of existing approaches, which is the motivation for this research.
3.2. **Route choice in uncongested transit assignment models**

3.2.1. Early approaches (before Spiess and Florian, 1989)

Dial (1967) presented one of the first transit assignment models. In his approach, he assumed a network where several lines served some of the stations. The stations are connected with so-called “trunklines” where each trunkline has the elements origin, destination, travel time and the set of lines serving this arc. The assumption is that the travel time is equal on all lines of this arc. Dial further approximated the waiting time or “interchange penalty” as half the headway of all services $l \in L$, leaving node $s$:  

$$w(s) = \frac{0.5}{\sum_{l \in L_s} f_l}$$  

(3-1)

The assumption is that travellers arrive randomly at the platform and that they will choose the vehicle arriving first. Implicitly Dial further assumes that the inter-arrival time between the services of different lines is constant. In other research discussed in the following the 0.5 is often replaced by 1, which assumes an exponential distribution of interarrival times, or more generally by a parameter $\alpha$.

Dial then explains that standard shortest path algorithms have to be adapted for the transit case because the Markov property is violated. The Markov property states that the route choice probabilities at a node are independent of the travellers’ origin. Further, in the example of Figure 3-1 passengers destined for C will take longer to get to B than passengers destined for D if Line II is more frequent. Dial therefore
adapts Moore’s shortest path algorithm (Moore, 1957). The adapted algorithm stores the waiting time for the transfer to the trunkline the route is using. If there is a through service at the trunkline exit the current waiting time is subtracted but a (larger) waiting time for the through service is added.

Figure 3-1 Simple transit network with Trunkline between nodes A and B.

A major critique of Dial’s work is that this algorithm can not deal with the situation where lines connect the same OD pair but have different travel times. In the example of Figure 3-2 Dial’s algorithm would assign all passengers to Line II for the journey between A and B and passengers destined for C would change at B.

Figure 3-2 Simple network with a fast (Line II) and a slow (Line I) service between A and B.

Fearnside and Drapper (1971) also based transit assignment on a road-assignment tool. In their model each node is represented by line specific nodes and boarding and alighting links. For example, node B as shown in above figures is represented as in Figure 3-3. An advantage is that this network representation allows the inclusion of waiting times and line specific transfer penalties which might reflect longer walking ways to node B_{II} compared to node B_{I}. The model was further developed in order to
be used in multi-modal studies. The disadvantage of this representation is the multiplicity of links and nodes, which Dial attempted to avoid.

![Diagram of a network with nodes A, B_I, B_C, and B_II, and links I and II, showing boarding and alighting.]  

**Figure 3-3** Interchange at Node B in Fearnside & Drapper (1971) network

Fearnside and Drapper assume that passengers have to decide for node B_I or B_II before boarding. This leads to an overestimation of waiting time if it is realistic to assume that passengers choose the first vehicle arriving from several possible (“common”) lines. As explained in the introduction transit networks with fast and slow lines (as in Figure 3-2) are good illustrations of the common line problem, where passengers include more than one route in their choice to minimise the travel time: In Figure 3-2 passengers for C might consider taking Line I as well as Line II from A. Especially if a passenger has just missed Line I, he/she will consider taking Line II to B in order to catch up with the slower Line I at this node.

The first papers on the common line problem “fixed” the routes of the passengers by calculating the shortest path to the destination according to link travel times and expected waiting times, which are reduced if several lines serve this path. Le Clerq (1972) developed an algorithm that searches for the shortest paths by looking at all possible interchanges from the service the traveller is currently using. The link travel
times and waiting times for each interchange are considered in the cost function. The algorithm therefore takes account of the fact that the Markov property is not holding for transit networks and also allows for different travel times of different services. In the example of Figure 3-2 the route choice will however still be fixed. Passengers destined to C will always take Line I if the expected average waiting time for interchange at B is larger than the additional on-board travel time for Line I. Similarly, if the waiting time at B is smaller then all passengers destined to C will take Line II and interchange at B.

3.2.2. Route-section and strategy-based approaches

Spiess and Florian (1989) showed, however, that passengers can often significantly reduce their travel time when they consider several paths to their destination. Following Lampkins and Saalman (1967), who explained that passengers will exclude lines from their choice set if they are obviously bad, e.g. travel time of slow line is longer than travel time plus headway on the fast line, Spiess and Florian explained that passengers develop so-called strategies if more than one transit line leads to their destination from their current location. In this case, passengers predetermined a set of attractive routes among the routes that bring them to their destination (possibly via very different paths) and choose the first vehicle arriving from this set of attractive lines. Therefore in fact not the traveller but the vehicle arriving first (from the set of attractive lines selected by the passenger) decides the route the traveller is taking. In the example of Figure 3-2, the passenger destined for C will minimise the travel time by choosing Line I or Line II at A whichever vehicle is arriving first. This is if the travel time for I is not larger than the travel time plus
the expected waiting time for Line II, otherwise the optimal strategy is to always wait for Line II.

Nguyen and Pallottino (1988) illustrated this problem by introducing the term hyperpath. A hyperpath consists of a set of paths that are potentially taken by the user depending on the arrival of the first vehicle at the passenger’s origin and the stations where he or she has to interchange. The probabilities of taking a path are given according to the line frequencies, so the problem is reduced to finding the shortest hyperpath, i.e. the set of paths that should be included in the traveller’s route choice. The underlying behavioural assumption is: “Before starting any trip, a passenger has chosen a fixed subset of transit lines for every stop he may encounter on his trip, and for every transit line the alighting point.” (Nguyen and Pallottino, 1988). The authors note that the sequential procedures by Dial (1971) are similar; however, the hyperpath concept does not suffer from “independence of irrelevant alternatives” disadvantage that besets logit assignment. The hyperpath model leads to the same solution as Spiess and Florian’s model, but with a more flexible network description that allows for more efficient computational techniques on larger networks. Cominetti and Correa (2001) for example use the hyperpath concept to adapt Dijkstra’s shortest-path algorithm to a “shortest hyperpath algorithm”.

The main challenge in the strategy approach is to find the set of routes $S$ that minimises the expected travel time $T_s$, which is given by (3.2):

$$T_s = E(W_x) + \sum_{i \in S} t_i H_i(S)$$  \hspace{1cm} (3-2)
where $E(W_s)$ is the expected waiting time, $t_i$ the expected travel time and $H_i(S)$ the probability that route $i$ is served first. Chiriqui and Robillard (1975) suggest a heuristic algorithm to solve this problem. Their algorithm is based on ordering the routes according to their in-vehicle travel time in ascending order. $T_s$ is first calculated taking the route section with lowest $t_i$ only. If adding other route sections reduces $T_s$ then these are added. If adding a route does not reduce $T_s$ further, the algorithm stops. The motivation behind this algorithm is that it is illogical for a passenger to let a bus of a given route go by and wait for a bus with longer in-vehicle travel time. It is obvious that this heuristic has its limit in large scale networks where the enumeration of all routes becomes difficult.

Chiriqui and Robillard (1975) and De Cea and Fernández (1993) use a model with dual network representation, also referred to as “route segments” model (Figure 3-4). The idea is that each stop that can be reached without interchange is represented as a separate link, a so-called “route section”. Therefore waiting times will only have to be considered at the boarding node of a “route-section”.

![Figure 3-4 Simple transit network with route sections (dual network)](image)
A route section can also include several lines with different travel times as shown in Figure 3-4 with $S_i$. In this case the expected in-vehicle travel time is calculated in proportion to the frequency of each line as in (3.3).

$$t_S = \frac{\sum_{i \in S} f_j t_j}{\sum_{i \in S} f_i}$$  \hspace{1cm} (3-3)

Because Chiriqui and Robilliard (1975) and De Cea and Fernandez (1993) do not explicitly use strategies this leads in general to a suboptimal solution. The slightly larger example network used by Spiess and Florian (1989), as well as De Cea and Fernandez (1993), illustrate this well (Figure 3-5 and Table 3-1). In the route section model passengers will choose the section that minimises the costs (and it is not considered that other routes might be quicker if they arrive earlier). Within the lines of a route section the flow is assigned to the specific lines according to their frequency. In Figure 3-5 this leads to $S_i$ not being used, whereas Spiess and Florian show that the optimal strategy includes using Line 1. Line 1 is for example quicker if it arrives immediately and passengers experience the expected waiting times for Lines 2 and 3. In general the approach using explicit strategies will lead to greater route dispersion.
a) Spiess and Florian (1989) example network

b) Same network with route sections

**Figure 3-5** Example network used in Spiess and Florian (1989) and De Cea and Fernandez (1993)

**Table 3-1** Differences in assignment results for network in Figure 3-5

<table>
<thead>
<tr>
<th>Link usage</th>
<th>Early work</th>
<th>Route Sections model</th>
<th>Strategy model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dial (1967); Fearnside and Drapper (1972)</td>
<td>(Chiriqui, 1975; De Cea and Fernandez, 1993)</td>
<td>(Spiess and Florian, 1989)</td>
</tr>
<tr>
<td>v₁</td>
<td>1</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>v₂(A-B)</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>v₂(B-C)</td>
<td>0</td>
<td>0.83</td>
<td>0.5</td>
</tr>
<tr>
<td>v₃(B-C)</td>
<td>0</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>v₃(C-D)</td>
<td>0</td>
<td>0.17</td>
<td>0.08</td>
</tr>
<tr>
<td>v₄</td>
<td>0</td>
<td>0.83</td>
<td>0.42</td>
</tr>
</tbody>
</table>

Chiriqui and Robillard (1975) point out that the “clever” passenger will reduce his travel time by observing the time he had to wait for the vehicle arriving first. If taking this vehicle means that his travel time is larger than the expected travel time
he might wait for a faster service on the basis that it is likely to arrive soon. This assumes an exponentially distributed waiting time with mean being the headway. If the arrival process is irregular this is of course not the case.

In general, the strategy will be more complex, the more information the passenger has and the more opportunities the network offers. Three levels of information and strategy complexity can be distinguished: If the passenger is bound to a specific line (e.g. because of ticket restrictions) he or she fixes the route before arriving at the platform (Level 1). The traveller might buy his or her ticket and make the route decision on the assumption that he or she has perfect information about the best route to his or her destination.

If, however, the passenger does not have to make the route decision before arriving at the platform, the route choice might be governed by the next vehicle arriving, e.g. several lines offering a service to the passenger’s destination or at least closer to the destination (Level 2). This is the common line problem as considered in the Spiess and Florian (1989) paper and the case most literature deals with. The passenger only knows the frequency of the services (and not the exact remaining waiting time until the arrival of the fast vehicle) and if the traveller sees the slower service arriving, he or she might prefer to take this vehicle, because the expected waiting time for the express service is too high to make it worthwhile waiting.

Nowadays, the passenger often has information about the arrival times of the subsequent vehicles as well. This third level of information is realistic for transit networks, where passengers can buy a network ticket and displays on the platforms
show the sequence and time left for the next trains arriving (metros and some buses in the U.K.). In this case the traveller might for example skip the next vehicle arriving, because he or she knows that an express-service is waiting to enter the station.

3.2.3. Linear Programming formulation and solution algorithm to the common line problem

In Spiess and Florian (1989) every route is represented with a link between each stop with the attributes travel time and frequency. Boarding links have no travel time but a frequency associated with them, alighting links have a cost of zero travel time and zero frequency.

With the assumption of exponentially distributed interarrival times the expected waiting time for each link can be calculated as in (3.4) with the sum of the frequency $f$ of all attractive outgoing links $A^+$ at a node $i$, and the link probability $P_a$ as in (3.5) with the frequency of this line compared to the frequency of all attractive lines at this node.

$$ W(A^+) = \frac{1}{\sum_{a \in A^+} f_a} $$  \hspace{1cm} (3-4)

$$ P_a(A^+) = \frac{f_a}{\sum_{a \in A^+} f_a} $$  \hspace{1cm} (3-5)

A non-trivial task is to find the optimal strategy that reduces the total journey time. Spiess and Florian present the problem as a linear program. The objective is to find the strategy that minimises the sum of the expected total travel time including waiting time for all flow from all origins to a destination. For each destination the
problem is formulated as in Eq. 3.6 where $v_a$ is the flow of link $a$, and $w_i$ is the total expected waiting time at node $i$.

$$\min \sum_{a \in A} c_a v_a + \sum_{i \in F} w_i$$  \hspace{1cm} (3.6)$$

The constraints belonging to the objective function (3.6) are firstly, flow conversation at each node, secondly the relationship between link flow and frequency stemming from the assumption that passengers board all attractive lines in proportion to their normal frequency (3.5) and thirdly the non-negativity of link flows.

Finally, they construct the dual formulation of the problem. The minimisation problem becomes a maximisation problem and the Lagrange multipliers $u_i$ can be interpreted as the expected total travel time from node $i$ to the destination but excluding the waiting time at $i$. The formulation of the dual problem is also the basis for the solution algorithm. The algorithm is composed of two steps; in the first the optimal strategy is found and in the second the traffic is assigned from all origins to the destination. The algorithm needs to be solved for each destination separately. It searches from the destination to all origins by taking the links closest to the destination first with increasing order. The key to finding the optimal strategy is then a comparison of the costs $u_i$ and $u_j + c_a$ for each link from $i$ to $j$. If $u_j + c_a \leq u_i$ the link $a$ will be added to the attractive links at node $i$. Since this comparison does not include the waiting time, this means that if the route arrives first at $i$, this service will be a shorter route than all options looked at so far. If the above comparison holds, the service frequency for the node will be increased by the frequency of the newly added service. This search for the optimal set of arcs is the same as the search for the

Spiess and Florian prove that the solution obtained by the algorithm is an optimal one by a) showing that \((u^*, v^*)\) are a feasible solution for the primal and dual problem and b) that the complementary slackness conditions are fulfilled. These state that for all used arcs the constraints of the primary problem are strict equalities. Optimality follows then directly from the linear programming theorem.

De Cea et al (1988) present a mathematical formulation of the route sections model as in Figure 3-5b which is very similar to above formulation. The dual variables in the model have the same meaning and the solution algorithm also follows an adaptation of Dijkstra’s shortest path algorithm. De Cea et al further explore the differences between results produced by the “strategy” and the “route section” model. Both approaches are tested on a large scale network with a high percentage of links that carry common lines. The results confirm those shown in the previous section, where the strategy approach leads to a wider and more equal distribution of flows on more routes. However, De Cea et al conclude that the differences in flows are not large but that the route section approach requires a significantly longer computation time. In reality the difference between the approaches will depend on factors like the fare scheme. If passengers have to pay each time they board there will be almost no difference but if there is one fare for the whole network the differences are likely to be larger since the route section approach will underestimate the route dispersion.
3.3. **The effective frequency approach**

3.3.1. **Problem description**

Spiess and Florian (1989) discuss the extension of their model to the case when link costs are depending on link flows, i.e. $c_a(v)$. This could be used to model discomfort in crowded vehicles. In this case their proposed solution algorithm needs to be adjusted since the problem cannot be solved destination by destination. The linear problem can be restated as a convex optimisation problem which is then solved with the linear approximation method. The method requires that the cost function must be continuously monotone increasing, which is not necessarily the case. A second limitation of this approach is that all passengers on board the vehicle suffer the same inconvenience, independent of where they boarded. This assumption neglects the fact that passengers boarding earlier are more likely to get a seat and therefore are less likely to experience inconvenience through crowding. A third limitation with this approach is that it assumes that the service frequencies are not affected by crowding.

Similarly, Lam et al (1999) propose to solve transit assignment with capacity problems through the introduction of “line specific overload delays”. The problem with their approach is also that it does not consider priority rules on transit systems. Passengers already on-board will not be deterred by passengers attempting to board later. The achievement of the Lam et al paper is rather that it formulates transit assignment as a stochastic user equilibrium problem. The problem is formulated as a mathematical programming problem and the Lagrangian multipliers of their solution are equivalent to the line specific overload delays. In Lam et al (1999) the frequency of the transit line is fixed, in Lam et al (2002) this is relaxed through the assumption
that the number of passengers boarding and alighting will influence the dwell time and hence the service frequency. However, in summary, flow dependent link costs are only suitable to model on-board inconvenience but not to model the problem of limited capacity.

3.3.2. Effective frequencies with practical capacities

Spiess and Florian mention that the aforementioned limitations can be overcome with the “effective frequency approach”. This approach is followed up by De Cea and Fernández (1993). The travel costs of transit arcs are assumed to be constant but the travel costs of waiting links are dependent on the link flows. Since the De Cea and Fernandez model is based on route sections there is a waiting link for each route section link. This one-to-one relationship makes it possible to collapse waiting links and transit links into a single link with in-vehicle cost plus waiting cost. The waiting cost term consists of two parts where the second term is a function of the route section flows $V_s$, the flow on other route sections competing with this flow $\tilde{V}_s$, and the “practical capacity” $K_s$ of this link (3.6). The in-vehicle cost $\bar{t}_s$ is calculated with Eq. (3.3).

$$c_s = \bar{t}_s + \frac{\alpha}{f_s} + \beta \varphi_s \left( \frac{V_s + \tilde{V}_s}{K_s} \right)$$  \hspace{1cm} (3-6)$$

$$w'_i = \frac{\alpha}{f_i} + \varphi \left( \frac{\tilde{V}_i}{K_i} \right)$$ \hspace{1cm} (3-7)$$

The condition on the flow function $\varphi_s$ is that it needs to ensure that $c_s$ is strictly monotone increasing with $V_s$. De Cea and Fernández suggest a power function with $n$
between 4 and 6 based on the experience for car networks. They note that $K_s$ is not a real capacity but that “as $n$ increases, $K_s$ acts more and more like a real capacity”. It is important to note that this approach does not guarantee that lines will not be overloaded. The competing passengers $\tilde{V}_s$ are made up of those passengers using all other route sections that contain lines that are part of section $s$ and alight after node $i(s)$, the origin of route section $s$.

The effective frequency is then defined as the inverse of the waiting time index $w_i^l (3.7)$. The idea behind this is that with more buses arriving full, the waiting time will increase, because it is harder to get onto the vehicle. If the vehicle arrives empty the effective frequency at node $i$, $f_i^l$, is equal to the nominal frequency $f_i$. On the other side a service will never be totally full because with increasing congestion $f_i^l \to 0$, but $f_i^l = 0$ will never be reached. De Cea and Fernández define the effective frequency as being the same for all passengers waiting. It is not considered that a high number of passengers wishing to board will reduce the chance for an individual to board.

De Cea and Fernández prove that the set of attractive lines monotonically increases with congestion and that the set of attractive lines in uncongested situations is also among the attractive lines in congested situations. Further, with this definition of effective frequency and $w_i^l$ dependent on the link flows the equilibrium problem is no longer linear. De Cea and Fernández conclude by presenting two algorithms to find an equilibrium solution that satisfies Wardrop’s principle (which states that the costs on all used paths is equal to the minimum cost and the flows on all paths with costs greater than the minimum cost is zero). In the first algorithm, congestion is only
considered at the route section level. The line frequencies are assumed to be constant but the costs still depend on the ratio \((V_s + \tilde{V}_s)/K_s\). This means that the user equilibrium condition has now an equivalent convex cost optimisation problem that can be solved with a Frank-Wolfe-type-algorithm. In this algorithm one calculates the minimum cost routes along the search direction \(\frac{\partial c}{\partial V}\) and assigns the demand to the newly obtained routes. If two subsequent iterations produce sufficiently close solutions the stopping criterion is fulfilled.

The second algorithm does assume effective frequencies before using the Frank-Wolfe type algorithm but loads the lines sequentially starting with the depot where it is known that the effective frequency equals the nominal frequency. Once the lines have been loaded \(\frac{\partial c}{\partial V}\) can be built and an improved solution can be found. This method is computational very demanding and De Cea and Fernández further do not give a formal proof of convergence for this intuitively correct solution method.

Both solution methods are demonstrated with the example network shown in Figure 3-5. In Chapter 3.2 the solution to the uncongested problem has already been presented. Results with the linearised approach show that if the network becomes mildly congested \(S_1\) becomes additionally attractive. In case of extreme congestion also route \(A \rightarrow S2 \rightarrow B \rightarrow S6 \rightarrow D\) is used. Line 1 is heavily overloaded in case of extreme congestion. Using the exact algorithm only leads to a slight reduction of the overloading on Line 1.
3.3.3. Effective frequencies with strict capacities

De Cea and Fernández (1993) acknowledge that “practical capacities” do not solve the problem of overcrowding. According to De Cea and Fernández, strict capacities will further not allow for explicit performance and convergence functions. However, they argue that this approach might be sufficient for strategic planning where one is not concerned about passenger loads for specific runs but is looking at the demand for the service over a longer time period.

Cominetti and Correa (2001) point out further shortcomings of the De Cea and Fernández approach. In particular they show that the above approach does not guarantee that Wardrop’s first principle is fulfilled. Cominetti and Correa prove that in some congested situations, costs can be minimised by a split between two strategies. Consider a network with one OD pair and two direct services between these where \( t_1 < t_2 \). The strategy to consider Line 2 only is obviously not optimal. Therefore passengers at the origin have the choice between two strategies: The first one is to wait for Line 1 and the second strategy is to take whichever line arrives first.

If the waiting time is not very high compared to the in-travel time in uncongested situations it is clearly better to wait for Line 1 even if Line 2 arrives first. Applying Spiess and Florian (1989) or De Cea and Fernández (1993) would yield the same result. In highly congested situations, the effective frequency approach of De Cea and Fernández would predict that it is better to use Strategy 2 because the effective frequency is much reduced compared to the nominal frequency (or in other words: “Since the network is congested and the fast service might be full, do not rely on this service but take whichever service is arriving first”). Cominetti and Correa (2001)
point out that it can however be easily shown that for certain demand regions a strategy mix in fact would reduce overall travel time. If in highly congested situations some passengers stick to Strategy 1 this would reduce overall travel time. In reality it might not be realistic to assume that some passengers would stick to Line 1 and not consider boarding Line 2, but this theoretical result shows that if in highly congested situations the operator would reduce the choice of some passengers, this can improve the overall performance of the network.

Cominetti and Correa (2001) then describe some conditions for this area of flow where a strategy-mix between two strategies is optimal. They further show that there are some areas where a flow increase does not result in an increase in travel cost as the demand can be accommodated by an uncongested line without cost increase.

Cominetti and Correa further use a formulation of the effective frequency based on queuing theory: If $v_i \to \bar{v}_i$ then $f_i' \to 0$ and $w_i \to \infty$ where $\bar{v}_i$ is the saturation flow and $w_i$ is the waiting time on the link. With this formulation of “strict capacities”, line loads will not exceed capacity. A further improvement of this paper is that they simplify Spiess and Florian’s formulation by using only arc flows and not strategy flows. Therefore it is not necessary to keep track of the different origins in the assignment. The assignment algorithm assumes that passengers travel on shortest hyperpaths and Dijkstra’s algorithm is used for finding shortest hyperpaths. As in De Cea and Fernández, the distribution of flows across links of a hyperpath is proportional to the effective frequency of this link.
Cepeda et al (2006) continue the work of Cominetti and Correa (2001) and prove some important characteristics of transit assignment with effective frequencies. Firstly they show that the ratio $v_a / f_a$ can be assumed to be an (unknown) variable and is not a constant that needs to be determined before calculating the network equilibrium. This allows applying the Cominetti and Correa (2001) approach for large scale networks. Secondly, they show that the global equilibrium is not trapped by a local equilibrium. Therefore it is possible to use a gap function that becomes zero if the solution is equal to the optimal one. The gap function is based on Wardrop’s first principle and shows the difference between the cost of all passengers using the cheapest hyperpath and the cost for all passengers with the current assignment. Based on these findings they propose a heuristic solution algorithm using the Method of Successive Averages that solves congested transit assignment for large-scale networks. Several case studies in major cities around the world are briefly described demonstrating the validity of this approach.

3.4. Approach with fail-to-board probabilities

One of the earliest models to take into account capacity constraints is the TRANSEPT model presented by Last & Leak (1976). The purpose of the model was to predict patronage on bus routes. Several case studies were conducted in complex networks and robust results achieved. The model is used in multi-modal studies estimating the mode choice between walking, car and bus with a logit model. It can be applied in order to evaluate the cost benefits for users and operators for different allocations of the buses.
In their model they use an intermediate level of network representation for bus routes to take advantage of the fact that many bus services use the same paths for large sections of their routes. Sections of routes where no bus routes start, end or divert are represented as one link. Interchanging between different bus routes is only possible at the end of such a link. TRANSEPT further assumes that all passengers take the shortest path, which means that the common line problem is not taken into account in so far as the choice of path might also depend on which vehicle is arriving first.

The choice between specific bus routes (on a common line) takes waiting times as well as capacity constraints into account. The headway is calculated as the sum of the frequencies and the average waiting time is calibrated with survey data. A logit model is used to distribute the demand between the routes according to their attractiveness and capacity constraints are then taken into account within an iterative process. The competition for space is modelled by loading the services from upstream to downstream. In this way the priority of those boarding further upstream is observed. This procedure allows calculating a probability of failure-to-board for each route and stop. This failure probability is used to increase the average waiting time. An iterative process is then used to load the network and “stabilise” the waiting times. Probability distributions are used to estimate the number of passengers boarding as well as alighting at each stop. The model requires that the overall capacity is sufficient for the demand as the waiting time would otherwise grow to infinity. If this is not the case, the model allocates additional buses to critical routes.
Last and Leek (1976) do not describe the mathematical formulation of their model in detail, so that it is not fully clear how they stabilise the waiting times. It is however clear that their procedure of “Looking for a stable solution” is not always realistic. Let us assume that the example network used before has the demand and capacity constraints as shown in Figure 3-6. In this case, the only stable solution would be if all passengers destined for C would take Line I at A. As explained before, this might however not be the case, because Line II will be more attractive to passengers for C if it arrives first. De Cea and Fernández (1993) further point out that loading the TRANSEPT network link by link means it is only feasible for radial networks.

![Diagram of the transit network with capacity constraints](image)

Figure 3-6 Transit network with capacity constraints

3.5. *Passenger distribution between different paths of one hyperpath (bus stop problem)*

Bouzaïene-Ayari *et al* (2001) point out that in the previous literature the distribution of passengers between the attractive lines of a strategy/hyperpath at a node is not always realistic. Instead of a priori using the nominal or effective frequency as basis for the distribution, Bouzaïene-Ayari *et al* generalise Eq. (3.5) and use formulation
(3.8) where $\xi_a$ is the so-called attraction factor for arc $a$ which is included in the set of attractive lines $A^i$ at node $i$.

$$P(A^i) = \frac{\xi_a}{\sum_{a \in A^i} \xi_a}$$ (3-8)

Instead of using $\xi_a = f_a$ (Spiess and Florian, 1989; De Cea and Fernández, 1993) or $\xi_a = f_a'$ (Cominetti and Correa, 2001) Bouzaïene-Ayari (1988) showed with a simulation of several bus stops that in heavily congested situation $\xi_a = (1-v_a)f_a/K_a$ is a better approximation.

Bouzaïene-Ayari et al however point out several shortcomings of all these models. First of all $\xi_a = f_a$ obviously does not allow for the possibility that one might not be able to board a vehicle because of overcrowding. Secondly, all above functions for $\xi_a$ assume that all transit lines have the same exponential distribution of waiting time. This implies high service irregularity for all attractive lines. This might however not be true and it should be considered that the regularity between the services might differ. At a node where passengers can board trams as well as buses, the trams might well be significantly more regular. To overcome this problem, Gendreau (1984) and Bouzaïene-Ayari et al propose to use an Erlang distribution for the waiting time of each line with a line specific shape parameter $k$. They state that the Erlang distribution has a number of well known properties which are widely used in queuing theory. Further, with $k \to 1$ an extremely irregular distribution and with $k \to \infty$ a deterministic distribution of the line regularity can be modelled.
Finally, Bouzaïene-Ayari et al describe that the models considering congestion, $\xi_a(v_a)$, are not necessarily consistent with the theory of strategies. If there is a mix of strategies, i.e. some passengers will always wait for the fast service, whereas others will take whichever service arrives first, then a model $\xi_a(v_a)$ might predict a higher waiting time for the latter strategy. They state that this is however irrational. The effective frequency approach leads to this because flow increases travel time even if flow is lower than capacity. Bouzaïene-Ayari et al’s critique of this being irrational is not justified if one considers inconvenience costs for travelling with the more crowded line: The passenger with the strategy “take whichever line comes first” might well encounter higher costs because less waiting time is outweighed by on-board vehicle crowding.

In order to overcome these shortcomings, Bouzaïene-Ayari et al propose that $\xi_a = 1/ W_a(v_a)$ where $W_a$ is the waiting time function for line $a$. Different waiting time distributions can be assumed where one needs to consider that those with strict capacity constraints ($W_a \rightarrow \infty$ for $v_a \rightarrow \bar{v}_a$) are more difficult to integrate into assignment models than those with practical capacity constraints ($W_a \rightarrow \infty$ for $v_a \rightarrow \infty$).

In general it can be concluded that Bouzaïene-Ayari et al’s work defines waiting time more accurately if good data are available. It is important to note that the work of Bouzaïene-Ayari only concerns the distribution of passengers at a bus stop. The paper does not consider the problem of finding the optimal set of links that should be included in the strategy/hyperpath.
3.6. **Discussion**

This chapter reviewed the major approaches to frequency-based transit assignment. It was first shown that the first assignment approaches, which were based on traffic assignment assumptions, did not consider that the Markov assumption does not hold for transit assignment. Passenger route choice at any given interchange will depend on the destination as well as the origin. Further improvements to frequency-based transit assignment did consider the common line problem. Chapter 2 discusses in which situations the common line problem is applicable and that passengers develop different strategies depending on the information about train arrival times. A solution to transit assignment with common lines based on linear programming is discussed following Spiess and Florian (1989).

The second part of this chapter reviews assignment that considers the consequences of capacity shortages. The most frequently used approach is the “effective frequency” one. Approaches with practical and with strict capacities have to be distinguished. Using practical capacities means that line-overloading is still possible. It further leads to situations where the Pareto-optimal solution is not equal to the network equilibrium. There are problems to find an equilibrium with the strict capacity approach as the network capacity might actually not be sufficient. Cepeda et al (2006) suggest using a dummy network that connects all destinations with walking links. Cominetti and Correa (2001) show that the problem can be described as a fixed point problem. But so far only heuristics are developed for the case where one allows
for the effective frequency influencing the route choice of passengers. Nevertheless, the approach has been shown to successfully model large scale overcrowded networks.

The effective frequency approach assumes strictly increasing cost functions for increasing demand. This can reflect discomfort caused by congestion. It does however not accurately reflect the cost of missing a vehicle due to overcrowding. The cost for this will increase rather more like a step function. One could however argue that, in heavily congested situations, a continuous function reflects the continuously increasing probability of missing a vehicle due to overcrowding. However, assuming that this risk of failing to board (an additional cost in heavily congested situations) is a continuous and not a different function from the discomfort caused through full vehicles is a major assumption. A further critique of the effective frequency approach is that it does not allow estimation of the number of passengers that are not able to board the vehicle. It is therefore also not suitable to model the build-up of congestion over time.

Chapter 2 reviewed schedule-based assignment approaches and in particular dynamic capacity constraint models. As mentioned in Chapter 3.3 De Cea and Fernández (1993) argue that for strategic planning one is not interested in passenger loads for specific runs. This is true for frequency-based services but one might still be interested in dynamic effects. For example providing sufficient capacity for the demand of the extended morning peak might still lead to severe congestion during the peak of the peak. Overcrowding within short time intervals are safety risks for the
whole service and lead to delays and hence reduced overall capacity during long periods of time.

In the following chapters an alternative to the effective frequency approach is described which is aimed at overcoming some of its shortcomings.

4.1. Introduction

The literature review highlighted several shortcomings of frequency-based assignment. Especially in congested situations when the demand might exceed the network capacity, and when the demand is changing over the assignment period the weaknesses of the currently available transit assignment tools are known. To overcome some of these shortcomings the following new method is developed. A main feature of this model is the explicit inclusion of strict capacity constraints, which is why the model is referred to as “CapCon”. These capacities are line specific and the approach does not rely on effective frequencies. This approach allows an explicit estimation of the number of passengers not being able to board which is also the basis for the dynamic extension described in Chapter 5.

This chapter first explains the network description and the notation used throughout this and the following chapters (4.2). Then the cost elements that are considered for the path choice are described, in particular the “risk of failing to board because of overcrowding”. In 4.3 it is explained how this cost can be described node specifically, which is the requirement for the optimal hyperpath search algorithm in 4.4. Chapter 4.5 explains how the path choice probabilities are used to calculate the network equilibrium and in 4.6 it is shown that there is a unique equilibrium for non-circular
as well as circular lines if one assumes passenger mingling. The network loading procedure becomes a deterministic user equilibrium assignment with the application of the MSA described in Chapter 4.7. If one does not consider common lines the hyperpath search becomes the search for the shortest path and also the network description can be significantly simplified. This is described in Chapter 4.8.

Sections 4.9 and 4.10 conclude this chapter. In Section 4.9 an application of the CapCon model to a small example network is described. The effects of introducing common lines and passenger risk averseness are shown. Finally, Section 4.10 gives a short summary and discusses the merits and shortcomings of this approach. The model presented in this chapter has also been published in several publications: Bell et al (2002) discuss the idea of applying absorbing Markov chains to network analysis and transit assignment, Schmöcker and Bell (2004) introduce capacity constrained transit assignment and Kurauchi et al (2003) focus on the path choice algorithm by replacing it with a search for the optimal hyperpath.

4.2. Network description and its notation

4.2.1. Network representation

Figure 4-1 shows the nodes and arcs of a station which has one platform and which is served by one line only. Besides the origin and the destination there are five node types. The stop node represents the bus stop or platform at which passengers wait for the service to arrive. To consider the capacity constraints failure nodes are introduced. The section on network loading will describe that all demand exceeding
the available capacity is not transferred further to the boarding node but instead to a fictional “bin” or excess demand node. It is further important to note that the network loading procedure assumes that at the boarding node passengers who board the service at this station and those who boarded upstream and did not alight mingle. At the alighting node passengers who stay on-board and those who alight at this station are separated, so that at the stop node alighting passengers, passengers who start their journey at this station and those passengers who failed to board at this stop node in the previous time interval mingle.

There are nine arc types which connect the nodes described above. Line arcs correspond to transit lines and connect a boarding node with the alighting node of the next station downstream. The cost of travelling on a line arc is assumed to be equal to the travel time between these two stations on this line. On-board wait arcs are used by passengers not alighting at this station. The cost of these arcs is equal to the dwelling time of the service at this stop. Access-walking arcs and egress-walking arcs connect the origin with the stop node and the stop node with the destination respectively. There is exactly one access arc and one egress arc in every hyperpath. The cost of walking from the origin to the first platform and the cost of walking from the last destination is currently not considered. Also alighting arcs, which connect alighting and stop nodes, and boarding-demand arcs, which connect a stop node with a fail node, do not have a cost associated with them. In order to reflect the reduced waiting time if passengers include common lines the waiting time is not associated with the boarding-demand arc or the boarding arc but instead with the stop node. For those passengers who get on the service, i.e. do not fail at the failure node, there is no additional cost, meaning that there is no cost associated with boarding arcs. The
amount of passengers exceeding the available capacity is transferred to their
destination via the failure arcs. Note that they are transferred to their respective
destination via the same failure arc, as the destination of the failure arc is redirected
to the destination of the passenger at the beginning of each hyperpath search as
explained in Section 4.4. This allows the formulation of the model in terms of flow
conservation, i.e. all demand is reaching its destination: Passengers who do not fail to
board anywhere without usage of failure arcs and those passengers who do fail to
board with the usage of exactly one failure link. Figure 4-1 illustrates that the excess
demand is removed from the downstream nodes. The cost of a failure arc is assumed
to be related to the passenger’s fear of failing to board and the probability of failing
to board which is discussed in Section 4.3.

Figure 4-1 Network representation of a station with 1 platform and 1 transit line

Figure 4-2 shows the arcs and nodes needed to represent a station with multiple
platforms and lines. It is important to note that failure nodes, alighting nodes and
boarding nodes (and hence also the arcs leading into or out of these nodes) are
transit line specific. However in general stop nodes are not line specific as there is for each platform or bus stop only one stop node. The figure further illustrates that the stop nodes belonging to the same station are connected via **transfer arcs**. There is one transfer arc connecting Platform 1 with Platform 2 and a second transfer arc representing walking in the opposite direction. The cost of the transfer arcs is set equal to the walking time between the two platforms.

**Figure 4-2** Station with 2 platforms and 2 lines serving platform 1
4.2.2. Notation (Glossary)

For the mathematical formulation of the assignment model the following notation will be used throughout this thesis. Four groups of variables can be distinguished: Global, node-specific, arc-specific and hyperpath-specific variables.

**Global**

\[ n: \] Number of nodes in network  
\[ n_o: \] Number of origins \( o \) in network  
\[ L: \] Set of Transit Lines (with \( l \in L \) and \( L_u: \) Set of transit lines served at platform \( u \))  
\[ cap_l: \] Capacity of a service on line \( l \)  
\[ R: \] Set of stations (with \( r \in R \))  
\[ U_l: \] Set of platforms served by line \( l \) (with \( u \in U \))  
\[ \theta_e: \] Sensitivity to boarding overcrowded services (risk averseness)

**Nodes**

\[ I: \] Set of Nodes (with \( i \in I \))  
\[ Out (a): \] The node arc \( a \) is leading out of  
\[ In (a): \] The node arc \( a \) is leading into.  
\[ E: \] Set of failure nodes (with \( e \in E \) and \( e_{ul} \) denoting a failure node associated with platform \( u \) and line \( l \))  
\[ \eta_i: \] 1 if node \( i \) is a failure node and 0 otherwise  
\[ S: \] Set of stop nodes (with \( s \in S \) and \( s_u \) denoting a stop node at platform \( u \))  
\[ \sigma_i: \] 1 if node \( i \) is a stop node and 0 otherwise  
\[ B: \] Set of boarding nodes (with \( b \in B \))  
\[ q_i: \] Failure probability at node \( i \)  
\[ sp^i: \] Available spaces at node \( i \) (vacancies for passengers wishing to board)  
\[ sp^r_i: \] Remaining spaces at node \( i \) (remaining vacancies after all passengers boarded)  
\[ y_d: \] Vector of demand from origins \( o \) to destination \( d \)  
\[ v: \] Passenger flow vector travelling via nodes \( i \) (\( v_{id}: \) and destined for \( d \))
\( \Pi_d: \) Matrix of transition probabilities from node \( i \) when destined for \( d \)

\( \Theta_d: \) Matrix of transition probabilities to an intermediate node \( j \)

\( \delta: \) Vector of transition probabilities from node \( i \) to destination \( d \)

**Arcs**

\( A: \) Set of Arcs (with \( a \in A \))

\( x: \) Passenger flow vector travelling via arc \( a \) (\( x_d: \) destined for \( d \))

\( wait_{ul}: \) On-board waiting arc on platform \( u \) for passengers staying on line \( l \)

\( line_{ul}: \) Line arc out of platform \( u \) on line \( l \)

\( board-d_{ul}: \) Boarding demand arc on platform \( u \) for line \( l \)

\( board_{ul}: \) Boarding arc on platform \( u \) for line \( l \) (The adjusted arc-flow considering capacity constraints)

\( on-board_{ul}: \) On-board wait arc on platform \( u \) for line \( l \)

\( c_{ai}: \) Cost of travelling on arc \( a \) connecting nodes \( i \) and \( j \)

\( f_{al}: \) Frequency of transit service on arc \( a \). Note that \( f_l \) and \( f_{l(a)} \) are also used because \( f_a \) is usually (but not necessarily) constant for all arcs served by the same line \( l \)

\( \pi_{ah}: \) Arc split probabilities. (Probability to choose arc \( a \) when traveller is at \( Out (a) \) and travelling on hyperpath \( h \))

**Hyperpath specific**

\( H_d: \) Set of all feasible hyperpaths to destination \( d \) (with \( h \in H_d \))

\( h: \) Hyperpath connecting intermediate node \( i \) to destination \( d \)

\( z_h: \) Flow on hyperpath \( h \) in the set of used hyperpaths

\( I_h: \) Set of nodes within hyperpath \( h \)

\( S_h: \) Set of stop nodes within hyperpath \( h \)

\( E_h: \) Set of failure nodes within hyperpath \( h \)

\( A_h: \) Set of arcs within hyperpath \( h \) (\( A_{hi}: \) Subset of arcs leading out of node \( i \))

\( g_h: \) Cost of travelling hyperpath \( h \)

\( w_{ih}: \) Expected waiting time at node \( i \) when travelling on hyperpath \( h \)

\( \alpha_{ah}: \) Probability of using arc \( a \) when travelling on hyperpath \( h \)
\( \beta_{hi} \): Probability of traversing node \( i \) when travelling on hyperpath \( h \)

\( P_h \): Set of elementary paths within hyperpath \( h \) (with \( p \in P_h \))

\( e_{ip} \): 1 if node \( i \) is element of path \( p \), 0 otherwise

\( \delta_{ap} \): 1 if arc \( a \) is element of path \( p \), 0 otherwise

\( \lambda_p \): Probability to choose elementary path \( p \)

4.3. The cost of a hyperpath

4.3.1. Elementary paths and hyperpaths

In Chapter 3 the common lines problem is explained. Common lines are respected in the following approach: It is assumed that the passengers select a set of lines from which they choose the first service arriving. Therefore, they are not choosing a shortest path \( p \), but a set of a paths that minimises the travel time or generalised costs. As explained in Chapter 3, following Nguyen and Pallottino (1988) this set of attractive paths \( P_h \) is called Hyperpath \( H \) and the set of paths that minimises the travel costs needs to be found. The following describes how the cost of such a hyperpath can be calculated. Note that the cost of a hyperpath is always equal or lower to the cost of any elementary path \( p \in H \) connecting origin \( o \) and destination \( d \).

4.3.2. Transition probabilities

In order to search for the shortest hyperpath the arc transition probabilities need to be defined. For every node \( i \) the arc transition probabilities \( \pi_{ah} \) satisfy:

\[
\sum_{a \in A_i} \pi_{ah} = 1, \quad \forall i \in I_h
\]  

(4-1)
In (4.1) and in all the following equations the subscript \( h \) indicates that the set of arcs or nodes concerns only those included in the hyperpath, i.e. \( A_{hi} \) is the set of arcs within hyperpath \( h \) that are leading out of node \( i \), further the arc transition probabilities should not be negative:

\[
\pi_{ah} \geq 0, \; \forall a \in A_{h} \quad (4-2)
\]

The following three equations define the transition probabilities at the various nodes of the network. At stop nodes \( \pi_{ah} \) is calculated proportional to the (nominal) frequencies of the lines that are included in the hyperpath:

\[
\pi_{ah} = f_a / \sum_{a \in A_{h}} f_a , \; \forall i \in S_{h} \quad (4-3)
\]

At failure nodes following relationship holds:

\[
\pi_{ah} = \begin{cases} 
1 - q_i & \text{if } \text{In}(a) \in B \\
q_i & \text{otherwise}
\end{cases}, \; \forall i \in E_{h} \quad (4-4)
\]

Since there are always exactly two arcs leading out of a failure node (one boarding arc and one failure arc) (4-4) also fulfils (4-1). Finally, at all other nodes there is never more than one arc included in a hyperpath so that:

\[
\pi_{ah} = \begin{cases} 
1 & \text{if } \text{Out}(a) \in I_{h} - S_{h} - E_{h} \\
0 & \text{otherwise}
\end{cases} \quad (4-5)
\]
4.3.3. The cost elements of a hyperpath

Let the probability of choosing any particular path $p$ of a hyperpath $h$, $\lambda_p$, be denoted as

$$\lambda_p = \prod_{a \in A_h} \delta_{ap} \qquad \forall p \in P_h \text{.}$$

(4-6)

with $\delta_{ap}$ equal to 1 if arc $a$ is an element of path $p$ and 0 otherwise. It follows therefore that

$$\sum_{p \in P_h} \lambda_p = 1$$

(4-7)

Further $\beta_{ih}$ is defined as the “probability of traversing node $i$ when travelling hyperpath $h$” and $\epsilon_{ip}$ is equal to 1 if node $i$ is an element of path $p$ and 0 otherwise so that

$$\beta_{ih} = \sum_{p \in P_h} \epsilon_{ip} \lambda_p \qquad \forall i \in I_h$$

(4-8)

$\alpha_{ah}$ is defined as the probability of using arc $a$ when travelling hyperpath $h$, so that

$$\alpha_{ah} = \sum_{p \in P_h} \delta_{ap} \lambda_p \qquad \forall a \in A_h$$

(4-9)

Using the definitions for $\beta_{ih}$ and $\alpha_{ah}$, enables the notation of the cost for travelling on hyperpath $g_h$ (4-10) by summing up the costs for all arcs and nodes that are part of the hyperpath. Note, that the cost term $c_a$ relates to the on-board travel time (for line arcs), dwelling times at stations (for on-board waiting arcs) and walking times (for transfer arcs) plus any other generalised costs, but excludes the waiting times. The waiting times at stop nodes $S_h$ are separately calculated with the second term in
(4-10) to represent the reduced expected waiting time $w_{ih}$ when the hyperpath includes several attractive paths. If the common line problem is not considered one could simply delete the second term and add the expected waiting time to the cost term $c_a$. The third term shows the costs of network overcrowding to the traveller. It is assumed that the traveller associates the probability of failing to board somewhere along the journey with an increased generalised cost. In this case it is not considered how much delay failing-to-board causes the traveller, but failing-to-board in itself is associated with a higher generalised costs. The risk averseness of passengers, or sensitivity of passengers to avoid boarding at overcrowded stations is expressed with the parameter $\theta_c$.

$$
\begin{align*}
g_h &= \sum_{a \in A_h} \alpha_{ah} c_a + \sum_{i \in S_h} \beta_{ih} w_{ih} - \theta_c \ln \left( \prod_{i \in E_h} (1 - q_i) \right) \\
&= \sum_{a \in A_h} \alpha_{ah} c_a + \sum_{i \in S_h} \beta_{ih} w_{ih} - \theta_c \ln \left( \prod_{i \in E_h} (1 - q_i) \right) \\
&= g_h - \theta_c \ln \left( \prod_{i \in E_h} (1 - q_i) \right)
\end{align*}
$$

The waiting time at a node is given by (4-11) which assumes an exponential distribution of the vehicle arriving times with the mean being their nominal frequency, but alternative assumptions are also possible (as discussed in Chapter 3.2)

$$
w_{ih} = \frac{1}{\sum_{a \in A_h} f_a}
$$

### 4.3.4. Bellman’s principle (The costs from a node)

In order to develop a search algorithm for an optimal hyperpath, it is advantageous to split the hyperpath costs into costs from each node, so that the costs from each node to the destination can be tested and the non-optimal nodes excluded. This requires
that the Bellman principle applies. This principle of optimality says that “an optimal path has the property that whatever the initial conditions and control variables (choices) over some initial period, the control (or decision variables) chosen over the remaining period must be optimal for the remaining problem, with the state resulting from the early decisions taken to be the initial condition” (Economics online, 2005)

In this case it means that the optimal path from node \( i \) is optimal for any traveller who has reached this node, independent of the origin. The following proves that Bellman’s principle applies to the cost formulation (4-10):

The problem of finding hyperpath \( h \) that minimises (4-10) is

\[
\min_{h \in \mathcal{H}} \sum_{a \in \mathcal{A}_h} \alpha_{ah} c_a + \sum_{i \in \mathcal{S}_h} \beta_{ih} w_{hi} - \theta_i \ln \left( \prod_{i \in \mathcal{E}_h} (1-q_i)^{\theta_i} \right)
\]  

(4-12)

subject to

\[
w_{ih} = \frac{1}{\sum_{a \in \mathcal{A}_i} f_a} \quad (4-11), \quad \beta_{ih} = \sum_{p \in \mathcal{P}_i} e_{ip} \lambda_p \quad (4-8) \quad \text{and} \quad \alpha_{ah} = \sum_{p \in \mathcal{P}_i} \delta_{ap} \lambda_p \quad (4-9)
\]

Let \( g_{h'} \) denote the cost of the sub-hyperpath \( h' \) from intermediate node \( i \) to \( d \) on hyperpath \( h \), and define the probability of using elementary path \( p_i \), \( \lambda_{p_i} \), as:

\[
\lambda_{p_i} = \prod_{a \in \mathcal{A}_{h'}} \pi_{ah}
\]

(4-13)

then clearly

\[
\sum_{p \in \mathcal{P}_{h'}} \lambda_{p} = 1
\]

(4-14)
and for all arcs $a$ connecting nodes $i$ and $j$

$$\lambda_{ip} = \pi_{ab}\lambda_{p'}, \forall a \in A_h$$  \hspace{1cm} \text{(4-15)}

Using $\lambda_{ip}$, $g_{hi}$ can be written as follows.

$$g_{hi} = \sum_{p \in P_{h,i}} \lambda_{p'} \left( \sum_{a \in A_{h,i}} \delta_{ai} c_a + \sum_{k \in S_{h,i}} e_{kp} w_{hk} - \theta \sum_{k \in E_{h,i}} e_{kp} \ln(1 - q_{h,i}) \right)$$

$$= \sum_{p \in P_{h,i}} \lambda_{p'} \left( c_{a_{p'}} + \sigma_i w_{h,i} - \eta_i \theta, \ln(1 - q_{h,i}) \right) + \sum_{a \in A_{h,i}} \pi_{ab} g_{hi}^{ab}$$  \hspace{1cm} \text{(4-16)}

where $\sigma_i$ is 1 if node $i$ is a stop node and 0 otherwise, $\eta_i$ is 1 if node $i$ is a failure node and 0 otherwise, and $a_{p'}$ denotes the arc on elementary path $p$ leading out from node $i$. Equation (4-16) shows that the costs can be calculated from each node $i$ as the costs from node $i$ to the downstream node $j$ plus the costs from node $j$ to the destination and hence the Bellman principle applies. The last line of (4-16) explains that this is valid for all node types: If $i$ is a stop node the waiting time at the node can be determined, if $i$ is a failure node, the passenger’s “safety margin” can be calculated and otherwise simply the arc cost of travelling from $i$ to $j$ is considered.

Finally, because the Bellman principle applies, (4-17) can be applied for each node which finds the minimum cost hyperpath. In the equation * indicates optimality and $A^*$ denotes a subset of arcs at node $i$ among $A_i$. 
\[
g^*_i = \begin{cases} 
0 & \text{if } i = d, \\
\gamma^*_i & \text{if } i \in \{B, A, o\} \\
1 + \min_{A' \subset A} \frac{\sum_{a \in A'} f_a g^*_i}{\sum_{a \in A'} f_a} & \text{if } i \in S \\
-\theta_e \ln(1 - q_i) + g^*_i & \text{if } i \in E
\end{cases}
\]  

(4-17)

4.4. Finding the optimal hyperpath

4.4.1. Introduction

The algorithm below finds an optimal hyperpath from all nodes to destination \(d\) and finds the arc split probabilities \(\pi_{ah}\) of matrix \(\Pi_d\). Optimal means in this case the hyperpath that minimises the generalised cost as defined in (4-10). The structure of the hyperpath search follows the one suggested by Nguyen and Pallotino (1988).

In the algorithm three sets of arcs are defined with all arcs being moved from set \(M_1\) to \(M_2\) or \(M_3\) during the procedure. Set \(M_1\) includes all not-yet checked arcs. All arcs that are found to minimise the travel cost, i.e. are part of one or more paths which are element of the optimal hyperpath are summed up in set \(M_2\). Those arcs that are checked and are not part of the optimal hyperpath are summarised in set \(M_3\).

After the initialisation Step 2 selects that arc among the arcs from set \(M_1\) with the least cost to the destination \(d\). In Step 3, if this arc is a boarding demand arc (out-node is a stop node and in-node is a Fail-to-board node) then it is checked whether the inclusion of this boarding demand arc can reduce the total cost to the destination under consideration of the reduced waiting time from the stop node through the inclusion of this arc.
If the origin of the arc is a failure node then it is checked whether the inclusion of this arc in the hyperpath can reduce the costs from the failure node under consideration of the person’s risk-averseness. Similarly for all other nodes: If the travel cost of this arc plus the cost to reach the destination from the arc’s in-node can reduce the costs from the out-node of the arc, the arc is included in the hyperpath (i.e. added to set $M_2$) and otherwise not (added to set $M_3$).

In Step 4 the costs of the failure arc are updated and set equal to the minimum cost found to the failure node so far. This is in order to ensure that the failure arc from node $i$ is taken with probability $q_i$, i.e. the choice is independent of the costs of the failure arc and only depends on the ratio between boarding demand and available spaces as will be further explained in Chapter 4.5.2 and in the section about network equilibrium (Chapter 4.6.2). Step 5 defines the termination of the algorithm and the final Step 6 defines the transition probabilities for all arcs in set $M_2$ with the relationships described in Eq. (4-3) to (4-5).

### 4.4.2. Hyperpath Search Algorithm

**Step 1** (Initialisation)

Set: $M_1:=A$, $M_2:=\emptyset$, $M_3:=\emptyset$;  
$g^i:=\infty$, $\forall i \in \{I-d\}$, $g^d:=0$;  
$\pi_a:=0$, $\forall a \in A$;  
$c_a:=\infty$, $\forall a \in A$ if ( $Out(a) \in E$ and $In(a) = d$ )

Set destination to be $In(a)$ in (2)
Step 2 (Finding a node of minimum cost from the destination)

Find \(a^*\) such that \(a^* = \min_{a \in M_1} \{ f_a + g^{in(a^*)} \} \);

Set \( M_1 := M_1 \setminus \{ a^* \}, M_2 := M_2 + \{ a^* \} \);

Step 3 (Updating node labels)

if \( Out(a^*) \in S \) and \( In(a^*) \in E \) then

Find a set of arcs \( A^* \subseteq \{ A_i \cap M_2 \} \), such that

\[
g^{Out(a)} := \min_{a \in A^*} \left( 1 + \sum_{a' \in A^*} f_a g^{In(a')} \right) \quad \forall a \in \{ A_i \cap M_2 \},
\]

if \( a \not\in A^* \) then Set \( M_2 := M_2 \setminus \{ a^* \}, M_3 := M_3 + \{ a^* \} \);
else if \( Out(a^*) \in E \) then

if \( g^{Out(a^*)} \geq -\theta_e \ln(1 - q_{Out(a^*)}) + g^{In(a^*)} \) then

\[
g^{In(a^*)} := -\theta_e \ln(1 - q_{Out(a^*)}) + g^{In(a^*)}
\]

else

Set \( M_2 := M_2 \setminus \{ a^* \}, M_3 := M_3 + \{ a^* \} \);
else

if \( g^{Out(a^*)} \geq c_a + g^{In(a^*)} \) then

\[
g^{Out(a^*)} := c_a + g^{In(a^*)}
\]

else

Set \( M_2 := M_2 \setminus \{ a^* \}, M_3 := M_3 + \{ a^* \} \);

Step 4 (Updating costs for failure arcs)

if \( Out(a^*) \in E \) and \( In(a^*) = d \) then

\[
c_a := g^{Out(a^*)};
\]

Step 5 (Iteration, Termination of Arc Search Loop)

Repeat (2) to (4) until \( M_1 = \phi \).
Step 6 (Calculating arc transition probabilities)

\[ \forall a \in M_2, \]

if \( Out(a) \in S \) then

\[ \pi_{ah} = \frac{f_a(a)}{\sum_{a \in A \setminus M_2} f_a(a)} \quad (4-3) \]

else if \( Out(a) \in E \) then

\[ \pi_{ah} = \begin{cases} 
1 - q_{Out(a)} & \text{if } In(a) \in B \\
q_{Out(a)} & \text{otherwise} 
\end{cases} \quad (4-4) \]

else

\[ \pi_{ah} = 1; \quad (4-5) \]

4.5. **Network Loading**

4.5.1. Arc and Node Volumes

As defined in the glossary, \( \Pi_d \) denotes a transition probability matrix for trips destined to \( d \), \( \Theta_d \) a matrix of transition probabilities (from node \( i \) and origin \( o \)) to an intermediate node \( j \) and \( \delta \) a vector of transition probabilities from node \( i \) to destination \( d \). Then

\[ \Pi_d = \begin{pmatrix} 
1 & 1 & 0 \\
\delta & \delta & \Theta_d \\
1 & n-1 
\end{pmatrix}^{n-1} \quad (4-18) \]

When \( \Pi_d \) is multiplied by itself \( k \) times, \( \Pi_d^k \) is obtained which denotes the probability of reaching any node within \( k \) moves.

\[ \Pi_d^k = \begin{pmatrix} 
1 & 1 & 0 \\
\delta^k & \delta^k & \Theta_d^k \\
1 & n-1 
\end{pmatrix}^{n-1} \quad (4-19) \]
When the transition is repeated *ad infinitum*, all traffic should be absorbed into the
destination, i.e. \( \lim_{k \to \infty} \Theta^k_d = 0 \).

By using this relationship, the probability \( \Pi^{k=\infty}_d \) denotes the probability that traffic
destined to \( d \) traverses from \( i \) to \( j \) which can be calculated as in (4-20) (see for
example Bell, 1995):

\[
\Pi^{k=\infty}_d = \Pi^0_d + \cdots + \Pi^k_d + \cdots = (I - \Pi_d)^{-1} - I = \left( \begin{array}{ccc}
0 & 1 & 1 \\
\delta^k & 1 & 1 \\
(1 - \Theta_d)^{-1} & 1 & 1
\end{array} \right)
\]

(4-20)

This series is convergent provided \( \Pi^{k}_d \to 0 \) as \( k \to \infty \). If the row sums of \( \Pi^{k}_d \) are
bounded by 1 and at least one row sum is strictly less than 1, and provided each
destination is reachable from any arc, it follows from the Perron-Frobenius Theorem
(see Cox and Miller, 1965) that all the eigen values of \( \Pi^{k}_d \) are strictly less than 1 and
so \( \Pi^{k}_d \to 0 \) as \( k \to \infty \). Provided \( \delta \) has at least one non-zero element, convergence of
the series is guaranteed.

In the case of all-or-nothing assignment without consideration of common lines, the
system is convergent after \( n \) steps (i.e. all trips have reached their destination,
meaning \( \Pi^{k+1}_d = \theta \)), where \( n \) is the number of arcs on the shortest path from the
origin to \( d \). In the case of deterministic or stochastic assignment with consideration
of common lines as in this case, the system is still convergent but it is likely to
require more than \( n+1 \) steps because not only the shortest path is used.

The submatrix \( \Theta_d \) can further be divided into two submatrices and two zero vectors:
\[ \Theta_d = \begin{pmatrix} 0 & \Theta_{1d} \\ 0 & \Theta_{2d} \end{pmatrix} \begin{pmatrix} n_o \\ n - n_o - 1 \end{pmatrix} \]  

(4-21)

Two vectors of zero indicate that there are no transitions from one origin to another origin and that there are no transitions from an intermediate node to an origin. \( \Theta_{1d} \) is the matrix of transition probabilities from all origins to all intermediate nodes, and \( \Theta_{2d} \) is the matrix of transition probabilities from one intermediate node to another intermediate node. It follows that in (4-20) the inverse of \( \mathbf{I} - \Theta_d \) can be calculated as in (4-22):

\[
(\mathbf{I} - \Theta_d)^{-1} = \begin{pmatrix} \mathbf{I} & \Theta_{1d}[\mathbf{I} - \Theta_{2d}]^{-1} \\ 0 & [\mathbf{I} - \Theta_{2d}]^{-1} \end{pmatrix} \]  

(4-22)

Let \( y_d \) be the vector of traffic produced at origins \( o \) destined to \( d \). Then the vector of traffic traversing intermediate node \( i \), \( v_d \), can be obtained from (4-23). In this and following equations the dash indicates matrix transposition.

\[
\mathbf{v}_d = \mathbf{y}_d \cdot \Theta_{1d} \left( [\mathbf{I} - \Theta_{2d}]^{-1} \right) \]  

(4-23)

Finally, arc traffic volumes \( x_d \) are calculated using \( \mathbf{v}_d \)

\[
\mathbf{x} = \sum_d x_d = \sum_d (\Theta_{1d} \Theta_{2d}) \begin{pmatrix} y_d \\ v_d \end{pmatrix} \]  

(4-24)
4.5.2. Ensuring strict capacity constraints

The key point to this approach is that it considers strict capacity constraints for each line at each boarding point. Therefore for each line arc (4-25) must be satisfied.

\[ f_i \cdot \text{cap}_i \geq x_{\text{line}_i} \quad \forall u \in U_i, l \in L \]  

(4-25)

The flow on a line arc consists of those staying on board plus those wishing to board. Those staying on board have priority over those wishing to board. This means that the fail-to-board probability \( q \) in (4-26) needs to be adjusted so that (4-25) is satisfied.

\[ x_{\text{line}_i} = x_{\text{on-board}_i} + x_{\text{board}_i} = x_{\text{on-board}_i} + (1 - q_{ul})x_{\text{board}-d_i}, \quad \forall u \in U_i, l \in L \]  

(4-26)

The adjustment is done with

\[ q_{ul} = 1 - \max \left( \min \left( \frac{f_i \cdot \text{cap}_i - x_{\text{on-board}_i}}{x_{\text{board}_i}}, 0 \right) \right), \quad \forall u \in U_i, l \in L. \]  

(4-27)

Note that the vacancies \( sp^r \) on line \( l \) for a service leaving stopnode \( s_u \) are:

\[ sp^r_{s_u} = f_i \cdot \text{cap}_i - x_{\text{on-board}_i} - (1 - q_{v_i})x_{\text{board}_i}, \quad \forall u \in U_i, l \in L \]  

(4-28)

4.6. The network equilibrium

4.6.1. Multiple hyperpaths to a destination

As explained in the literature review Cominetti and Correa (2001) showed that for the effective frequency approach in fact a split between two hyperpaths can minimise the user costs. The same effect can also be found for the CapCon approach.
Consider the simplest network with one origin and one destination connected by two different lines. Let Line 1 have a smaller travel time but operate with lower frequency than Line 2. Clearly, if these are common lines, Strategy S2 “Consider Line 2 only” is not optimal. However, if the passengers are risk-averse ($\theta > 0$) then it can be found that the user cost is minimised if passengers split between Strategy S12: “Take Line 1 or 2 whichever comes first” and Strategy S1 “Only consider Line 1”.

In Figure 4-3 this is illustrated. Assume following line characteristics: $c_1 = 60\text{sec}$, $c_2 = 90\text{sec}$, $f_1 = 180\text{sec}$, $f_2 = 60\text{sec}$, $cap_1 = cap_2 = 50\text{pas/hour}$; and a demand of 150 passengers/hour with $\theta = 100$. If all passengers take Strategy S12 the costs for Strategy S12 (62.95) is lower than if all passengers take Strategy S1 (113.86). Therefore Strategy S12 is the strategy with the lowest cost of the three possible strategies {S1, S2, S12}. However, the user cost can be minimised if around 33% of the passengers choose S1 and 66% choose S12. In this case the cost is only 43.3 and the fail to board probabilities are 0.32 for Line 1 and 0.34 for Line 2 respectively.
In conclusion, to find the user equilibrium does in some circumstances require the consideration of multiple strategies. The interpretation of such an equilibrium is as follows: In the above case, it means that some passengers at the platform / bus stop limit themselves to a specific service (in the above case 33% of the passengers limit themselves to Line 1) even if the other service is arriving first. The passenger costs are reduced because these passengers will not be disappointed by failing to board Service 2 if this arrives earlier and is already full.

Figure 4-3 Example for a user equilibrium consisting of two hyperpaths
4.6.2. Characterisation of the network equilibrium

The network equilibrium can be found when the following two complementary slackness conditions apply:

Let $H^*$ be defined as the set of optimal hyperpaths to destination $d$. Firstly, the user equilibrium implies that for all destinations $H^*$ is empty or the cost difference $g'$ between the used hyperpaths $h$ (to destination $d$) and all other (unused) hyperpaths $h^* \in H$ is zero (Wardrop principle). This can be expressed with (4-29) where the cost difference $g'$ is defined as in (4-30) and $z_h^*$ is the flow on the hyperpath $h$ in the set of optimal hyperpaths $H^*$. $\min g_d$ is the cost of the minimum cost hyperpath in $H^*_d$.

$$z_h^* \cdot g_h^*(z^*, q^*) = 0, \ g_h^*(z^*, q^*) \geq 0, \ \forall h \in H_d, \ \forall d \in D \tag{4-29}$$

$$g_h^*(z^*, q^*) = g_h - \min g_d, \ \forall h \in H_d \tag{4-30}$$

Secondly, for each platform it must be true that either the fail-to-board-probability is zero or there are no spaces left when the vehicle is leaving the platform. This is expressed with the complementary slackness condition (4-31).

$$q^* \cdot sp^*(z^*, q^*) = 0, \ sp^*(z^*, q^*) \geq 0, \ \forall 0 \leq q \leq 1 \tag{4-31}$$

In summary, the assignment has to find $(z^*, q^*)$ such that (4-29) and (4-31) are fulfilled.

4.6.3. Existence of a unique network equilibrium

Finding the equilibrium $(z^*, q^*)$ is a fixed point problem. This is because the failure probability $q$ can be regarded as an endogenous variable, because the flow $z_h$ of a
hyperpath and the failure probability depend on each other: The hyperpath flow clearly depends on \( q \) because the failure probability directly influences the number boarding at each station. However, \( q \) depends on the line capacity and the available spaces at the station which depends on the flow. Hence, in turn \( q \) depends on the flow \( z \). The existence of a fixed point is intuitive since any excess demand simply implies non-zero failures to board. The uniqueness of the fixed point for non-circular lines is dealt with in the proof below.

**Result:** The fixed point for \( q \) is unique for non-circular lines.

**Proof:** Suppose not. Without loss of generality suppose there are two values for node \( j \), namely \( q_j \neq q'_j \). This would imply different numbers of available spaces at node \( j \), namely \( sp^a_j \neq sp^a'_j \). Suppose, also without loss of generality, that \( sp^a_j > sp^a'_j \). This implies that more could have boarded at node \( j-1 \), the upstream node to \( j \), if \( q_{j-1} > 0 \). This would imply a contradiction. Alternatively, if \( q_{j-1} = 0 \), then more could have boarded at \( j-2 \), which is upstream of arc \( j-1 \). This would imply a contradiction. This argument continues until either a contradiction is encountered or the start arc of the line is reached. If the start arc is reached without contradiction, and therefore without more being able to board at any arc upstream of \( j \), then \( sp'_j \) is not feasible, also implying a contradiction. Hence \( q_j \neq q'_j \) necessarily implies a contradiction.

### 4.6.4. A Note on Circular Lines

The above proof only holds for non-circular lines, as circular lines have no start arc. However, circular lines are common practice for several bus and metro networks. Below example in Table 4-1 shows a circular line consisting of four stops. The
capacity of the line is assumed to be 100 passengers and the demand is symmetric with 50 passengers from each station to each station. The example illustrates that there are several feasible fixed points that fulfil the constraint (4-31) for capacity constraint assignment.

Table 4-1 Different feasible solutions to a circular line assignment problem

(Example 1)

<table>
<thead>
<tr>
<th>Solution</th>
<th>j</th>
<th>spaces at dest=j+1</th>
<th>board at dest=j+2</th>
<th>dest=j+3</th>
<th>q(j)</th>
<th>fail to board</th>
</tr>
</thead>
<tbody>
<tr>
<td>a possible solution</td>
<td>A</td>
<td>50</td>
<td>20</td>
<td>10</td>
<td>0.67</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>50</td>
<td>10</td>
<td>20</td>
<td>0.67</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>40</td>
<td>10</td>
<td>20</td>
<td>0.73</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>50</td>
<td>10</td>
<td>20</td>
<td>0.67</td>
<td>100</td>
</tr>
<tr>
<td>priority to few-stop boarders</td>
<td>A</td>
<td>100</td>
<td>50</td>
<td>50</td>
<td>0</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>50</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>100</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>50</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0.67</td>
</tr>
<tr>
<td>priority to many-stop boarders</td>
<td>A</td>
<td>25</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>50</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>25</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>50</td>
<td>0</td>
<td>25</td>
<td>0</td>
<td>0.67</td>
</tr>
<tr>
<td>Mingling</td>
<td>A</td>
<td>50</td>
<td>16.67</td>
<td>16.67</td>
<td>16.67</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>50</td>
<td>16.67</td>
<td>16.67</td>
<td>16.67</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>50</td>
<td>16.67</td>
<td>16.67</td>
<td>16.67</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>50</td>
<td>16.67</td>
<td>16.67</td>
<td>16.67</td>
<td>0.67</td>
</tr>
</tbody>
</table>

A solution that encourages short distance usage of the line obviously increases the line performance in terms of minimisation of the total number of passengers failing to board. Solution “priority to few-stop boarders” is therefore the system-optimal one, whereas Solution “priority to many-stop boarders” lets most passengers fail. Various other solutions are possible as also shown in above table.

However, there is only one feasible solution if one assumes passenger mingling, as is done in the network loading procedure described above. The first three solutions in above table assume that to some destinations more passengers are able to board than to others, even though it is assumed that the demand to each destination is equal. For the four node network following constraints must hold:
At Node A: \( \text{cap} = (b_{AB} + b_{AC} + b_{AD}) + (b_{DB} + b_{DC} + b_{CB}) \)

At Node B: \( \text{cap} = (b_{BC} + b_{BD} + b_{BA}) + (b_{AC} + b_{AD} + b_{DC}) \)

... 

If we define \( b_{ij} \) as the passengers getting on-board at \( i \) and alighting at \( j \) in general following constraints must hold:

\[
\text{cap} = \left( b_{i1,j1} + b_{i2,j2} + \ldots + b_{ij} \right) + \left( b_{i1,j1} + b_{i1,j2} + \ldots + b_{i2,j1} + b_{i2,j2} + \ldots + b_{ij} \right) \forall i \in I, \quad (4-32)
\]

Therefore, for a circle with \( n \) nodes there are \( n^*(n-1) \) variables for \( n \) constraints which explains why multiple solutions are likely. However, in the case of mingling further following constraints must hold:

\[
\frac{y_{ij}}{b_{ij}} = \frac{y_{ik}}{b_{ik}} \Rightarrow b_{ij} = R_{ij,ik} b_{ik} \quad \forall i, j, k \in I, i \neq j \forall i \in I, \quad (4-33)
\]

where \( R_{ij,ik} \) is the boarding-demand from station \( i \) to \( j \) divided by the boarding-demand from station \( i \) to \( k \).

Therefore for the 4 node network we have following four constraints for four variables:

At Node A: \( \text{cap} = b_{AB} \left[ \frac{1}{R_{AB,AC}} + \frac{1}{R_{AB,AD}} \right] + \left( b_{BC} \times 0 \right) + b_{CD} \left[ \frac{1}{R_{CD,CA}} \right] + b_{DA} \left[ \frac{1}{R_{DA,DB}} + \frac{1}{R_{DA,DC}} \right] \)

At Node B: \( \text{cap} = b_{AB} \left[ \frac{1}{R_{AB,AC}} + \frac{1}{R_{AB,AD}} \right] + b_{BC} \left[ \frac{1}{R_{BC,BD}} + \frac{1}{R_{BC,BA}} \right] + b_{CD} \times 0 + b_{DA} \left[ \frac{1}{R_{DA,DC}} \right] \)

... 

In Matrix form this can be written as:
(4.34) \[
\begin{pmatrix}
  a_{11} & 0 & a_{13} & a_{14} \\
  a_{21} & a_{22} & 0 & a_{24} \\
  a_{31} & a_{32} & a_{33} & 0 \\
  0 & a_{42} & a_{43} & a_{44}
\end{pmatrix}
\begin{pmatrix}
  b_{AB} \\
  b_{BC} \\
  b_{CD} \\
  b_{DA}
\end{pmatrix}
= \begin{pmatrix}
  \text{cap} \\
  \text{cap} \\
  \text{cap} \\
  \text{cap}
\end{pmatrix}
\]

with \( \alpha_{11} := \left[1 + \frac{1}{R_{AB,AC}} + \frac{1}{R_{AB,AD}} \right] \); \( \alpha_{13} := \left[1 + \frac{1}{R_{CD,CA}} \right] \); \( \alpha_{14} := \left[1 + \frac{1}{R_{DA,DR}} + \frac{1}{R_{DA,DC}} \right] \)

There will be a unique solution if the above matrix \( \mathbf{A} \) with elements \( \alpha_{ij} \) is non-singular. The matrix depends on the demand distribution from each node and non-singularity is not necessarily the case. For example Table 4-2 shows an example where the above matrix is not unique even if mingling is assumed.

**Table 4-2 Different feasible solutions to a circular line assignment problem**

(Example 2)

<table>
<thead>
<tr>
<th>Solution</th>
<th>j</th>
<th>spaces at(j) dest=(j+1)</th>
<th>board at (j) dest=(j+1) dest=(j+2)</th>
<th>(q(j))</th>
<th>fail to board</th>
</tr>
</thead>
<tbody>
<tr>
<td>a possible solution</td>
<td>A</td>
<td>65</td>
<td>0</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>35</td>
<td>0</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>65</td>
<td>0</td>
<td>65</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>35</td>
<td>0</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>another possible solution</td>
<td>A</td>
<td>55</td>
<td>0</td>
<td>55</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>45</td>
<td>0</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>55</td>
<td>0</td>
<td>55</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>45</td>
<td>0</td>
<td>45</td>
<td>0</td>
</tr>
</tbody>
</table>

However, it needs to be emphasised that these cases are very rare:

Firstly, if at only station of the circular line the fail-to-board probability equals zero there is a unique solution. In this case the proof for non-circular lines can be applied, as the station with \(q_i = 0\) can be assumed to be the line terminal. It is difficult to find networks where each station of a circular is overcrowded.
Secondly, even if all stations are overcrowded, an analysis of the determinant of the matrix shows that there is a tendency for it being non-zero. The determinant for the 4 node circle is:

\[
\text{Det}A = \\
+ \alpha_{11} \alpha_{22} \alpha_{33} \alpha_{44} + \alpha_{11} \alpha_{24} \alpha_{32} \alpha_{43} - \alpha_{11} \alpha_{24} \alpha_{33} \alpha_{42} \\
+ \alpha_{13} \alpha_{24} \alpha_{32} \alpha_{44} + \alpha_{13} \alpha_{24} \alpha_{31} \alpha_{42} - \alpha_{13} \alpha_{24} \alpha_{31} \alpha_{44} \\
+ \alpha_{14} \alpha_{21} \alpha_{33} \alpha_{42} + \alpha_{14} \alpha_{22} \alpha_{31} \alpha_{43} - \alpha_{14} \alpha_{21} \alpha_{32} \alpha_{43}
\]

If the demand is equally spread between the stations, there are twice as many positive terms as negative terms. Further, the larger the circle the more \( \alpha_{ij} \) will be non-zero, meaning that \( \text{Det} (A)=0 \) becomes less likely.

Thirdly, it might be a reasonable assumption to assume that often \( y_{ij} > y_{ik} \) if \( k \) is much further away than \( j \) (often there are quicker arcs than riding the circular line nearly for a full circle or the circle might be operated in both directions). In that case, we expect for the first column of the 4 node network:

\[
\alpha_{11} = 1 + \frac{b_{ac}}{b_{ab}} + \frac{b_{ad}}{b_{ab}} = 1 + \text{“a value smaller 1” + “a very small value”}
\]

\( \alpha_{12} = 0; \)

\[
\alpha_{13} = \frac{b_{be}}{b_{cd}} = \text{“a very small value”}
\]

\[
\alpha_{14} = \frac{b_{DB}}{b_{DA}} + \frac{b_{DC}}{b_{DA}} = \text{“a value smaller 1” + “very small value”}
\]

In general we can expect for the four node network: \( \alpha_{ij} > \alpha_{i,j-1} > \alpha_{i,i-2} > \alpha_{i,j+1} = 0 \)
And for a network with \( n \) nodes: \( \alpha_{ii} > 1; \alpha_{ii} > \alpha_{i,j-1} > \alpha_{i,j-2} > \ldots > \alpha_{i,i+1} = 0. \) If this assumption holds for all stations, the Levy-Desplanques theorem holds which states that strict diagonal dominance guarantees non-singularity:

\[
\text{If } \sum_{j=1}^{n} a_{ij} < a_{ii}; \forall i \in I \quad \Rightarrow \quad \text{A is non-singular} \quad (4-35)
\]

In conclusion, for circular lines there is a possibility for multiple solutions but a unique solution is much more likely under the assumption of passenger mingling.

### 4.7. Method of Successive Averages (MSA)

#### 4.7.1. Introduction

A general procedure to solve fixed point problems is the Method of Successive Averages (MSA). The MSA has been widely used and found to be successful for solving network equilibrium assignment problems. Bell and Cassir (2002) used it for example to determine the arc fail probabilities in the vulnerability analysis of road networks which resembles to some degree the problem described here. It is further worth noting that Capeda et al (2006) successfully use the MSA to determine the equilibrium in large scale congested transit networks with strict capacity constraints. The following iterations, if convergent, lead to \( q^* \).
4.7.2. MSA algorithm

Step 1 (Initialisation for MSA iterations)

Set \( n := 0; \)

Set \( q_i^{(0)} = 0, \forall i \in E; \)

Step 2 (Minimum Cost Hyperpath Search)

Calculate \( \pi_{ab} \) with the “Minimum Cost Hyperpath Search” algorithm for given \( q^{(n)} \) as in Section 4.6 and \((I-\Pi_d)^{-1}\) as in Equations (4-18) to (4-22)

Step 3 (Markov Chain Flow Assignment)

Calculate \( v^{(n)} \) with (4-23)

Calculate \( x^{(n)} \) with (4-24)

Step 4 (Correction of \( q^{(n)} \) and \( x^{(n)} \)).

Set \( n \leftarrow n+1 \)

Adjust \( q^{(n)} \) with (4-27)

\[
x_a^{(n)} \leftarrow \left(1/n\right)x_a^{(n)} + \left(1-1/n\right)x_a^{(n-1)}; \forall a \in A
\]

Step 5 (Iteration, Termination of MSA loop)

Return to Step 2 until \( q^{(n)} \) and \( x^{(n)} \) cease to change significantly.

In the above algorithm it should be noted that \( x \) ceases to change if \( q \) ceases to change. This is because the arc costs in each hyperpath search iteration are changing only if \( q \) changes. Therefore if \( q \) ceases to change, the hyperpath search will result in the same arc split probabilities and hence the same node and arc flows.

4.7.3. Convergence of the MSA

The number of iterations needed until convergence of the MSA depends significantly on the network design, the demand, the risk-averseness \( \theta_e \) and the tightness of the
convergence criteria. The convergence criteria in Step 5 of the MSA can be expressed as:

\[ |x^{n+1}_a - x^n_a| < \epsilon \quad \forall a \in A \quad (4-36) \]

which is equivalent to \( |q^{n+1}_i - q^*_i| < \epsilon \quad \forall i \in I \) where \( \epsilon \) is to be defined. The smaller \( \epsilon \) the more iterations are carried out with the MSA. For practical purposes \( \epsilon = 0.1 \) seems to be sufficient according to experiences with the London network. For illustration purposes an even tighter convergence criteria is used in some of the following case studies.

If the network is not congested the algorithm will terminate immediately. If \( \theta_e = 0 \) the algorithm will in most cases also terminate rapidly, as convergence is immediately reached once a solution is found that observes the capacity constraint (4-25). If \( \theta_e > 0 \) often a large number of iterations are needed as convergence can be slow. Below illustrates the convergence of the MSA algorithm for the simplest network with one origin and one destination connected by two lines. The parameters are the same as in Figure 4-3 so that the user equilibrium is found with a combination between the two strategies S12 (“Take whichever service arrives first”) and S1 (“Consider only Line 1”). The figure shows that the MSA converges after around 40 iterations. After this point \( q_1 = 0.32 \) and \( q_2 = 0.34 \) and in two out of three iterations the MSA finds strategy S12 to be the least cost strategy and in 1 out of 3 iterations S1 is the least cost strategy. This corresponds to the user equilibrium shown in Figure 4-3.
4.8. **Assignment without consideration of common lines**

4.8.1. Introduction

As described in Chapter 2.2 transit assignment without the consideration of common lines is interesting in its own right because there are several applications where the problem does not apply. In this case one does not need to search for the optimal hyperpath but the search for the shortest path is sufficient. The problem becomes equivalent to passengers choosing exactly one boarding arc (or a walking arc) from each stop node. The above network notation and hyperpath search can still be used by doing one of the following two changes:

- The network is defined in such a way that all lines are platform specific. In this case the passenger always needs to walk in order to reach different lines, so that the common line problem does not apply.
• In Step 3 of the hyperpath search algorithm in Section 4.6 the number of active outgoing arcs from a stop node is restricted to 1:

\[
\text{if } \text{Out}(a^*) \in S \text{ and } \text{In}(a^*) \in E \text{ then }
\]

\[
\text{if } g^{\text{Out}(a^*)} \geq \frac{1}{f_a} + g^{\text{In}(a^*)} \text{ then }
\]

\[
g^{\text{Out}(a^*)} := \frac{1}{f_a} + g^{\text{In}(a^*)}
\]

\[
\text{else }
\]

\[
\text{Set } M_2 := M_2 - \{a^*\}, M_3 := M_3 + \{a^*\};
\]

One can see that the search for the shortest path simplifies the problem significantly. Therefore, alternatively, the capacity constrained transit assignment problem can also be solved with a very simplified network representation. In this case the network can be described without the necessity to introduce fail nodes and fail arcs. As will be described in Chapter 7, the number of nodes influences the network performance greatly, so that this network description allows a significant reduction in calculation time. Further boarding nodes and alighting nodes do not need to be introduced if the stop nodes are assumed to be line specific. The walking arcs to the platforms can reflect the expected waiting time to board a line.

### 4.8.2. Calculation of Transition Probabilities in simplified network

The approach with the simplified network requires the introduction of a variable \( \text{line}_{ij} \) which is defined as

\[
\text{line}_{ij} = \begin{cases} 
1 & \text{if } j \text{ is the downstream node of } i \text{ on the same line} \\
0 & \text{otherwise} 
\end{cases}
\]
The calculation of matrix $\Pi_d$ does not require the hyperpath search anymore but can be carried out as in (4-38) if deterministic user equilibrium assignment is assumed as throughout this chapter. The calculation of the least cost path can be done with a standard shortest path algorithm, e.g. Dijkstra or Floyd.

\[
\pi_{ij} = \begin{cases} 
line_{ij} + (1 - \line_{ij})(1 - q_{ij}) & \text{if } j \text{ follows } i \neq j \text{ on the least cost path from } i \text{ to } d \\
0 & \text{otherwise}
\end{cases} 
\]

(4-38)

The inversion of the Matrix $\Pi_d$ and calculation of node and arc volumes can then be carried out in the same way as described for the case with consideration of common lines. So far, the node fail probabilities have not been calculated. This can be done with the Correction Algorithm below.

### 4.8.3. Calculation of fail-to-board probabilities (Correction Algorithm)

The boarding demand and the available spaces at a node can be determined with (4-39) and (4-40). In (4-39) the boarding demand is determined through the node volume arriving through transition from any other neighbouring node which is not served by the same transit line. Similarly, the available spaces are determined through the subtraction of all those passengers who are already on-board before node $j$ and who alight after node $j$.

\[
\text{board-} d_j = \sum_d \sum_i \text{\textit{v}}_{ijd} \ \pi_{ijd} (1 - \line_{ij}), \ \forall i,j \in D, \forall d \in D
\]

(4-39)

\[
\text{sp}^a_j = \text{Max} \ \{0, \text{\textit{cap}}_j - \sum_d \sum_i \sum_m \text{\textit{v}}_{ijd} \ \line_{ij} \ \pi_{ijd} \ \pi_{jmjd} \ \line_{jm} \}, \\
\forall i,j,m \in D, \forall d \in D
\]

(4-40)
board-\(d_i \leq sp^a_i \leq cap_i, \forall i \in I\) \hfill (4-41)

The boarding demand can then be adjusted so that (4-41) is observed through the following algorithm.

**Step 1** (Initialisation): \(q = 0\)

**Step 2** (Calculate \(\Pi_d\) for all \(d\)): Apply Shortest Path algorithm

**Step 3** (Calculate node and arc volumes): Network loading with Markov Chain Flow Assignment

**Step 4** (Calculate board-\(d\) and \(sp^a\)): Solve (4-38) and (4-39)

**Step 5** (Update \(q\) for each platform):

\[
\begin{align*}
\text{If } board-d_i > 0 & \quad \text{then } q_i \leftarrow \min \{1, \max \{0, 1 - (1 - q_i) sp^a_i / v_i\}\} \\
\text{else if } y^d_i > 0 & \quad \text{then } q_i \leftarrow 0 \\
\text{else } q_i & \leftarrow 1
\end{align*}
\]

**Step 6** (Iteration, Termination): Return to Step 2 until \(q\) ceases to change

The rationale for the Correction Algorithm is the same as for the calculation of \(q\) in (4-27): Firstly, the algorithm ensures that if more passengers could board at a given node then more would, and secondly, the OD-composition of those failing to board is the same as that for those boarding (passenger mingling). Note that Step 4 implies that after termination if \(sp^a_j > board-d_j\), then \(q_j = 0\). In words, if more could board at node \(j\) then more would, which is the same characteristic as expressed in (4-31). Experience suggests that the algorithm terminates rapidly (in the case of the example in Chapter 4.9 after three iterations).
4.8.4. Risk averse assignment

If risk-averse users are assumed, the cost function can be assumed similar to (4-10). In this case the hyperpath consist of a single path. Therefore, if one assumes, deterministic assignment the parameter $\alpha_{ap}$ is 1 if arc $a$ is on the shortest path $p$ to destination $d$ and 0 otherwise. The second term in the cost function (4-10) is the waiting time at boarding nodes. Because in this network description the waiting time for a line can be expressed through the walking arcs to a node, the second term is not needed anymore but the costs are also expressed through the first term in (4-42). (Consequently $\beta_{ip}$ is also not anymore needed.)

$$g_p = \sum_{a \in I_p} \alpha_{ap} c_a - \theta_v \ln \left( \prod_{i \in I_p} (1 - q_i)^{\alpha_a} \right)$$ (4-42)

With this cost formulation and the correction algorithm described in the previous section, the risk-averse assignment with the MSA as described in Section 4.7 can also be used for this network description. Step 2 in the MSA (Hyperpath search) should only be replaced with a shortest path algorithm (e.g. Dijkstra) and Step 3 with the Correction Algorithm as described in Section 4.8.3.
4.9. Case Study on a small example network

4.9.1. Network description

The effects of risk-averse passenger behaviour and the effects of the common line problem can best be illustrated with a small example network. The following network is used:

Line 1
Headway: 6 minutes
Capacity: 10 pas/veh

Line 2
Headway: 12 minutes
Capacity: 20 pas/veh

Figure 4-5 Example network with 2 transit lines

With common lines the network is presented with the description of a station as explained in Section 4.2 and illustrated in Figure 4-1 so that in total following arcs are needed to present the network (Figure 4-6):
If one ignores the common line problem the same network can be presented in a much simpler way as shown in Figure 4-7.
Let us assume a demand of 100 passengers per hour from each station to each station downstream, meaning that 300 passengers originate from Station A, 200 from Station B and 100 from Station C. Therefore at Station A and B capacity problems are guaranteed as there is only a total line capacity of 200 passengers from A, and 100 passengers at B. Because of passengers staying on-board at C and because of the priority rules the following results show that also passengers at C will encounter capacity problems.

The MSA termination criterion is set at \( eps = 0.01 \), meaning that the algorithm will only terminate if the averaged flows do not change by more than 0.01 passengers on any line between two iterations. Table 4-3 shows the results for different levels of
risk averseness $\theta_e$ and the results with and without consideration of common lines.

In all cases the capacity constraints are fully met. The number of passengers boarding never exceeds the available capacity at any line and platform. The number of iterations needed increases significantly with the introduction of risk-averseness ($\theta_e > 0$ compared to $\theta_e = 0$). In general one can observe that the number of iterations increases with further increase in $\theta_e$. For large scale case studies, a less strict setting of the MSA termination criterion is probably sufficient as this reduces the number of iterations and hence run time significantly and the changes in line flows are minimal.

In this network for example setting the termination criterion to 0.1 instead reduces the iterations until convergence from 140 to 38 (for $\theta_e = 0$, with common lines).

Table 4-3 Results for different $\theta$ and with/without consideration of common lines

<table>
<thead>
<tr>
<th>Station</th>
<th>Line</th>
<th>$q$</th>
<th>Board Demand</th>
<th>Alight</th>
<th>Av.Cap</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0.40</td>
<td>166.7</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>0.25</td>
<td>133.3</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0.70</td>
<td>60.0</td>
<td>60</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>66.7</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>0.25</td>
<td>33.3</td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station</th>
<th>Line</th>
<th>$q$</th>
<th>Board Demand</th>
<th>Alight</th>
<th>Av.Cap</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0.36</td>
<td>157.5</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>0.30</td>
<td>142.5</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0.68</td>
<td>63.4</td>
<td>63.6</td>
<td>63.6</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>66.6</td>
<td>68.3</td>
<td>68.3</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>0.1</td>
<td>29.9</td>
<td>29.9</td>
<td>29.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station</th>
<th>Line</th>
<th>$q$</th>
<th>Board Demand</th>
<th>Alight</th>
<th>Av.Cap</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0.35</td>
<td>153.7</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>0.32</td>
<td>146.3</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0.68</td>
<td>65.0</td>
<td>65</td>
<td>65</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>66.6</td>
<td>67.6</td>
<td>67.6</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>0.05</td>
<td>31.7</td>
<td>31.7</td>
<td>31.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station</th>
<th>Line</th>
<th>$q$</th>
<th>Board Demand</th>
<th>Alight</th>
<th>Av.Cap</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0.40</td>
<td>150.7</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>0.33</td>
<td>149.3</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0.67</td>
<td>66.2</td>
<td>66.4</td>
<td>66.4</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>66.5</td>
<td>66.7</td>
<td>66.9</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>0.01</td>
<td>33.2</td>
<td>33.3</td>
<td>33.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station</th>
<th>Line</th>
<th>$q$</th>
<th>Board Demand</th>
<th>Alight</th>
<th>Av.Cap</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0.34</td>
<td>150.7</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>0.33</td>
<td>149.3</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0.67</td>
<td>66.2</td>
<td>66.4</td>
<td>66.4</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>66.5</td>
<td>66.7</td>
<td>66.9</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>0.01</td>
<td>33.2</td>
<td>33.3</td>
<td>33.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station</th>
<th>Line</th>
<th>$q$</th>
<th>Board Demand</th>
<th>Alight</th>
<th>Av.Cap</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0.34</td>
<td>150.7</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>0.33</td>
<td>149.3</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0.67</td>
<td>66.2</td>
<td>66.4</td>
<td>66.4</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0</td>
<td>66.5</td>
<td>66.7</td>
<td>66.9</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>0.01</td>
<td>33.2</td>
<td>33.3</td>
<td>33.3</td>
</tr>
</tbody>
</table>
4.9.2. Effect of Risk-averseness

If passengers are not risk averse ($\theta_e = 0$) the fail-to-board probability $q$ has no influence on the perceived cost and all passengers stick to the shortest route found according to in-vehicle travel time and waiting time. When common lines are not considered all passengers at Station A destined for C and D will attempt to board Line 2 and all passengers destined for B will attempt to board Line 1. Therefore $q_{L2}$ is much higher than $q_{L1}$ at Station A. If passengers are more risk-averse the above table shows that $q_{L1}$ and $q_{L2}$ become more similar because passengers will spread out over both lines and avoid the heavily congested lines. The effect can be observed whether the common lines are considered or not.

In this network the risk-averseness has further the effect that the lines are better utilised as illustrated in Figure 4-8. The number of unutilised line spaces is calculated as the sum of the remaining spaces $y^r$ for all platforms and all lines. This does however not necessarily have the effect that the total number of passenger decreases as also shown in the figure below. In case common lines are not considered the total number of passengers failing to board is in fact lowest for $\theta_e = 10$. If $\theta_e$ further increases more passengers fail to board at Station B which outweighs the small decrease in passengers failing to board at C. This is because with higher risk-averseness more passengers at A are encouraged to take the longer Line 1 (via B) and therefore the crowding at B becomes worse.
4.9.3. Effect of common lines

Comparing the results for $\theta = 0$ with and without consideration of common lines (CL) the consideration of two paths clearly reduces the fail-to-board probability at Station A. However, the effect for passengers at Station B is negative because with CL also passengers destined for C and D will board Line 1, meaning that the available spaces at B are reduced. This effect becomes even more pronounced for $\theta = 10$ where the total number of passengers failing to board increases. In general, as one might expect, one can conclude that a design of stations and fare systems that allows the consideration of common lines will help passengers at stations with a choice between lines but the effect for passengers at smaller stations might in fact be an adverse one.

Further, Figure 4-8 shows that the consideration of common lines has a positive effect on line space utilisation. Even if $\theta = 0$ the common lines guarantee that all
lines are nearly fully utilised. The example further shows that the higher the risk-averseness is the similar the results become for consideration with or without common lines. This is also what one might expect because (in congested situations) risk-averseness as well as the possibility for passengers to consider common lines, tends to spread the line usage over all available lines. The interpretation for this is that risk-averse passengers will also consider walking between lines, which means that the fact whether there are common lines or not becomes secondary.

4.10. Discussion

4.10.1. Summary
This chapter introduced a new approach to frequency-based transit assignment that considers strict capacity constraints and allows for the consideration of common lines. Alternatively the chapter also presented a very similar approach that does not consider common lines but allows a much simpler network description. The larger case study in Chapter 8 will show that the multiplicity of nodes can lead to run-time issues which can be overcome if one uses the simple network description.

Central to the approach is the introduction of the fail-to-board probability $q$. In case one considers common lines this requires the introduction of a line-specific fail node. Any demand that exceeds the available capacity of a service at this platform is assigned to the virtual “bin” from this fail node.

The generalised cost function can be described node specifically for the fail nodes and all other nodes so that a hyperpath search similar to the one presented in Nguyen
and Pallotino (1988) can be applied. The hyperpath search presented here finds the set of paths that minimises the travel cost including the cost associated with the risk of failing to board.

The network loading procedure with the transition probabilities found through the hyperpath search then conforms to the Markov property: At every node the path choice is independent of the previous path. Therefore the inversion of the matrix containing the transition probabilities can then be used to obtain the node and arc volumes.

It is shown that there is a fixed point solution to the assignment even if the network contains circular lines, because the approach assumes passenger mingling. Finding the fixed point and user equilibrium might require a split between two hyperpaths which was also found by Cominetti and Correa (2001). This is observed in the current approach through network loading with the Method of Successive Averages (MSA) which finds the equilibrium by averaging between (possibly different) hyperpaths found in several iterations.

### 4.10.2. Distribution between Attractive Lines in Congested Situations

A limitation of this approach in the case of common lines might be that it allows passengers to fail to board even if one of their attractive lines is not congested. Let’s assume that passengers at a stop node include Line 1 with $f_1$ and Line 2 with $f_2$ into their route choice. Further if we assume that $q_1 = 0$ and $q_2 > 0$ then the above approach assumes that $f_2 / (f_1 + f_2) * q_2$ passengers fail to board. The example below of
a network with two lines connecting one OD pair where both lines are equally attractive except that Line 1 has a lower capacity illustrates this point.

\[
\begin{align*}
& f_{L1} = f_{L2} ; \quad c_{L1} = c_{L2} \\
& cap_{L1}: 20 \text{ pas} \\
& cap_{L2}: 100\text{pas} \\
& OD \text{ demand: 80 pas}
\end{align*}
\]

Figure 4-9 Two line network with different capacities of the lines

There is sufficient total capacity for all passengers, however if we assume a risk-averseness of \( \theta = 0 \) the lines will be equally attractive and 20 passengers will fail to board even though there are still empty spaces on Line 2 which is also included in all passengers’ choice set:

\[
\begin{align*}
board_{dL1} &= 40\text{pas} \quad \Rightarrow q_{L1} = 0.5; \quad board_{L1} = 20\text{pas} \\
board_{dL2} &= 40\text{pas} \quad \Rightarrow q_{L2} = 0; \quad board_{L2} = 40\text{pas}
\end{align*}
\]

It should be noted that this is only a problem if one assumes a low level of risk-averseness. The higher the risk-averseness the more the capacity of the lines will determine their attractiveness. With higher risk-averseness passengers will split between two strategies with some passengers using Strategy “Take Line 1 or 2 whichever comes first” and some using strategy “Only consider Line 2”. This will lead to more passengers using Line 2 and less passengers failing to board Figure 4-10).
Another solution to this problem is a change of the path choice probability $\pi_{ah}$ in (4-3) to the effective frequency as introduced by De Cea and Fernandez (1993). One could utilise the fail-to-board probability to determine an effective frequency that guarantees that no passengers are failing to board in case there is sufficient total capacity:

$$\tilde{f}_a(x_a) = f_a \left[1 - q_{Out(a)}\right], \forall a \in A_a \mid ln(a) \in B$$

(4-43)

where $\beta$ is a calibration parameter explaining how sensitive the line frequency is to overcrowding. Cepeda et al (2006) for example use $\beta=0.2$ for their similar definition of effective frequency. The effective frequency $\tilde{f}_a$ will then be used as in (4-3) to determine $\pi_{ah}$:

$$\pi_{ah} = \tilde{f}_a \sum_{a_i \in A_a} \tilde{f}_a, \forall i \in S_h$$

(4-44)
Figure 4-11 Effect of Introducing Effective Frequencies

Figure 4-11 shows that introducing a high enough beta will fill up all the available line spaces. Therefore, introducing effective frequencies might lead to more realistic results if one assumes a low level of passenger risk-averseness. However, the figure also shows that in contrast to Figure 4-10 the flow on Line 1 is reduced with increasing $\beta$. One would however expect that all spaces on Line 1 are filled as both lines arrive with the same frequency and only those passengers who can not board Line 1 will now attempt boarding Line 2.

As shown in Cominetti and Correa (2001) the effective frequency approach also requires in some demand areas a split between two strategies. For the MSA convergence this means a larger number of iterations are needed similar to the effect of using a higher level of risk-averseness (see Section 4.7.3). For large $\beta$ however, the required iterations are reduced. This is because the increased path choice
probability for the less congested Line 2 means that all passengers can use Strategy “Take Line 1 or 2 whichever comes first”.

4.10.3. Further Assumptions made in the CapCon approach

Further to the issue of distribution between attractive lines discussed in the previous section, the following model assumptions also have some limitations:

Firstly, passenger mingling is assumed. This means that passengers with different destinations have the same chance of boarding the service. In reality this might however not be true. Especially on long platforms passengers to different destination might gather at different parts of the platform which are closer to their respective station exit when alighting the service. Therefore the chance of failing to board might be to some degree destination specific.

Secondly, deterministic user equilibrium is assumed: It is assumed that passengers perceive the crowding the same and also have the same level of risk-averseness. In reality this is however not likely to be the case. A distinction between commuters and occasional users such as tourists might be useful. The difference might be two-fold: Commuters might be less risk-averse, because travel time is likely to outweigh the discomfort caused by travelling on an overcrowded service. Further, commuters might perceive the available capacity to be higher than tourists, since tourists might abstain from boarding at lower line loading than commuters.
Thirdly, this chapter did not discuss the possibility of walking between stations which might be a choice of passengers in order to avoid overcrowding. Similar to transfer arcs between platforms transfer arcs between neighbouring stations could however be introduced in this model. These arcs are not of interest if one wants to replicate current line loadings with OD data from ticket gates, it might however be of interest to allow walking between stations in order to estimate the impact of changes to the service attributes.

Finally, some other possible limitations are discussed in the following chapters, in particular the treatment of excess demand and the formulation of the generalised cost function.
5. Dynamic Network Loading

5.1. Introduction

Even in highly congested transit networks, periods of extreme overcrowding often last for much less than the full peak period. In London for example London Underground operates their am peak service from 7am to 10am and the am peak is roughly thought to be between 8.00 and 9.30am. However, at Victoria station usually only between 8.30 and 9.00 is the congestion so severe that some passengers do not get onto trains because of overcrowding.

The CapCon approach presented in the previous chapter allows an explicit estimation of the number of passengers not being able to board which is the basis for the dynamic model described here. This chapter extends the approach by considering time dependent fail-to-board probabilities. The modelling period is divided into several time intervals and it is assumed that passengers failing to board in one time interval remain on the platform and therefore they have to be added to the demand arriving in the following time interval. Figure 5-1 is a simple illustration of this with a typical profile of a morning peak demand: During the first two time intervals the service can cope with the demand through an increase in service capacity from off-peak to peak service. (The capacity increase could be through higher frequency or through larger vehicle). A further increase in the demand at the peak of the peak exceeds however the service capacity. Therefore passengers $mq_3$ fail to board and will attempt to board the service in the next time interval. The new demand in time
interval 4 is lower than that of time interval 3 however because of the additional
demand from the previous time interval the congestion might actually be worse
shortly after the peak demand enters the stations. In the fifth time interval the
demand is so far reduced that the service can cope with the new demand plus the left-
over passengers from the previous time interval.

Note that in the following it will not only be assumed that passengers with different
destinations mingle as explained in the previous chapter but it will also be assumed
that passengers arriving in different time intervals mingle. This means that in time
interval \( n \) the excess demand of interval \( n-1 \) has no priority over the passengers
starting their journey in \( n \).

![Diagram of the dynamic approach](image-url)

**Figure 5-1** Simple illustration of the dynamic approach

The static approach presented in the previous chapter assumed that all passengers
travelling via the same arcs are competing for space. In this extension it is further
considered that some journeys can not be completed within a single time interval and
hence passengers traverse not all nodes of their path within one time interval. It is
assumed that long trips are continued from the last node they reached in the previous
time interval. How many trips are unfinished and how far passengers can travel in one time interval is obviously dependent on the length of a time interval chosen.

In summary the demand for each time interval is made up of three groups

a) Passengers starting their journey in this time-interval

b) Passengers who failed-to-board in the previous time-interval

c) Passengers with long journeys that can not be completed in one time interval

This chapter firstly explains the changes in the network notation and the additional notation required for the dynamic approach. Section 5.3 is then the centre of this chapter as it explains how the excess demand and how long journeys are reassigned in the subsequent time interval. Combining the common lines problem and dynamic effects brings the difficulty that passengers might be at different nodes at the end of the time-interval only depending on which vehicle arrived first. Therefore Section 5.4 describes a solution where this problem is solved by firstly calculating the travel times between all nodes for a previously determined hyperpath. The adjustments to the static MSA algorithm presented in the previous chapter are discussed in Section 5.5. Introducing time intervals further allows the calculation of expected delays and the formulation of a generalised cost function that considers these expected delays (Section 5.6). The approach is then illustrated in Section 5.7 with a case study on the same small network that is used in the previous chapter. Finally, the assumptions made in the dynamic model are discussed in Section 5.8.

Publications related to the dynamic model presented in this chapter can be found in the Smeed price winning paper Schmöcker (2006) and in Schmöcker et al (2006).
5.2. **Network description and its notation**

5.2.1. **Network representation**

The dynamic extension of the model is based on the idea that the bin is only a temporary state. It is assumed that passengers who fail-to-board within time interval $T_I$ will mingle with the passengers starting their trip at the platform within the next interval $T_I+1$. Figure 5-2 illustrates that trips that failed to board are re-assigned to the previous stop node in the following time interval. Therefore the excess demand node in Figure 4-1 is not required any more. Figure 5-3 then illustrates the nodes and arcs needed for the description of a station with multiple lines and consideration of dynamic effects.

![Diagram](image)

**Figure 5-2** Re-assigning demand that failed to board
5.2.2. Additional notation

In addition to the notation presented in the previous chapter following notation is required for the description of the dynamic extension:

- **$TI$:** Time interval
- **$TID$:** Duration of a time interval
- **$mq_{di}$**: Number of passengers who failed to board at node $i$ and are destined for $d$
- **$mu_{di}$**: Number of passengers who could not finish their trip and ended their journey at node $i$ destined for $d$
- **$TT_d$:** Travel time matrix (for a given $H_d$) indicating travel time between nodes $i$ and $j$ for passengers travelling to $d$ (with $tt_{ij} \in TT_d$)
- **$\Delta_d$:** Binary matrix indicating whether node $j$ can be reached from node $i$ within one time interval for passengers travelling to $d$ (with $\delta_{ij} \in \Delta_d$)
**K^t_{d}:** Binary matrix indicating whether arc \((i,j)\) is a final node for passengers starting their journey in this time interval from node \(r\) and travelling to \(d\) (with \(\kappa^t_{ij} \in K_d\))

**\(\theta_d:\)** Sensitivity to delays through failure to board

**\(\theta_I:\)** User expectation of overcrowding in subsequent time intervals

In the dynamic formulation of the assignment the flows and passenger path choice is changing in the time intervals \(TI\). Therefore all of the flow dependent variables introduced in the previous chapter are now time dependent. However, because in the model formulation the variables always refer to the current time interval \(TI\) (unless indicated), for simplicity a superscript \(TI\) is not added and the variables are not introduced. An exception is the demand vector \(y_d\). The vector is re-defined in this chapter as below. Further note, that in the dynamic model \(y_{d}^{TI}\) is not only limited to origin nodes.

\[ y_{d}^{TI} : \text{Vector of demand from nodes } i \text{ to destination } d \text{ (during time interval } TI) \]

### 5.3. Dynamic Network Loading

#### 5.3.1. Dealing with excess boarding demand

With \(q_i\) and the flow on the boarding demand arcs, \(x_{\text{board,d}}\), the number of passengers who failed to board at node \(i\) and are destined to \(d\), \(m_{qd}\), can be determined. It is assumed that passengers who failed to board will stay on the platform but might reconsider their hyperpath and choose from a different set of lines in the next time
interval. Therefore they are added to the boarding demand at the stop node in the
next time interval.

In (5-1), \( mq_{ds(u)} \) is the sum of those who failed to board at the failure nodes of lines \( l \)
that are served at the same platform \( u \) (with \( s_u \) indicating the stop node of platform
\( u \)):

\[
mq_{ds(u)} = \sum_{l \in L_u} q^*_e x^*_{d, \text{board}, -d, u}, \forall u \in U
\]  

(5-1)

\( d \) is the destination of the travellers and \( ^* \) indicates optimality, i.e. \( mq_{ds} \) is calculated
after the equilibrium is found. Chapter 7 and Figure 7-1 discuss and illustrate the
flow-chart for the dynamic capacity constrained assignment. The figure shows that
the equilibrium for one time interval is calculated before moving on to the next time
interval. Therefore the fail-to-board trips are added to the demand of the next time
interval as formulated in (5-2):

\[
y_d^{TI+1} \leftarrow y_d^{TI+1} + mq_d
\]  

(5-2)

Equations (5-1) and (5-2) as well as the flow chart Figure 7-1 show that for this
description of the network equilibrium it is not necessary to define the fail-to-board
probabilities or the node volumes time-dependently. As will also be discussed in
Chapter 7 this is important for efficient computation of large scale networks. Only
the demand for the subsequent time interval needs to be updated according to the
results of the current time interval. Note, that in this dynamic formulation the vector
\( y_d \) is not restricted to origin nodes. For \( TI \geq 2 \) passengers might also “originate” from
other nodes as they continue their journey.
5.3.2. Dealing with trips unfinished within a time-interval

Especially if short simulation time intervals are chosen, it is likely that some trips will not have reached their destination within one time interval. The approach described in Chapter 4 assumes however that passengers will pass all nodes they visit on their journey to \( d \) within the current time interval. In the following, the network loading is refined so that trips will only be assigned to those nodes that can be reached within the time interval. These trips are then picked up in the following time interval at the final node the trip reached in the previous time interval and continued towards their destination.

In the static formulation of network loading the vector of traffic traversing intermediate nodes \( i \), \( \mathbf{v}_d \), is obtained by (4-23):

\[
\mathbf{v}_d^{\text{static}} = (\mathbf{I} + \Pi_d^2 + \Pi_d^4 + \Pi_d^6 + \ldots) \mathbf{y}_d^{\text{static}} = \left( \left[ \mathbf{I} - \Pi_d \right]^{-1} \right)' \mathbf{y}_d^{\text{static}}
\]

For a dynamic formulation let us define the matrix \( \Delta_d \) with elements \( \delta_{ij} \), which takes the value of 1 if node \( j \) is reachable from node \( i \) in one time interval and otherwise zero.

\[
\delta_{ij} = \begin{cases} 
1 & \text{if } tt_{ij} \leq TID \\
0 & \text{otherwise}
\end{cases} \quad (5-3)
\]

\( \Delta_d \) is clearly dependent on \( TID \), further the subscript \( d \) is required because the travel time between nodes \( i \) and \( j \) is dependent on the hyperpath \( h \). Then the passenger volume that could reach node \( j \) can be expressed as \( \Delta_d' \mathbf{y}_d \) and the node volumes of passengers destined to \( d \) can be written as:
\[ v_d = \left( [I - \Theta_d]^{-1} \right)^t \Delta_d y_d \]  

(5-4a)

Note that, for simplicity the notation with \( \Theta_1d \) and \( \Theta_2d \) introduced in the static model is avoided in this chapter, but clearly (5-4) could also be written as:

\[ v_d = (\Theta_d [I - \Theta_2d]^{-1})^t \Delta_d y_d \]  

(5-4b)

In this study, all passengers are assigned at the beginning of the time interval and will be re-assigned from the furthest tail node they can reach within the time interval. To formulate this, let us define \( K_{r_d} \) with elements \( [\kappa_{r_i}] \) as an arc connecting \((i, j)\) satisfying following criteria:

\[
\kappa_{r_{ij}} = \begin{cases} 
1 & \text{if } tt_{ri} \leq TID \text{ and } tt_{rj} > TID \\
0 & \text{otherwise}
\end{cases}
\]  

(5-5)

where \( r \) is the start node of trips in this time interval, which can be an origin but also another node if the trip is continued over several time intervals. Note that because of common lines there might be several final nodes from each start node \( r \) that can be reached within one time interval. An algorithm to find the travel time matrix \( TT \) is described in the next section. With this definition of the “final nodes reachable in one time interval” the number of passengers who could not finish their trip and who have to be reassigned in the following time interval are calculated as in (5-6):

\[ m_d = \sum_r \left( [I - \Pi_d^*]^{-1} \right)^t K_{r_d}^* y_d \]  

(5-6)

The star in (5-6) indicates that \( m_d \) is calculated only after the equilibrium solution has been found. Similar to the fail-to-board trips, the unfinished trips are added to the demand of the next time interval as formulated in (5-7):
\[ y_{d}^{t+1} \leftarrow y_{d}^{t+1} + mu_{d} \] (5-7)

Note that travel times are associated with transit arcs (travelling on-board between stations), on-board waiting arcs (passengers waiting on-board while the train is stopping in a station), transfer arcs (to walk between platforms) and boarding arcs (passengers waiting for a service to arrive). Therefore the final nodes \( j \) of the arcs \((i, j)\) in \( K'_{d} \) can be stop nodes, boarding nodes or alighting nodes but not failure nodes origins or destinations.

### 5.4. Finding the travel time of a hyperpath

#### 5.4.1. Introduction

In the following, an algorithm is described which finds the user equilibrium travel time between any pair of nodes \( i \) and \( j \) for passengers destined for \( d \) with the optimal set of hyperpaths \( H_{d} \). For the calculation of the unfinished trips only a travel time matrix is needed for those node pairs that are part of any \( h \in H_{d} \). Therefore the following algorithm calculates travel times only for those arcs where \( \pi_{ij} > 0 \), which simplifies the problem.

Steps 1 to 3 initialise \( TT^{1} \) so that it contains the travel time for node transitions considering direct arc travel times but the travel time required for transitions that are not possible with a single arc transition are set to a high value. The number of paths \( np_{ij} \) that traverse node \( j \) for passengers from \( i \) is set to 1 if there is a direct arc between nodes \( i \) and \( j \) and 0 otherwise. (The matrix \( NP \) is used for the adjustment of
inaccuracies in the number of unfinished trips described in the following section). Steps 4 to 7 are then iterated until all passengers have reached their destination or until there are no more moves possible with travel times lower than or equal to $TID$ (termination criteria in Step 8). In Step 5, the matrix $NP$, is updated if there are new paths found from $i$ to $j$ with one more arc transition and Step 6 finds the travel times for these paths. In Step 7 the newly found moves are then added to the matrix $TT$ so that this matrix at the end contains the travel time for all transitions that are possible within one time interval.

5.4.2. Algorithm to find Travel Time Matrix

Step 1 (Initialisation)

Set $TT^1 := [tt_{ij}^1] := 999; \forall tt_{ij} \in TT$;

$NP := [np_{ij}] := 0; \forall np_{ij} \in NP$

Step 2 (Create $TT^1$ matrix)

$\forall a \in A_h$

If $a \in BoardingArcs$ : \hspace{1cm} $tt_{ij}^1 := w_{ih}$ (with $w_{ih}$ as in 4-11).

Else if $a \in FailureArcs$ : \hspace{1cm} $tt_{ij}^1 := 999$

Else \hspace{1cm} $tt_{ij}^1 := traveltime_a$

Step 3 (Initialise iterations)

Set $TT := TT^1$ and $TT_{\text{new}} := TT^1$

Set $np_{ij} := 1 \ \forall tt_{ij}^1 \neq 999$

Step 4 (Adjust $TT_{\text{old}}$; Set all $tt_{ij}^{\text{new}}$ back)

Set $TT_{\text{old}} := TT_{\text{new}}$

$TT_{\text{new}} := 999 \ \forall tt_{ij}^{\text{new}} \in TT_{\text{new}}$;
Step 5 (Update the number of paths that use a node)

Calculate NP defined as:

\[ np_{ij} \leftarrow \sum_{k=1}^{N} [(np_{ij} + 1) \cap (tt_{ik}^{old} \neq 999) \cap (tt_{kj}^{old} \neq 999)] \]

Step 6 (Find nodes reached in the subsequent move)

Calculate TT\(^{\text{new}}\) with TT\(^{1}\) and TT\(^{\text{old}}\), defined as:

\[ tt_{ij}^{\text{new}} \leftarrow \min_k [(tt_{ik}^{old} + tt_{kj}^{1}) \cap (tt_{kj}^{old} \neq 999) \cap (tt_{kj}^{1} \neq 999)] \]

Step 7 (Update TT with moves found in previous step)

Add all \( tt_{ij}^{\text{new}} \neq 999 \) to TT:

\[ tt_{ij} \leftarrow \text{Min}(tt_{ij}, tt_{ij}^{\text{new}}) \]

Step 8 (Terminate Loop)

Repeat Steps 4 to 7 until all \( tt_{ij}^{\text{old}} > TID \)

5.4.3. Adjustments for inaccuracies in \( \mu_d \)

Travel times between nodes are not necessarily unique. As the route choice and hence travel time depends on which service arrives first there is not necessarily a unique backnode. At equilibrium the same node might be reached by passengers using different paths or hyperpaths with slightly different travel times. If the travel times on different paths between the same nodes are close to the time interval duration with \( tt_{path1} \leq TID \) and \( tt_{path2} > TID \) this can lead to errors in the calculation of the number of unfinished trips at a particular node. In the above algorithm, the travel times are set to the minimum value if passengers on several paths with different travel time traverse the same node. The adjustment described in this section can then minimise the error and ensure flow conservation.
Figure 5-4 illustrates the problem of possible passenger double counting: In the figure the hyperpath for passengers from the same origin consists of two paths that are merging at node B. The final node reached by passengers travelling on path 2 is Node A whereas passengers travelling on path 1 are able to travel until Node C in this time interval. If one assumes the travel time to Node C to be the minimum of the travel time on Path 1 and Path 2 then the node volume of unfinished trips in (5-6) will be overestimated by the number of passengers from Path 2. This is because the number of unfinished trips at A will be equal to the number of passengers on Path 2 and the number of unfinished trips at C will be equal to the sum of passengers on Paths 1 and 2 (double counting of passengers on Path 2). The following correction avoids the double counting:

\[
mu_{ij} \left\{ \begin{array}{ll}
mu_{ij} - \left( v_{od} + m_{q,dr} + m_{u,dr} \right) \pi_{ij} & \text{if C1 to C4 all apply} \\
mu_{ij} & \text{otherwise} \end{array} \right.; \quad \forall \ r, j \in I \quad (5-8)
\]

With 

- C1: \( np_{ij} > 1 \)
- C2: there are two arcs \( a=(i,j) \) with \( a' \neq a \) and \( \kappa_{ij} = \kappa_{i'j'} = 1 \) (\( j \) and \( j' \) are both final nodes from \( r \))
- C3: \( j \) and \( j' \) are nodes not far from each other (\( tt_{ij} \) is “relatively small”)

Figure 5-4 Illustration of a hyperpath with several final nodes and merging paths
\[ C4 := \pi_{ij'} \leq \pi_{ij} \]

Four conditions need to apply if the number of unfinished trips needs to be corrected for double counting: Firstly, C1 ensures that the final node \( j \) (in Figure 5-4 this is Node C) which is considered for correction is used by several paths. Secondly, C2 ensures that there are several final nodes found. Thirdly, the second final node \( j' \) (Node A) and \( j \) should be relatively close to each other (C3). (In the case study a threshold of \( t_{jj'} < \frac{TID}{4} \) is chosen). This is to ensure that no correction happens in case Path 2 uses a very different route. If constraint C3 is set too tight one might miss some double counting, if it is set too loose too many corrections might be made resulting in too few passengers being reassigned in the subsequent time interval. And fourthly, the transition probability to reach node \( j \) should not be smaller than the transition probability to reach node \( j' \) (C4). This means that node \( j \) is the final node for more passengers from \( r \) than node \( j' \). In all the case studies, however, there are only a few cases where the problem of double counting occurs and with C1 to C4 the corrections are carried out correctly (evident through flow conservation).

5.5. MSA for dynamic network loading

5.5.1. Introduction

In this section the MSA described in the previous chapter is extended with a time loop to handle the dynamic network loading. The additional loop is set outside the MSA presented in Chapter 4.7.2. This means that first the equilibrium solution is found for a time interval, then the number of trips that failed to board and the number
of trips that are unfinished are calculated (Step 4). Therefore $\text{mq}_d$ and $\text{mu}_d$ are only calculated once for every time loop when the equilibrium is found and these vectors are then used to update the demand of the subsequent time interval.

5.5.2. MSA with Time Loop

Step 1 (Initialisation)

Set $TI := 0$;

$\text{mq}_d := 0; \text{mu}_d := 0 \ \forall s \in S$

Step 2 (Time Loop)

Load $y_d^{TI}$ (Trips starting in this time interval)

Step 3 (MSA as in Chapter 4.7.2 but with dynamic assignment)

Initialisation

Dynamic Assignment

Minimum Cost Hyperpath Search

Calculate Travel Time Matrix and $\Delta_d$ with (5-3)

Markov Chain Flow Assignment with $v^{(n)}$ as in (5-4)

Correction of $q^{(n)}$ and $x^{(n)}$

Iteration until termination of MSA Loop

Step 4 (Find fail-to-board passengers and unfinished trips)

Calculate $\text{mq}_d$ with (5-1)

Calculate $\text{mu}_d$ with (5-5) and (5-6)

Step 5 (Add $\text{mu}_d$ and $\text{mq}_d$ to origin demand of $TI+1$)

Update $y_d^{TI+1}$ with (5-2) and (5-6)

Step 6 (Termination of time loop)

$TI \leftarrow TI+1$

Return to Step 2 until $TI = Tl_{\text{end}}$
5.6. Alternative formulation of the generalised cost function

5.6.1. Expected delays

In Equation 4-10 the generalised costs of travelling on a hyperpath are expressed as

\[
g_h = \sum_{a \in A_h} \alpha_{ah} c_a + \sum_{i \in S_h} \beta_{ih} w_i - \eta \ln \left( \prod_{i \in S_h} (1 - q_i)^{\beta_{ih}} \right)
\]

As explained in the previous chapter the third term of (4-10) describes the increase in perceived generalised cost because of a fail-to-board-probability somewhere along the hyperpath.

The approach is similar to the measurement of “terminal reliability” as in Bell and Iida (1997) because it considers the possibility of failing somewhere along the journey. The formulation is however not directly related to the increase in travel time through failing to board. The dynamic extension of the approach allows an estimation of delays caused to the passenger through capacity constraints and therefore also a formulation of the generalised costs that considers the probability of being delayed.

The fail-to-board probability \( q_i \) causes an expected delay of \( TID^* q_i \) for passengers traversing node \( i \) in this time interval where \( TID \) is the duration of the time interval. The passenger must further consider that there is a probability that he also fails to board in the subsequent time interval. Therefore the probability of failing to board in
the current time interval but being able to board in time interval \( TI +1, TI +2, TI +3, \ldots \) is \( q(1-q), q^2(1-q), q^3(1-q), \ldots \) which causes an expected delay of

\[
TID * (q(1-q) + 2q^2(1-q) + 3q^3(1-q) + ... ) = TID \frac{q}{1-q}
\]

\[ (5-9) \]

Proof:

Set \( S = q(1-q) + 2q^2(1-q) + 3q^3(1-q) + ... \)

Then \( Sq = q^2(1-q) + 2q^3(1-q) + 3q^4(1-q) \) and

\[
S - Sq = q(1-q) + q^2(1-q) + q^3(1-q) + ... = (1-q) \frac{q}{(1-q)} = q,
\]

and hence \( S = \frac{q}{1-q} \)

Therefore the generalised cost becomes

\[
g_h = \sum_{a \in A_h} \alpha \cdot c_a + \sum_{i \in S_h} \beta \cdot w_{ih} - \theta_d \sum_{i \in S_h} \beta \cdot \left( \frac{TID \cdot q_i}{1-q_i} \right)
\]

\[ (5-10) \]

where \( \theta_d \) is the person’s value of likely delays through overcrowding. If one would weight the first two terms of the generalised cost (on-board travel time and waiting time) with a parameter \( \theta_o \) for “normal travel time” it is probably realistic to assume \( \theta_o < \theta_d \) because of “delay frustration”. Therefore in this case where \( \theta_o \) is omitted a value \( \theta_d > 1 \) is reasonable.
5.6.2. Considering variation of q over time

Equation (5-10) assumes that the passenger expects the failure probability $q$ to stay constant over the time intervals. This is often not a realistic assumption, as passengers might know that the network congestion will decrease (or increase) over time.

In order to reflect user expectations, the expected delay can be adjusted as in (5-11) with $\theta_T$:

$$g_h = \sum_{a \in A_h} \alpha_{ah} c_{d} + \sum_{i \in S_h} \beta_{ih} w_{ih} - \theta_d \sum_{i \in F_a} \beta_{ih} \frac{TID \cdot q_i}{(1 - q_i) \theta_T}$$  \hspace{1cm} (5-11)

Table 5-1 shows an interpretation for $\theta_T$. When modelling the congestion during a peak period one could therefore vary the value of $\theta_T$ in different time intervals. A reasonable assumption for $\theta_T$ might be illustrated in Figure 5-5. The figure shows a monotonically decreasing value for $\theta_T$. Before the morning peak the value is likely to be larger than 1, as passengers know that congestion is likely to increase in the subsequent time intervals. Until the beginning of the peak-period passengers will not expect the congestion to decrease with the following service arrivals, only at the end of the peak period and thereafter will passengers expect that subsequent services are emptier than in current conditions.
Table 5-1 Interpretation of $\theta_T$

<table>
<thead>
<tr>
<th>$\theta_T$</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_T = 0$</td>
<td>Passenger is sure to be able to board in next time interval</td>
</tr>
<tr>
<td>$0 &lt; \theta_T &lt; 1$</td>
<td>Passenger is optimistic that the delay is decreasing in subsequent time intervals</td>
</tr>
<tr>
<td>$\theta_T = 1$</td>
<td>Passenger expects the network congestion to stay similar in subsequent time intervals</td>
</tr>
<tr>
<td>$1 &lt; \theta_T &lt; \infty$</td>
<td>Passenger expects the delay to worsen in the following time intervals</td>
</tr>
</tbody>
</table>

Figure 5-5 Change of $\theta_T$ during the simulation period

A second, but probably less severe, restriction of Equation (5-10) as well as (5-11) is that the formulation includes the duration of the time interval only but not the headway of the services. This is because the failure probability $q$ also relates to the service capacity during a time interval. However, when failing to board a service with low frequency, the frustration is probably higher. Therefore one could consider introducing $\theta_l$ as a line specific parameter.
5.7. Case Study on a small example network

5.7.1. Introduction

In the following the dynamic extension is illustrated with the same network as in Chapter 4.9 (Figure 4-5). The demand is also assumed to be the same, i.e. 100 passengers travel from each station to each station downstream (600 passengers in total, split into 100 passengers for OD pairs: A→B, A→C, B→C, B→D, C→D). In order to illustrate the treatment of the excess demand it is assumed that there is no new demand after one hour. However, the simulation period is extended to 3 hours to let all passengers arrive at their destination.

In the following first the effect of choosing different time interval durations (TID) and secondly the effect of risk-averseness assuming the generalised cost function (5-10) is modelled. In all scenarios it is assumed that the common lines problem applies.

5.7.2. Modelling different time interval durations

The three hour simulation period is modelled with TID equal to 60min, 30min and 15min. In case of TID = 30min (15min) it is assumed that the demand is equally spread between the two (four) intervals of the first hour. In order to allow a better comparison of these results with the static results the expected delay is not considered in the demand function.
Table 5-2  \( TID = 60\text{min} \) a) Fail-to-board probabilities and boarders, b) Line flows

<table>
<thead>
<tr>
<th>Station</th>
<th>Line</th>
<th>1st hour</th>
<th>2nd hour</th>
<th>3rd hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>L1</td>
<td>0.4</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>L2</td>
<td>0.25</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>L1</td>
<td>0.7</td>
<td>60</td>
<td>0.45</td>
</tr>
<tr>
<td>C</td>
<td>L1</td>
<td></td>
<td>65.7</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>L2</td>
<td>0.25</td>
<td>25</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5-3  \( TID = 30\text{min} \) a) Fail-to-board probabilities and boarders, b) Line flows

<table>
<thead>
<tr>
<th>Station</th>
<th>Line</th>
<th>q</th>
<th>Board</th>
<th>q Board</th>
<th>q Board</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>L1</td>
<td>0.4</td>
<td>50</td>
<td>0.2</td>
<td>50</td>
</tr>
<tr>
<td>A</td>
<td>L2</td>
<td>0.25</td>
<td>50</td>
<td>0.41</td>
<td>50</td>
</tr>
<tr>
<td>B</td>
<td>L1</td>
<td>0.7</td>
<td>30</td>
<td>0.82</td>
<td>30</td>
</tr>
<tr>
<td>C</td>
<td>L1</td>
<td>0</td>
<td>33.3</td>
<td>0.04</td>
<td>34.8</td>
</tr>
<tr>
<td>C</td>
<td>L2</td>
<td>0.25</td>
<td>12.5</td>
<td>0.27</td>
<td>13.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Line Arc</th>
<th>0-30min</th>
<th>30-60min</th>
<th>60-90min</th>
<th>90-120min</th>
<th>120-150min</th>
<th>150-180min</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 A→B</td>
<td>100.1</td>
<td>63.2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>L1 B→C</td>
<td>100.0</td>
<td>100.1</td>
<td>63.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>L1 C→X</td>
<td>96.4</td>
<td>44.0</td>
<td>31.9</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>L2 A→C</td>
<td>100.1</td>
<td>36.6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>L2 C→D</td>
<td>99.7</td>
<td>27.8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5-4  \( TID = 15\text{min} \) a) Fail-to-board probabilities, b) Line flows

<table>
<thead>
<tr>
<th>Station</th>
<th>Line</th>
<th>0-15</th>
<th>15-30</th>
<th>30-45</th>
<th>45-60</th>
<th>60-75</th>
<th>75-90</th>
<th>90-105</th>
<th>105-120</th>
<th>120-135</th>
<th>135-150</th>
<th>150-165</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>L1</td>
<td>0.4</td>
<td>0.57</td>
<td>0.66</td>
<td>0.72</td>
<td>0.6</td>
<td>0.32</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>L2</td>
<td>0.25</td>
<td>0.41</td>
<td>0.63</td>
<td>0.86</td>
<td>0.6</td>
<td>0.36</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>L1</td>
<td>0.5</td>
<td>0.8</td>
<td>0.86</td>
<td>0.88</td>
<td>0.88</td>
<td>0.57</td>
<td>0.84</td>
<td>0.75</td>
<td>0.63</td>
<td>0.41</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>L1</td>
<td>0</td>
<td>0.25</td>
<td>0.15</td>
<td>0.18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>L2</td>
<td>0</td>
<td>0.25</td>
<td>0.37</td>
<td>0.37</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Line Arc</th>
<th>0-15</th>
<th>15-30</th>
<th>30-45</th>
<th>45-60</th>
<th>60-75</th>
<th>75-90</th>
<th>90-105</th>
<th>105-120</th>
<th>120-135</th>
<th>135-150</th>
<th>150-165</th>
<th>165-180</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 A→B</td>
<td>0.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
</tr>
<tr>
<td>L1 B→C</td>
<td>0.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
</tr>
<tr>
<td>L1 C→X</td>
<td>0.0</td>
<td>16.7</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>13.1</td>
<td>7.7</td>
<td>7.7</td>
<td>8.3</td>
<td>11.1</td>
<td>12.5</td>
<td>8.8</td>
</tr>
<tr>
<td>L2 A→C</td>
<td>0.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>25.0</td>
<td>16.0</td>
<td>1.4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>L2 C→D</td>
<td>8.3</td>
<td>6.3</td>
<td>25.4</td>
<td>25.2</td>
<td>20.9</td>
<td>18.1</td>
<td>17.5</td>
<td>9.8</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

123
In Table 5-2 the fail-to-board probability and the number of passengers failing to board during the first hour are identical to the results gained with the static model. Those passengers that failed to board are then reassigned in the second and third hour. In the second time interval the demand at Station B is still higher than the available capacity \( q_{B,L1}^{T=2} = 0.45 \) so that 63.9 passengers only arrive during the third time interval at their destination even though they started their journey during the first time interval. The line loads in Table 5-2 confirm that Line 1 leaving Station B is still fully occupied in the second time interval and is also still used in the third time interval.

Modelling shorter time intervals allows a more detailed analysis of the overcrowding in the network. The models with \( TID = 30 \) and especially with \( TID = 15 \) reveal the gradual increase and later decrease of the congestion at Station B better. In case of \( TID = 30 \) it can be seen that the overcrowding at B is worse after 30min \( q_{B,L1}^{T=2} = 0.82 \) compared to the first time interval \( q_{B,L1}^{T=1} = 0.7 \). This effect is hidden if one models too long time intervals. Note further that in the model with \( TID = 15 \) the fail-to-board probability at Station C for Line 2 is 0 in the first time interval in contrast to 0.25 for the models with \( TID = 30 \) and \( TID = 60 \). This is because it takes more than 15min for passengers from A to arrive at C so that the model recognises that the first group of passengers from Station C do not compete for space with passengers from Station A.

A comparison of the line flows for \( TID = 30 \) (Table 5-3) and \( TID = 15 \) (Table 5-4) further reveals a weakness of the model. According to the results with TID=30 all passengers have arrived after 180min. According the results with TID =15 however,
the last passengers only arrive after 195min. The differences occur because of the “setting back” of passengers in case the arc can not be fully traversed within one time interval. The trips are continued in the following time interval from the last node reached in the previous time interval. In the network used in this case study traversing the arc between Station A and C takes 15min so that some significant errors might occur. The error could be reduced if the long arc is divided into several shorter arcs. If one models however networks in inner cities the arc travel times are usually relatively short compared to the duration of a time interval so that the error is not as large as in this example. Note that this approach requires that time interval duration $TID$ is larger than the longest arc in the network.

5.7.3. OD reliability

Figure 5-6 illustrates the number of passengers that have arrived at the destination after a certain time period. Plotting the same graph for $TID=30$ or $TID=60$ gives a similar result, but does not reveal many details, for example the curves for OD demand $B\rightarrow C$ and $B\rightarrow D$ are the same. The figure illustrates that even though passengers from $A\rightarrow D$ have a longer journey all passengers have already arrived earlier than passengers between $A\rightarrow C$. (Because some passengers travelling to C also use the more heavily overcrowded Line 1 whereas all passengers travelling $A\rightarrow D$ use the less crowded Line 2. The figure illustrates that in particular passengers from $B$ suffer delays through overcrowding as the failure probability for these passengers is highest.
5.7.4. Delay expectation and risk averseness

In the previous sections 5.7.2 and 5.7.3 the risk-averseness of users ($\theta_e$ in the generalised cost function (4-10)) or the expectation of delays ($\theta_d$ in the generalised cost function (5-11)) has not been considered ($\theta_d = \theta_e = 0$). As shown already in the case study in Chapter 4.9 considering these additional costs leads to rerouting of passengers in order to avoid overcrowded platforms. Table 5-5a shows the fail to board probabilities if passengers consider that there might be delays and if they value these delays in the same way as normal travel time ($\theta_d = 1$) for all time intervals with $TID=15$. For comparison, Tables b) and c) further show $\mathbf{q}$ with the assumption of risk-averseness ($\theta_e=10$ and $\theta_e=100$). One can see that the results for $\theta_d = 1$ and $\theta_e = 100$ are fairly similar. In both cases the costs of potentially failing-to-board dominate the route choice decisions. A calibration of the parameter might well find that $0 < \theta_d < 1$ is a more realistic assumption because passengers might expect to get on the next train. $\theta_d = 1$ assumes, however, full passenger mingling between all passengers.
arriving in a certain time interval, i.e. the chance of boarding one of the following trains is assumed to be the same as the chance of boarding the first arriving train.

Table 5-5 Fail-to-board probabilities for a) $\theta_d=1$; b) $\theta_e=10$ and c) $\theta_e=100$

<table>
<thead>
<tr>
<th>Station</th>
<th>Line</th>
<th>0-15</th>
<th>16-30</th>
<th>30-45</th>
<th>46-60</th>
<th>60-75</th>
<th>75-90</th>
<th>90-105</th>
<th>105-120</th>
<th>120-135</th>
<th>135-150</th>
<th>150-166</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>L1</td>
<td>0.37</td>
<td>0.52</td>
<td>0.62</td>
<td>0.68</td>
<td>0.62</td>
<td>0.16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>L2</td>
<td>0.29</td>
<td>0.47</td>
<td>0.58</td>
<td>0.86</td>
<td>0.47</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>L1</td>
<td>0.5</td>
<td>0.81</td>
<td>0.87</td>
<td>0.53</td>
<td>0.88</td>
<td>0.57</td>
<td>0.83</td>
<td>0.7</td>
<td>0.66</td>
<td>0.22</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>L1</td>
<td>0</td>
<td>0.26</td>
<td>0.13</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>L2</td>
<td>0.14</td>
<td>0.16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station</th>
<th>Line</th>
<th>0-15</th>
<th>16-30</th>
<th>30-45</th>
<th>46-60</th>
<th>60-75</th>
<th>75-90</th>
<th>90-105</th>
<th>105-120</th>
<th>120-135</th>
<th>135-150</th>
<th>150-166</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>L1</td>
<td>0.4</td>
<td>0.56</td>
<td>0.65</td>
<td>0.71</td>
<td>0.67</td>
<td>0.25</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>L2</td>
<td>0.25</td>
<td>0.41</td>
<td>0.63</td>
<td>0.61</td>
<td>0.41</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>L1</td>
<td>0.5</td>
<td>0.8</td>
<td>0.86</td>
<td>0.9</td>
<td>0.88</td>
<td>0.87</td>
<td>0.84</td>
<td>0.73</td>
<td>0.61</td>
<td>0.36</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>L1</td>
<td>0</td>
<td>0.26</td>
<td>0.16</td>
<td>0.19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>L2</td>
<td>0.26</td>
<td>0.36</td>
<td>0.33</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station</th>
<th>Line</th>
<th>0-15</th>
<th>16-30</th>
<th>30-45</th>
<th>46-60</th>
<th>60-75</th>
<th>75-90</th>
<th>90-105</th>
<th>105-120</th>
<th>120-135</th>
<th>135-150</th>
<th>150-166</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>L1</td>
<td>0.34</td>
<td>0.51</td>
<td>0.61</td>
<td>0.67</td>
<td>0.41</td>
<td>0.03</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A</td>
<td>L2</td>
<td>0.32</td>
<td>0.49</td>
<td>0.69</td>
<td>0.66</td>
<td>0.49</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>L1</td>
<td>0.5</td>
<td>0.81</td>
<td>0.77</td>
<td>0.8</td>
<td>0.56</td>
<td>0.83</td>
<td>0.65</td>
<td>0.46</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>L1</td>
<td>0</td>
<td>0.2</td>
<td>0.08</td>
<td>0.23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>L2</td>
<td>0.14</td>
<td>0.15</td>
<td>0.23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Finally, Figure 5-7 illustrates the impact of delay expectation on the reliability of the three origins. In the figure, the demand to different destinations is summarised. The figure reflects the change in $q$ for $\theta_d=1$ compared to $\theta_d=0$. The two curves for passengers from A and C are very similar, meaning that the consideration of delays can not significantly reduce the expected travel time. However, the graph shows that the slight re-routing of passengers from A has a positive impact on passengers from B.
5.8. Discussion

5.8.1. Summary

This chapter presented an extension of the model presented in the previous chapter that allows modelling the build-up of crowding over a peak-period. It is explained how passengers who failed-to-board in a previous time interval are re-assigned in the subsequent time interval and how trips that can not be completed in a single time interval are continued from the last node they reached in the previous time interval. Further an improved formulation of the generalised cost function is presented and the approach is illustrated with a small example network. It is shown that this approach is suitable to estimate the delays through overcrowding experienced for passengers from different origins.

This is probably the first approach to model dynamic network loading with a frequency-based model and the consideration of strict capacity constraints. The price
for keeping the model as a frequency-based one (with simpler input data compared to a schedule-based model) is that several assumptions have to be made:

5.8.2. Mingling in the dynamic case

Also for this dynamic approach passenger mingling is assumed (as it is inherent to any approach using Markov Chains). This means that not only passengers with different destinations mingle but also that passengers arriving from different arcs at a node during the time interval mingle. In particular it is assumed that passengers who failed-to-board in a previous time interval mingle with passengers arriving (newly) in this time interval. This is in many cases not realistic as passengers who failed-to-board in a previous time interval will have a higher chance of boarding than passengers arriving anew. On the other side the assumption of FIFO as in most schedule-based models might also not be fully realistic. As explained in the case study this means that probably an assumption of $0 < \theta_d < 1$ is realistic.

5.8.3. “Setting back” of unfinished trips

Trips that can not be completed within a single time interval are continued from the last node they reached in the previous time interval. This means that parts of the arc leading out of the last node reached within a time interval might be travelled twice by passengers. At first partly in the previous time interval and secondly in the subsequent time interval the arc is traversed in full again by the passengers. This obviously leads to some errors in particular if the network includes arcs with long travel times. The errors and the problem of long arcs in the network can be minimised if these are split into several shorter arcs. Note also that this approach
requires that the time interval duration is longer than the time it takes to traverse any arc in the network. If this is not the case the arc can not be traversed by passengers.

Further, in some (rare) circumstances the errors through the setting back of passengers can violate the capacity constraints. For example Figure 5-8 illustrates part of a network where two paths of the optimal hyperpath merge. Passengers on path 1 are slightly slower than passengers on path 2, so that the final node reached by passengers on path 1 is N1 and the final node reached by passengers on path 2 is N2. If one assumes a capacity of 100 for each line the capacity constraints are not violated if in the next time interval passengers choose the same hyperpath and if the demand is not increased as illustrated in part b) of Figure 5-8. If however, passengers change their strategy in the subsequent time interval it might happen that the capacity constraints on an arc are violated as can be see in part c) of the figure. In Figure 5-8c the demand is unchanged however passengers risk averseness led more passengers to choose path 2 in TI=2 compared to TI=1. Therefore the number of passengers assigned to arc III is in total 110 passengers (70 passengers who continue their journey on path 1 and 40 passengers from path 2). Note that the capacity constraint on the subsequent arc IV is not violated anymore.

In all case studies a slight violation of capacity constraints through the setting back of passengers has only been found on a few arcs.
5.8.4. Choice of Time Interval Duration

The above problem also leads to the question of the optimal choice of the time interval duration $TID$. On one side, the case study showed that a too long $TID$ does not reveal adequately the build up of congestion. Further the choice of a long $TID$ makes it more likely that passengers with the same OD who travel on different paths have merged at the end of the time interval leading to the problem of double counting as described in Section 5.4.3. On the other side, a shorter $TID$ leads to larger errors in
the calculation of unfinished trips. In practice the choice of TID will also depend on
the available data as OD data are often collected for longer time intervals and
dividing these OD matrices into smaller time intervals can also lead to significant
errors. A further consideration might be that choosing shorter time intervals also
increases the computational effort required.

5.8.5. Split between paths with equal costs
The model finds the user equilibrium but does not consider stochastic effects. This
means that at the equilibrium the arc flows are unique but the path flows are not (see
for example Bell and Iida, 1997). For the calculation of the remaining passengers and
the number of unfinished trips at a node the path flows are however utilised in this
approach. It is assumed that there is a split between all paths with the same cost and
that are part of the optimal hyperpath according to the frequency of each path, which
is probably the best assumption in the absence of any further information.

5.8.6. Failure probability in future time intervals
A significant limitation of the current model might be that the passenger’s route
choice is only made according to the current network congestion. Especially if the
trip is long and covers several time intervals this is not a realistic assumption. A
rough solution for this problem is presented in Section 5.6 with the introduction of
the parameter $\theta_T$. A better solution might be to introduce a time-loop instead of
calculating the equilibrium for each time interval separately (and moving on to the
next time interval without any feedback) as done in the current approach. Introducing
a time loop might not only change the route choice of passengers but would also
allow introducing a model of optimal departure time.
Such an approach would however significantly increase the memory requirements and computation time. In the current model only the demand vector $y_d^{TI}$ needs to be stored time-dependently which means that the memory requirement of this current approach is almost independent of the number of time intervals saving a significant amount of memory space which is important for larger models.
6. Further congestion effects – Model refinements

6.1. Introduction

Having presented the “dynamic CapCon model” in the previous two chapters this chapter discusses two model improvements. Firstly, extensions to the generalised cost functions are discussed in Chapter 6.2, as the cost function so far only considers in-vehicle time, platform waiting time and the fear of failing to board (4-10) or, alternatively, the delay through failing to board (5-10). In particular one might want to consider that crowding also increases the perceived cost of travelling on the vehicle because of less comfort.

Secondly, crowding might not only occur on the platform and on the vehicle but already before at the ticket gate or at the escalator. The ticket gates might even be used by the transit operator in order to control the passenger inflow to the platform and avoid platform congestion. In a similar way as capacity constraints are introduced to lines, the CapCon model could also be extended to consider constraints for the inflow to the platforms. This is presented in Chapter 6.3.

There are two publications relating to the model extensions presented in this chapter. Firstly, Schmöcker et al (2005) discuss a case study where first the crowding costs are changed through additional seats in the vehicle and then the platform access control is added to the model. The impact of these measures on the overall performance is evaluated. Secondly, Kurauchi et al (2006) attempt to optimise the network capacity utilisation of the London Underground network through platform
access control. In their paper the authors use an approach based on Genetic Algorithms to find the optimal settings for the platform access capacity.

6.2. On-Board Congestion

6.2.1. Adjustment of the generalised cost function

The generalised cost function as defined in the previous chapter (Equation 5-12) can be extended to include several other aspects of the (generalised) cost experienced by passengers. Shimamoto et al (2005) for example applied the CapCon model to evaluate possible public transport fare strategies and included therefore the fare into the generalised cost.

In order to model the discomfort caused to passengers through crowded vehicles and to distinguish this from the impact of capacity constraints let us define

\( \theta_c \): Sensitivity to on-board congestion

\( c_a(x_a) \): Cost of travelling on arc \( a \) which is a function of the flow \( x_a \).

In this case (5-10) is replaced by Equation (6-1) where the cost of using a transit arc is depending on the arc volume \( x_a \).

\[
g_h = \sum_{a \in A_h} \alpha_{ah} c_a(x) + \sum_{i \in S_h} \beta_{ih} w_{ih} - \theta_c \sum_{i \in E_h} \beta_{ih} \frac{TID \cdot q_i}{(1 - q_i)}
\]  

(6-1)

In general, the cost function for arc \( a \) is clearly depending on the flow, the capacity and \( \theta_c \) as shown in (6-2):

\[
c_a = t_a + \theta_c \cdot \text{CF}(x_a, \text{cap}_{h(a)})
\]  

(6-2)
where CF is a cost function to be specified. There are various possible cost functions to define the increase in travel cost through congestion. Most widely used for traffic and transit assignment are BPR-like cost functions (Bureau of Public Roads, 1964) as in (6-3) which assumes costs are increasing with flow according to a power function in overcrowded vehicles. Besides $\theta_c$, which has the unit time, the dimensionless power function needs to be calibrated according to user behaviour. In (6-3) and below example the power function four is chosen. Note that the capacity is arc specific so that the capacity for all other arcs than transit arcs can be set to infinity (no crowding effect).

$$c_a(x) = c_a + \theta_c \left( \frac{x_a}{c a p_a} \right)^4$$  \hspace{1cm} (6-3)

Including flow depending arc flows in the MSA means that at the beginning of each hyperpath search the arc flows are adjusted according to the average arc flows found in the previous iterations.

### 6.2.2. Example: Comparison between crowding and fail-to-board

Applying cost function (6-3) to the small network example used in previous chapters leads to the results shown in Table 6-1. In this example it is assumed that users do not consider the fail-to-board probability (risk-averseness $\theta_e = 0$). Figure 6-1 compares these results with the results found under consideration of risk-averseness $\theta_e$ (and crowding costs $\theta_c = 0$) as illustrated in Chapter 4.9 (Table 4-3, with common lines).
The figure shows that the total number of passengers failing to board is actually lower when considering the crowding effects and the unutilised space is lower if one considers failure to board. The reason for this is that the cost function (6-2) for full arcs is constant, independent of the excess boarding-demand. Therefore passengers from Station A traveling to C and D do not consider Line 1 anymore (as Line 2 is shorter and both lines are full, so that the crowding cost is the same for Line 1 and 2) as shown in Table 6-1. In conclusion, applying the crowding cost function if one does not allow for line flows above capacity does not reflect the extra costs experienced by passengers through excess demand.

**Table 6-1** Results for different crowding sensitivities $\theta_c$ with risk-averseness $\theta_e=0$

<table>
<thead>
<tr>
<th>Station</th>
<th>Line</th>
<th>$\theta_e$</th>
<th>$\theta_c$</th>
<th>q</th>
<th>Board</th>
<th>Demand</th>
<th>Alight</th>
<th>Avg.Cap</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0.70</td>
<td>60</td>
<td>200.0</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td>66.7</td>
<td>66.7</td>
<td>70</td>
<td>70</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>0.25</td>
<td>0</td>
<td>0.25</td>
<td>25</td>
<td>33.3</td>
<td>25</td>
<td>25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station</th>
<th>Line</th>
<th>$\theta_e$</th>
<th>$\theta_c$</th>
<th>q</th>
<th>Board</th>
<th>Demand</th>
<th>Alight</th>
<th>Avg.Cap</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0.01</td>
<td>0</td>
<td>0.01</td>
<td>100</td>
<td>100.9</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>0.50</td>
<td>0</td>
<td>0.50</td>
<td>100</td>
<td>199.1</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0.50</td>
<td>0</td>
<td>0.50</td>
<td>99.2</td>
<td>200.0</td>
<td>96.4</td>
<td>99.2</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0.23</td>
<td>0</td>
<td>0.23</td>
<td>51.6</td>
<td>66.7</td>
<td>49.2</td>
<td>51.6</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>33.3</td>
<td>33.3</td>
<td>50.6</td>
<td>48.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station</th>
<th>Line</th>
<th>$\theta_e$</th>
<th>$\theta_c$</th>
<th>q</th>
<th>Board</th>
<th>Demand</th>
<th>Alight</th>
<th>Avg.Cap</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>100</td>
<td>100.2</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>0.50</td>
<td>0</td>
<td>0.50</td>
<td>100</td>
<td>199.8</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0.50</td>
<td>0</td>
<td>0.50</td>
<td>99.8</td>
<td>200.0</td>
<td>99</td>
<td>99.8</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0.02</td>
<td>0</td>
<td>0.02</td>
<td>57.2</td>
<td>58.6</td>
<td>56.4</td>
<td>57.2</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>48.8</td>
<td>48.8</td>
<td>50.9</td>
<td>50.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Station</th>
<th>Line</th>
<th>$\theta_e$</th>
<th>$\theta_c$</th>
<th>q</th>
<th>Board</th>
<th>Demand</th>
<th>Alight</th>
<th>Avg.Cap</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>100</td>
<td>100.2</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>A</td>
<td>2</td>
<td>0.50</td>
<td>0</td>
<td>0.50</td>
<td>100</td>
<td>199.8</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>0.50</td>
<td>0</td>
<td>0.50</td>
<td>99.8</td>
<td>200.0</td>
<td>99</td>
<td>99.8</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>0.02</td>
<td>0</td>
<td>0.02</td>
<td>57.2</td>
<td>58.6</td>
<td>56.4</td>
<td>57.2</td>
</tr>
<tr>
<td>C</td>
<td>2</td>
<td>0.00</td>
<td>0</td>
<td>0.00</td>
<td>48.8</td>
<td>48.8</td>
<td>50.9</td>
<td>50.4</td>
</tr>
</tbody>
</table>
Figure 6-1 Effect of $\theta_e$ and $\theta_c$ on utilisation of available line spaces and total number of passengers failing to board

Figure 6-2 shows the results if one considers crowding costs as well as passenger risk-averseness. As expected the results are in between those considering $\theta_c$ and $\theta_e$ only. The convergence of the MSA is slow as the hyperpath costs in the iterations depend on changing line arc costs as well as changing failure node costs.
Figure 6-2 Effect of combined $\theta_e$ and $\theta_c$ on utilisation of available line spaces and total number of passengers failing to board

6.2.3. Perceived costs of sitting and standing in a crowded service

The crowding cost function presented above does not allow for the fact that not all passengers will experience the on-board congestion in the same way. In particular, whether the passenger is able to find a seat will significantly influence the perceived costs as for example found by London Underground (London Underground Limited, 1988). If one assumes that the chance of getting a seat is the same for all passengers (i.e. independent of the boarding point), the following cost function can be assumed.
\[ c_a(x) = \begin{cases} 
\frac{c_{a,sit}(x)}{seats_a} \left( \frac{seats_a}{x} \right) + c_{a,stand}(x) \left( \frac{x - seats_a}{x} \right) & \text{if } (x > seats_a) \\
0 & \text{if } (x \leq seats_a) 
\end{cases} \quad (6-4) \]

where \( c_{a,sit}(x) \) and \( c_{a,stand}(x) \) are cost functions to be calibrated with \( c_{a,sit}(x) < c_{a,stand}(x) \).

The above function does not recognise that passengers who board earlier are more likely to get a seat than passengers boarding later.

A small change to the CapCon network description further allows to model the chance of initially finding a seat. In Figure 6-3 a “Fail-to-Sit node” is introduced similar to the “Fail-to-Board” node. Passengers who fail to find a seat could then be assigned a penalty to reflect their higher generalised costs. The probability of finding a seat \( q_{sit} \) can be calculated similar to the fail-to-board probability \( q \) in (4-27).

\[ q_{sit}^u = 1 - \max \left\{ \min \left( \frac{seats_{ij} - x_{on-board,ui}}{x_{boarding-demand,ui}}, 1 \right), 0 \right\}, \forall u \in U, l \in L. \quad (6-5) \]

Because the approach assumes mingling it must be assumed that the chance of finding a seat is the same for all passengers wishing to board at this stop during this interval. With this assumption the cost function can then be adjusted to:
\[ g_h = \sum_{a \in A_h} \alpha_{ah} c_a(x) + SPenalty \sum_{i \in B_h} \beta_{ih} q_i^{sit} + \sum_{i \in B_h} \beta_{ih} w_{ih} - \theta_d \sum_{i \in B_h} \beta_{ih} \frac{TID \cdot q_i}{(1 - q_i)^\theta} \]  

(6-6)

where \( SPenalty \) is the cost for not finding a seat and \( \beta_{ih} \) in the second term of the cost function is the probability to pass boarding node \( B_h \). The introduction of a fail-to-sit penalty does not reflect that those passengers who initially fail to find a seat might however find a seat later through passengers alighting. This would require changes to the network description. To some degree one might however argue that it is mainly the chance of being able to sit from the point of boarding that influences the perceived generalised cost for commuters.

6.3. **Fail To Access**

6.3.1. **Introduction**

The congestion in transit networks does often not only occur on the platform or on the service but passengers might experience delays even earlier. Delays in buying tickets through long queues, or delays on the stairs or the escalators leading to the platform, are often experienced.

In some cases these delays are even desired by the operator as it means a temporarily reduced demand of passengers wishing to board. Passengers trying to enter stations after large events like football matches or concerts are familiar with the experience that police and station staff are controlling the inflow to the platform. It is even a daily experience at London’s Victoria Station that the access gates to the station are closed for several minutes during the morning peak to avoid overcrowding of the
platforms. As pointed out in the introduction, Britain’s Rail Safety and Standards Board further recommends that the throughput of ticket gates should be adjusted to control the crowding of platforms (RSSB, 2005). The following describes an extension to the CapCon model that allows modelling the effect of platform access control on the line loadings.

6.3.2. Access to Fail Nodes

The basic idea for modelling access demand to platforms is the same as for modelling fail-to-board probabilities or fail-to-sit probabilities. Fail-to-Access nodes are introduced in between the origin node and the stop node (Figure 6-4).

![Figure 6-4 Nodes and Arcs for modelling platform access control](image-url)
Further following definitions are added:

\( \text{cap}_{ru} \): Maximum number of passenger entering platform \( u \) at station \( r \) during one time interval.

\( \theta_p \): Sensitivity to using overcrowded platforms (risk averseness)

\( \text{access-d}_{ru} \): Access demand arc to platform \( u \) at station \( r \).

\( \text{access}_{ru} \): Access arc to platform \( u \) at station \( r \).

\( q^a_{ru} \): Fail-to-access probability to platform \( u \) at station \( r \).

6.3.3. Ensuring strict capacity constraints (for platform access)

With the above definitions the fail-to-access probability \( q^a \) can be determined. Constraint (6-7) must be fulfilled so that the number of passengers accessing the platform is adjusted with \( q^a \) as in (6-8) and (6-9) which follows the adjustment of the fail-to-board probability in Chapter 4.5.2. Note that in this case the formulation is even simpler as there is no equivalent to the on-board passengers (compare (6-8) and (6-9) with (4-26) and (4-27)).

\[
\text{cap}_{ru} \geq x_{\text{access}_{ru}} \quad \forall u \in U \tag{6-7}
\]

\[
x_{\text{access}_{ru}} = (1 - q^a_{ru}) x_{\text{access-d}_{ru}} \tag{6-8}
\]

\[
q^a_{ru} = 1 - \min \left( \frac{\text{cap}_{ru}}{x_{\text{access}_{ru}}}, 1 \right) \tag{6-9}
\]

It should be noted that (6-7) and (6-9) do not consider that alighting passengers occupy platform capacity. This could be changed by subtracting the number of alighting passengers from the access capacity. However, as discussed before it might
be the main objective of the operator to control the number of passengers competing for the escalator and for boarding in which case the above formulation is more relevant.

6.3.4. Dealing with excess demand at platform access arcs

Having obtained the equilibrium through the MSA, $mq^u_{do}$, is the destination specific excess demand of passengers failing to access platform $u$ at station $r$ which is determined with (6-10). Because these passengers are in the same position as those passengers who have not yet started their journey, the access demand can be added to the demand from origin $o$ in the following time interval (6-11). The assumption here is again passenger mingling, i.e. those passengers who failed to access in the previous time interval mingle with passengers who start their journey in the subsequent time interval.

$$mq^u_{do} = \sum_{user} q^u_{ru} x^*_{d,over_{ru}} \forall r \in R \quad (6-10)$$

$$y^{Tr+1}_d \leftarrow y^{Tr+1}_d + mq^a_d \quad (6-11)$$

The user equilibrium can be determined with the same procedure described in the previous two chapters with two minor modifications. Firstly, the hyperpath search algorithm described in Chapter 4.4.2 can be applied if the Fail-to-Access nodes are added to the Fail-to-Board nodes (in Step 3, 4, and 6). Secondly, the MSA algorithm for dynamic network loading as in Chapter 5.5.2 is adjusted in Steps 2 and 4 as shown below:
**Step 2** (Time Loop)

*Load* $y^{TI}_d$ (Trips starting in this time interval)

and add platform access demand as in (6-12) if $TI > 1$

**Step 4** (Find fail-to-board passengers and unfinished trips)

*Calculate* $mq_d$ with (5-1), $mu_j$ with (5-6) and $mq^d$ with (6-11).

### 6.3.5. Example: Restricting platform access

The impact of platform access restrictions can also be illustrated with the same example network used in Chapter 4 and 5. In the following the access to the platform of Station A is restricted. The demand from A is 300 passengers per hour so that obviously a platform access capacity $cap_A \geq 300$ pass/hour has no effect.

Figure 6-5 shows the impact of reducing the access to the platform at Station A equal or below 300 pass/hour for a risk-averseness of $\theta_e = 10$ and assuming that the common line problem applies. If $cap_A = 300$ then 240.1 passengers are failing to board which corresponds to the results in Table 4-3. The figure shows that the number of passengers failing to access the platform increases linearly with the introduction of the platform access constraint. This is obvious as there are no platforms to choose from (if A would have several platforms and if passengers are risk-averse, they might reroute to lines departing from other platforms in order to avoid failing to access).
If the platform access constraint allows more than 250 pass/hour to access the platform, the total number of passengers failing to be served from A does not increase. This is because the number of passengers failing-to-board decreases at the same rate as the number of passengers failing-to-access the platform increases. If less than 250 passengers are allowed to access the platform the total number of passengers failing to be served increases because Lines 1 and 2 are not completely full anymore.

![Figure 6-5 Effect of Introducing a Platform Access Constraint at Station A](image)

In conclusion, the operator can reduce the number of passengers accessing the platform to 250 pass/hour without reducing the service quality. The operator might choose to do so in order to reduce platform congestion but obviously at the price of increasing the congestion in other parts in and around the station. However, an operator might choose this strategy as it will help to keep the dwell times more stable. Further, it reduces the risks of accidents through passengers pushing to get onto the next train.
6.4. **Discussion**

6.4.1. **Summary of Congestion Costs**

This chapter presented two extensions to the CapCon model in order to model different aspects of passenger crowding. Firstly, the modelling of on-board crowding was discussed. This should be modelled with arc costs depending on the level of line loading. Modelling in this way this further reflects the fact that the amount of discomfort experienced by travelling on-board an overcrowded vehicle is also dependent on the length of the journey. Further, previous research suggests that standing and sitting in a crowded vehicle should be distinguished. Two approaches for this are mentioned in this chapter. One could either average the cost experienced by sitting and crowding over all passengers on-board as in (6-4) which ignores however that passenger boarding earlier have a higher chance of finding a seat. Alternatively one could calculate the Fail-to-Sit-probability and assign a penalty to those passengers failing to sit at the boarding node as in (6-5) and (6-6). This approach regards the priority of those boarding earlier, but does not reflect that passengers who initially stand might find a seat later because of passengers alighting.

Secondly, the impact of constraints for passengers wishing to access the platform is discussed. The main objective of modelling such a constraint is that transit operators might want to know by how much the number of passengers waiting on the platform can be reduced for safety and dwell time reasons but without reducing the total number of passengers boarding.
Figure 6-6 illustrates three different impacts of overcrowding modelled with the CapCon approach: Failing-to-board a service, failing to access a platform and failing to find a seat at the point of boarding.

![Diagram of different crowding impacts]

**Figure 6-6** Illustration of the different crowding impacts

Modelling these congestion impacts with the dynamic model means that the node volume in a time interval consists of four passenger groups:

a) The demand created by passengers starting their journey in this time interval  

b) Passengers who failed to access the platform within the last time interval  

c) Passengers who failed to board a service within the last time interval  

d) Passengers who did not finish their journey in the last time interval

The first two passenger groups are combined into one by Eq. (6-11) as passengers who failed to access the platform in the previous time interval and those who start their journey in this time interval are assumed to mingle. The calculation of the third and fourth group is explained in Chapter 5 and the sum of these groups is then loaded to the network as discussed in Section 5.3.
6.4.2. Further adjustments to the generalised cost function

Besides congestion-related costs, the generalised cost function can be further extended to distinguish different elements of the perceived journey costs. There are several publications on the values for waiting time and walking time between platforms for frequency-based transit networks; see for example Szplett and Wirashenge (1984); Parveen, M. and Shalaby, A. (2004) or the guideline for transit modelling by the Department for Transport (2005). New York City Transit weighs for example waiting time with 1.25 and walking time between platforms with 1.5 compared to the time spent travelling on-board (NYCT, 2001). These parameters $\theta_{\text{walk}}$ and $\theta_{\text{wait}}$ can be introduced in the cost function as in (6-13).

$$
g_h = \sum_{a \in \{\text{LineArcs}\}} \alpha_{ab} c_d(x) + \theta_{\text{walk}} \sum_{a \in \{\text{On-board\ WaitArcs}\}} \alpha_{ab} c_d(x) + \theta_{\text{wait}} \sum_{i \in S_h} \beta_{ih} w_{ih} - \theta_d \sum_{i \in E_h} \beta_{ih} \frac{TID \cdot q_i}{(1-q_i) \theta_e}$$

(6-12)

Finally, adding the fare to the generalised cost is one of the most important factors of the generalised costs. How the fare should be added depends however significantly on the fare system. If the fare is line-specific, it should be added at the boarding node (node-specific cost, as waiting time and fail-to-board probability). If the fare is depending on the length of the journey it should be added to the arc-specific costs (as arc travel time and on-board crowding). In other networks, like London, the fare is zone specific in which case it is better to add an OD specific constant to the generalised costs reflecting the fare.
7. Software implementation

7.1. Introduction

The previous three chapters presented the formulation of the dynamic capacity constrained transit model. This chapter explains how this is implemented into software. The programming language used for software implementation is Delphi 7®, which allowed an object oriented program structure and the creation of a visual interface.

The next section will summarise the structure of the software and firstly explain which input data the model requires and how these input data can then be efficiently used to create the network description used in previous chapters. This is followed by a flow chart to explain the whole programme structure.

The programme structure highlights that the calculation of the equilibrium for each time interval can become computational expensive. In fact for large case studies like the London network used in the following chapter, one might encounter unacceptably long run times. Section 7.3 therefore explains how the run-time of the software can be significantly reduced by exploiting the sparseness of the transition matrix.
7.2. **Software structure**

7.2.1. **Automatic creation of nodes and arcs**

The network representation as defined in Chapter 4 (Figure 4-2) and Chapter 6.3 (with fail-to-access nodes in Figure 6-4) requires a multiplicity of nodes and arcs which would make it very tedious to create such a network manually. All the nodes can however be created by the software once the user has defined the stations, platforms, walking arcs and transit lines. Table 7-1 shows the model input data and Table 7-2 explains how the nodes can then be created by the software. Once the nodes have been created also arcs can be created by the software itself (Table 7-3).

<table>
<thead>
<tr>
<th>Input File</th>
<th>Data description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Station Data</td>
<td>Defines stations, including the number and names of their platforms</td>
</tr>
<tr>
<td>Platform Data</td>
<td>Access restrictions for each platform (if required)</td>
</tr>
<tr>
<td>Transit Data</td>
<td>Frequency, capacity of lines (plus number of seats if required)</td>
</tr>
<tr>
<td>Transfer Arc Data</td>
<td>Defines walking arcs (From: Station, Platform, To: Station, Platform) and the walking time</td>
</tr>
<tr>
<td>OD Data</td>
<td>Defines the OD-Matrix for all time intervals</td>
</tr>
<tr>
<td>Line Data</td>
<td>Defines the Station, Platform which the line serves including the travel time (one file for each line)</td>
</tr>
</tbody>
</table>
### Table 7-2 Node Creation

<table>
<thead>
<tr>
<th>Node Type</th>
<th>Creation Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop</td>
<td>Create one for each platform (as defined in Station Data)</td>
</tr>
<tr>
<td>Fail-To-Access</td>
<td>Create one for each platform (as defined in Station Data)</td>
</tr>
<tr>
<td>Origin</td>
<td>Create one for each station (as defined in Station Data)</td>
</tr>
<tr>
<td>Destination</td>
<td>Create one for each station (as defined in Station Data)</td>
</tr>
<tr>
<td>Boarding</td>
<td>Create one for each transit arc. Station and Platform are the same as the</td>
</tr>
<tr>
<td>Fail-To-Board</td>
<td>origin of the transit arc (as defined in Line Data)</td>
</tr>
<tr>
<td>Alighting</td>
<td>Create one for each transit arc. Station and Platform are the same as the</td>
</tr>
<tr>
<td></td>
<td>destination of the transit arc (as defined in Line Data)</td>
</tr>
</tbody>
</table>

### Table 7-3 Arc Creation

<table>
<thead>
<tr>
<th>Arc Type</th>
<th>Creation Rule</th>
<th>Origin Node</th>
<th>Destination Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line</td>
<td>Defined as in Line Data files</td>
<td>Boarding</td>
<td>Alighting</td>
</tr>
<tr>
<td>Transit</td>
<td>Defined as in Transfer Arc Data files</td>
<td>Stop</td>
<td>Stop</td>
</tr>
<tr>
<td>On-Board wait</td>
<td>Create one for each line and stop node if the line has an alighting and</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>boarding node. (Node is neither Origin nor Destination of this line)</td>
<td>Alighting</td>
<td>Boarding</td>
</tr>
<tr>
<td>Boarding-demand</td>
<td>Create one for each Fail-to-Board Node</td>
<td>Stop</td>
<td>Fail-to-Board</td>
</tr>
<tr>
<td>Boarding</td>
<td>Create one for each Fail Node</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alighting</td>
<td>Create one for each Alighting Node</td>
<td>Alighting</td>
<td>Stop</td>
</tr>
<tr>
<td>Access</td>
<td>Create one for each Fail-To-Access Node</td>
<td>Origin</td>
<td>Fail-to-Access</td>
</tr>
<tr>
<td>Egress</td>
<td>Create one from each platform to the destination belonging to this station</td>
<td>Stop</td>
<td>Destination</td>
</tr>
<tr>
<td>Fail-to-Board</td>
<td>Create one for each Fail-to-Board Node</td>
<td>Fail-to-Board</td>
<td>“Bin”</td>
</tr>
<tr>
<td>Fail-to-Access</td>
<td>Create one for each Fail-to-Access Node</td>
<td>Fail-to-Access</td>
<td>“Bin”</td>
</tr>
</tbody>
</table>
7.2.2. Flow Chart

Figure 7-1 illustrates the structure of the software. Firstly the network is loaded and the variables are initialised, after which the calculation consists of three loops. The outermost loop is the time step loop. As discussed in Chapter 5 the current software structure calculates the equilibrium solution for each time step separately. If one wants to include cost estimates from future time steps in the calculation of the route choice for the current time interval, a further loop is required where the costs found in subsequent time intervals is considered in the MSA.

The second loop is the MSA for each time interval which requires iterations until the convergence criteria (4-37) is fulfilled. A more detailed structure of the MSA is shown in Chapter 4.7. In the current software the arc volumes are averaged and not the fail-to-board probabilities. Alternatively, one could also average the fail-to-board-probability as the problem is a fixed-point-problem as explained in Chapter 4.6.

The innermost loop is the destination loop. For each destination, the hyperpath search algorithm is applied and the travel time matrix is calculated. The computationally most expensive parts of the software are the calculation of the travel time matrix and the inversion of the transition matrix as both parts require several matrix multiplications. In order to shorten the runtime therefore at the beginning of each destination loop, it is checked whether there is any demand that travels to the destination of concern. If not, the hyperpath search and network loading procedure for this destination can be skipped.
Load Input Data

Create Nodes and Arcs

Initialise variables, Set TI = 0, itera=0, d=0

Set TI = TI +1

Add fail-to access passengers from previous time interval to origin demand

Set itera = itera+1

Set d= d+1

If there is demand with dest[d]?

Yes

Hyperpath Search (calculate \( I_d \))

Calculate Travel Time Matrix

Find Final Nodes

Create \( I_d^T \) (inverse \( I_d \))

Create \( T_d^L \) (inverse \( T_d \))

Load Trips to Nodes and Arcs (OD trips, unfinished trips and fail-to-board trips)

if d is last destination

No

Yes

MSA for arc volumes

Update Fail-to-board probabilities

Check convergence

No

Yes

Output of Line Loads, Fail-to-board probabilities

Calculate Failed-to-board, Failed-to-access and unfinished trips

If TI is last time interval

No

Yes

End

**Figure 7-1** Flow chart of Software Structure
7.3.  Model Run Time Issues

7.3.1. Introduction

In the previous three chapters the CapCon approach has only been tested on a small hypothetical network. For such a network there are no run time issues, however in order to apply the approach to a large scale network as in the case study in the next chapter one needs to consider efficient computation. The multiplicity of arcs and nodes in the CapCon network and the software structure means that this can become a significant problem.

The computational effort required for the hyperpath search depends on the number of arcs in the network, whereas network loading is dependent on the number of nodes. It is found that the network loading process can become especially computational expensive since it requires the inversion of \((I - \Pi_d)\) where \(\Pi_d\) are the destination-specific transition matrices and \(I\) is an identity matrix of appropriate dimension (Eq. 4-20).

Matrix inversion is a \(O(n^3)\) process and therefore the required computational effort increases rapidly for larger networks. For example inverting \((I-\Pi_d)\) for the Zone 1 network of London’s Underground network (62 Stations, 800 nodes) requires 35sec when applying the Gauss-Jordan elimination method, but the inversions of the transition matrix for the full London network (262 Stations, 4442 nodes) takes around 10min. These calculation times are clearly not acceptable, if one considers that inversion is required for each destination, each MSA iteration and each time
interval. However, sparse matrix techniques can be applied to substantially reduce the computation time.

### 7.3.2. Transition matrix structure

The structure of the transition matrix $\mathbf{\Pi}_d$ is extremely sparse, as all elements $\pi_{ij}$ of row and node $i$ are 0 except if node $j$ can be reached via an arc from node $i$ and further this arc is also part of any hyperpath from any origin to the destination $d$. In the London network for example there are never more than three lines served from any platform so that the maximum number of non-zero elements per row is three: At a stop node all three lines might be part of the hyperpath, at all other nodes the number of active arcs will be two or less.

In total for the London network it is found that the $(\mathbf{I}-\mathbf{\Pi}_d)$ matrices have a degree of sparseness of 99.8%. This means that only 0.2% of the matrix elements are non-zero. In the literature this is categorised as an extremely sparse matrix (Duff *et al*, 1986) and using a “standard” matrix inversion algorithm without any further optimisation techniques clearly means wasted computational effort. The matrix has further following attributes:

- All elements off the main diagonal are equal to 1,
- None of the other non-zero elements is smaller than -1,
- The non-zero elements of the matrix are relatively randomly spread over the matrix.
As described in the following, it is found that the multiplication of matrices with this structure can be significantly simplified by utilising the fact that each row has a maximum of three non-zero entries. Therefore the matrix inversion is firstly split up into several matrix multiplications.

7.3.3. Dividing the inversion process

Following Strassen (1969) the computational cost for matrix inversion can be slightly reduced by applying following steps:

- \((\Pi_{AA})^{-1} = \text{Invert } \Pi_{AA}\)
- \(R1 = \Pi_{BA}^* (\Pi_{AA})^{-1}\)
- \(\Delta = \Pi_{BB} - R1*\Pi_{AB}\)
- \(\Delta^{-1} = \text{Invert } \Delta;\)
- \(\Pi^{-1}_{BB} = \Delta^{-1}\)
- \(\Pi^{-1}_{BA} = -\Delta^* R1\)
- \(R2 = (\Pi_{AA})^{-1} \Pi_{AB}\)
- \(\Pi^{-1}_{AB} = - R2 * \Delta\)
- \(\Pi^{-1}_{AA} = (\Pi_{AA})^{-1} - R2^* \Pi^{-1}_{BA}\)

The matrix inversion is divided into multiplications of six smaller matrices plus the inversion of two smaller matrices. These steps can be applied recursively so that for each recursion the two inversions are replaced by 6 multiplications. For very large matrices this can reduce the computational effort from an \(O(n^3)\) process to an \(O(N^{\log_{2.7} N}) = O(N^{2.81})\) process. However, even for the full London network with more than 4000 nodes the improvements are only around 5%.
7.3.4. Matrix reorganisation

Considering the structure of the transition matrix allows for a further simplification of the inversion process. The network structure in Figure 4-2 shows that there are no arcs and therefore no possible direct transitions between any of the three failure node types as well as Origins and Destinations. This means that the first Quadrant of the matrix can be defined as an Identity matrix if the matrix is reordered in such a way that the failure nodes, origins and destinations are the first entries in the transition matrix. This is illustrated in Figure 7-2 where the above mentioned node types are defined as Group A and all other node types are in Group B.

![Diagram of Matrix Reorganisation](image)

**Figure 7-2** Structure of the Matrices (I-Πₙ) after reordering of nodes

The matrix inversion as set out by Strassen (1969) can now be simplified to one matrix inversion and four matrix multiplications, because inversion and multiplication with the Identity matrix \((\Pi_{AA})^{-1}\) is not necessary. The multiplication of the two matrices R1 and R2 as in 7.3.3 is therefore not necessary:

- \(\text{Delta} = \Pi_{BB} - (\Pi_{BA}^* (\Pi_{AA})^{-1})^* \Pi_{AB} = \Pi_{BB} - \Pi_{BA}^* \Pi_{AB}\)
- Invert Delta;
- \(\Pi_{BB}^{-1} = \text{Delta}^{-1}\)
\[
\begin{align*}
\Pi^{-1}_{BA} &= -\Delta \ast (\Pi_{BA} \ast (\Pi_{AA})^{-1}) = -\Delta \ast \Pi_{BA} \\
\Pi^{-1}_{AB} &= -((\Pi_{AA})^{-1} \ast \Pi_{AB}) \ast \Delta = -\Pi_{AB} \ast \Delta \\
\Pi^{-1}_{AA} &= (\Pi_{AA})^{-1} - ((\Pi_{AA})^{-1} \ast \Pi_{AB}) \ast \Pi^{-1}_{BA} = I - \Pi_{AB} \ast \Pi^{-1}_{BA}
\end{align*}
\]

Applying these simplifications however still does not lead to a significant reduction in run time. For the Zone1 London network the inversion of \(\Delta\) requires 5 sec but the matrix multiplications require 28 sec. (Compared to this the inversion of the full Matrix \((I-\Pi_d)\) requires 36 sec). The shift of computational effort from the inversion to the matrix multiplications is however utilised in the following section.

### 7.3.5 Sparse matrix multiplication

Matrix multiplication is a computational expensive process. The general form for multiplying two square matrices \(A\) and \(B\) with dimension \(N\) is

\[
\text{For } i = 1 \text{ to } N \\
\quad \text{For } j = 1 \text{ to } N \\
\quad \quad \text{For } k = 1 \text{ to } N \\
C[i,j] = C[i,j] + A[i,k] \ast B[k,j]
\]

and hence requires \(N^3\) multiplications. If it is however known that matrix \(A\) has a maximum of only \(m\) non-zero entries per row, this can be simplified to a \(mN^2\) process. For this \(\text{IndA}\) is a matrix that contains the indices of the non-zero entries of \(A\) as illustrated in Figure 7-3 for a matrix with dimension 4 and \(m=2\).

\[
A = \begin{pmatrix}
  x & x \\
  x & x \\
  x & x
\end{pmatrix} \quad \quad \text{IndA} = \begin{pmatrix}
  1 & 3 \\
  3 & -1 \\
  1 & 4 \\
  2 & -1
\end{pmatrix}
\]

**Figure 7-3** Definition of IndA Matrix
Then the matrix multiplication can be simplified to

\[
\begin{align*}
\text{For } i = 1 \text{ to } N \\
\text{For } j = 1 \text{ to } N \\
\text{For } k = 1 \text{ to } m \\
C[i,j] &= C[i,j] + A[i,\text{IndA}[i,k]] \times B[\text{IndA}[i,k],j] \quad \forall \text{IndA}[i,k] \neq -1
\end{align*}
\]

For the London Zone 1 network this means a reduction in computational time from 28 sec to 5 sec.

In the same way as for matrix multiplication, the above technique can also be applied to the multiplication of matrices with vectors, as in Equation 4-23. However, the cost saving is only $mN$ compared to $N^2$ which does not lead to a notable reduction in model run-time.

The same technique can also be applied for the calculation of the travel matrices. Savings in calculation over 80% can be achieved for the computation of $TT^{\text{new}}$, $TT$ and $NP$ (Steps 5 to 7 in Algorithm 5.4.2).

### 7.3.6. Optimised matrix inversion

In summary, the calculation time for the inversion of the transition matrix $\Pi_d$ for the inner London network could be reduced from 36 sec to 3 sec. This could be achieved firstly by restructuring the transition matrix, and secondly by applying the matrix inversion as set out by Strassen in an iterative way (i.e. for each inversion of the submatrices (Delta and R2) the Strassen procedure is again applied. The matrix multiplications in the Strassen procedure are optimised through sparse matrix
multiplications. The main savings in calculation time are through these sparse matrix multiplications.

**7.3.7. Enumeration of moves**

Instead of optimising the matrix inversion process as described in the previous sections, the inversion can also be avoided by enumerating the moves of the passengers to reach their destination. Equation 4-20 explained that the inversion process is equal to summing up the transition matrices \( \Pi_0, \Pi_1, \Pi_2, \ldots \) until all passengers have reached their destination after \( k \) moves.

\[
\Pi_d^{k \to \infty} = \Pi_0^d + \cdots + \Pi_d^k + \cdots = (I - \Pi_d)^{-1} - I \tag{4-20}
\]

With the algorithm below, the inversion can be calculated “step-wise”. The termination rule, \( \Pi_d^k = 0 \), means that all passengers have reached their destination.

**Step 1 (Initialisation) :** \textbf{Set} \textbf{Inv} and \( \Pi_d^k \) equal to \( \Pi^0 \);

**Step 2 (Update \( \Pi_d^k \)) :** \textbf{Calculate} \( \Pi_d^k \leftarrow \Pi_d^k \ast \Pi^0 \)

**Step 3 (Update \( \Pi_1 \)) :** \textbf{Calculate} \( \Pi_1 \leftarrow \Pi_1 + \Pi_d^k \)

**Step 4 (Termination) :** Terminate if \( \Pi_d^k = 0 \), otherwise return to Step 2

The algorithm requires one matrix multiplication and one matrix inversion for each iteration, which would make it computational very expensive. However, the same sparse matrix multiplication as described in Section 7.3.5 can be used since \( \Pi^0 \) is a sparse matrix. Whether this algorithm performs better than the matrix inversion depends on the network structure. For the London inner zone network around 60 iterations for destinations at the edge of the network and a few less iterations for destinations that are located in the centre of the network are needed. The algorithm
will be more interesting for networks with “long-distance arcs” or if passengers only travel short distances as then the termination criterium $\Pi_k^d = 0$ will be fulfilled already after a few iterations.

Further, the algorithm becomes interesting if short time intervals are modelled. In this case, the number of moves the passengers can make within one time interval is limited. For example, if all line arcs are two minutes or longer and the duration of the time interval is 10 minutes, the passengers can travel a maximum of 5 line arcs. This means a maximum of 9 node transitions (Boarding Node → Alighting Node → Boarding Node → Alighting Node → …) if the passenger is not interchanging and a few more moves if passengers are interchanging between lines (Boarding Node → Alighting Node → Stop Node → Fail-to-Board Node → Boarding Node → …).

### 7.4. Summary and discussion

The complex network representation of the CapCon network means that the node and arc creation can not be done by the user for large networks. Chapter 7.2 discussed an efficient solution for the automatic creation of nodes and arcs.

The network representation and the software structure consisting of three nested loops for which the computationally expensive network loading process needs to be carried out inside the inner most loop further means that the computational effort increases rapidly for large networks. Chapter 7.3 explained however that the
computational effort can be significantly reduced through a combination of matrix reorganisation and the application of sparse matrix multiplication techniques.

For the Inner London network case study presented in the next chapter this means that one iteration of the MSA takes around 3min (72 destinations and 2.5sec for hyperpath search and network loading for each destination). This is still a long run time if one considers that the MSA might require 50 or more iterations for convergence and that this equilibrium needs to be calculated for several time intervals. In the following case study four time intervals are congested and require several iterations so that the total calculation time was 10hours with a Pentium 4 PC.
8. Crowding and Capacity Problems in London

8.1. Introduction

The previous chapters applied the CapCon model only to a small example network. This was useful in order to illustrate the theory of capacity constrained assignment. In this chapter the assignment model is applied to a larger network, in this case the London Underground network. While the case study might be of interest in itself, the main objective is to illustrate that the model presented can also be used for practical applications. The results of the assignment are compared with observed data but some discrepancies have to be expected because of some rough estimation in the model due to unavailable data.

Firstly, the available data are described which also shows the assumptions that have to be made in the assignment. Then the static model is applied and compared to assignment results by London Underground. Finally, the demand data is split into shorter time intervals and the dynamic model is applied. A summary of the dynamic assignment model and this case study is also published in Schmöcker and Bell (2006).
8.2.  *Data description*

8.2.1.  *Network data*

For this case study only the inner zone of the London network is used because this is the part of the network where congestion is most severe and passengers have choices to avoid congestion by using other routes. (On several lines outside the inner zone crowding is also severe however passengers often have no route choice in this part of the network so that it is of less interest for this case study.) Besides run time issues as discussed in the previous chapter, taking only the inner zone has the further advantage that fare effects can be ignored (if the whole network is taken into account, one needs to consider that some passengers will choose to travel longer paths in order to avoid the premium fare of Zone 1). A third benefit of reducing the network to the inner zone is that the complex schedule for some of the outer branches of the lines can be ignored. In particular Northern Line and the District Line services terminate at different stations but also on the other lines often a service does not start or end at the terminals shown in Figure 8-1. London Underground made only data on the frequency of the services in the central part of the network available for this project.
The inner London network consists of 56 Stations which are served by 11 transit lines. The Northern Line however consists of two separate branches in the city centre (and even more branches in the outer zone) and also the District Line consists of several branches.

Modelling the inner zone requires us to model three District Line branches as separate lines: Services from West London (Richmond, Ealing Broadway) destined for East London (Tower Hill, Upminster), services from Wimbledon destined for Edgware Road and services travelling from Wimbledon eastbound via Victoria. All branches of the District Line travelling via Victoria merge at Earl’s Court. The Northern line consists of several branches in North London (Edgware, Mill Hill East, High Barnett) and also in South London several stations are used as terminals (in
particular Kennington and Morden). Further to this, services to all destinations can travel via two different branches in the central part of London, the “City branch” via Bank and the “West End branch” via Charing Cross. Modelling the inner zone requires us to model these two different branches as different lines. This means that in total 14 transit lines are modelled with 297 line arcs. Further, 102 transfer arcs are modelled to allow for walking between different platforms of the same station.

At several stations in the network, different lines depart from opposite sites of the same platform. In this case it is assumed that the common line problem applies. For example at Baker Street Station this is the case for the southbound Jubilee Line and the southbound Metropolitan Line. It is assumed that some passengers who are waiting on the platform will take whichever line arrives first (if both lines are included in their optimal hyperpath). In some cases where the platform is narrow this assumption is reasonable, in other cases this assumption might be more risky, for example some platforms are wider and the view between the two sides of the platform is obstructed. At other stations again, the common line issue fully applies as different lines depart from the same platform, at South Kensington for example two District Line branches and the Circle Line.

London Underground made the following network and service data available for this case study: Run time on arcs, estimated walking times between different platforms, scheduled service frequency and capacity of the different trains serving the lines. The impact of passenger crowding on this data is not known.
The capacity of a service is of course difficult to measure as it will be perceived differently by different passengers. At some degree of crowding some passengers will abstain from attempting to board and rather wait for a subsequent service, whereas others will still attempt to board this service. London Underground distinguishes between “Realistic crush capacity” and “Absolute crush capacity”. The former assumes 4 passengers per m² and the latter 7 passengers per m². In the following, realistic crush capacity is used as London Underground assumes this reflects better passenger behaviour in central London. The “absolute crush capacity” is the guideline given by the manufacturer.

8.2.2. Demand Data

An OD matrix for the whole network for different times of the day is obtained by London Underground through the Rolling Origin and Destination Survey (RODS). This survey is based on a questionnaire that is distributed to around 20% of all passengers at a number of stations. The questionnaire asks passengers about their origin, destination, trip purpose, points of interchange as well as ticket types used (London Underground Limited, 2005; Maundrey, 2005). Each year around 10% of the stations are covered, with some busier stations being sampled more often. This means that within a 15 year cycle all stations are covered at least once. The return rate of questionnaires is 25% so that in total 5% of all passengers are interviewed. The results are then multiplied according to the return ratio in order to obtain the network OD matrix. In order to develop a robust estimate the OD matrix was only constructed for larger time intervals (early morning, am peak, midday, pm peak, evening and late evening). For this case study the data for the morning peak are taken. This is defined by London Underground as 0700 to 1000.
8.2.3. Reduction of network OD matrix to Inner Zone OD matrix

OD data were available for the whole network only, so that the OD matrix for the inner zone needs to be estimated. Reducing the OD matrix to the inner zone means that five types of trips need to be taken into account:

- Passenger trips starting and finishing in the inner zone (inner zone trips)
- Passenger trips with origin and destination outside the inner zone, and which do not pass through the inner zone (outer zone trips).
- Passenger trips originating outside the inner zone and with destination inside the inner zone (inbound trips)
- Passenger trips with origin inside the zone and destination outside the inner zone (outbound trips)
- Passenger trips with origin and destination outside the inner zone, but which pass through the inner zone (through trips)

These trip types can be distinguished through an assignment of all trips to the whole network. For this purpose the whole network has been coded in the CapCon network but the capacity constraints have not been set active (in order to have no passengers failing to board). The outer zone trips do not have to be considered and the inner zone trips can be added to the new OD matrix without any difficulties. The original am peak OD matrix for the full London network consisted of 846934 trips. Because of the deletion of the outer zone trips this is reduced to 636904 trips for the inner zone OD matrix.
In order to reflect the origins and destinations of the inbound, outbound and through trips 18 new line terminals are added to the inner zone network as shown in Figure 8-2. For example trips originating in North London and entering the inner zone via the Piccadilly Line are now assumed to originate at “Piccadilly North” and similarly trips leaving the inner zone say on the westbound Central Line are assigned “Central West” as a new destination.

Figure 8-2 The London network used for this case study

For the through trips following problem has to be considered: In most cases all passengers with the same OD pair are entering and leaving the zone via the same routes. In some cases this is however not the case because of common lines. Some passengers travelling from Mile End to Ealing Broadway for example might travel on the Central Line (in which case their origin and destination in the reduced OD matrix should be “Central East” and “Central West”, see Figure 8-2) whereas others travel via the District Line (in which case the trips should be recorded in the reduced OD matrix with origin “District/H&City East” and destination “District West”). This is illustrated in Figure 8-3 where passengers from an Origin O might travel via
different routes to their destination $D$ and in particular they enter the inner zone at different stations $O^*_i$ and leave the network at different destinations $D^*_j$.

![Diagram showing different routes to their destination and inner zone stations](image)

**Figure 8-3** Reduction of the full OD matrix to the OD matrix for the inner zone

The hyperpath search and assignment procedure described in Chapter 4 explained that for each destination a transition matrix $\Pi_d$ is calculated with elements $\pi_{ij}$ denoting the probability of passing a node $j$ when the passenger is at node $i$ and is destined for $d$. Therefore the OD trips for the through trips for the reduced network can be calculated by (8-1) where $i$ and $j$ are the new line terminals.

$$\begin{align*}
(o^*_i, d^*_j) &= \sum_{OD} \sum_{j \in D^*_j} \sum_{i \in O^*_i} O_{ij} \cdot \pi_{O_i, O_j, D} \cdot \pi_{O^*_j, D^*_j | D} \\
&= \sum_{OD} \sum_{j \in D^*_j} \sum_{i \in O^*_i} O_{ij} \cdot \pi_{O_i, O_j, D} \cdot \pi_{O^*_j, D^*_j | D} 
\end{align*}$$

(8-1)

In order to analyse the fit of the reduced OD matrix, the line flows on the arcs entering and exiting the inner zone are compared with the observed values of London Underground (Table 8-1) for these arcs. This is because the CapCon flows on these arcs comprise the new OD matrix (plus inner zone trips). Table 8-1 shows that the fit is acceptable for this case study with a standard deviation of 1700pas for the three
hour period. The standard deviation is given by (8-2) where $x_i$ are the estimated flows and $x_o$ are the estimated flows by London Underground.

$$s = \sqrt{\frac{1}{N} \sum_{i} \left( x_i - x_o \right)^2}$$ (8-2)

It should further be noted, that the CapCon flows should be expected to be higher than the observed flows (confirmed in Table 8-1). This is because the assignment is based on the OD matrix with all trips that originate between 7 and 10am. Some of these trips will however not have been completed or even entered the inner zone by 10am, which explains why the CapCon results are expected to be slightly higher. Some errors in the CapCon flows might further be due to errors in the initial OD matrix.

<table>
<thead>
<tr>
<th>Line (IN)</th>
<th>LuL Flows</th>
<th>CapCon</th>
<th>Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bakerloo West</td>
<td>13935</td>
<td>15384</td>
<td>1449</td>
</tr>
<tr>
<td>Metropolitan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>North</td>
<td>20338</td>
<td>22606</td>
<td>2268</td>
</tr>
<tr>
<td>Jubilee North</td>
<td>25954</td>
<td>26133</td>
<td>179</td>
</tr>
<tr>
<td>Northern North</td>
<td>38518</td>
<td>39916</td>
<td>1398</td>
</tr>
<tr>
<td>Piccadilly North</td>
<td>27631</td>
<td>27373</td>
<td>-456</td>
</tr>
<tr>
<td>Victoria North</td>
<td>42435</td>
<td>42988</td>
<td>553</td>
</tr>
<tr>
<td>Central East</td>
<td>40362</td>
<td>43794</td>
<td>3432</td>
</tr>
<tr>
<td>District, H.&amp;City East</td>
<td>22751</td>
<td>18454</td>
<td>1642</td>
</tr>
<tr>
<td>Jubilee East</td>
<td>22881</td>
<td>24840</td>
<td>1959</td>
</tr>
<tr>
<td>Northern South (Bank)</td>
<td>19477</td>
<td>18946</td>
<td>-531</td>
</tr>
<tr>
<td>Bakerloo South</td>
<td>3805</td>
<td>3375</td>
<td>-431</td>
</tr>
<tr>
<td>Northern South (Waterloo)</td>
<td>11670</td>
<td>11632</td>
<td>-38</td>
</tr>
<tr>
<td>Victoria South</td>
<td>29117</td>
<td>30797</td>
<td>1680</td>
</tr>
<tr>
<td>District South</td>
<td>19558</td>
<td>20529</td>
<td>967</td>
</tr>
<tr>
<td>Piccadilly West</td>
<td>23591</td>
<td>24499</td>
<td>908</td>
</tr>
<tr>
<td>District West</td>
<td>15087</td>
<td>18265</td>
<td>3178</td>
</tr>
<tr>
<td>Central West</td>
<td>19550</td>
<td>21061</td>
<td>1511</td>
</tr>
<tr>
<td>H.&amp;City West</td>
<td>6697</td>
<td>6925</td>
<td>28</td>
</tr>
<tr>
<td>Total</td>
<td>403757</td>
<td>417513</td>
<td>13756</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Line (OUT)</th>
<th>LuL Flows</th>
<th>CapCon</th>
<th>Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bakerloo West</td>
<td>4610</td>
<td>5086</td>
<td>476</td>
</tr>
<tr>
<td>Metropolitan</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>North</td>
<td>3971</td>
<td>4021</td>
<td>50</td>
</tr>
<tr>
<td>Jubilee North</td>
<td>6916</td>
<td>6737</td>
<td>-176</td>
</tr>
<tr>
<td>Northern North</td>
<td>16235</td>
<td>16191</td>
<td>-44</td>
</tr>
<tr>
<td>Piccadilly North</td>
<td>5467</td>
<td>3862</td>
<td>-1605</td>
</tr>
<tr>
<td>Victoria North</td>
<td>8439</td>
<td>12149</td>
<td>3710</td>
</tr>
<tr>
<td>Central East</td>
<td>7983</td>
<td>8492</td>
<td>509</td>
</tr>
<tr>
<td>District, H.&amp;City East</td>
<td>4997</td>
<td>3784</td>
<td>-1213</td>
</tr>
<tr>
<td>Jubilee East</td>
<td>23560</td>
<td>25255</td>
<td>1695</td>
</tr>
<tr>
<td>Northern South (Bank)</td>
<td>7998</td>
<td>7912</td>
<td>-86</td>
</tr>
<tr>
<td>Bakerloo South</td>
<td>3079</td>
<td>2665</td>
<td>-414</td>
</tr>
<tr>
<td>Northern South (Waterloo)</td>
<td>1809</td>
<td>1223</td>
<td>-586</td>
</tr>
<tr>
<td>Victoria South</td>
<td>11683</td>
<td>13084</td>
<td>1401</td>
</tr>
<tr>
<td>District South</td>
<td>8534</td>
<td>9537</td>
<td>3</td>
</tr>
<tr>
<td>Piccadilly West</td>
<td>11692</td>
<td>9617</td>
<td>-2075</td>
</tr>
<tr>
<td>District West</td>
<td>11213</td>
<td>14698</td>
<td>3495</td>
</tr>
<tr>
<td>Central West</td>
<td>10183</td>
<td>9082</td>
<td>-1101</td>
</tr>
<tr>
<td>H.&amp;City West</td>
<td>3063</td>
<td>4278</td>
<td>1215</td>
</tr>
<tr>
<td>Total</td>
<td>151431</td>
<td>156663</td>
<td>5232</td>
</tr>
</tbody>
</table>
8.3. 3 hour model runs

8.3.1. Comparison with flows estimated by London Underground

In a first assignment the observed line loads by London Underground are compared with the line loads when applying the static CapCon model. London Underground estimated the line loads through the information from the RODS survey. Not all passengers gave information about their interchange points so that the route choice for these journeys had to be estimated. London Underground did this by assuming passengers take the shortest path. If there is a choice between lines the split between them is estimated by a logit function. The procedure is not nearer described in any available document and needs improvement as mentioned by Maundrey (2005).

As one might expect the errors for the arc flows are larger compared to the errors for the boarder node in- and outflows because the errors from the estimation of the calculation of the reduced OD matrix are carried over. This is reflected in a larger standard deviation of 2834pas. Figure 8-4 illustrates the fit of observed and estimated flows. Shimamoto (2006) estimated the model fit with nearly the same data set for different values of walking time (on transfer arcs) and waiting time at platforms. The best model fit was obtained by assuming a walking time and waiting time value of 1.5 times the value of in-vehicle travel time. Shimamoto’s results show however only a very marginally improvement in model fit (less than 1%) and the improvement could not be confirmed with this data set.
8.3.2. Comparison of arc flows with and without consideration of common lines

In order to quantify the effect of the common line problem the assignment has been carried out with and without consideration of common lines. The run without consideration of common line effects increases the std deviation from 2834pas to 3652pas, meaning that the model fit is worse. Table 8-2 shows that an assignment without common lines concentrates the line loads more on the line with the highest frequency, whereas an assignment with common lines distributes the line loads more. This explains why assignments without consideration of hyperpaths (i.e. schedule-based assignment in high frequency networks) sometimes overestimate the line loads of busy lines.

Figure 8-4 Observed and assigned flows for the morning peak period
Table 8-2 Example of differences in arc flows for assignment with and without common lines

<table>
<thead>
<tr>
<th></th>
<th>From</th>
<th>To</th>
<th>Visitor Estimate</th>
<th>With Common Lines</th>
<th>Without Common Lines</th>
</tr>
</thead>
<tbody>
<tr>
<td>King's Cross</td>
<td>Farringdon</td>
<td>Metropolitan</td>
<td>15606</td>
<td>19629</td>
<td>26347</td>
</tr>
<tr>
<td>King's Cross</td>
<td>Farringdon</td>
<td>Circle</td>
<td>6166</td>
<td>4457</td>
<td>1214</td>
</tr>
<tr>
<td>King's Cross</td>
<td>Farringdon</td>
<td>H. &amp; City</td>
<td>7525</td>
<td>6626</td>
<td>3268</td>
</tr>
<tr>
<td>Mansion House</td>
<td>Cannon Street</td>
<td>Circle</td>
<td>3253</td>
<td>2527</td>
<td>927</td>
</tr>
<tr>
<td>Mansion House</td>
<td>Cannon Street</td>
<td>District</td>
<td>12072</td>
<td>10933</td>
<td>12544</td>
</tr>
</tbody>
</table>

The effect of common line consideration is of course dependent on the network structure as well as the assignment parameters. Consideration of stochastic effects will reduce the common line effects, as does the consideration of line crowding. If the assignment considers that passengers are avoiding crowded lines, this will also lead to a better distribution of line loads.

Figure 8-5 illustrates the effect of common lines in the London network. As expected, the effect is largest on line sections with parallel lines or lines that are summarised as “trunk lines” in Dial (1967). However it is important to note that some effects can also be found on other line sections, e.g. the two Northern Lines branches.
Modelling the three hour peak period leads to results showing no stations with demand above capacity. Table 8-3 shows the line loads of the most congested line sections within Zone 1 of the London network (excluding arcs from the new terminals). Several line loads of the Victoria Line are shown, which is the line carrying most passengers during the morning. Further the highest line load of each of the other lines is listed. The table shows the flows as estimated by London Underground (LuL) and as estimated by the model described in this paper. In general the line loads between the model and London Underground estimates correspond well. The highest congestion is encountered on the westbound Central Line. But also on this line there is sufficient spare capacity on all line sections according to the 3 hour assignment. On none of the line sections is there more than 80% of the (realistic crush) capacity used. Because there are no passengers failing to board, the setting of the risk-averseness parameter in the generalised cost function becomes irrelevant.
8.4. Dynamic runs (15minute model)

8.4.1. Demand assumption

In the following it is assumed that the demand matrix is peaked within the morning peak period. The three hour demand is divided into twelve 15min time intervals. Estimated line loads of London Underground show the highest line flows between 0830 and 0900 (Figure 8-6) so that the following distribution of the 3hour morning peak demand might be a rough estimation in the absence of OD data for smaller time intervals than 3 hours.
8.4.2. Effect of risk-averseness

The following illustrates the assignment results for the dynamic assignment with twelve 15min intervals, firstly, assuming that passengers do not consider platform crowding (θ_d=0) and secondly, with the assumption of risk-averse passengers (θ_d > 0). The conversion of the model was relatively quick as the line flows and failure probabilities became stable after around 20 iterations. In the following results the number of iterations was fixed at 60.
In contrast to Table 8-3, the line loads on several arcs are now reaching the available capacity during several time intervals. Table 8-4 illustrates that the model estimate of demand at a number of stations will exceed the available capacity between 0800 and 0900. At Euston station for example, 708 passengers would like to board the Victoria Line, however there is only space for 696 passengers to board (212 spaces through passengers alighting at Euston), meaning that 12 passengers (2% of the boarding demand) will not be able to board the line. At London Bridge the capacity shortage is even more severe with 10% of the passengers waiting to board not being able to do so. The next three stations in Table 2 are all line terminals. This means that capacity problems already occur before the services enter the Inner Zone of the London network. Especially on the District Line services coming from Wimbledon (in the reduced network model “District South”) are already full when they arrive at Earl’s Court. Passengers at previous stations might fail-to-board because the train is full on arrival and nearly none of the passengers are alighting before Earls’ Court in the morning peak. The table also shows that the line load and boarding demand is equally split between the two District Line branches, as in this case passengers will board whichever train comes first and, if necessary change to their preferred branch later at a less congested station.

Table 8-4 Stations where demand exceeds capacity ($\theta_d=5\text{, }0830-0845$)

<table>
<thead>
<tr>
<th>Station</th>
<th>Line</th>
<th>q</th>
<th>Board</th>
<th>Alight</th>
<th>Boarding Demand</th>
<th>Excess Demand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euston</td>
<td>Victoria (southb.)</td>
<td>0.02</td>
<td>696</td>
<td>212</td>
<td>708</td>
<td>12</td>
</tr>
<tr>
<td>London Bridge</td>
<td>Northern (northb., via Bank)</td>
<td>0.10</td>
<td>450</td>
<td>469</td>
<td>545.7</td>
<td>55.4</td>
</tr>
<tr>
<td>District South</td>
<td>District, (eastb., via Victoria)</td>
<td>0.18</td>
<td>1231</td>
<td>0</td>
<td>1502.7</td>
<td>271.7</td>
</tr>
<tr>
<td>District South</td>
<td>District, (northb., to Edgware Rd)</td>
<td>0.18</td>
<td>1231</td>
<td>0</td>
<td>1502.7</td>
<td>271.7</td>
</tr>
<tr>
<td>Central East</td>
<td>Central (westb.)</td>
<td>0.11</td>
<td>5547</td>
<td>0</td>
<td>5206.6</td>
<td>669.6</td>
</tr>
</tbody>
</table>
Figure 8-8 shows the total number of passengers failing to board during the 12 time intervals for varying $\theta_d$. During the early time intervals, no passengers fail to board; only during the peak of the peak capacity problems occur. Clearly the capacity problems are highest between 0830 and 0845. After 0900, the network can serve all the demand again. The difference between the curves shows that if one assumes risk-averse passengers, far less passengers fail to board.

In the case of $\theta_d=1$, the sum of the total line loads during the time interval 0830 to 0845 is 5570 passengers higher than if $\theta_d=0$ is assumed, showing that some passengers are taking detours in order to avoid overcrowded stations. In case of $\theta_d>0$ the total line flows are increasing in the time intervals between 0800 and 0900am. This is because more risk-averse behaviour will lead passengers to choose longer routes in order to avoid failing to board. Between 0900 and 0930am the total line flows are however lower because less passengers failed to board in previous time intervals and therefore there are less passengers in the network. The total line flows
over the whole simulation period are increased for all $\theta_d > 0$ illustrating the effect of passengers taking longer but less congested routes. It is perhaps surprising to see the total line loads for $\theta_d=20$ being reduced compared to $\theta_d=10$ and $\theta_d=5$. The reason is however the complex interaction between the route choice of passengers from different OD pairs: Risk-averseness leads some passengers to take longer routes, which allows other passengers to take more direct routes without risking to fail-to-board.

![Figure 8-9](image)

**Figure 8-9** Comparison of total line flows for different $\theta_d$

### 8.4.3. Comparison of capacity constrained assignment and assignment with crowding effects

In order to illustrate the benefit of the approach presented the line loadings under consideration of capacity constraints and when these are not considered are compared. Additionally, the line loadings are calculated when capacity constraints are only indirectly considered through crowding effects. In this assignment no passengers are assumed to fail to board but passengers are deterred from crowded lines through a crowding penalty. Crowding penalties as in Eq. (6-5) are assumed
with parameters estimated by London Underground (London Underground Limited, 1988). The additional costs for sitting and standing in a crowded car are found to be:

\[ c_{a,sit}(x) = c_a \left[ 1 + 0.17y(x) + 0.385y(x)^2 \right] \]  
\[
(8-3)
\]

\[ c_{a,stand}(x) = c_a \left[ 1.4 + 0.84y(x) + 0.44y(x)^2 \right] \]  
\[
(8-4)
\]

with \( y(x) \) defined as \( y(x) = \left[ \frac{x - seats}{cap - seats} \right] \). The line loads resulting from this assignment are called “Crowding function” in Table 8-5 and the assignment without capacity constraints and without crowding costs is referred to as “No crowding costs”.

As expected the line loads on some lines are above the capacity in the assignment with “no crowding costs”. Table 8-5 shows those arcs that have line flows higher than capacity in the assignment without consideration of crowding costs and capacity constraints. The table shows that the capacity constraint assignment observes the capacity constraints. Further it illustrates that the consideration of crowding costs does not guarantee the observation of capacity constraints. In Table 8-5 four arcs with a line load above capacity are highlighted. One arc (London Bridge to Bank/Monument on the Northern Line) could be avoided by passengers, the three other arcs are outgoing arcs from the line terminals where passengers cannot avoid boarding the crowded line by choosing an alternative route.
Table 8-5 Selected line loads for assignment with and without capacity constraints during 0830-0845

<table>
<thead>
<tr>
<th>Out Station</th>
<th>In Station</th>
<th>Line</th>
<th>Line capacity</th>
<th>( \text{CapCon} ) ( \text{thetas}_{d1} )</th>
<th>Crowding function</th>
<th>No crowding costs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euston</td>
<td>WarrenSt</td>
<td>Victoria (southb.)</td>
<td>5831</td>
<td>5832</td>
<td>5667</td>
<td>6077</td>
</tr>
<tr>
<td>DisSouth</td>
<td>EarsCourt</td>
<td>District (eastb.)</td>
<td>1231</td>
<td>1232</td>
<td>1320</td>
<td>1320</td>
</tr>
<tr>
<td>DisSouth</td>
<td>EarsCourt</td>
<td>District (northb.)</td>
<td>1231</td>
<td>1232</td>
<td>1320</td>
<td>1320</td>
</tr>
<tr>
<td>CentEast</td>
<td>LiverpoolSt</td>
<td>Central (westb.)</td>
<td>5547</td>
<td>5550</td>
<td>5763</td>
<td>5763</td>
</tr>
<tr>
<td>LiverpoolSt</td>
<td>BankMon</td>
<td>Central (westb.)</td>
<td>5547</td>
<td>5426</td>
<td>5229</td>
<td>5564</td>
</tr>
<tr>
<td>LondonBr</td>
<td>BankMon</td>
<td>Northern (northb.)</td>
<td>3103</td>
<td>3101</td>
<td>3475</td>
<td>3491</td>
</tr>
<tr>
<td>BankMon</td>
<td>Moorgate</td>
<td>Northern (northb.)</td>
<td>3103</td>
<td>2769</td>
<td>2637</td>
<td>3103</td>
</tr>
</tbody>
</table>

8.5. **Capacity reducing effects and an assignment example**

8.5.1. Consideration of service irregularities

The assignment so far assumed that the service operates normally with the scheduled frequency. However, train cancellations and service irregularities often occur in London leading to longer waiting times and capacity reductions. The website of London Underground publishes regularly performance statistics which show among other things the percentage of operated train kilometres compared to the scheduled train-kilometres. Table 8-6 shows this statistics for the period July and August 2005.

Adeney and Schmöcker (2004) looked into accident causes and the resulting delays for several metros in an RTSC benchmarking project. They report that after the cause of the incident has been resolved, capacity problems near the location of the incident and knock-on effects might cause problems all along the line through longer dwell times and the resulting “bunching effect”. This in fact means a reduction in capacity. Discussions with London Underground revealed that even small peak-hour delays are often un-recoverable until after the peak period and lead to enormous capacity

183
problems and delays for the whole line. Further, delays often spread to other lines that are connected with the primarily affected line through one or more major interchanges.

Table 8-6 Scheduled peak service km operated in July and August 2005; Source: London Underground (2005b)

<table>
<thead>
<tr>
<th>Line</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bakerloo</td>
<td>96.3%</td>
</tr>
<tr>
<td>Central</td>
<td>92.0%</td>
</tr>
<tr>
<td>Circle&amp;H.&amp;City</td>
<td>85.4%</td>
</tr>
<tr>
<td>District</td>
<td>91.1%</td>
</tr>
<tr>
<td>Jubilee</td>
<td>95.9%</td>
</tr>
<tr>
<td>Metropolitan</td>
<td>96.1%</td>
</tr>
<tr>
<td>Northern</td>
<td>93.3%</td>
</tr>
<tr>
<td>Piccadilly</td>
<td>93.3%</td>
</tr>
<tr>
<td>Victoria</td>
<td>94.8%</td>
</tr>
<tr>
<td>W&amp;H&amp;City</td>
<td>96.0%</td>
</tr>
<tr>
<td>Network</td>
<td>93.2%</td>
</tr>
</tbody>
</table>

8.5.2. London Underground's research on crowding behaviour

In order to better understand the impact of crowding on passenger behaviour London Underground carried out their own study. Their research consisted of platform observations and in-depth interviews with passengers travelling from two different stations. The stations chosen for the survey are Seven Sisters on the Victoria Line and Kennington on the Northern Line. The interesting point about these two stations is that the morning peak service consists of alternate empty and crowded trains. Therefore commuters who know about this timetable can trade off between boarding a crowded service and waiting for an emptier service. The site observations were carried out by counting the number of passenger on-board the incoming train, the number of passengers alighting, the number of passengers boarding and the number of passengers staying behind to wait for an emptier train. In addition passengers on
the platform were interviewed for around two minutes to understand if they are aware of the timetable and to find out how often they let a crowded train go.

Figure 8-10 illustrates the observations at Seven Sisters. The Victoria Line trains have eight carriages and at Seven Sisters the platform entrances are opposite doors 2 and 7. The figure shows that there is some passenger mingling but that a higher percentage of passengers is waiting at the doors then at other places along the platform. Further, the proportion of passengers not boarding is significantly lower at the cars near the platform. London Underground’s explanation is that “this suggests that people who intend to move are also likely to move to a more comfortable area of the platform.” It might further be that a larger percentage of these passenger are aware of the schedule and therefore have a better estimate of the benefit and waiting time increase associated with letting a train go.

![Figure 8-10 Differences in Passenger behaviour along the platform (Data taken from London Underground Limited (1988), rearranged)](image)

The findings of London Underground lead to following conclusions for this thesis: Firstly, capacity is often not fully utilised in trains. Some passengers will prefer to
stay behind even though the train is not nearly fully. Also, some carriages will be fuller than others so that some passengers will fail-to-board even though not the whole train is full. The findings further suggest that the assumption of passenger mingling is to some degree true but the fail-to-board probability might also be correlated with the origin and destination of passengers.

8.5.3. Assignment with lower capacity

A third explanation for more congestion than modelled in this case study so far is that at some days the demand will be higher than estimated in the OD matrix. The RODS survey is always carried out on a “typical autumn day” therefore ignoring additional demand that might occur during peak seasons and before and after public holidays.

In order to show model results with higher congestion through service delays, unutilised space or days with higher than usual demand in the following it is assumed that the service capacity and service headway on all lines is reduced to 80%.
Figure 8-11 illustrates at which stations passengers will encounter capacity problems. It can be seen that in this case the capacity problems are severe and passengers at several stations would fail to board. Especially several lines passing through the King’s Cross area, District Line services in the South and West of Central London and the westbound Central Line would suffer. Besides this, the main interchange points between commuter rail and underground are also sensitive to a reduction in capacity. London Bridge, Waterloo station, Victoria station and the area around Paddington all show high fail-to-board probabilities. It is interesting to observe that the most crowded stations are mainly at the edge of the Zone 1 network. This can be explained through the characteristics of the morning demand which are mainly trips from the outer zones into the inner zone.

Figure 8-12 shows the fail-to-board probabilities over the full simulation period for different $\theta_d$. Compared to the previous case, the number of passengers failing to
board is significantly increased and the congestion lasts for a much longer time period. The convergence of the results for higher \( \theta_d \) can also be seen again, indicating that a higher level of risk-averseness will not change the results any further.

![Graph showing total passengers failing to board for different values of \( \theta_d \)](image)

**Figure 8-12** Total Number of passengers failing to board (reduced capacity)

### 8.6. Discussion

In this chapter an application of the CapCon network to the London network is illustrated. It is shown that the approach can be applied to a larger network. However a computation time of several hours for this network size means also that the size of the network is limited. As mentioned in the previous chapter, the computation time is mainly dependent on the number of nodes in the network. It should however also be noted that Shimamoto (2006) successfully applied the static CapCon model to the Kyoto City o bus network, a network with several thousand nodes.
As the model is based on transition probabilities it is for example easy to calculate OD-specific transition probabilities and reliability measures. In this case study, the node transition probabilities have been used to calculate the reduced OD matrix for London’s inner zone.

The results for the London network illustrate the need to consider dynamic effects. Capacity problems do occur in the London network but these are not revealed when one models larger time intervals. Also modelling the interval with the highest congestion only in a static model does not show much congestion, only through the forward-carrying of unfinished trips and trips that failed-to board are capacity problems shown.

The assignment results are plausible as they show congestion around the most important interchange points in the network. However, because of the described data limitations the results should be treated with due scepticism.
9. Conclusions

This study presents a new approach to capacity constrained frequency-based transit assignment that is thought to be an alternative to the effective frequency-based approach. The important difference between these two approaches is that the “CapCon” approach explicitly considers that passengers are failing-to-board and that the excess demand is enumerated and removed from any downstream arcs. This is also the basis for the dynamic extension of the model. In order to determine the excess demand, platform specific fail-to-board nodes and probabilities are introduced. Fail-to-board probabilities have already been used by Last and Leek (1976). However, in their study it is assumed that passengers are “cooperative” so that the available network capacity is fully utilised. In the CapCon model this is instead replaced by the consideration of priority rules and the formulation as a user equilibrium problem which is more realistic.

Further, it is considered that in many situations the common line problem applies and passengers decide for an optimal strategy (hyperpath) instead of deciding for an optimal path. It is shown with the London network that this might indeed have an influence on the assignment results. It is however important to consider in which situations the common line problem applies. Current trends in frequency-based assignment research assume it is as a necessity to assume that it applies. However infrastructure design, count-down information, complex fare structures (not least due to privatisation) all mean that the common lines problem does not always apply. Therefore, this research presented a network notation with and without consideration of common lines. If common lines do not need to be considered the formulation of
the problem can be simplified and the number of nodes significantly reduced leading to savings in memory and run time.

The assignment is carried out under the assumption that passengers might be “risk-averse”. This means that passengers will consider the chance of failing-to-board and reroute (if there is a feasible alternative) in order to avoid the possibility of being delayed through crowding at the platform. This notation of risk-averseness is probably more realistic than the assumption of passengers knowing exactly whether or not they can board the next arriving vehicle, as assumed in schedule-based assignment models with the assumption of FIFO.

In the schedule-based models of Wong and Tong (2001) and Papola et al (2005) it is assumed that from a certain arrival time onwards passengers will have to let the first train go because of overcrowding. In this model this is replaced by the assumption of passenger mingling. This assumes that all passengers on the platform have the same chance of boarding the next service, irrespective of origin or arrival time at the platform. In some cases the assumption of mingling is realistic (e.g. long platforms) in other circumstances (e.g. bus stops) FIFO is a better assumption.

It should be noted that the assumption of FIFO in schedule-based approaches as well as the assumption of mingling in the CapCon model stem from the requirement to find a stable solution and can not be changed easily. It is shown that the CapCon user equilibrium is only a unique solution if mingling is assumed. This is in particular the case if the network has circular lines.
The assumption of mingling is also important for the dynamic extension of the model. In the dynamic version, the passengers who failed to board are re-assigned from the node they failed to board. Further, the passengers who could not complete their journey within one time interval are re-assigned from the last node they reached in the previous time interval. For both groups of passengers it is assumed that they split over the different paths that are part of their hyperpath according to the arc split probabilities. Other assumptions might give the same arc flows, but with deterministic user equilibrium assignment (DUE) only the assumption of passenger mingling assures that passengers split according to arc split probabilities. This is because in DUE the arc flows are unique, but the path flows are in general not if we do not impose the additional constraint of passenger mingling.

The dynamic extension might be further criticised for setting back passengers to the last node they have reached during the last time interval. Ways to minimise the errors resulting from this have been described. It is further thought that this first approach to dynamic frequency based assignment is nonetheless a significant advance for strategic planning. As shown with the London case study the analysis of capacity problems requires a dynamic approach. The steady-state analysis of larger time periods will underestimate the peaked characteristics of capacity problems and modelling the peak of the peak only can not reflect the build up of congestion through passengers who started their journey slightly before the peak modelling period.

Through the London case study it was demonstrated that the approach can be used for larger networks. The computation time required (several hours for the dynamic
model) is long but feasible for strategic planning purposes. Some methods that improved the computation time have been presented but further work could be done in this area. For example, as the computational time is strongly dependent on the number of nodes, if the planner has already a rough idea at which stations capacity problems might occur one could reduce the number of nodes by deleting the failure nodes from other stations. Also, in areas of the network where the common line problem does not apply the network representation can be simplified as discussed in Chapter 4. Moreover, Bellei et al (2005) also use the MSA to find the user equilibrium (though for dynamic traffic assignment). They note that less iterations are needed if the Bather method, which is a slight modification of the MSA, is applied (Bottom and Chabini, 2001).

In Chapter 6, the thesis discussed first the consideration of flow dependent arc costs and secondly the introduction of fail-to-access nodes. With these two model refinements it is possible to model all three main effects of crowding on the perceived passenger cost: Firstly, the inconvenience through on-board crowding (crowding problem), secondly, delays and crowding on the platform (capacity problem) and thirdly delays and crowding before the passenger even gets onto the platform (access capacity problem).

It was further discussed that the on-board crowding will be perceived differently if passengers have a seat or not. If the costs are averaged according to the proportion of passengers standing and sitting this does not consider priority rules, i.e. passengers who will board the train at the terminal are more likely to gain a seat. The thesis discussed an idea to increase passenger costs if the line load at their boarding point is
larger than the number of seats in the train, which means that passengers might fail to get a seat initially. This does not consider that passengers might gain a seat later and the approach could be refined for example through the introduction of separate arcs for those passengers sitting and those passengers standing.

Probably a more important area of further work is the inclusion of stochastic elements. The work of Lam et al (1999b, 2002) showed that a formulation of the problem in terms of stochastic user equilibrium assignment is feasible and also leads to a unique solution. Including stochastic elements into the CapCon approach would be a significant improvement because the assumption that all passengers have the same level of risk-averseness is not realistic. Some passengers will re-route in order to avoid crowding effects, while others will stick to the shortest path independent of the passenger crowding.

Further it needs to be considered that the crowding of the vehicles will be perceived differently by passengers. Some passengers will still board the service whereas others will perceive the service as being so full that they rather wait for the next one. The surveys by London Underground (London Underground Limited, 1988) showed this effect even if the arriving train was only half occupied. The “typical commuter” for example might still board a service that a “typical tourist” might not board anymore. On the other side the commuter might want to avoid the uncertainty of being delayed at an interchange point and take a slightly longer route. The tourist might not be well informed about other routes. In conclusion one might want to introduce not a fixed vehicle capacity but instead a perceived vehicle capacity. The commuter will perceive the vehicle capacity to be larger than the tourists who rather
waits for a not so crowded service. On the other side the risk-averseness (parameter $\theta_e$ or $\theta_d$ in the model) of the commuter might be higher because firstly he wants to make sure that he arrives in time and secondly because his knowledge about the network is better.

Further, departure time choice is currently not included in the model. In fact the model does not currently consider a feedback from future time intervals to the current time interval. The main argument against this feedback loop is that it would significantly increase the required memory and computation time of the model. In the case studies it is simply assumed that passengers choose their route according to the network condition in the current time interval even if their journey will take longer than one time interval. As journeys through the congested city centre of London only take around 30min this might not be a very significant restriction. Further, an improvement to this assumption is presented in Chapter 6 where the parameter $\theta_r$ is introduced in order to model passengers expectation about whether the congestion will increase or decrease during subsequent time intervals. Compared to road traffic the in-vehicle travel time stays relatively constant which also reduces the need to introduce a feed-back from future time intervals on the route choice and departure time choice in the current time interval.
References


Study Institute (ASI) Course on ‘Advanced modelling for transit operations and service planning’, Session F, paper F1, Hong Kong.


Schmöcker, J-D. and Bell, M.G.H. (2006). The build up of Capacity Problems during Peak Hour. A dynamic Frequency-based model. Accepted for publication in the selected proceedings of the 2nd International Workshop on Schedule-Based Approach to Dynamic Transit Modelling (SBDTM), Ischia (Naples, Italy), May 29-30.


