Linköping University

Methods for Capacity Allocation in Deregulated Railway Markets

Abderrahman Ait Ali

Supervised by Jonas Eliasson

Department of Science and Technology
Division of Communications and Transport Systems
Linköpings universitet, SE-601 74 Norrköping, Sweden

Norrköping 2020

Methods for Capacity Allocation in Deregulated Railway Markets
Abderrahman Ait Ali

Supervisor: Jonas Eliasson
Co-supervisors: Anders Peterson and Maria Börjesson

Linköping Studies in Science and Technology. Dissertation No. 2101

Cover illustration is a graphical timetable from *RailSys* simulation software.

ISBN 978-91-7929-771-8
ISSN 0345–7524

Printed by LiU-Tryck, Linköping, Sweden 2020

Abstract

Faced with increasing challenges, railways around Europe have recently undergone major reforms aiming to improve the efficiency and competitiveness of the railway sector. New market structures such as vertical separation, deregulation and open access can allow for reduced public expenditures, increased market competition, and more efficient railway systems.

However, these structures have introduced new challenges for managing infrastructure and operations. Railway capacity allocation, previously internally performed within monopolistic national companies, are now conferred to an infrastructure manager. The manager is responsible for transparent and efficient allocation of available capacity to the different (often competing) licensed railway undertakings.

This thesis aims at developing a number of methods that can help allocate capacity in a deregulated (vertically separated) railway market. It focuses on efficiency in terms of social welfare, and transparency in terms of clarity and fairness. The work is concerned with successive allocation of capacity for publicly controlled and commercial traffic within a segmented railway market.

The contributions include cost benefit analysis methods that allow public transport authorities to assess the social welfare of their traffic, and create efficient schedules. The thesis also describes a market-based transparent capacity allocation where infrastructure managers price commercial train paths to solve capacity conflicts with publicly controlled traffic. Additionally, solution methods are developed to help estimate passenger demand, which is a necessary input both for resolving conflicts, and for creating efficient timetables.

Future capacity allocation in deregulated markets may include solution methods from this thesis. However, further experimentations are still required to address concerns such as data, legislation and acceptability. Moreover, future works can include prototyping and pilot projects on the proposed solutions, and investigating legal and digitalisation strategies to facilitate the implementation of such solutions.

Keywords: railway capacity; capacity allocation; train timetable; cost benefit analysis; deregulated market.

Sammanfattning

Med ökande utmaningar har järnvägar runt om i Europa genomgått stora reformer som syftar till att förbättra järnvägssektorns effektivitet och konkurrenskraft. Nya marknadsstrukturer såsom vertikal separering, avreglering och öppet tillträde för flera operatörer kan möjliggöra minskade offentliga kostnader, ökad marknadskonkurrens och effektivare järnvägssystem.

Denna omreglering av järnvägsmarknaderna har dock skapat nya utmaningar för hanteringen av järnvägsinfrastruktur och drift. Tilldelning av järnvägskapacitet, vilket tidigare sköttes inom nationella monopolföretag, måste nu göras av en infrastrukturförvaltare (*infrastructure manager*). Förvaltarens kapacitetstilldelning till olika (ofta konkurrerande) licensierade järnvägsföretag (*railway undertakings*) måste samtidigt vara transparent, rättvis och leda till ett effektivt kapacitetsutnyttjande.

I denna avhandling utvecklas metoder som kan användas av en infrastrukturförvaltare för att tilldela kapacitet i en avreglerad järnvägsmarknad. Den fokuserar på samhällsekonomiskt effektiva utfall men även transparens, tydlighet och rättvisa.

Avhandlingens bidrag omfattar samhällsekonomiska analysmetoder som gör det möjligt för regionala kollektivtrafikmyndigheter att bedöma den samhällsekonomiska effektiviteten för deras trafikering och skapa ett effektivt utbud. Med dessa metoder som utgångspunkt beskrivs en marknadsbaserad och transparent tilldelningsprocess för kapacitet där infrastrukturförvaltare prissätter kommersiella tåglägen för att lösa kapacitetskonflikter med offentligt kontrollerad trafik. Dessutom utvecklas optimeringsmetoder för att estimera passagerarefterfrågan och för att skapa effektiva tågtidtabeller.

Framtida kapacitetstilldelning på avreglerade marknader kan inkludera lösningsmetoder från denna avhandling. Ytterligare experiment krävs dock fortfarande för att hantera problem såsom data, lagstiftning och godtagbarhet. Dessutom kan framtida arbete omfatta prototyper och pilotprojekt av de föreslagna lösningarna och undersöka lagliga och digitaliseringsstrategier för att underlätta implementeringen av sådana lösningar.

Nyckelord: spårkapacitet; kapacitetstilldelning; tågtidtabell; samhällsekonomisk analys; avreglerad marknad.

Acknowledgements

Thanks to the support of many people, this doctoral thesis is the result of a life changing positive experience. I would like to acknowledge you here, person by person, but I will surely be unable to mention you all. *I hereby thank you ALL from the heart of my heart.*

First and foremost, Jonas, no words can describe your involvement, guidance and support to start, do and finish this journey. You have been the supervisor and the friend that I have wished to have. Without you, much of this experience would not be a reality. *Stort TACK Jonas!*

Maria and Anders, my co-supervisors, you have been immensely helpful. Maria, you have been supportive from the beginning until the end. Anders, thank you for stepping up to help me finish this journey. *Tackar!*

Per Olov Lindberg, Jan-Eric Nilsson and Martin Aronsson, my first tutors, you have helped me begin this journey. Jan-Eric together with PO, your experience and expertise made my first research work more rigorous. Martin, discussing with you have always been insightful. *Tack alla!*

Jennifer Warg, Emanuel Broman, Victoria Svedberg, Sara Gestrelius, Emma Solinen, Carl-William Palmqvist, Johan Högdahl, Ingrid Johansson, Niloofar Minbashi and Félix Vautard, the future of (Swedish) railway research, it has been very enjoyable to work and/or discuss with you. Jenny, you have always been helpful. Emanuel, it has been nice to share most of this journey with you. *Tack alla för allt!*

Hans Dahlberg, Mattias Haraldsson & Jan-Erik Swärdh, Jan Lundgren, the project partners from Trafikverket, VTI and Linköping University (LiU), respectively. Hasse, you have been an enthusiast project leader from the beginning. Mattias, Jan-Erik and Jan, you have helped make my work environment more productive and enjoyable. *Tack ska ni ha!*

I will not forget to express my gratefulness to Yves Crozet for kindly accepting to be my opponent for the final defence, to Karin Brundell-Freij for the final seminar, and to Tomas and Mats for the KTS start seminar. I also express my gratitude to all the members of the examination board, namely Gunnar Isacsson, Jan Persson and Siri Pettersen Strandenes.

I was lucky that my journey went through different workplaces, i.e., KTH, VTI (Stockholm), LiU (Norrköping) and IFSTTAR-LVMT (Paris). My former colleagues at KTH (Alyn, Anders, Athina, Behzad, Bibbi, Bolle, David, Dimas, Erik, Gerhard, Hans, Hugo, Isak, Jiali, Joel, Jonas, Joram,

Josef, Juan, Markus, Masoud, Matej, Oskar, Roberto, Soumela, Tasos, Todor, Wei, Wilco, Yusak), the new ones at LiU (Alan, Anna, Antzela, Christiane, Clas, Ghazwan, Joakim, Kalle, Martin, Mats, Nikki, Nikos, Nils, Therese, Tomas, Viveka) and VTI (Ajsuna, Ary, Chengxi, Disa, Ida, Inge, Jiali (again), Johanna, Kristofer, Lisa, Noor, Roger, Tomas (again), Ulrika), and my hosts at IFSTTAR-LVMT (Martin, Nicolas, Paola, Sophie), you have all made my journey more joyful.

My dearest friends (Abdessamad, Ahmed, André, Anass, Aymen, Driss, David, Hafid, Habib, Hicham, Othmane, Rachid, Salah, Taoufiq, Yassine, Yassir, Youssef), I am grateful to have you all. My mom, dad and closest relatives, you have always been supportive. Amina, thank you for your continuous and overwhelming love.

ⵙⵔⵓ ⴰⵣⵎⵔ ⵓⵜⵜⴰⵣ
Do good and you will find it!

Stockholm, 2020-10-01
Abdou (with ❤)

Preface

This doctoral thesis is the culmination of research conducted between 2015 and 2020 at the Royal Institute of Technology (KTH), the Swedish National Road and Transport Research Institute (VTI) and Linköping University (LiU) in Sweden. Funded by the Swedish Transport Administration (Trafikverket), the thesis is part of the SamEff project (*Samhällsekonomiskt effektiv tilldelning av järnvägskapacitet)* which stands for societally efficient railway capacity allocation.

The thesis consists of five papers appended to an introductory essay.

List of included papers

Paper 1 (P1): Ait-Ali and Eliasson (2019). A Survey of Railway Deregulation in Europe. Submitted for journal publication.

Paper 2 (P2): Ait-Ali et al. (2020). Pricing Commercial Train Path Requests Based on Societal Costs. Published in Transportation Research Part A: Policy and Practice, Volume 132, February 2020, Pages 452-464.

Paper 3 (P3): Ait-Ali et al. (2020). Are commuter train timetables consistent with passengers' valuations of waiting times and in-vehicle crowding? Submitted for journal publication.

Paper 4 (P4): Ait-Ali et al. (2020). Disaggregation in Bundle Methods: Application to the Train Timetabling Problem. Published in Journal of Rail Transport Planning & Management, 100200.

Paper 5 (P5): Ait-Ali and Eliasson (2020). The Value of Additional Data for Public Transport Origin-Destination Matrix Estimation. Submitted for journal publication.

List of related (but not included) papers

Related paper 1 (RP1): Ait-Ali et al. (2017). Measuring the Socio-economic Benefits of Train Timetables Application to Commuter Train Services in Stockholm. Published in Transportation Research Procedia, 27, 849-856.

Related paper 2 (RP2): Warg et al. (2019). Assessment of Commuter Train Timetables Including Transfers. Published in Transportation Research Procedia, 37, 11-18.

Related paper 3 (RP3): Ait-Ali and Eliasson (2019). Dynamic Origin-Destination Estimation Using Smart Card Data: An Entropy Maximisation Approach. Published in arXiv:1909.02826.

Conferences

Parts of the work in this thesis include journal and conference papers. Most of these papers were disseminated and presented at local and international conferences and seminars. **Table 1** lists the international conferences where papers have been disseminated and presented.

Table 1. International conferences and disseminated papers.

Paper(s)	International conference
P3	European Conference of Society for Benefit-Cost Analysis, 26th – 27th November 2019, Toulouse – France
P2, RP3	8th International Conference on Railway Operations Modelling and Analysis, 17th-20th June 2019, Norrköping – Sweden
RP1	21st EURO Working Group on Transportation Meeting, 17th-19th September 2018, Braunschweig – Germany
P5	29th European Conference on Operational Research EURO 2018, 9th -11th July 2018, Valencia – Spain
RP2	20th EURO Working Group on Transportation Meeting, 4th - 6th September 2017, Budapest – Hungary
P4	7th International Conference on Railway Operations Modelling and Analysis, 4th – 7th April 2017, Lille - France

Author contribution statement

Ait-Ali, A., the author of this thesis, is the main contributor in the included papers. He has conducted the research, literature review, model development, experimentation as well as documentation. **Eliasson, J.,** the main supervisor, has provided research ideas, support, advice, and extensive review of this thesis, all the included and related papers. **Warg, J.,** co-author of P2, has collaborated with the main author in problem formulation, experimental design and writing the paper. **Lindberg**, **P. O.,** co-author of P4, has provided the main research idea. All co-authors have helped in result analysis and/or paper review.

Terminology

The following glossary presents definitions (in alphabetical order) of the main terminology (italicised when first used) that is adopted in this thesis.

The definitions are based on a number of references from railway and economics. Most of the railway-related definitions are borrowed from the glossary of terms by RNE (2017). Definitions of economics-related terms are mainly from the book by Wetzstein (2013).

Swedish translations are checked using Sweden's national term bank database (Rikstermbanken, 2019).

Annual timetable (*årlig tågplan*): yearly constructed schedule listing the times and the locations at which certain events, e.g., arrivals and departures, are expected to take place (same as the working timetable).
Commercial train services (*kommersiella tågutbud*): train services that are operated on a profit-maximising basis, e.g., freight, long distance passenger trains (in contrast to subsidised train services).
Competitive tendering (*konkurrensutsatt upphandling*): process of bidding to win the rights to run train services, i.e., for-track competition.
Concession (*koncession*): management contract giving the right to operate a service over a defined period (typically several years) subject to meeting certain requirements, often awarded by competitive tendering.
Consumer surplus (*konsumentöverskott*): benefit that is received by the consumers of a product or a service from the difference between the price and the willingness-to-pay.
Corner solution (*hörnlösning*): an optimal solution in a point where several linear constraints meet, making its location independent of certain input parameters.
Cost benefit analysis (*kostnads-nyttoanalys*): approach to calculate and compare the benefits and costs of a certain project or policy.
Deregulation (*avreglering*): process of removing barriers to entry in the market, and thus increase competition.
Dispatching, traffic control (*trafikledning*): directing and facilitating the movement of trains in a certain area and period of time.
EU directive (*EU-direktiv*): legal act of the EU that needs to be transposed into national law in the member states without dictating how.
EU regulation (*EU-förordning*): legal act of the EU that becomes immediately enforceable as law in all member states simultaneously.
Framework agreement (*ramavtal*): setting out capacity allocation rights over a period longer than one working timetable.

Franchising (*franchise*): exclusive right to operate a service under a higher degree of specification (compared to concession, e.g. setting fare levels and financial risks) and may involve payments between the transport authority and the franchisee.
Freight traffic (*godstrafik*): railway traffic transporting goods (in contrast to passenger traffic).
Gamification *(spelifiering)*: use of game principles and design to solve problems in non-game contexts, e.g., to improve productivity or for learning.
Grandfather right (*hävdvunnen rättighet, oöversatt*): rights and rules favoring incumbents at the expense of new entrants.
Headway (*tågseparation*): time or distance between two consecutive trains.
Incumbent operator (*etablerad operatör*): national railway undertaking(s) or operator(s) traditionally owning rolling stock, responsible for production, operations, maintenance and infrastructure (before the vertical separation and the deregulation).
Infrastructure manager (*infrastrukturförvaltare*): body responsible for administering rail infrastructure and managing its facilities.
Monopoly (*monopol*): when an actor is the only supplier of a certain service or product in a market.
Nationalisation (*nationalisering*): process of converting private assets to public ones owned by the state (in contrast to privatisation).
Network statement (*järnvägsnätsbeskrivning*): document which sets out in detail the general rules and procedures for allocating railway capacity, including information required for capacity applications.
Open access (*öppet tillträde*): process by which non-incumbent operators can also access the infrastructure, enabling them to run services complementing or competing with others, i.e., on-track competition.
Passenger traffic (*persontrafik*): railway traffic transporting passengers (in contrast to freight traffic).
Privatisation (*privatisering*): process of converting state-owned public assets to private ones (in contrast to nationalisation).
Producer surplus (*producentöverskott*): the monetary value that is gained by the producers of a product due to the difference between the price and their production cost or willingness-to-sell.
Public service obligation (*trafikeringsplikt*): responsibility of the railway undertaking to maintain a certain level of public services, e.g., number of train departures or frequency, ticket prices.
Public utility (*allmännyttig tjänst*): service provided on a regulated public infrastructure such as electricity, water and telecommunication.
Publicly controlled (or subsidised) train services (*subventionerade tågtjänster*): train services where timetables and fares are

determined by a public agency, presumably to maximise social welfare (in contrast to commercial train services, where a profit-maximising company decides timetables and fares).
Railway capacity allocation (*järnvägskapacitetstilldelning):* process where capacity is granted to a railway undertaking (or other applicants) by the relevant capacity allocation body (infrastructure manager).
Railway regulator (*regulator*): independent, official regulatory body for rail; its duties and powers are set out in the national legislation.
Railway undertaking (*järnvägsföretag*): any licensed public or private entity, the principal business of which is to provide services for the transport of goods and/or passengers by rail.
Reserve capacity (*reservkapacitet*): capacity kept available within the final working timetable allowing quick and appropriate responses to ad hoc requests.
Rolling stock (*rullande materiel*): collective term for the railway fleet describing all the vehicles on a track (in contrast to fixed stock or infrastructure).
Social cost (*samhällsekonomisk kostnad*): total cost incurred by the society including consumer, producer and external costs. (same as the societal cost, in contrast to the social surplus).
Track access charges (*banavgifter*): fees that are paid to the infrastructure manager by an operator for running trains on its infrastructure.
Train path (*tågläge*): definition of a train's route in terms of time and space with details of locations at which it will pass, including any activities that the train will perform, e.g., train crew, locomotive changes.
Train timetabling (*tidtabelläggning*): process of consultation and planning to define expected train movements taking place on the infrastructure during a certain period time.
Transaction costs (*transaktionskostnader*): costs related to the economic interaction between separate entities.
Vertical separation (*vertikal separation*): separation of infrastructure management and railway operations (e.g., train services).

Contents

List of Tables

List of Figures

All the figures in this thesis are the author's own work, unless otherwise noted.

List of Acronyms

CBA	Cost Benefit Analysis
CERRE	Centre on Regulation in Europe
Cx	Contribution x (x is a number from 1 to 10)
EC	European Commission
EM	Entropy Maximisation
EU	European Union
GDP	Gross Domestic Product
ICT	Information and Communications Technology
IM	Infrastructure Manager
IP	Integer Program
ITF	International Transport Forum
KPI	Key Performance Indicator
MIP	Mixed Integer Program
MPK	*Marknadsanpassad Planering av Kapacitet* (English: Market-adapted Planning of Capacity)
OD	Origin Destination
OECD	Organisation for Economic Cooperation and Development
PPP	Public Private Partnerships
PSO	Public Service Obligation
PT	Public Transport
PTA	Public Transport Authority
Px	Paper x (x is a number from 1 to 5)
RKTM	*Regional kollektivtrafikmyndighet* (in Swedish) (English: Regional Public Transport Authority or PTA)
RMSE	Root Mean Square Error
RNE	RailNetEurope
RPx	Related Paper x (x is a number from 1 to 3)
RQx	Research Question x (x is a number from 1 to 5)
RU	Railway Undertaking
SEK	Swedish Krona (1 Euro is around 10 SEK)
SERA	Single European Railway Area
SJ	*Statens Järnvägar* (English: Swedish Railways)
SL	*Storstockholms Lokaltrafik* (in Swedish) (English: Greater Stockholm Local Transit)
TTP	Train Timetabling Problem
TTR	Train Timetable Redesign
UIC	*Union Internationale des Chemins de fer* (in French) (English: International Union of Railways)
WTP	Willingness-To-Pay

Chapter 1

Introduction

Do not say a little in many words but a great deal in a few
Pythagoras, Greek philosopher

1. Introduction

States long adopted a *laissez-faire* policy in early railways, allowing private companies to build, operate, maintain, and hence own railway systems, i.e., *privatised railways*. Some developments (e.g., passenger trains, fierce competition between investors or *railway mania* and the industrial revolution) made governments pay increasing attention. Many railways were therefore *nationalised*, and thus operated by monopolistic state-owned companies providing both *passenger and freight traffic*.

During the late 20th century, national railways have been facing increasing challenges due to efficiency and cost problems, and competition from other modes. Several railway markets, mainly in the European Union (EU), have been subsequently reformed by splitting their monopolistic national railways into infrastructure management and train services. This splitting, also called *vertical separation*, allows for opening the railway market to competition. New (domestic or foreign) railway companies may provide train services, a process often referred to as *deregulation* which reduces state market control.

By deregulating their railways, governments aim to reduce public expenditures, increase service quality, and improve system efficiency. For this to succeed, there is still need for instruments to intervene, i.e., regulating the deregulation. An important element in this context is the allocation of railway capacity which faces new challenges due to the deregulation. In other words, the previously closed internal capacity allocation, within monopolistic national railway companies, needs to be replaced with a more transparent and (still) efficient allocation of available capacity to the different (possibly competing) companies in the market. This task is the main problem that this thesis attempts to address.

This first chapter introduces more relevant information to understand the research context and motivation of this work. It also presents the structure of the thesis, and finally, states its delimitation.

1.1. Research context

With decreasing efficiency and increasing spending, state-controlled railways came under pressure, and a trend of deregulation reforms emerged which allowed private actors in the market once again (Laurino et al., 2015). Sweden was first to start deregulating its national market (as early as 1988) after vertically separating railway services from infrastructure management (Hansson and Nilsson, 1991).

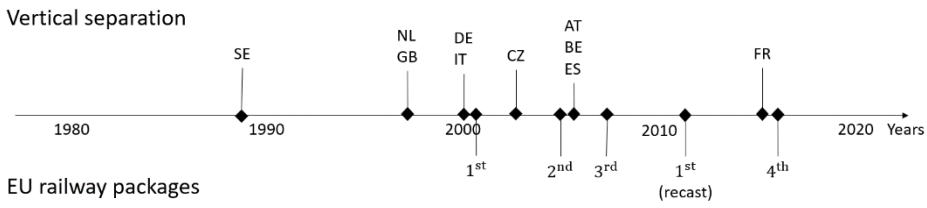


Figure 1. Timeline of vertical separation and EU railway packages.

Following the 1991/440/EEC first *directive* (EC, 1991), several EU member states adopted vertical separation as illustrated in the timeline presented in **Figure 1**. The directive allows one of three alternatives: accounting, organisational or institutional separation. The first type guarantees separate financial accounts, the second is about independent units within one larger institution, and the third refers to the complete separation as in Sweden. This resulted in various market structures throughout Europe but all have at least a vertical separation in accounting (Monami, 2000, Nash, 2008).

Following further EU directives and *regulations* (grouped as *railway packages*), train services in different market segments have been gradually opened for competition (EC, 2001). Further calls from the European Commission (EC) aimed, among other things, to establish a Single European Railway Area (SERA) as stipulated by the 34/EC SERA directive, a recast of the 1st railway package (EC, 2012). The directives have also aimed to promote competition, interoperability, transparency and efficiency, see **Appendix 1: EU directives**.

In the context of railway capacity allocation, transparency means that all the process is comprehensive, clear and above all non-discriminatory to any of the market players. However, efficiency may be interpreted in various ways depending on the national railway legislation. In Sweden, the objective of *railway capacity allocation* is to achieve maximal *socioeconomic or societal efficiency* (*samhällsekonomisk effektivitet* in Swedish) meaning that the net *social surplus* is maximised including benefits for all *consumers* and *producers* as well as all external effects, see **Appendix 2: Swedish railway law**.

The contributions of this thesis attempt to address the problem of socioeconomically efficient and transparent capacity allocation in vertically separated and deregulated railways, e.g., Sweden.

1.2. Thesis outline

Four chapters form the thesis. This 1st chapter introduces the research setting by presenting the research context, the thesis outline and delimitation. Chapter 2 presents the relevant background information and terminology on railway capacity, its allocation and market deregulation. It also provides a review of the literature including related existing research and experiments with focus on Europe and Sweden. The conducted research is described in the 3rd chapter which starts with the gaps and challenges, and the research questions follow with a presentation of the methodology. A discussion of the conducted research concludes the chapter. The contributions and future works in chapter 4 conclude the thesis.

Relevant excerpts from the European and Swedish legislation can be found in the two appendices. Finally, all the included papers are appended to this thesis.

1.3. Delimitation

The scope of this thesis is delimited in several dimensions. First and foremost, the focus is mostly on the efficiency of the capacity allocation, rather than its transparency. Second, specific allocation contracts such as *franchising*, *concessions* and *framework agreements*, although important, are not studied in detail but only briefly mentioned. However, we study situations of capacity conflicts between *publicly controlled and commercial train services* regardless of the allocation contracts. Moreover, certain market segments (e.g., infrastructure maintenance) and allocation steps (e.g., ad hoc) are only briefly discussed. Furthermore, *dispatching* or real time *traffic control* aspects (e.g., timetable robustness and train punctuality) are not considered, but these are well developed in the literature (Andersson et al., 2013), and can therefore be included in a later stage of the allocation. Last but not least, legal issues are only briefly mentioned and discussed.

Chapter 2

Literature Review

‘Många vet mycket, ingen vet allt’

Many know much, but nobody knows everything
Swedish proverb

2. Literature Review

In this 2nd chapter, background information on railway capacity, its allocation and market deregulation are presented while introducing relevant terminology. Related existing research and experiments are also briefly reviewed focusing on the European and Swedish context.

2.1. Railway capacity

In the railway sector, capacity has different meanings depending on the context where it is used. Although no unique definition exists, railway capacity is an important concept that can be defined and analysed based on specific aspects (Petersen, 1974). For instance, it is highly affected by factors such as infrastructure (number of tracks and network design), type of train traffic (e.g., freight, high-speed or commuter trains) and other operational factors (Forsgren, 2003, Abril et al., 2008).

One definition, also used in Sweden, is from the 406 code by the International Union of Railways (IUC) which states that capacity of any infrastructure is the number of possible paths in a time window (UIC, 2004). Such a number may depend on additional factors such as the path mix (traffic heterogeneity), service quality and other considerations for constructing train timetables (Goverde and Hansen, 2013).

In the context of railway capacity allocation, RailNetEurope (RNE), in its glossary of terms, refers to capacity as the actual *train path* which describes the infrastructure needed for running a train between two places over a given period of time, i.e., time-space taken up in the *annual timetable* by the passage of the train including safety margins (RNE, 2017). Later in this thesis, we will see that train paths include certain flexibility and can be adjusted during allocation.

Railway capacity at certain parts of the infrastructure may depend on (or affect) that of other parts in the network. For instance, (primary) delays in one place may cause (secondary) delays in others, or improved accessibility on some parts may induce demand on others. Such network effects indicate that capacity analysis is combinatorial in nature, and that most related problems are hard to solve using state-of-the-art solvers, e.g., *train timetabling* (Caprara et al., 2002).

2.2. Capacity allocation

The allocation of railway capacity refers to the process where train path requests are granted by the relevant capacity allocation body, often called *infrastructure manager* (IM), to capacity applicants or train operators, also called *railway undertakings* (RUs). The allocated capacity can be used for running freight or passenger trains as well as for infrastructure maintenance. The IM is responsible for the allocation based on specific conditions and rules, compiled in the national *network statement*. Such allocation is repeated on a yearly basis to construct a new annual timetable specifying when and where trains run (RNE, 2017).

Unlike road traffic with an ad hoc allocation of capacity (queues can possibly be building up, i.e., road congestion), railway capacity must be planned and allocated beforehand (van Wee et al., 2013). Thus, capacity congestion in railways may emerge when the available capacity is not enough to include all the requested train paths. Capacity allocation is therefore fundamental in the railway sector for prior planning of the traffic, and for solving capacity conflicts, if any.

When allocating capacity, the IM has certain flexibility to adjust and reschedule the original train path requests. Thus, allocating capacity means including the (adjusted) requested train paths in the annual timetable. Each request represents a plan for a certain service, such as a freight or passenger trains. Such services differ in many ways, e.g., speed, distance, publicly controlled or commercial, and therefore express varying requirements. Furthermore, how capacity is allocated may also depend on the structure of the market, e.g., existing (or dominant) complementary or competing train services, degree of market competition and deregulation (Gibson, 2003).

2.3. Market deregulation

Railways are often referred to as examples of natural monopoly due to their substantial initial fixed costs (De Palma and Monardo, 2019). Other examples can be found in *public utility* networks such as gas, electricity and water. In natural monopolies such as railways, it is more practical to have a *monopoly* that provides the railway network. Hence, the early monopolistic and highly regulated national railways.

With the emergence of market deregulation trends in the railway sector, new structures appeared which vary from one country to the other due to various reasons, e.g., political, economic and geographical (Laurino et

al., 2015). One fundamental difference between these markets is their degree of *vertical separation* (or integration) which refers to the division of responsibilities between infrastructure management and rail services. Another important difference is the level of deregulation, i.e., the horizontal relationship between the different actors in a market segment with similar roles and responsibilities (e.g., RUs).

These fundamental differences lead to four main structures as illustrated in **Figure 2** where each arrow indicates the movement from one structure to another, either separation or integration in the horizontal or vertical dimension. In the same figure, a railway company (large grey box) may be responsible for rail services and/or the network (smaller white boxes).

Contrasting market structures (and segments) have distinct characteristics, and therefore need different capacity allocation principles (Gibson, 2003). In vertically integrated markets (i.e., top in **Figure 2**), capacity allocation is internally administered, and reduced to the so-called train timetabling problem (TTP) where the monopolistic company constructs a feasible train timetable that maximises the company's objective function (Brännlund et al., 1998, Caprara et al., 2002). This is different for vertically separated markets (i.e., bottom in **Figure 2**) with separate IMs. Railway deregulation allows for the presence of actors (RUs) other than the *incumbent operator(s)*. In order to allocate capacity, the IM needs to accommodate different (sometimes conflicting) train path requests from RUs, and settle all the possible disputes. RUs are usually required to pay *track access charges* for their respective allocated paths (Freebairn, 1998, Bouf et al., 2005).

Each structure has pros and cons (Mizutani et al., 2015, Abbott and Cohen, 2017). On the one hand, integration is better for reducing *transaction costs* between separate entities which are working together (Merkert, 2012, Merkert and Nash, 2013). On the other hand, separation, if managed well, can increase competition and thereby productivity and service quality. However, if competition is not well regulated, market inefficiencies may emerge, for instance due to anti-competitive practices by certain RUs leading to market outcomes that are far from maximising social welfare (Broman and Eliasson, 2019).

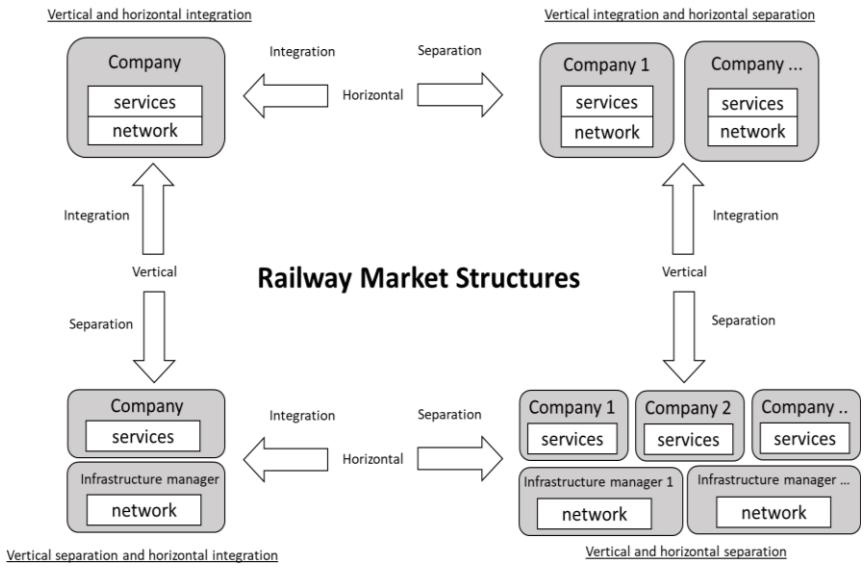


Figure 2. Overview of the main railway market structures.

In what follows, the thesis mostly focuses on capacity allocation in vertically separated (and deregulated) markets with open-access of the kind mostly found in EU markets, and in particular the Swedish railway system (Jensen and Stelling, 2007, Alexandersson and Rigas, 2013). Most European railways are vertically separated and deregulated with horizontal separation in services, see **Figure 3**. This structure aims at stimulating competition by allowing new (possibly foreign) companies to provide services alongside, often in competition with, the incumbent (or the previous monopolistic national company), if any. Interoperability is thus required for licensed companies to provide services across the European SERA market (Crozet et al., 2012).

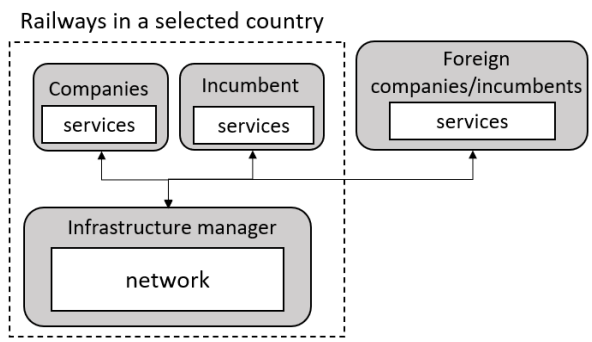


Figure 3. Example of a deregulated market structure in Europe.

2.4. European context

As part of the European reforms, several member states adjusted their market structures and national railway legislation to EU policy guidelines (Monami, 2000, Nash et al., 2014). The EC has introduced several railway packages as guidelines to help implement the deregulation (EC, 1991, EC, 2001, EC, 2012). An overview of the packages and main topics of the corresponding directives is presented in **Table 2**.

Table 2. EU packages and main topics in the directives.

Package	Year	Main topics
1st	2001	Vertical separation for market deregulation: cross-border freight, access charges, licensing
2nd	2004	Integrated European railway area: safety, interoperability, national freight
3rd	2007	International passenger: open access, subsidised services, interoperability
4th	2016	Domestic passenger services: interoperability, governance, licensing

A timeline of these packages was also previously presented in **Figure 1**. The 1st package (initiated with directive 91/440 from 1991) was an early attempt to set certain guidelines for market deregulation and capacity allocation (EC, 2001). Accordingly, all member states are required to have at least vertical separation in terms of accounting. A recast established, among others, principles for interoperability in the SERA markets (EC, 2012). The packages that followed focused on the successive deregulation of different market segments, e.g., cross-border freight (2001), national freight (2004), international passenger (2007) and domestic passenger (2016) as part of the more recent 4th railway package (EC, 2016).

In European deregulated markets, the IM publishes the national network statement on a yearly basis providing guidelines on how capacity is allocated for the licensed RUs. The allocation generally starts one year (noted X-12) before adopting the new annual timetable. The IM receives capacity (train path) requests which are formulated by capacity applicants (RUs). A draft of the annual timetable is prepared by the IM for coordination with RUs to settle potential capacity conflicts.

When unresolved through negotiations and voluntary compromises, capacity conflicts are settled unilaterally by the IM using predetermined priority criteria. Lines (and time periods) where such conflicts occur are

declared congested, and capacity analysis is conducted by the IM for reinforcement plans to improve the capacity supply.

Once the draft is published, the late train path requests are received and allocated depending on the available *reserve capacity*. This ad hoc allocation of capacity continues during the year, even after the start day of the annual timetable, i.e., between X and X+12. The allocation is supervised by the *railway regulator*, often an independent governmental body. An overview of the allocation is illustrated in **Figure 4**.

2.5. Swedish capacity allocation

The Swedish railway market was managed by the Swedish State Railways SJ (*Statens Järnvägar*) until 1988, when infrastructure management was separated from operations and transferred to the newly created Swedish Rail Administration (Banverket) leading to one of the first vertically separated railway markets in the world. In 2001, SJ was split into several state-owned companies: SJ (passenger), Green Cargo (freight), Jernhusen (stations) and Euromaint (maintenance). In 2010, Banverket was integrated with the Swedish Road Administration (Vägverket) to form the Swedish Transport Administration (Trafikverket).

Trafikverket allocates capacity in the Swedish network similarly to many European deregulated markets, see **Figure 4** for an overview of the different steps of the capacity allocation. One of the main differences is in the settlement of capacity conflicts that remain after the coordination with applicants which settles most of the capacity conflicts. Unlike many IMs which use simple and general priority lists, Trafikverket uses priority criteria based on *cost benefit analysis* (CBA) rules aiming to reflect which train path requests that yield the highest social welfare (Trafikverket, 2020). This Swedish CBA-based prioritisation appears to be more developed than the basic rule-of-thumbs criteria that are used for conflict settlement by many European IMs.

Trafikverket uses the CBA-based prioritisation to unilaterally settle the remaining conflicts only if the coordination process fails. Depending on the train category, different weights are used for certain variables such as the scheduled travel distance and time, train connections and cancellation. These weights are estimated using econometric studies to reflect their social welfare effects (Trafikverket, 2016a).

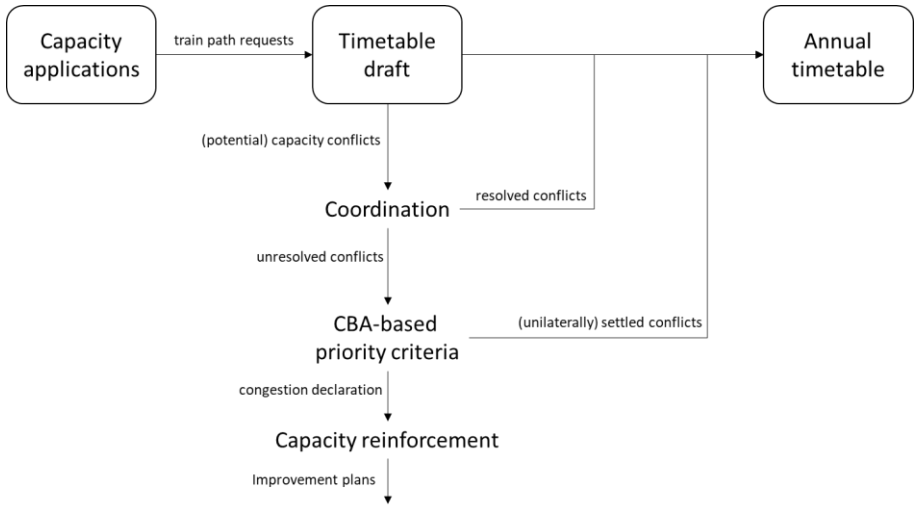


Figure 4. Overview of Swedish railway capacity allocation.

In case of (potential) unfair prioritisation or discrimination, RUs can report complaints and appeal to the regulator, i.e., Swedish Transport Agency (Transportstyrelsen).

2.6. Existing research and experiments

The problem of capacity allocation in deregulated markets has been studied extensively in other fields, such as airport slots (Rassenti et al., 1982, Gilbo, 1993), public utilities such as energy (gas and electricity), telecommunications, and water (McMillan, 1994, McAfee and McMillan, 1996). However, few academic works or experiments have been conducted in the railway sector.

Several research papers look at important components to consider for railway capacity allocation such as train timetabling and access charges (Gibson, 2003). Some others discuss the challenges of railway deregulation (Crozet et al., 2012) and the potentials of market-based solutions such as (combinatorial) auction (Nilsson, 2002, Borndörfer et al., 2006, Perennes, 2014). Most studies develop specific algorithms to allocate and/or price railway capacity (Lusby et al., 2011). These studies rarely consider the context of deregulation and the various market segments.

In a doctoral thesis, Pena-Alcaraz (2015) studies the capacity allocation in a deregulated and vertically separated market (called shared railway). The author investigates a capacity allocation solution that combines

problems of RUs (e.g., train timetabling) and IMs (e.g., capacity pricing), and the market outcome for different pricing strategies. However, no considerations are given to the social welfare in the allocation. These welfare aspects are considered in another doctoral thesis by Perez Herrero (2016) who uses an economic approach to study railway capacity in light of the market deregulation. Although no capacity allocation model is proposed (or studied), the author highlights the use of (optimal) congestion pricing of capacity as an instrument to improve the social welfare of capacity allocation outcomes.

Aspects relating to railway deregulation and capacity allocation have also been the subject of several reports from international organisations and forums. Such reports attempt to summarise and analyse their prospects and challenges. An early publication from OECD (2005) gives a comprehensive summary of the structural reforms that have happened in all the member countries. Focusing on EU countries, Crozet et al. (2012), in a policy report for CERRE, looks at how vertical separation can increase railway efficiency, and identifies key issues and regulatory recommendations for the introduction of competition to the market. This is later discussed by Crozet (2016a) at the International Transport Forum (ITF). Several CERRE follow-up studies deal with more specific aspects such as the liberalisation of passenger rail services in France (Crozet, 2016b), Germany (Link, 2016), Great Britain (Smith, 2016) and Sweden (Nilsson, 2016), or the levying of track access charges in France (Crozet, 2018), Germany (Link, 2018), Sweden (Nilsson, 2018) and Great Britain (Nash et al., 2018).

At the EU level, Train Timetable Redesign (TTR) is an initiative that attempts to redesign the international timetabling process (capacity allocation) in Europe to improve the competitiveness of cross-border (freight) train services. The TTR initiative introduces the concept of rolling planning which allows for ad hoc capacity requests in addition to the traditional annual requests. For instance, it is possible to safeguard bands of train paths (i.e., reserve capacity) and continuously allocate them for freight traffic. Certain pilot lines on cross-border European freight corridors are used for further experimentation (RNE, 2019).

At the national level, Trafikverket initiated a development project for market-adapted planning of capacity in Sweden, locally called MPK. The project aims to create a new (more flexible) approach for railway capacity allocation, and to develop new (digital) supporting tools (Gestrelius et al., 2020). Important contributions include a digital portal for capacity application which allows to access (and manage), for instance, train path

requests, capacity restrictions and track access charges (Trafikverket, 2016b). The project also attempts to implement the concept of incremental allocation (*successiv tilldelning*) which was previously studied and presented, e.g., by Aronsson et al. (2012). In such allocation, the annual timetable is initially flexible, and is incrementally constructed starting from (long term) delivery commitments to (more specific) production plans.

Chapter 3

Conducted Research

'العلم يؤتى ولا يأتي'

You have to approach science; it will not come to you
Arabic proverb

3. Conducted Research

Following the literature review, this 3rd chapter presents the conducted research. It describes the literature gaps and challenges, and formulates the research questions. A presentation of the research methodology and the developed models follows. Discussion of the conducted research concludes the chapter.

3.1. Challenges and research gaps

While studying the Swedish allocation process, Eliasson and Aronsson (2014) show that even the relatively well-developed Swedish conflict resolution model has some flaws. First, CBA calculations rely on certain variables (such as fares, demand, running costs) that are difficult or impossible to observe for commercial train services. For private, commercially driven traffic, such data is highly sensitive business information, and often unknown at the time of capacity allocation. On the other hand, such data is usually available for *publicly controlled traffic* (e.g., subsidised regional or commuter services). Second, the weights that are used in the CBA-based priority model are static. This means that certain train categories are always prioritised over others leading to so-called *corner solutions*. Thus, in case no interactions between complementary services exist, the diminishing returns to scale is not captured. In reality, the marginal societal benefit of higher frequency (or shorter *headway*) on a train service decreases, but this is not captured by the CBA-based priority criteria.

With these challenges in mind, the current CBA-based conflict settlement might lead to inefficient capacity allocation outcomes. Deregulated markets are more prone to capacity conflicts, especially with limited infrastructure capacity and increasing demand. These inefficiencies can therefore intensify in deregulated European markets such as Sweden, and hence the importance of a more efficient (and transparent) capacity allocation and conflict resolution.

Any adjustment to the current capacity allocation should abide by the legislation. On the one hand, EU policies provide guidelines related to capacity allocation and access charges. In **Appendix 1: EU directives**, the Articles state that conflict settlement can make use of access charges which may be included as an additional charge for scarcity. Such charges can be used to allocate capacity to the most important services to society in a fair and non-discriminatory manner. On the other hand, the Swedish

legislation (*Järnvägslagen*) also provides certain general guidelines for allocating infrastructure. In **Appendix 2: Swedish railway law**, the Clause states that Trafikverket is required to assess the capacity needs of the different types of services (including reserve capacity) and that, in case of unsettled conflicts after the coordination process, it is required to allocate capacity with the help of charges or priority criteria that yield (socioeconomic) efficient utilisation of the infrastructure.

The use of (CBA-based) priority criteria to settle capacity conflicts is generally aligned with the legislation. However, the use of such criteria in deregulated markets faces challenges, e.g., relevant data availability, and may lead to inefficient outcomes which goes against the guidelines. Both European and Swedish legislations allow for using a market-based capacity allocation, i.e., scarcity charges or pricing, an option that has been previously used to allocate capacity, e.g., for airport slots and public utilities. Thus, the lack of models and applications for railway capacity allocation in deregulated markets.

Parts of the conducted research consist of several methods that can help allocate capacity in deregulated markets, e.g., Sweden. Such methods can address many discussed challenges that relate to efficiency and transparency. Moreover, this work helps reduce the described existing gap in the research literature, e.g., capacity allocation and market deregulation in the railway sector.

3.2. Research questions

Several challenges and research gaps appear in the light of the literature review. This thesis attempts to answer a number of research questions (RQs) to help address some of the main cited challenges, and to fill in the mentioned research gaps.

To start with, this thesis reviews and analyses existing capacity allocation practices in a number of European railway markets.

RQ1. *What capacity allocation is used in current deregulated markets?*

Based on the answers to RQ1, a market-based and transparent capacity allocation is proposed. The focus is on improving the efficiency of existing capacity conflict solutions in important market segments.

RQ2. *How can capacity conflicts be more efficiently resolved between commercial and subsidised traffic?*

Answering RQ2 requires dealing with a number of other related RQs. First and foremost, capacity conflicts with commercial traffic are solved based on existing conventional CBA guidelines. An efficient capacity conflict resolution (to answer RQ2) therefore relies on the assumption that subsidised traffic supply is efficient according to these guidelines.

RQ3. *Is subsidised traffic supply efficient according to CBA guidelines?*

A second related RQ focuses on ways to use mathematical optimisation to improve RUs' traffic supply, e.g., train timetables.

RQ4. *How can mathematical optimisation be used to further improve the traffic supply?*

The third and last related RQ deals with demand data (origin destination or OD matrices), an important input data for more accurate policy decisions, e.g., more efficient traffic supply.

RQ5. *How much demand data is needed for more accurate policy decisions?*

In the remainder of this thesis, we will present the conducted research, results and contributions to address the presented RQs.

3.3. Research methodology

A top-down approach is followed to conduct this research, see **Figure 5** for an overview of the components of the methodology and the corresponding included papers.

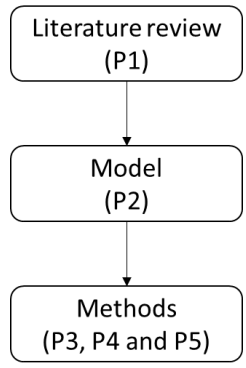


Figure 5. Flowchart of the adopted top-down research methodology.

A survey, in P1, of different European deregulated railway markets focuses on how capacity is currently allocated especially in case of conflicts. Conclusions from the survey, and the study by Eliasson and Aronsson (2014), help identify the need to develop a more efficient capacity allocation model for deregulated markets such as Sweden. The proposed model, in P2, focuses on allocating capacity in two important market segments, i.e., publicly controlled and commercial traffic.

Thereafter, the conducted research work aims at designing, implementing, and experimenting with several methods to help successively allocate capacity between publicly controlled and commercial traffic. This work includes methods, for instance, to construct efficient train timetables (P4), to estimate relevant input data such as passenger demand (P5) and CBA cost parameters (P3). More details about these methods are presented later in this chapter.

Various research methods are also adopted at different stages of this work. **Table 3** gives an overview of these methods and the corresponding included papers.

Table 3. Research methodology and adopted methods.

	Research methodology		
Research methods	**Literature**	**Model**	**Methods**
Qualitative text analysis	P1		
Cost benefit analysis (CBA)		P2	P3
Mathematical programming			P4 and P5
Passenger flow simulation		P2	P3 and P5
Data analysis	P1	P2	P3, P4 and P5

Qualitative analysis of official text material (e.g., network statements) is mostly used for the early literature review in P1. Quantitative methods are later used for the allocation model in P2. For instance, CBA is used to assess train timetables in P3, and to price capacity requests in P2. Mathematical programming is applied to model train timetabling in P4 and OD estimation in P5. Methods for passenger flow simulation are useful to assess timetables in P2 and P3, and estimate OD matrices in P5.

Extensive data analysis is used to study and compare different European railways in P1, and to describe the infrastructure and operations in P2 using the (microscopic) railway simulation software *RailSys* (Radtke and Bendfeldt, 2001). Train timetables are exported and manually adjusted to construct train path requests for testing in P2 and P4. Moreover, OD estimation methods in P5 use extensive passenger demand data from smart cards. Finally, CBA cost parameters, used in P2 and P3, are based on detailed trip valuation data from national and local guidelines (Trafikverket, 2016a, SLL, 2017).

3.4. Market-based capacity allocation

The literature review indicates that countries in Europe are increasingly adopting a deregulated market structure for both passenger and freight traffic. These reforms are driven at the EU level by an attempt to, among other things, stimulate competition. However, the incumbents are often favoured in capacity allocation, and still dominate most markets.

Traditional capacity allocation requires adaptations to best serve the new deregulated markets focusing on transparent and efficient allocation of capacity. Adaptations such as (CBA-based) priority criteria are however not always able to capture the marginal social benefit of certain services, e.g., private-commercial traffic due to limited data availability. The conducted research aims at studying possible improvements to this capacity allocation (presented in **Section 2.5**) in the light of the cited challenges and issues brought by the deregulation of railway markets, e.g., Sweden.

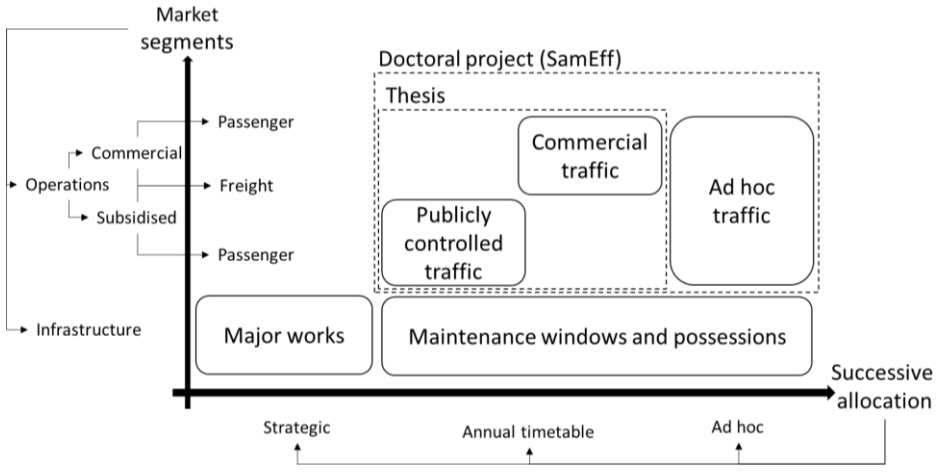


Figure 6. Successive capacity allocation in a segmented market.

For this, we consider a segmented deregulated market, and study successive allocation of capacity over these segments. **Figure 6** presents a simplified overview of the allocation of capacity in a segmented deregulated market such as Sweden. The horizontal axis represents the different steps of the allocation whereas the vertical one refers to the market segments. In such successive allocation, capacity is consecutively allocated to different segments, i.e., publicly controlled, then commercial traffic and finally ad hoc requests. Thus, it must not to be confused with incremental allocation (*successiv tilldelning*), a concept that was previously mentioned in **Section 2.6**.

Although the boundaries are not always clear, railway markets have several different segments depending, for instance, on the services, funding, and regulations. The main focus of this thesis is on publicly controlled (local or regional commuter) and commercial traffic segments as well as their interactions. Both passenger and freight services are included in the segment for commercial traffic. Interested readers are referred to the study by Froidh and Nelldal (2015) on the different types of traffic supply in Sweden after the deregulation.

Several years before the annual timetable, new infrastructure investments are decided based on government's transport strategic plans. The core step in capacity allocation is the construction of the annual timetable which is the scope of the doctoral project. We distinguish between publicly controlled and commercial traffic since these have different characteristics. On the one hand, publicly controlled (or subsidised) services cover mainly the operation market segment of unprofitable local

and regional passenger traffic, some trains (e.g., freight postal or passenger night services) can sometimes also be included in this segment but are not studied here. On the other hand, commercial services cover profitable market segments for freight and passenger traffic.

The capacity needed for publicly controlled traffic are applied for by the regional public transport authorities (PTAs or RKTM in Sweden), presumably based on social welfare considerations. Hence, an ideal reference timetable for these regional and local train services should aim for maximising the total societal welfare. Given relevant data, some methods (e.g., in P3) in this thesis can be used to achieve that.

For commercial traffic, licensed operators, both state-owned and private, may apply for train paths in the annual timetable. Which train paths (i.e., ideal timetable) to apply for is the results of their business plans which aim at maximising their profit, i.e., revenue minus operation costs. These operators often compete for capacity with each other and with other operators, including PTAs (Alexandersson et al., 2018). In this thesis, we focus on the inter-segment capacity conflicts between publicly controlled and commercial trains. Conflicts between commercial services, although important, fall mainly outside the scope of this thesis. Such conflicts can be resolved with methods such as auction (Affuso, 2003, Perennes, 2014).

The last step is the ad hoc allocation (or short-term planning) of late path requests. One way to allocate these train paths is to use a dynamic pricing (or yield management) scheme. Train paths are priced based on the capacity demand and supply (i.e., reserve capacity). This step is not further explained as it falls slightly outside the scope of this thesis. Such dynamic capacity pricing models are studied by Svedberg (2018), and later by Aronsson (2019) who looks at the overall supply of reserve capacity for ad hoc allocation.

Another important part of the last step, falling outside the scope of the thesis, is the allocation of capacity for infrastructure maintenance. Interested readers are referred to the doctoral thesis by Lidén (2018) for more details on how such capacity can be planned and allocated together with train services.

3.5. Subsidised traffic

Local and regional commuter trains are examples of services that are often part of the publicly controlled traffic. In deregulated markets, the

PTAs are responsible for this type of traffic often through concessions or *public service obligation* (PSO) contracts with RUs. These contracts are increasingly awarded based on *competitive tendering* (for-track competition) considering several key performance indicators (KPIs) such as costs, punctuality, sustainability, and innovation. Other special types of contracts also exist but not studied here, e.g., Public Private Partnerships (PPP).

Ideally, the PTA specifies the traffic supply aiming at maximising the societal welfare. That is to say that out of all the possible traffic plans (e.g., frequencies), it chooses the plan yielding the highest societal net welfare. Thereafter, the RU (awarded the contract) should execute the traffic plan in the best possible way under the conditions stated in the contract. An overview of the capacity allocation cycle for publicly controlled traffic is presented in **Figure 7** showing the scope of some of the included papers. Note that unlike the less detailed traffic plans (specified by the PTAs), train timetables (operated by the RUs) are more detailed translations of the traffic plan which should additionally consider all the operational constraints (e.g., crew, fleet, network infrastructure) for feasible and safe operations.

The PTAs face the challenging task of specifying a traffic plan that is as efficient as possible in terms of societal welfare. Based on this plan (and contract), the RUs have the (internal) task of finding an operational timetable that is commercially efficient, i.e., economically optimised. There are different ways to do this. The traditional method is to update, often manually, a reference traffic plan (e.g., from last or previous years) based on new information about population growth and urban development, etc. This is often done with the help of expert planners who have accumulated years of experience.

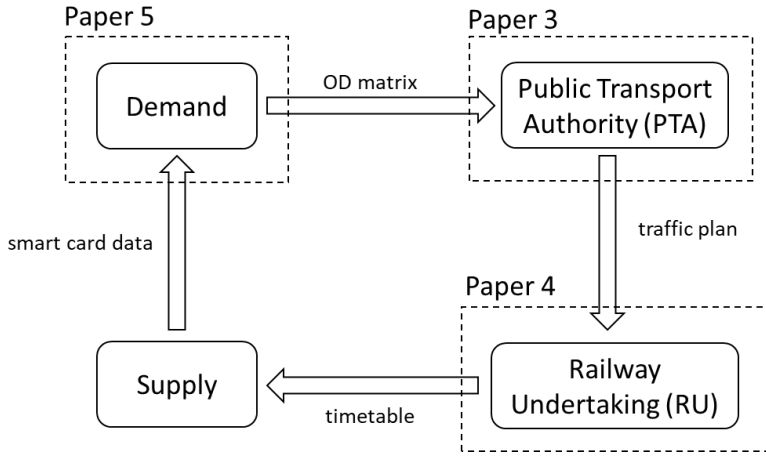


Figure 7. Overview of capacity allocation for publicly controlled traffic.

However, this thesis includes (e.g., in P3) a combination of methods from operations research and (micro-)economics to allocate capacity for publicly controlled traffic. In order to study the societal efficiency of a certain traffic plan for a publicly controlled service, CBA methods allow the PTAs to quantify and compare the welfare effects of these plans. Thereafter, timetabling optimisation methods (e.g., mathematical programming and Lagrangian relaxation), such as in P3, can be used to find the optimal plan that the RUs can execute, such as in P4, under the operational conditions and the infrastructure constraints.

Although not included in this thesis, it is possible to combine the two mentioned steps in one optimisation model, interested readers are referred to the related (i.e., part of the same project) licentiate thesis by Svedberg (2018). However, such train timetabling problems are more complex and thus harder to solve (Svedberg et al., 2015). This thesis contributes thus with models for both steps separately.

Note that the societal efficiency of the traffic plans using CBA requires the availability of relevant data, e.g., passenger demand and operating costs. Such data can be made available in publicly controlled traffic segments, e.g., smart cards as in P5, but not (necessarily) in others such as commercial traffic.

3.6. Commercial traffic

One of the main challenges of allocating capacity in deregulated markets is to, transparently and efficiently, solve capacity conflicts between the different (often competing) applicants. These applicants can be from different market segments, running different or complementary services. In this thesis, we focus on conflicts arising between the commercial (freight or passenger) traffic and the previously discussed publicly controlled ones. A situation which is increasingly common in heterogenous networks such as Sweden's where both types of traffic are steadily growing since the 1990s (Nilsson, 2016).

As discussed earlier, an alternative to the currently widely used priority criteria is a market-based allocation. In this case, it is important to correctly price the infrastructure capacity (i.e., train paths) to reflect marginal societal costs for more efficient capacity allocation outcomes (Perez Herrero, 2016). The capacity is hence allocated based on the prices for train path requests and the applicants' willing-to-pay (WTP).

As an illustration, a case study (from P2) looks at capacity conflicts between publicly controlled Stockholm commuter services and an inter-regional commercial passenger train. The idea is to study the loss in societal welfare for commuter services when scheduling the commercial train path. **Figure 8** presents different rescheduling scenarios to solve capacity conflicts.

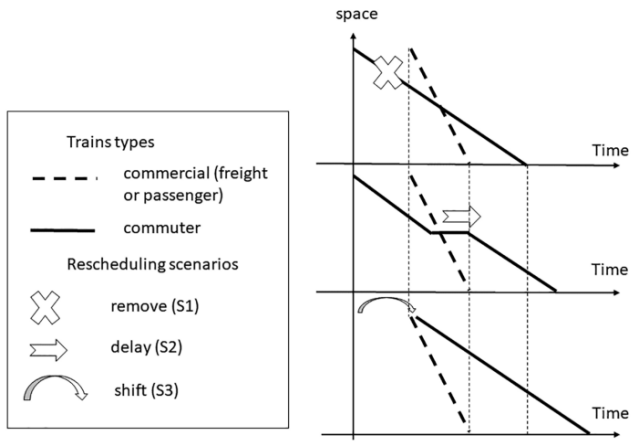


Figure 8. Illustration of different train path adjustments.

For different time periods of a typical working day, **Figure 9** presents the prices of the commercial train paths. Such prices change significantly depending on the time (peak or off-peak) and/or the rescheduling scenario (how train path requests are adjusted). An important assumption in this case study is that the reference commuter timetable is optimal, i.e., yields maximal total societal welfare. Thus, any rescheduling of this timetable will generate a loss in the total societal welfare, i.e., non-negative price for the commercial train path.

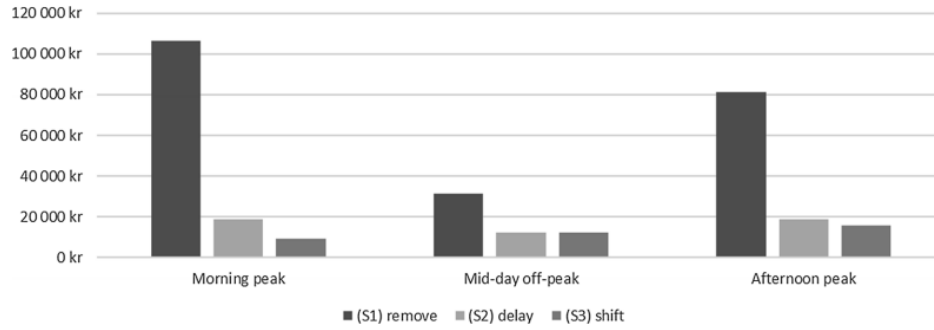


Figure 9. Example of commercial train path pricing (in SEK).

The price for commercial train paths is based on the marginal loss of total societal welfare for publicly controlled traffic. It serves as a reservation price for the path to be allocated. However, some paths can be requested by more than one applicant (e.g., in open access lines). In this case, the price can be used as an initial reservation price in an auction process between these applicants (Kuo and Miller-Hooks, 2015, Stojadinović et al., 2019).

3.7. Discussion

Several methods for allocating capacity in deregulated railway markets are previously briefly discussed. Further discussion of the conducted research is presented in this section.

Railways have high initial investment costs, for instance, to acquire the necessary *rolling stock* (Murillo-Hoyos et al., 2016). This can become a substantial economic burden and barrier to entry, especially for potential new entrants, often in need for several years to become profitable. Such considerations are not modelled in this thesis, however, one way to mitigate these entry barriers is the allocation of capacity over multiple years, also called traffic agreements (RNE, 2017). An additional solution

consists in setting up (state-owned) leasing companies assuming the ownership of the rolling stock.

An important issue relates to information asymmetry. In reality, the incumbent(s) in some railways may have substantial information advantage in the market compared to new entrants. In addition, the presence of cross-subsidisation may further magnify this advantage. Such market imperfections (or distortions) often lead to inefficient market outcomes. In this case, the regulators should intervene to ensure, for instance, transparent access to information, ticket sales channels and fair access charges for the new entrants. However, these asymmetric regulations, in favour of new entrants (on the expense of the incumbent), can threaten interoperability, e.g., within the harmonised SERA markets (Montero, 2019).

Another important issue is data quality and availability, both are assumed for most of the presented methods. For instance, demand data is used to calculate the loss in the total societal welfare for publicly controlled traffic which is in turn used to price commercial train paths. Better data quality is therefore important for a more efficient capacity pricing and allocation. However, this requires, among others, the development of suitable infrastructure for information and communications technologies (ICTs), e.g., through digitalisation. In this context, cybersecurity should be as important for safety as infrastructure maintenance.

Different types of congestion are mentioned in the conducted research. On the one hand, capacity congestions (on the track) are used to motivate the need for capacity allocation. These occur when the IM rejects train paths requests from one or more RUs due to capacity shortage, i.e., the capacity constraints (as in P4) are binding, and the corresponding multipliers (or shadow prices) can be interpreted as the marginal value of increased capacity, e.g., useful for planning future infrastructure investments. On the other hand, the cost functions (in P2 and P3) accounts for in-vehicle congestion costs as part of the consumer surplus (for subsidised passenger traffic). Such congestion is incurred by train passengers in the form of an increased perceived travel time (or sometimes denied boarding). With these two congestion perspectives in mind, it is important to incentivise RUs to efficiently use allocated train paths together with their rolling stock (e.g., number of wagons and train formation) in order to reduce both capacity and in-vehicle congestion.

As mentioned before, apart from the subsidised versus commercial traffic, other types also exist in different deregulated railways. For instance,

some RUs can be subsidised (for welfare maximisation), and at the same time be allowed to provide commercial traffic (and maximise profit). Another example is when PTA's traffic goes beyond the boundaries of the region, and thus compete with other regional subsidised or commercial services. There are multiple examples of such situations (in Sweden), and often require ad hoc solutions between the stakeholders, e.g., agreements between PTAs and/or RUs. Thus, these situations have not been studied in this thesis. Interested readers are referred to the study by Alexandersson et al. (2018) on such situations around the greater Stockholm region in Sweden.

Besides, political considerations can undermine the efficiency and transparency of the capacity allocation. One can suspect that *grandfather rights* and political lobbying are factors that may affect capacity allocation, although such practices are illegal and not allowed to affect capacity allocation decisions by the IM (Gestrelius et al., 2020). Another related issue that needs to be addressed is equity in the supply and accessibility of transport services (Rubensson, 2019). Finally, transferring profits abroad by foreign RUs should also be highlighted as these benefits are exported, i.e., not accounted for in the national gross domestic product (GDP).

Differentiation exists already in certain track access charges, e.g., passage or emission charges (Nilsson, 2018). However, the introduction of a demand-responsive market-based pricing can face resistance before being gradually accepted, e.g., road congestion pricing in Stockholm (Eliasson, 2008). Such resistance would come from the different stakeholders such as commercial freight and passenger RUs. Ways to soften the transition exist, e.g., gradual implementation through further developments and prototyping with the different proposed methods. One possible experiment is to use *gamification*, i.e., capacity allocation as a game between the different stakeholders (Meijer, 2015). Other more practical experiments could be conducted on specific national corridors as with several freight corridors in the TTR project (RNE, 2019).

Finally, European and national legislations do not seem to be ahead of the developments in the railway market, especially when it comes to capacity allocation, and how it should be implemented in deregulated markets. Thus, legal grounds need to be clarified and developed ahead of the new developments, e.g., (market-based) capacity allocation.

Chapter 4

Contributions & Future Works

‘C'est par la logique qu'on démontre, c'est par l'intuition qu'on invente’

It is by logic that we prove, but by intuition that we discover
Henri Poincaré, French polymath

4. Contributions and Future Works

This 4th chapter provides a summary for each included paper. It describes the papers' contributions to answer the corresponding research questions. The chapter ends with conclusions and insights for future works.

4.1. Summary of the papers

Summary of P1

A Survey of Railway Deregulation in Europe.

This review paper describes the railway deregulation in Europe by studying market organisation in several selected countries. It focuses mostly on deregulated markets which underwent major reforms following a number of EU directives (EC, 1991, EC, 2001, EC, 2012). These reforms introduced new market structures and more importantly new challenges for capacity allocation. One of the goals of these reforms is to introduce (or increase) competition, both for passenger and freight services. Thus, the need for capacity allocation that is transparent from a procedural perspective, clear and non-discriminatory. Such allocation also needs to be efficient from a market perspective, ensuring the best societal value.

The paper reviews aspects related to capacity allocation such as solving capacity conflicts on markets where several operators exist and compete for capacity. For each selected country, an extensive desk survey reviews a brief history of the market, existing actors, resolution of capacity conflicts and principles for calculating access charges, if any.

Based on the review, few if any countries use capacity allocation methods that are transparent and efficient. As to transparency, it is difficult for outsiders or new entrants, to understand the priority criteria (often second-best solutions) when capacity conflicts occur. Moreover, the actor responsible for capacity allocation (the IM) is sometimes related to the incumbent (often dominant) RU, and new entrants have reasonable concerns for discrimination in the market. As to efficiency, such aspects are rarely considered in the allocation even if the purpose of a competitive market is to ensure, in the long run, that train services which give the best value for money to consumers should get priority in allocating capacity, especially in case of conflicts. However, the

review indicates that efficiency considerations are surprisingly almost absent. Priority criteria have (at best) a vague relation to consumer demand and market efficiency. A vast majority of priority criteria and decision rules instead relates to simple administrative or technical criteria, e.g., longer train paths, passenger services over freight services, and timetable robustness.

The survey shows that most countries still have some way to go in opening the market for competition and benefiting from it. In particular, rules for allocating capacity need to be more transparent (from a procedural perspective) and more efficient (from a societal perspective).

Summary of P2

Pricing Commercial Train Path Requests Based on Societal Costs.

The paper describes how commercial train path requests can be priced based on societal costs. It considers the case where the railway market is deregulated. Inevitably, capacity conflicts arise, and in such cases the IM needs to prioritise between conflicting path requests. Depending on the ownership of the infrastructure, the IM can have different objectives. One of these is to allocate capacity in a way that maximises total societal benefits. The paper describes an approach to resolve conflicting capacity requests between commercial trains (maximising profit) and publicly controlled traffic (maximising welfare).

The model presented in this paper allows to calculate the societal costs (i.e., loss in social welfare) caused by changing the commuter train timetable to accommodate the commercial train path. Such costs include in-vehicle and waiting times, transfers, (in-vehicle) crowding and operating costs, and are used to price the commercial train path. The societal costs (or train path prices) are calculated using origin destination matrices (based on smart card data, studied in more details in P5 and RP3), time valuations (e.g., value of travel time and waiting time, the latter is studied in more details in P3) and parameters for operating costs.

The railway network in Stockholm is used as the case study. The results show that accommodating additional train paths in the busy commuter timetable comes at a high societal cost – much higher than the current charges (called *passageavgift* or passing fees) intended to partly reflect scarce capacity in transport hubs such as Stockholm. We also show that it is possible to substantially reduce the costs of

changes in commuter train timetables by choosing the best rescheduling alternative.

The main contribution of this paper is a method to calculate a reservation price for a commercial train path request by estimating the societal costs (i.e., loss of benefits) of the changes needed in a baseline commuter train timetable to accommodate this path request. If the commercial operator is willing to pay this reservation price, it is awarded the path and the commuter train timetable is adjusted; if not, the request is declined, and the commuter trains are given priority.

Summary of P3

Are Commuter Train Timetables Consistent with Passengers' Valuations of Waiting Times and In-vehicle Crowding?

This paper is an attempt to check if commuter train timetables are consistent with valuations of certain trip parameters such as waiting time and in-vehicle crowding that are estimated from passenger preferences and used in CBA. It is a follow-up study to P2 where such consistency is assumed, i.e., subsidised traffic supply is efficient according to CBA guidelines. Thus, this study compares passengers' valuations (i.e., traveller perspective) with the ones implied by the service frequencies in the PTA's traffic plan for commuting services (i.e., government's perspective).

Stockholm's commuter train services (*pendeltåg*) are used as a case study to compare the societally optimal and SL's actual frequencies where SL (*Storstockholms Lokaltrafik*) is the PTA in the region of Stockholm. Such comparison allows to estimate SL implicit valuations of the studied trip parameters (i.e., waiting time and crowding). Using an analytic CBA model, similar to the one presented in P2, this paper is a numerical study of the optimal frequency (or headway) on certain highly frequented lines. The results suggest that SL's timetables are not quite consistent with passengers' valuations.

In order to explain this inconsistency, i.e., SL frequencies being slightly higher than optimal, this study further estimates SL implicit valuation. For instance, these valuations for waiting times are found to be twice as high as the ones used in CBA guidelines, which are often estimated based on passenger (stated or revealed) preferences. Moreover, we find that the optimal frequencies are more sensitive to the waiting time

valuation than to that of crowding, implying lower levels of crowding if trains are assumed to be punctual (or robust timetables).

Even if the presented results remain inconclusive, due to different assumptions (e.g., no delays, frequency granularity and fixed operating costs), the work in this study provides an example of models that can be used by PTAs for more efficient (and transparent) allocation of capacity, especially if such capacity is shared with other subsidised and/or commercial services in a deregulated railway market.

Summary of P4

A disaggregate bundle method for train timetabling problems.

P4 studies the train timetabling problem (TTP) which refers to finding a feasible train timetable minimising a certain objective function. TTP is difficult to solve using the-state-of-art optimisation algorithms in a tractable period of time since it is NP-hard (Caprara et al., 2002). Therefore, solving the TTP often means finding a good quality solution within a given period of computation time.

This paper studies the existing TTP model by Brännlund et al. (1998), discretised in time and space, and formulated as a mathematical program, more specifically Integer Programs (IP). Alternative formulations exist such as Mixed Integer Programs (MIP), mostly used for continuous models in time and/or space (Forsgren et al., 2013).

The study derives an alternative solution method using lagrangian relaxation, called disaggregate bundle method, and compares its computational performances with the (standard) aggregate method, also used by Brännlund et al. (1998). The comparison is based on a real-world timetabling scenario from the Iron Ore line (*Malmbanan*) in Northern Sweden.

Numerical results indicate that the proposed solution method tends to give shorter execution times compared to the existing standard method. Moreover, the disaggregate method generates larger sets of possible train paths, a useful feature for constructing better feasible timetables.

Additional outputs also include optimal shadow prices (or multipliers), useful for analysing capacity scarcity in space and time, and for calculating the marginal value of new infrastructure (or maintenance)

investments. Hence, this study shows that the proposed approach has the potential to improve lagrangian-based solution methods for solving the TTP.

Summary of P5

The Value of Additional Data for Public Transport Origin-Destination Matrix Estimation.

With the increasing amount of data, generated from public transport (PT) systems such as smart cards, this final paper focuses on (dynamic) OD estimation in a railway PT network. The aim is to study the value of additional PT data when used for the estimation in an entry-only network, i.e., only the origin counts are known (from smart cards).

Using the principle of entropy maximisation (EM), the study compares the estimation quality or relative root mean square error (RMSE), when combinations of different data types are known and used for estimating the dynamic OD matrix. The RMSE of two policy-relevant estimated outputs are studied, namely the flows at exit stations and at links. Combinations of data types such as the number of alighters, average travel distance and certain link flows are valuated and compared for different time periods of the day (i.e., morning and afternoon peak hours, and midday off-peak).

The study uses extensive travel demand data (e.g., based on smart cards) from the Piccadilly line in London. The results indicate that, although inexpensive, certain data can be more valuable and considerably improve the quality of the estimation compared to more expensive and detailed data. The marginal value of such detailed additional data may be lower, especially when other data is already used in the estimation. Such results are inconclusive, and require further validation using, for instance, other estimation models, error metrics and additional data sources in other case studies.

4.2. Main contributions

In order to answer the research questions, the thesis has several main contributions which we present in this section.

RQ1. What capacity allocation is used in current deregulated markets?

C1. *Overview of deregulated railway markets in Europe:*
Capacity allocation aspects are reviewed in **[P1]**. With focus on European deregulated markets, the paper provides details on the current legislation, organisation, competition, capacity allocation and track access charges.

RQ2. How can capacity conflicts be more efficiently resolved between commercial and subsidised traffic?

C2. *Currently used ways to solve capacity conflicts:*
Focusing on European deregulated markets, **[P1]** reviews existing ways to solve capacity conflicts, and analyses their (dis-)advantages.

C3. *Pricing commercial train paths using marginal societal costs:*
[P2] describes a more efficient market-based allocation where commercial train path requests are priced based on their marginal societal costs on commuter traffic.

RQ3. Is subsidised traffic supply efficient according to CBA guidelines?

C4. *Assessment of societal costs for commuter traffic:*
Based on existing CBA guidelines, **[P2]** presents a method to calculate the societal costs of changes in the subsidised commuter traffic.

C5. *PTA's implicit valuation for waiting time and crowding:*
By comparing the current (PTA's) and optimal traffic supply, **[P3]** shows how PTA's implicit valuation for waiting and in-vehicle crowding can be inferred.

RQ4. How can mathematical optimisation be used to further improve the traffic supply?

C6. *Optimal frequencies for commuter traffic:*
In the absence of a closed form for the optimum, **[P3]** uses simulation-based optimisation to find the numerical values of the optimal frequencies for commuter traffic.

C7. *Improved method to solve lagrangian-based TTP models:*
As a first step to solve the TTP, **[P4]** studies an improved variant of bundle methods (called disaggregate) to find good quality solutions for the (relaxed) timetabling optimisation problem.

C8. *EM-based estimation models for dynamic OD matrices:*
An important input data to improve (passenger) traffic supply is the dynamic OD matrix. In **[P5]**, an EM-based estimation model is presented and used to find dynamic OD matrices based on smart card and additional data.

RQ5. How much demand data is needed for more accurate policy decisions?

C9. *Optimal traffic supply based on OD data:*
[P3] shows how dynamic OD data can be used to make more accurate policy decisions regarding commuter traffic supply.

C10. *Value of additional data for better dynamic OD estimates:*
More accurate policy decisions are based on more accurate OD estimates. **[P5]** shows that certain additional data can substantially improve the accuracy of dynamic OD estimates.

These contributions are of interest to different stakeholders within the railway market. **Table 4** provides a mapping of the main contributions, the papers, and the main stakeholders with examples from Sweden.

Table 4. Main contributions and interested stakeholder(s).

Stakeholder (Sweden)	RQ1	RQ2		RQ3		RQ4			RQ5	
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Regulator (Transportstyrelsen)	P1	P1								
IM (Trafikverket)	P1	P1	P2	P2			P4			
PTA (SL)				P2	P3	P3		P5	P3	P5
RU (SJ)							P4	P5	P3	P5
RU (Green Cargo)							P4			

4.3. Conclusions and future works

Deregulation of railway markets brought new issues and challenges to capacity allocation. This thesis deals with the problem of how to efficiently and transparently allocate available railway capacity to different competing applicants. For that, we present methods to successively allocate capacity in a segmented market.

The contributions of this thesis focus on CBA solutions that allow PTAs to efficiently create traffic plans for publicly controlled traffic. For the

IM, pricing train paths are studied to solve capacity conflicts and allocate capacity for commercial traffic. Moreover, various methods are developed for other stakeholders, e.g., data collection and train timetabling.

The case studies indicate that the proposed solutions have an important potential to be integrated in future capacity allocations, especially with the increasing scarcity in capacity. However, experimentation and prototyping are still needed. These can help address further concerns such as data, legislation and acceptability.

Future research work can build upon the methods presented in this thesis to develop and implement solutions for other segments such as infrastructure maintenance and ad hoc allocation (e.g., reserve capacity and real time traffic management). Other future works may also include further testing of the proposed solutions. For instance, gamification of the allocation process or more practical pilot projects on specific national corridors (as in the TTR initiative) can reveal further relevant insights. Such future works can also investigate digitalisation strategies, and look at how the legislation can be in line with the new developments.

References

ABBOTT, M. & COHEN, B. 2017. Vertical integration, separation in the rail industry: a survey of empirical studies on efficiency. *European Journal of Transport and Infrastructure Research,* 17**,** 207-224.

ABRIL, M., BARBER, F., INGOLOTTI, L., SALIDO, M. A., TORMOS, P. & LOVA, A. 2008. An assessment of railway capacity. *Transportation Research Part E-Logistics and Transportation Review,* 44**,** 774-806.

AFFUSO, L. 2003. Auctions of rail capacity? *Utilities Policy,* 11**,** 43-46.

AIT-ALI, A. & ELIASSON, J. 2019. Dynamic Origin-Destination Estimation Using Smart Card Data: An Entropy Maximisation Approach. *Preprint arXiv:1909.02826*.

AIT-ALI, A., ELIASSON, J. & WARG, J. Measuring the Socio-economic Benefits of Train Timetables Application to Commuter Train Services in Stockholm. 20th EURO Working Group on Transportation Meeting, EWGT 2017, 2017 Budapest. 849-856.

AIT-ALI, A., LINDBERG, P. O., ELIASSON, J., NILSSON, J.-E. & PETERSON, A. 2020a. A disaggregate bundle method for train timetabling problems. *Journal of Rail Transport Planning & Management***,** 100200.

AIT-ALI, A., WARG, J. & ELIASSON, J. 2020b. Pricing commercial train path requests based on societal costs. *Transportation Research Part A: Policy and Practice,* 132**,** 452-464.

ALEXANDERSSON, G., BONDEMARK, A., HENRIKSSON, L. & HULTÉN, S. 2018. Coopetition between commercial and subsidized railway services – The case of the greater Stockholm region. *Research in Transportation Economics,* 69**,** 349-359.

ALEXANDERSSON, G. & RIGAS, K. 2013. Rail liberalisation in Sweden. Policy development in a European context. *Research in Transportation Business & Management,* 6**,** 88-98.

ANDERSSON, E. V., PETERSON, A. & TÖRNQUIST KRASEMANN, J. 2013. Quantifying railway timetable robustness in critical points. *Journal of Rail Transport Planning & Management,* 3**,** 95-110.

ARONSSON, M. 2019. Reservkapacitet i tågplaneprocessen : Förstudie. *RISE Rapport*.

ARONSSON, M., FORSGREN, M. & GESTRELIUS, S. The Road to Incremental Allocation & Incremental Planning Content and Potential. 2012.

BORNDÖRFER, R., GRÖTSCHEL, M., LUKAC, S., MITUSCH, K., SCHLECHTE, T., SCHULTZ, S. & TANNER, A. 2006. An Auctioning Approach to Railway Slot Allocation. *Competition and Regulation in Network Industries,* 1**,** 163-196.

BOUF, D., CROZET, Y. & LÉVÊQUE, J. Vertical separation, disputes resolution and competition in railway industry. Thredbo 9, 9th conference on competition and ownership in land transport, 5-9 september 2005, Lisbonne., 2005 Lisbon Lisbon Technical University, 14 p.

BROMAN, E. & ELIASSON, J. 2019. Welfare effects of open access competition on railway markets. *Transportation Research Part A: Policy and Practice,* 129**,** 72-91.

BRÄNNLUND, U., LINDBERG, P. O., NÕU, A. & NILSSON, J.-E. 1998. Railway Timetabling using Lagrangian Relaxation. *Transportation Science,* 32**,** 358-369.

CAPRARA, A., FISCHETTI, M. & TOTH, P. 2002. Modeling and solving the train timetabling problem. *Operations Research,* 50**,** 851-861.

CROZET, Y. 2016a. Introducing competition in the European rail sector. *Discussion Paper prepared for the Roundtable on Assessing regulatory changes in the transport sector*.
CROZET, Y. 2016b. Liberalisatin of passenger rail services - France.
CROZET, Y. 2018. Case Study – France: logic and limits of full cost coverage. *In:* CERRE (ed.) *Track access charges: reconciling conflicting objectives.* CERRE & University of Lyon (LAET).
CROZET, Y., NASH, C. & PRESTON, J. 2012. Beyond the quiet life of a natural monopoly: Regulatory challenges ahead for Europe's rail sector. *Policy paper, CERRE, Brussels, December,* 24.
DE PALMA, A. & MONARDO, J. 2019. Natural Monopoly in Transport.
EC 1991. Council Directive 91/440/EEC of 29 July 1991 on the development of the Community's railways. European Commission.
EC 2001. Directive 2001/14/EC on the allocation of railway infrastructure capacity and the levying of charges for the use of railway infrastructure and safety certification. EU Parliament.
EC 2012. Directive 2012/34/EU on establishing a single European railway area. EU Parliament.
EC 2016. Fourth railway package of 2016. European Commission.
ELIASSON, J. 2008. Lessons from the Stockholm congestion charging trial. *Transport policy,* 15**,** 395-404.
ELIASSON, J. & ARONSSON, M. 2014. Samhällsekonomiskt effektiv tilldelning av järnvägskapacitet: några synpunkter på Trafikverkets nuvarande process. *Working papers in Transport Economics*. CTS - Centre for Transport Studies Stockholm (KTH and VTI).
FORSGREN, M. 2003. Computation of Capacity on railway Networks. *SICS Research Report*.
FORSGREN, M., ARONSSON, M. & GESTRELIUS, S. 2013. Maintaining tracks and traffic flow at the same time. *Journal of Rail Transport Planning & Management,* 3**,** 111-123.
FREEBAIRN, J. 1998. Access prices for rail infrastructure. *Economic Record,* 74**,** 286-296.
FROIDH, O. & NELLDAL, B. L. 2015. The impact of market opening on the supply of interregional train services. *Journal of Transport Geography,* 46**,** 189-200.
GESTRELIUS, S., PETERSON, A. & ARONSSON, M. 2020. Timetable quality from the perspective of a railway infrastructure manager in a deregulated market: An interview study with Swedish practitioners. *Journal of Rail Transport Planning & Management***,** 100202.
GIBSON, S. 2003. Allocation of capacity in the rail industry. *Utilities Policy,* 11**,** 39-42.
GILBO, E. P. 1993. Airport capacity: representation, estimation, optimization. *IEEE Transactions on Control Systems Technology,* 1**,** 144-154.
GOVERDE, R. M. P. & HANSEN, I. A. Performance indicators for railway timetables. 2013 IEEE International Conference on Intelligent Rail Transportation Proceedings, 30 Aug.-1 Sept. 2013 2013. 301-306.
HANSSON, L. & NILSSON, J. E. 1991. A new Swedish railroad policy: Separation of infrastructure and traffic production. *Transportation Research Part a-Policy and Practice,* 25**,** 153-159.
JENSEN, A. & STELLING, P. 2007. Economic impacts of Swedish railway deregulation: A longitudinal study. *Transportation Research Part E-Logistics and Transportation Review,* 43**,** 516-534.

KUO, A. & MILLER-HOOKS, E. 2015. Combinatorial auctions of railway track capacity in vertically separated freight transport markets. *Journal of Rail Transport Planning & Management,* 5**,** 1-11.
LAURINO, A., RAMELLA, F. & BERIA, P. 2015. The economic regulation of railway networks: A worldwide survey. *Transportation Research Part A: Policy and Practice,* 77**,** 202-212.
LIDÉN, T. 2018. *Concurrent planning of railway maintenance windows and train services*. Doctoral thesis, comprehensive summary, Linköping University Electronic Press.
LINK, H. 2016. Liberalisatin of passenger rail services - Germany.
LINK, H. 2018. Case Study – Germany. *In:* CERRE (ed.) *Track access charges: reconciling conflicting objectives*. German Institute for Economic Research (DIW Berlin).
LUSBY, R. M., LARSEN, J., EHRGOTT, M. & RYAN, D. 2011. Railway track allocation: models and methods. *OR Spectrum,* 33**,** 843-883.
MCAFEE, R. P. & MCMILLAN, J. 1996. Analyzing the Airwaves Auction. *Journal of Economic Perspectives,* 10**,** 159-175.
MCMILLAN, J. 1994. Selling Spectrum Rights. *Journal of Economic Perspectives,* 8**,** 145-162.
MEIJER, S. 2015. The Power of Sponges: Comparing High-Tech and Low-Tech Gaming for Innovation. *Simulation & Gaming,* 46**,** 512-535.
MERKERT, R. 2012. An empirical study on the transaction sector within rail firms. *Transportmetrica,* 8**,** 1-16.
MERKERT, R. & NASH, C. A. 2013. Investigating European railway managers' perception of transaction costs at the train operation/infrastructure interface. *Transportation Research Part a-Policy and Practice,* 54**,** 14-25.
MIZUTANI, F., SMITH, A., NASH, C. & URANISHI, S. 2015. Comparing the Costs of Vertical Separation, Integration, and Intermediate Organisational Structures in European and East Asian Railways. *Journal of Transport Economics and Policy,* 49**,** 496-515.
MONAMI, E. 2000. European passenger rail reforms: A comparative assessment of the emerging models. *Transport Reviews,* 20**,** 91-112.
MONTERO, J. J. 2019. Asymmetric regulation for competition in European railways? *Competition and Regulation in Network Industries,* 20**,** 184-201.
MURILLO-HOYOS, J., VOLOVSKI, M. & LABI, S. 2016. Rolling stock purchase cost for rail and road public transportation: random-parameter modelling and marginal effect analysis. *Transportmetrica A: Transport Science,* 12**,** 436-457.
NASH, C. 2008. Passenger railway reform in the last 20 years - European experience reconsidered. *Reforms in Public Transport,* 22**,** 61-70.
NASH, C., CROZET, Y., LINK, H., NILSSON, J.-E. & SMITH, A. 2018. Track access charges: reconciling conflicting objectives - project report. *In:* CERRE (ed.). CERRE.
NASH, C. A., SMITH, A. S. J., VAN DE VELDE, D., MIZUTANI, F. & URANISHI, S. 2014. Structural reforms in the railways: Incentive misalignment and cost implications. *Research in Transportation Economics,* 48**,** 16-23.
NILSSON, J.-E. 2002. Towards a welfare enhancing process to manage railway infrastructure access. *Transportation Research Part A,* 36**,** 419–436.
NILSSON, J.-E. 2016. Liberalisation of passenger rail services - Sweden.
NILSSON, J. E. 2018. Case Study – Sweden: Track access charges and the implementation of the SERA directive - promoting efficient use of railway infrastructure or not? *In:* CERRE (ed.) *Track access charges: reconciling*

conflicting objectives. VTI Swedish National Road and Transport Research Institute.
OECD 2005. Structural Reform in the Rail Industry. *Competition Policy Roundtables*.
PENA-ALCARAZ, M. M. T. 2015. *Analysis of Capacity Pricing and Allocation Mechanisms in Shared Railway Systems.* PhD, MIT - Massachusetts Institute of Technology.
PERENNES, P. 2014. Use of combinatorial auctions in the railway industry: Can the "invisible hand" draw the railway timetable? *Transportation Research Part A: Policy and Practice,* 67**,** 175-187.
PEREZ HERRERO, M. 2016. *Rail capacity constraints : an economic approach*. PhD, Université Lumière Lyon 2.
PETERSEN, E. R. 1974. Over the road transit time for a single track railway. *Transportation Science,* 8**,** 65-74.
RADTKE, A. & BENDFELDT, J. Handling of railway operation problems with RailSys. Proceedings of the 5th world congress on rail research. , 2001 Cologne.
RASSENTI, S. J., SMITH, V. L. & BULFIN, R. L. 1982. A Combinatorial Auction Mechanism for Airport Time Slot Allocation. *The Bell Journal of Economics,* 13**,** 402-417.
RIKSDAG. 2004. *Järnvägslag* [Online]. Infrastrukturdepartementet Available: https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/jarnvagslag-2004519_sfs-2004-519 [Accessed 2019].
RIKSTERMBANKEN. 2019. *Sweden's National Term Bank* [Online]. Available: http://www.rikstermbanken.se/ [Accessed 2019].
RNE 2017. Glossary of Terms Related to Network Statements. 7 ed.
RNE. 2019. *TTR General Introduction* [Online]. Available: http://ttr.rne.eu/general/general-introduction/ [Accessed 2019].
RUBENSSON, I. 2019. *Making Equity in Public Transport Count*. Doctoral thesis, comprehensive summary, KTH Royal Institute of Technology.
SLL 2017. Dokumentation av SAMS 3.0. Stockholm.
SMITH, A. 2016. Liberalisatin of passenger rail services - Britain.
STOJADINOVIĆ, N., BOŠKOVIĆ, B., TRIFUNOVIĆ, D. & JANKOVIĆ, S. 2019. Train path congestion management: Using hybrid auctions for decentralized railway capacity allocation. *Transportation Research Part A: Policy and Practice,* 129**,** 123-139.
SVEDBERG, V. 2018. *Towards optimal railway track utilization based on societal benefit*. Licentiate thesis, monograph, Linköping University Electronic Press.
SVEDBERG, V., ARONSSON, M. & JOBORN, M. Timetabling based on generalised cost. 2015.
TRAFIKVERKET 2016a. English summary of ASEK recommendations.
TRAFIKVERKET. 2016b. *Market-adapted planning of capacity (MPK) - approaches and tools for the future* [Online]. Available: https://www.trafikverket.se/en/startpage/operations/Operations-railway/improved-capacity/market-adapted-planning-of-capacity-mpk---approaches-and-tools-for-the-future/ [Accessed 2020].
TRAFIKVERKET 2020. Network Statement 2021. Swedish Transport Administration.
UIC 2004. UIC Code 406 - Capacity. *In:* RAILWAYS, I. U. O. (ed.).
VAN WEE, B., ANNEMA, J. A. & BANISTER, D. 2013. *The transport system and transport policy: an introduction*, Edward Elgar Publishing.

WARG, J., AIT-ALI, A. & ELIASSON, J. 2019. Assessment of Commuter Train Timetables Including Transfers. *Transportation Research Procedia,* 37**,** 11-18.
WETZSTEIN, M. E. 2013. *Microeconomic theory: Concepts and connections, second edition*, Taylor and Francis.

Appendix 1: EU directives

The following Points are excerpts from the SERA directive (EC, 2012).

Point 1 of article 39 on capacity allocation states that:

> *1. Member States may lay down a framework for the allocation of infrastructure capacity subject to the condition of management independence laid down in Article 4. Specific capacity-allocation rules shall be laid down. The infrastructure manager shall perform the capacity-allocation processes. In particular, the infrastructure manager shall ensure that infrastructure capacity is allocated in a fair and non-discriminatory manner and in accordance with Union law.*

Point 4 of article 31 on the principles of track access charges states that:

> *4. The infrastructure charges referred to in paragraph 3 may include a charge which reflects the scarcity of capacity of the identifiable section of the infrastructure during periods of congestion.*

Points 3 and 4 from article 47 on conflict resolution guidelines in congested infrastructures:

> *3.Where charges in accordance with Article 31(4) have not been levied or have not achieved a satisfactory result and the infrastructure has been declared to be congested, the infrastructure manager may, in addition, employ priority criteria to allocate infrastructure capacity.*

> *4. The priority criteria shall take account of the importance of a service to society relative to any other service which will consequently be excluded.*

Appendix 2: Swedish railway law

The following Clause is an excerpt from the Swedish railway law (*Järnvägslagen*) on the general guidelines for railway capacity allocation (Riksdag, 2004).

Clause 3 of the 6th chapter from 2004:519 [in Swedish]:

> *En infrastrukturförvaltare ska bedöma behovet av att organisera tåglägen för olika typer av transporter, inklusive behovet av reservkapacitet. Om ansökningarna om infrastrukturkapacitet inte kan samordnas, ska förvaltaren tilldela kapacitet med hjälp av avgifter eller i enlighet med prioriteringskriterier som medför ett samhällsekonomiskt effektivt utnyttjande av infrastrukturen.*

Included Papers

The papers associated with this thesis are the following:

P1. *A Survey of Railway Deregulation in Europe.*
- Submitted for journal publication

P2. *Pricing Commercial Train Path Requests Based on Societal Costs.*
- Published in
  - Transportation Research Part A: Policy and Practice, Volume 132, February 2020, Pages 452-464
  - https://doi.org/10.1016/j.tra.2019.12.005

P3. *Are commuter train timetables consistent with passengers' valuations of waiting times and in-vehicle crowding?*
- Submitted for journal publication

P4. *Disaggregation in Bundle Methods: Application to the Train Timetabling Problem.*
- Published in
  - Journal of Rail Transport Planning & Management, 100200
  - https://doi.org/10.1016/j.jrtpm.2020.100200

P5. *The Value of Additional Data for Public Transport Origin-Destination Matrix Estimation.*
- Submitted for journal publication

Paper P1

A Survey of Railway Deregulation in Europe.

Ait-Ali A.[1,2], Eliasson J.[2] (2019)
[1] VTI Swedish National Road and Transport Research Institute, Transport Economics (TEK), Stockholm
[2] Linköping University, Department of Science and Technology (KTS), Norrköping

Submitted for journal publication

Abstract

Most railway markets in Europe have been reorganized to allow competition between different operators. Several European countries have thus vertically separated their railway markets, separating infrastructure management from provisions of train services. This allows several train operators to compete for passengers and freight services. Different ways have emerged for vertical separation, capacity allocation and track access charges. This paper reviews, compares and discusses important deregulation aspects, using examples from different European countries to show different possible solutions. The study describes how competition has been introduced and regulated, with a particular focus on describing the different ways capacity is allocated and how conflicting capacity requests by different train operators are resolved. It also reviews the related issue of how access charges are constructed and applied. We conclude that few countries have so far managed to create efficient and transparent processes for allocating capacity between competing train operators. Although allowed by the legislation, market-based allocation is absent or never used, and incumbent operators still dominate certain markets. In order to foster more competition in European railway markets, capacity allocation processes need to become more transparent as well as efficient.

Keywords: railway markets; vertical separation; competition; capacity allocation; access charges.

1. Introduction

In the past, railway markets in most European countries were organized as single monopolistic companies controlling both infrastructure and railway services. In recent decades, however, many countries have introduced competition in railway markets by vertical separation, i.e., separating the responsibility for infrastructure from the provision of railway services for freight and passengers. Such developments have been further stimulated by the European legislation (EC, 1991, EC, 2001, EC, 2012, EC, 2016).

These recent reforms in Europe have brought new and various types and variants of market organizations, capacity allocation and track access charging. Allowing different (often competing) operators on the same track means that their capacity requests may come into conflicts. The process of allocating capacity and resolving conflicting capacity requests is therefore central for the functioning of these railway markets. This is highlighted in several studies in the literature (Gibson, 2003) as well as by the European legislation (EC, 2001). Ideally, the conflict resolution process needs to be both transparent, i.e., clear and non-discriminatory, and efficient, i.e., lead to societally and economically optimal outcomes. Such capacity conflicts may occur also in other vertically separated and deregulated markets, such as telecommunications (Klein, 1999) and air transportation (Gilbo, 1993), but these are not nearly as complex and have been more extensively researched compared to the railway sector.

This review provides an updated overview of the European railway deregulation focusing on capacity allocation and track access charges. Both are crucial instruments in such deregulated railway markets where different operators can compete for capacity. Based on the analysis of publicly available documents, we perform an up-to-date analytical comparison (in selected markets) of how competition was introduced and regulated, how capacity is allocated between competing train operators, and how track access charges are levied. The survey aims to add to the existing literature by describing, comparing and discussing various existing approaches in Europe. The current review is also one of relatively few studies that is the result of extensive desk research based directly on the national network statements, i.e., official documents providing descriptions of, among others, capacity allocation and access charges.

The paper starts with this introductory section. Section 2 presents the main existing related surveys in the literature, some general information on railway market organizations and European Union (EU) policy. The main part of this paper is in section 3 where we review the railway deregulation in a number of markets, selected to illustrate a range of different market organizations and capacity allocation processes. Section 4 concludes the review.

2. Existing Surveys

Structural reforms of European railway markets date back to the first European directive (EC, 1991), but the questions about market organization and capacity allocation are older. A number of existing surveys have reviewed, compared and/or analyzed aspects of these reforms, e.g., market organization, competition, capacity allocation and access charges. In this section, we present some related studies, and show how our study contributes to the existing literature. **Table 1** provides a comparison between this paper and the main existing surveys in terms of the main reviewed aspects as well as the studied markets.

In the late 1980s and after the pioneering vertical separation and deregulation in Sweden, Hansson and Nilsson (1991) described the market reorganization, institutional aspects and practical problems inherent to the new reforms. After directive 91/440/EEC (EC, 1991), a few other countries (e.g., the UK and Germany) followed Sweden shortly after and new market organizations emerged. Monami (2000) compared different organizational models in certain markets (i.e., Belgium, France, Germany, the UK), and identified key dimensions to describe each model and how they are connected.

In 2001, the directive 2001/14/EC set guidelines for railway capacity allocation and track access charges (EC, 2001). Since then, railway market reforms have been successively implemented in many other member states of the European Union (EU), and more studies have followed covering more aspects of the reforms. The Organisation for Economic Cooperation and Development (OECD) published a comprehensive summary of the structural reforms that have happened in all the OECD member countries (OECD, 2005). Another extensive survey by Laurino et al. (2015) reviewed the market organization in 20 countries worldwide with different regulatory approaches to deal with the monopolistic nature (i.e., natural monopoly) of railway infrastructure.

Table 1. Comparative overview of existing surveys.

Reference (chronologically)	Main aspect(s)	Market(s)
Hansson and Nilsson (1991)	Market organization	SE
Monami (2000)	Market organization	BE, FR, DE, GB, SE
Link (2004)	Access charges, competition	DE
Crozet (2004)	Access charges	EU
OECD (2005)	Market organization	OECD
Bouf et al. (2005)	Capacity allocation (conflicts)	FR, GB
Jensen and Stelling (2007)	Performances	SE
UIC (2009)	Access charges (noise)	EU
Friebel et al. (2010)	Performances	EU
Crozet et al. (2012)	Competition (passenger)	EU
Van de Velde et al. (2012)	Performances	EU
OECD (2013)	Market organization, competition	OECD
Alexandersson and Rigas (2013)	Market organization, capacity allocation	EU
Nash et al. (2013)	Competition, performances	SE, GB and DE
Nash et al. (2014)	Performances (costs)	EU
Laurino et al. (2015)	Market organization	20 countries (worldwide)
Nash et al. (2016)	Competition (passenger)	SE, FR, GB, DE
Crozet (2016a)	Competition	EU
Abbott and Cohen (2017)	Efficiency	EU
Nash et al. (2018)	Access charges	SE, FR, GB, DE
This paper (2020)	**Market organization, competition, capacity allocation, access charges**	**EU, CH, GB (US and JP are also mentioned)**

There are many reviews on competition and/or efficiency in railway markets. In Germany, Link (2004) discussed the problems facing on-track competition in the regional passenger market by analyzing the effects of access conditions and charges. In a similar Swedish study, Alexandersson and Rigas (2013) studied the European effects of the Swedish policy for opening access to the passenger market since 1988 until the complete deregulation in 2012. In a related comparative study, Nash et al. (2013) reviewed the introduction of competition in Britain, Germany and Sweden. At the European level, Crozet et al. (2012), in a policy report for the Centre on Regulation in Europe (CERRE), discussed vertical separation as a first attempt to increase railway efficiency. Based on analysing lessons from opening markets to competition, the authors identified key issues and policy recommendations to tackle the regulatory challenges for the implementation of competition in the European market. Similar topics

were also discussed within the OECD's policy competition roundtables on the recent development in railway markets (OECD, 2013). A later study by CERRE focused on the liberalization of passenger rail services, the authors reviewed and analyzed the markets in France (Crozet, 2016b), Germany (Link, 2016), Great Britain (Smith, 2016) and Sweden (Nilsson, 2016). Closely related, a discussion paper from the International Transport Forum (ITF) by Crozet (2016a) described how competition was introduced in several European countries.

A number of authors have focused on the efficiency and performances of different market organizations, e.g., Jensen and Stelling (2007), Asmild et al. (2009), Friebel et al. (2010), Van de Velde et al. (2012), Nash et al. (2014) and Abbott and Cohen (2017). Although important, these aspects fall mostly outside the scope of the current survey.

Reviews of capacity allocation and conflict resolution processes are scarcer, but there are some, mostly covering a relatively small number of countries each. Bouf et al. (2005) focuses on conflict resolutions in the allocation process in France and Britain. The study looked specifically at the dispute/conflict resolution systems between the infrastructure manager and railway undertakings as a result of the vertical separation.

As to access charges, Crozet (2004) reviewed the charging systems in several European countries a few years after the 2001 directive (EC, 2001). The author attempted to find some best practices for infrastructure charging easily transferable between countries. A more specific report from the International Union of Railways (IUC) focused on noise-related access charges in Europe (UIC, 2009). In another recently published study by CERRE, case studies reviewed how track access charges are levied in four European railways, i.e., France (Crozet, 2018), Germany (Link, 2018), Sweden (Nilsson, 2018) and Great Britain (Smith and Nash, 2018).

3. Railway deregulation in Europe

In this section, we review several aspects that relate to the railway deregulation. In particular, we present the European reforms and legislation, and analyze the market organization, vertical separation, competition and track access charges. These aspects are reviewed and discussed in the context of the following European countries: Austria (AT), Belgium (BE), Czech Republic (CZ), France (FR), Germany (DE),

Great Britain (GB, Northern Ireland is omitted), Italy (IT), the Netherlands (NL), Spain (ES), Sweden (SE) and Switzerland (CH).

The markets illustrate the various ways railway markets are organized and capacity is allocated, and have been selected as examples for various reasons. Some were pioneers in railway deregulation (i.e., SE, GB and NL) and market opening (e.g., CZ) while others have important European cross-border traffic (e.g., AT, BE and CH) or extensive high-speed networks (e.g., FR, DE, IT and ES). These reasons all mean that these railway markets have to deal with potentially complex issues regarding regulation and competition, and hence their regulatory processes and framework provide interesting insights and conclusions as European railway markets become increasingly deregulated and potentially more open for competition between several operators. Note that some markets are selected for more than one cited reason. In certain discussions, Japan (JP) and the United States (US) are also briefly mentioned to contrast with their special market organization. **Figure 1** presents a map showing the selected railways and some of the selection reasons.

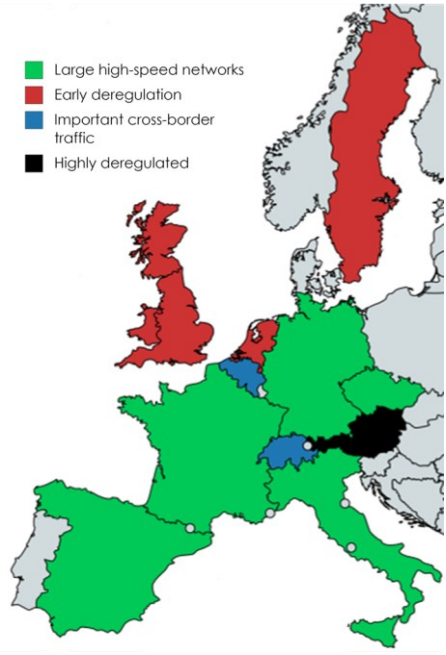


Figure 1. Selected European railways.

The selected markets (including JP and US) are used as case studies that serve for illustration, analysis and discussion, and were also selected based on data availability to illustrate the broader range of market organization and competition, capacity allocation and access charges. For that, country-specific documents are used as the primary source material, e.g., latest national network statements describing (among other things) capacity allocation and access charges. These are complemented with updated information from recent comparative studies. Secondary references, e.g., reports and data from inter-governmental organizations and academic papers, are also used but to a less extent.

A historical overview is first presented including the main developments toward the deregulation of the European railway. Second, different market organizations are described and discussed before focusing on vertical separation. A review of competition and capacity allocation follows. This section is concluded with a brief discussion on track access charges and their use to solve capacity conflicts.

3.1. Historical overview

Early railways were built, operated, maintained and owned by private companies. Railway networks continued their expansion thanks to the many private investors during the industrial revolution (sometimes called the *railway mania*). Further developments, such as increasing passenger traffic and fierce competition between investors, made governments pay increasing attention to rail transport. A combination of railways' growing societal importance, decreasing profitability for railway companies and a strive to take advantage of various economies of scale made most (although not all) European countries nationalize large parts of the railway networks and establish national railway monopolies during the early 20th century.

During the late 20th century, the railway sector faced new challenges such as declining rail modal share due to increasing competition from other modes. Decreasing efficiency and increasing government expenditures with poor performances meant that state-controlled railways came under pressure (OECD, 2005), and a trend of deregulation reforms emerged to allow private actors in the market once again (Laurino et al., 2015). Sweden was the first country to initiate such deregulation (as early as 1988) after the vertical separation between railway operations and infrastructure management (Hansson and Nilsson, 1991). SJ (the government agency that managed the national railways

until 1988) became a railway undertaking (e.g., operator) whereas the infrastructure management was transferred to Banverket (the Swedish Rail Administration). In 2001, SJ was further split into several state-owned companies (e.g., the passenger operator SJ and the freight operator Green Cargo). In 2010, Banverket was merged with Vägverket (the Swedish Road Administration) to form Trafikverket (the Swedish Transport Administration).

After the Swedish deregulation and to stop the decline in the rail sector and increase its competitiveness, several EU reforms and policies have been introduced as early as 1991 in the form of different directives and regulations grouped in successive railway packages. The European Commission (EC) first introduced directive 91/440/EEC about vertical separation distinguishing between three alternatives, i.e., accounting, organizational and institutional (EC, 1991) separation. The first type guarantees separate financial accounts, the second is about independent units within one larger institution and the third is complete separation. The directive required at least a separation in terms of accounting. A timeline in **Figure 2** shows the history of European vertical separation as well as EU railway packages.

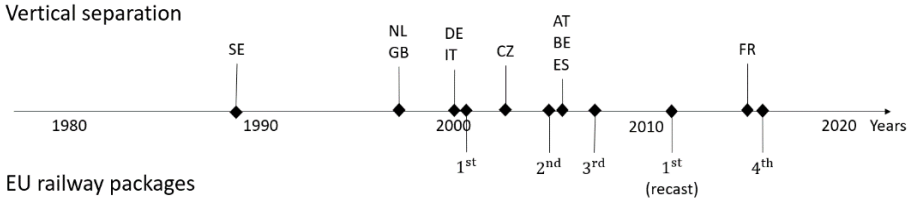


Figure 2. Timeline of the history of European railway packages and vertical separations.

Several EU member states followed to vertically separate their national railways, and to deregulate a number of rail services such as international and long-distance passenger, freight, and maintenance (Monami, 2000, Nash, 2008). These reforms came as a response to calls from the European Commission (EC) to, among other things, promote transparent access and efficient utilization of existing rail infrastructure (EC, 2001) in the Single European Railway Area or SERA (EC, 2012). **Table 2** presents different railway packages, directives and the corresponding treated issues.

Table 2. European railway legislation and its main topics (and/or requirements).

Package	Year	Name	Main topics (and/or requirements)
	1991	440/EEC	Vertical separation (at least accounting)
	1995	18/EU	Licensing railway companies
1st	2001	12/EU	Open access for cross-border freight
		14/EU	Capacity allocation and access charges
2nd	2004	49/EC	Railway safety
		50/EC	Interoperability (mainly high speed)
		51/EC	Open access for domestic freight services
		881	European Railway Agency (ERA), safety and interoperability
3rd	2007	58/EC	Open access for international passenger services
		59/EC	Harmonized license for train drivers
		1370	Open access for subsidized passenger services
		1371	Rights of rail passengers, e.g., delay compensation
1st (recast)	2012	34/EC (SERA)	Harmonization of the track access charges
	2015	909	Modalities for the calculation of track access charges
4th	2016	797	Interoperability
		798	Safety (recast of 2004/49/EC)
		2370	Open access for domestic passenger services
		2338	Public service obligations (PSO contracts)

The first directives focused on the fundamental reforms of the market reorganization and regulation, e.g., vertical separation and licensing. The 1st package of 2001 required all EU markets to be vertically separated (at least in accounting), making their markets ready for open access or new entrants and hence deregulation. The following packages aimed at the successive deregulation of different market segments, e.g., cross-border freight (2001), domestic freight (2004), international passenger (2007) and domestic or national passenger (2016) in the recent fourth railway package. Thus, open access for passenger services (except for international services) has only been recently required, and competitive tendering is not yet a requirement (EC, 2016).

Several legislations provided guidelines for deregulation aspects such as safety, interoperability and licensing, and more importantly capacity allocation and access charges. In the SERA directive, capacity allocation is treated in the 1st point[1] of Article 39 stating that it is the responsibility of the infrastructure manager to allocate capacity in a fair and

[1] *1. [...] The infrastructure manager shall perform the capacity-allocation processes. In particular, the infrastructure manager shall ensure that infrastructure capacity is allocated in a fair and non-discriminatory manner and in accordance with Union law.*

non-discriminatory manner (EC, 2012). Another important aspect of the allocation process relates to solving conflicts between capacity applicants. In the same directive, the 4th point[2] of Article 31 as well as 3rd and 4th points[3] of Article 47 treat access charges and congested infrastructure where capacity conflicts occur. The first points state that access charges can include an additional charge for scarcity. If conflicts persist, the two other points suggest the use of priority criteria to allocate capacity to the most important services to society.

3.2. Market organizations

The structure of railway markets can be characterized according to the extent of vertical and/or horizontal separation (or integration). **Figure 3** illustrates the main market reorganizations based on the two dimensions. The arrows indicate the structural reforms (vertical or horizontal, separation or integration) that are needed to move from one market organization to another.

The vertical dimension involves the division of responsibility between infrastructure management and railway services. The responsibilities of the former include tasks such as development and maintenance, traffic control and capacity allocation, sometimes also real estate and stations. As for railway services, these include running trains and related tasks such as ticket sales. Licensed operators, also called railway undertakings, are allowed to provide train services.

A common structure is vertical (and horizontal) integration, see top left in **Figure 3**. One actor, often a state-owned company, is responsible for the entire national railway system, being at the same time the infrastructure manager and a monopolistic operator. Another variant with a long history is one with several distinct railway networks or sub-markets where each is vertically integrated, see top-right in **Figure 3**. An example of the latter is Japan after the 1987 privatization of the Japanese National Railways (JNR) into Japan Railways Group JRG. The group consists of six (horizontally separate) private passenger

[2] *4. The infrastructure charges [referred to in paragraph 3] may include a charge which reflects the scarcity of capacity of the identifiable section of the infrastructure during periods of congestion.*

[3] *3.Where charges [in accordance with Article 31(4)] have not been levied or have not achieved a satisfactory result and the infrastructure has been declared to be congested, the infrastructure manager may, in addition, employ priority criteria to allocate infrastructure capacity.*
4. The priority criteria shall take account of the importance of a service to society relative to any other service which will consequently be excluded.

companies (the government is the sole shareholder), organized by regions (e.g., JR Hokkaido, JR East). Each one is responsible for both infrastructure management and railway operations (vertical integration) in their respective regions (Trafikanalys, 2014). Another example is the United States which is dominated by freight services, while passenger services have a relatively small market share. The freight market includes many private operators which generally own the infrastructure they use (vertical integration) but are separated (horizontally in infrastructure) into several distinct railway systems.

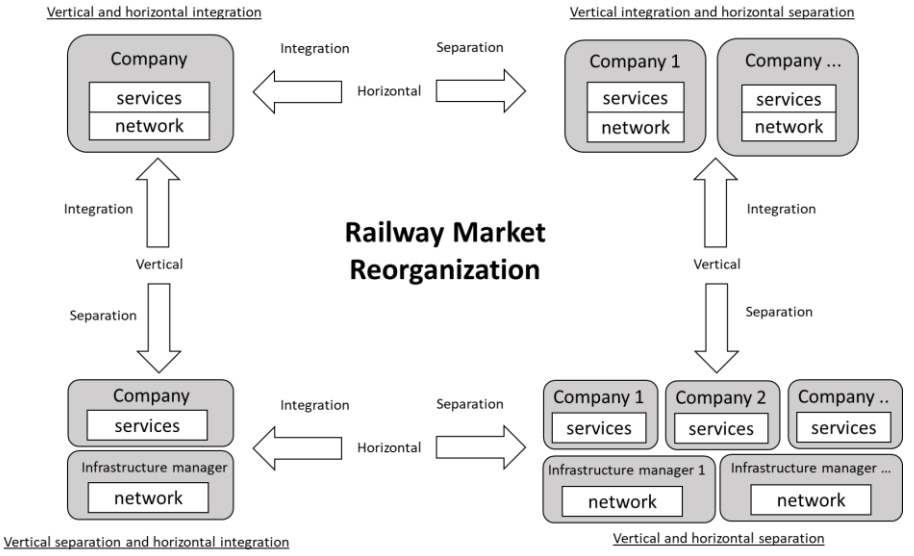


Figure 3. Railway market reorganizations in horizontal and vertical dimensions.

As a result of the reforms, vertical separation became the main market organization in Europe, see bottom in **Figure 3**. Although not illustrated in the figure, various forms of vertical separation exist which lead to players that have different legal status, e.g., state-owned or holding companies, subsidiaries or governmental agencies. A more detailed discussion about vertically separated market organizations is presented later in the section.

The horizontal dimension concerns the relationship between different market players with similar roles or responsibilities, such as different infrastructure managers or different railway undertakings (Yeung, 2008). In a horizontally separated market (i.e., to the right in **Figure 3),** there may be several railway operators providing competing or

complementary services (separation in services), or several infrastructure managers with responsibilities for different parts of the network (separation in infrastructure). There are various ways to allocate capacity if the market is also vertically separated (bottom-right in **Figure 3**), e.g., franchising, competitive tendering or open access. To foster competition in such markets, no company must be discriminated or favored, and capacity allocation is regulated, as in Europe, by an independent rail regulator.

Table 3 presents examples of market organizations in different countries. In contrast to the European markets, Japan and the US have different structures, where passenger (in JP) and freight (in the US) companies are vertically integrated (Trafikanalys, 2014). However, the state-owned (horizontally separated) freight (in JP) and passenger (in the US) companies have certain rights regarding access to the infrastructure capacity (Talebian et al., 2018). Although horizontally separated, Switzerland is one of few remaining vertically integrated railways in Europe.

Table 3. Examples of market organization in different countries

	Vertical	**Horizontal**
Integration	JP (passenger), US (freight), CH	JP (freight), US (passenger)
Separation	AT, BE, CZ, FR, DE, GB, IT, NL, ES, SE, JP (freight), US (passenger)	AT, BE, CZ, FR, DE, GB, IT, NL, ES, SE, JP (passenger), US (freight), CH

All EU member states have reorganized their railway markets by vertically (and horizontally) separating their monopolistic national railways. Although stipulated by the same European legislation, the market reorganizations (or vertical separation) were not always similar in different European markets.

3.3. Vertical separation

Already in the late 1980s, Sweden began to vertically separate their railway markets into infrastructure management and railway services (bottom-left in **Figure 3**). This was a first step towards opening the market for competing new entrants. The vertical separation was later adopted by several EU policies, and aimed to foster competition and interoperability (EC, 2001). The study of the effects of the reforms is outside the scope of this survey but a number of studies exist on the reforms in freight markets (Ludvigsen, 2009), on interoperability (Abbott and Cohen, 2017) and on transaction costs (Merkert, 2012).

A typical example of vertical separation is when a government agency is responsible for infrastructure management, while one (incumbent) company (or more) is responsible for providing railway services. This corresponds to the bottom-right in **Figure 3** with a single infrastructure manager as represented in **Figure 4**. The railway undertakings, responsible for operations, can include the existing incumbent (if any), new national companies and/or other players from abroad (incumbents or new entrants). These companies (privately or publicly owned) can operate passenger and/or freight services (commercial or subsidized). Note that the distinction is sometimes blurred, and no general rules exist for which services should be subsidized. In Europe for instance, commuter and regional services are often subsidized by local or central public transport agencies, but intercity, long-distance, high-speed, international and freight train services are generally market-based.

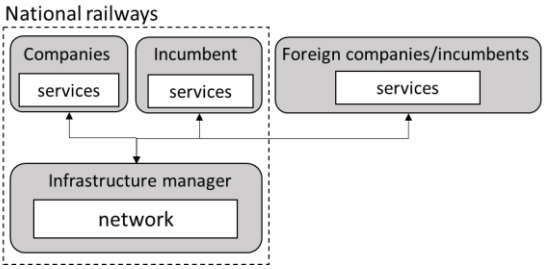


Figure 4. Vertical (and horizontal) separation in European markets.

Most of the selected countries have adopted a vertically (and horizontally) separated railway market organization. The exception is Switzerland which mostly remained vertically integrated. It has adopted, however, a horizontal separation between several railway companies since there are several networks and operators mostly owning their infrastructure, the largest of which is SBB controlling around 58% of the Swiss railway network (SBB, 2020).

There are currently two equally frequent forms of vertical separation, from the three possible alternatives that are allowed by the legislation. **Table 4** presents the vertical separation and infrastructure management in selected markets. The institutional (or complete) separation is mostly found in countries with early deregulation and high level of market competition, e.g., SE and GB. Many other countries have adopted a separation where the infrastructure management (together

with the incumbent operator) is a subsidiary of a parent or holding company, e.g., DE and FR.

Note that Switzerland, Japan and the US are not included in the table since all are vertically integrated. In Japan and the US, infrastructure management (including capacity allocation) is the responsibility of the state-owned companies, for passengers in JP, and private companies for freight in the US. Switzerland has, however, a not-for-profit agency, Trasse, which is the infrastructure manager and thus allocates capacity for licensed railway companies (e.g., SBB, BLS and SOB).

Besides the various types of vertical separation, **Table 4** shows that there are different forms for managing the infrastructure. In institutional vertical separation, the infrastructure manager may be a state-owned company which is not a subsidiary or part of any other parent or holding company unlike organizational vertical separation. Another form for infrastructure management is an independent government agency which has no commercial or business interest. Such form is found in the Netherlands with ProRail (ProRail, 2020) and in Sweden with Trafikverket (Trafikverket, 2020b).

Table 4. Vertical separation and infrastructure management in selected European markets.

Vertical separation	Infrastructure manager	Countries
Organisational	Subsidiary of a holding company	AT, IT
	Subsidiary of a parent company	BE, FR, DE
Institutional	State-owned company	CZ, GB, ES
	Governmental agency	NL, SE

In markets with organizational separation, the infrastructure manager is usually either a subsidiary of a holding company which is built to solely hold shares such as in Austria and Italy, or otherwise of a larger parent company which has other activities inside (or outside) the rail industry such as in Belgium, France and Germany. Both variants may lead to conflicts of interests when it comes to solving capacity conflicts, since the parent or holding company controlling the infrastructure manager may also have companies in the market. Link (2004) concludes that failing to find an appropriate organizational framework could be an obstacle for fostering competition and system efficiency.

In addition to the infrastructure manager, vertically separated markets include important players such as the incumbent and the regulator. Incumbent companies may remain after the vertical separation of the

national railways, whereas the independent regulator exists to ensure that there are no discriminatory practices in the market, for example the Office of Rail and Road (ORR) in Britain. In some cases, no incumbent remains after the separation – this is for example the case in Britain. A list of the existing incumbent operator(s) and their relation to the infrastructure manager is given in **Table 5** for the selected markets. Note that GB is not included as no incumbent exist in the market.

As mentioned before, the incumbent is completely independent from the infrastructure manager after institutional vertical separation, also called complete separation. However, in the case of organizational separation, dependencies may remain, meaning that conflicts of interests may emerge. The independent regulator must ensure that the incumbent and the infrastructure manager have no anti-competitive practices during the allocation of capacity to the different players in the market.

Table 5. The incumbent(s) in selected markets and their link to the infrastructure manager.

Country	Incumbent(s) in the market	Link to the infrastructure manager
Austria	ÖBB	Same holding company (ÖBB)
Belgium	SNCB	Same parent company (SNCB)
Czech	ČD	None
France	SNCF Voyageurs (passenger), SNCF Logistics (freight)	Same parent company (SNCF)
Germany	DB Bahn (passenger), DB Schenker (freight)	Same parent company (DB)
Italy	Trenitalia	Same holding company (FSI)
Netherlands	NS (passenger)	None
Spain	Renfe	None
Sweden	SJ (passenger), Green Cargo (freight)	None

3.4. Competition

The new market organization is not an ultimate goal in itself, but a means to foster more competition in the market in order to increase the efficiency and quality of the railway sector. In this context, introducing vertical separation (and hence competition) is not even a necessary condition for good quality train services, e.g., CH. However, such separation is a necessary step towards a more competitive market as promoted by the European legislation.

Even in vertically separated markets, the incumbent operator(s) can still hold large shares of the freight and/or passenger market. In some cases, incumbents in their own home market are themselves new entrants or have subsidiaries in other markets (e.g., the German DB and the French SNCF). The presence of a dominant incumbent as the main player in the market can sometimes prevent competition on the track (e.g., open access) as well as the upcoming competition for the track (e.g., competitive tendering). This is since benefits of scale make it difficult for new entrants to compete often due to high initial investment costs. Other entry barriers might also exist such as access to ticket sales platforms and information asymmetry. The market shares (in train-km) are presented in **Figure 5** for both freight and passenger incumbents in the selected markets. Note that GB and NL (freight) are not included.

Figure 5. Market share (based on train-km) of the incumbent for freight and passenger services (IRG, 2019).

Analyzing **Figure 5** indicates that most of the competition is still happening in the freight segment of the railway market. One reason could be that freight markets were deregulated first, before passenger markets. It is also important to mention that some railway systems (e.g., NL) are dominated by passenger traffic which could also play an important role in fostering competition regardless of the market organization. Competition in passenger markets is generally less intense. Some countries have almost no competition for passenger services (e.g., FR, ES). The exception in this survey is Britain (with mostly franchising contracts and no incumbent) and Sweden which were both among the first countries to deregulate their markets. Germany has, to some extent, competitive freight market, and has also recently

increased the shares of new entrants with more open access contracts for passenger services, e.g., Flixtrain.

Most of the competition in the passenger market segment is on profitable commercial lines, e.g., international long distance and high-speed lines. Competition on these lines is often in the form of open access for train path(s). Publicly controlled subsidized passenger services (e.g., regional) are also expected to have more competition in the form of competitive tendering for long term contracts. There is a substantial grey area here and drawing a clear line between these two types of services is often difficult. As a rule, intercity services are usually commercial, whereas local and regional services are often subsidized. In this context, the fourth European railway package from 2016 aims to increase competition in rail passenger markets by adopting open access for commercial lines and competitive tendering for subsidized ones (EC, 2016).

An important feature of vertically separated markets with high competition is that solving capacity conflicts should be done in a both transparent and (socio)economically efficient way. This is the aim, at least, in the EU (as well as GB and CH), as opposed to first ensuring the benefits of the infrastructure owners (e.g., JP and the US). Most of the reviewed countries with a low degree of competition have a market organization in which the infrastructure manager is somehow linked to the incumbent operator (e.g., BE, FR and IT). As mentioned before, this conflict of interest may discourage new entrants and decrease competition. This is particularly salient in the case of capacity conflicts in the allocation process where both new entrant(s) and the incumbent request conflicting train path(s) from an infrastructure manager which is owned by the same parent/holding company (that controls the incumbent). As described in the next sections, certain countries (e.g., FR) have more general (less precise) rules for capacity allocation and conflict resolution criteria than others (e.g., BE and IT). Such more general rules tend to increase the uncertainty for the new entrants. One way to avoid this could be to develop and use clearer conflict resolution procedures. Alexandersson and Rigas (2013) also conclude that tools to address issues related to capacity allocation and access charges must be further developed.

Another issue that hinders competition is the large initial costs and financial losses related to acquiring the necessary rolling stock and operating services (Murillo-Hoyos et al., 2016). New entrants often need several years to become profitable, see for example the case study by

Tomeš et al. (2016) regarding RegioJet in CZ. One way to help new entrants is to use framework agreements in the capacity allocation process for long term allocation over several annual timetables. Most of the reviewed countries already have it in their allocation process.

3.5. Capacity allocation

In vertically separated markets, the capacity allocator (which is usually the same as the infrastructure manager) is often a subsidiary of a (state-owned) company, or sometimes a governmental agency. **Table 6** lists examples of how this can be organized. This contrasts with an allocation body that is part of one vertically integrated railway company as in Japan (for passenger) and the US (for freight). In this case, capacity allocation is done internally within the integrated company, and capacity conflicts never become explicit or public. In such markets, no railway regulator is needed to oversee the market.

Table 6. The capacity allocation body (infrastructure manager) and the regulator in selected countries.

Country	Infrastructure manager		Railway regulator	
	Name	Legal status	Name	Legal status
Austria	ÖBB-Infra	Subsidiary of a holding	Schienen-Control	State-owned company
Belgium	Infrabel	Subsidiary of a parent	Regul	Governmental agency
Czech	SŽDC	State-owned company	Drážní Inspekce	Governmental agency
France	SNCF Réseau	Subsidiary of a parent	Arafer	Governmental agency
Germany	DB Netze	Subsidiary of a parent	BNetzA	Governmental agency
Great Britain	Network Rail	State-owned company	ORR	Parliamentary agency
Italy	RFI	Subsidiary of a holding	ART	Parliamentary agency
Netherlands	ProRail	Governmental agency	ACM	Governmental agency
Spain	Adif	State-owned company	CNMC	Parliamentary agency
Sweden	Trafikverket	Governmental agency	Transportstyrelsen	Governmental agency
Switzerland	Trasse	Nonprofit company	BAV	Independent commission

However, in the studied vertically separated markets the regulator is required to be independent in order to supervise the work of the infrastructure manager, and to make sure that the capacity allocation is non-discriminatory. **Table 6** indicates that the legal status of the regulator is generally similar across the studied markets. Slight differences

exist as some regulators are under the control of the government (executive) whereas others are controlled by the parliament (legislature).

Although vertically integrated, Switzerland (which is not a member of the EU) ensures certain compliance with EU policies. The capacity allocation is performed by a nonprofit company (Trasse), which is under the supervision of an independent commission of experts (BAV). Thus, both the infrastructure manager and the regulators are completely independent bodies.

The capacity allocation in the selected vertically separated markets follows similar steps as presented in **Figure 6** which summarizes the Swedish allocation of railway capacity. The process starts with the infrastructure manager receiving applications for capacity from licensed railway undertakings. Based on the conditions and terms specified in the national network statement, the process generally starts with operators submitting train path requests with all information needed to construct a proposed timetable. The infrastructure manager proposes a draft of the annual timetable. Minor conflicts can usually be resolved by small adjustments of path requests, so the framework often specifies certain time intervals with which the infrastructure manager can unilaterally do adjustments (without negotiating with the applicants). Major conflicts are usually solved in a coordination process where the different applicants conduct informal discussions with the infrastructure manager to settle conflicts. These applicants can further apply for settlement of disputes if there are remaining conflicts (after the coordination process). These conflicts are usually resolved by the infrastructure manager taking unilateral decisions (without further consultation of the applicants) based on conflict resolution procedures (often priority criteria or decision rules). At this stage, the infrastructure managers are often obliged to declare the infrastructure as congested, and to conduct capacity analysis and implement reinforcement plans to improve the supply of capacity and its utilization for the next timetables. Applicants can appeal the capacity allocation decisions to the independent national railway regulator. Such appeals can be used against any discriminatory behavior from the capacity allocation body.

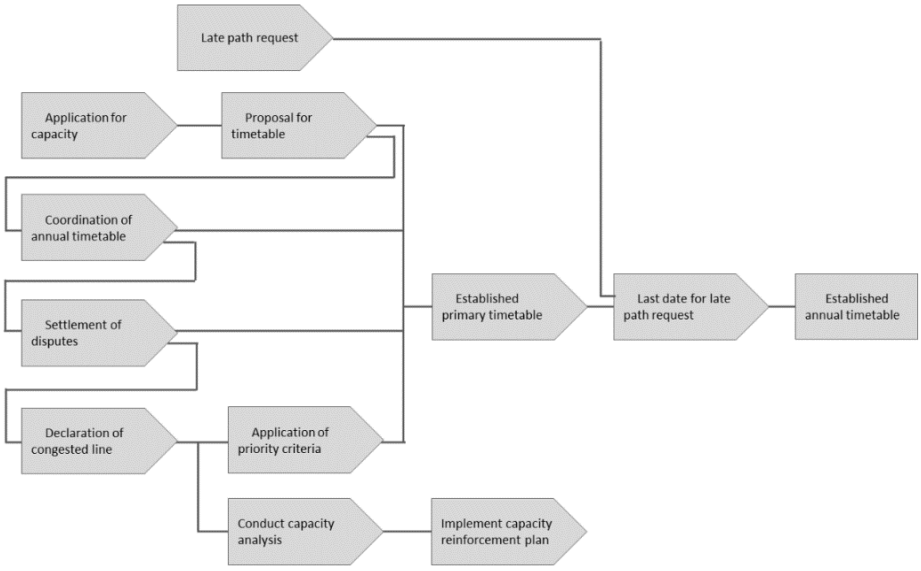


Figure 6. Overview of capacity allocation in the Swedish railways (Trafikverket, 2020b).

Although capacity processes are, to a large extent, similar in the reviewed countries, specific procedures to solve capacity conflicts have been relatively developed in some markets more than in others. Major variations can be found when it comes to the settlement of disputes and/or the application of priority criteria. To illustrate these differences, **Table 7** presents a comparative overview of the priority criteria that are applied to settle capacity disputes.

The table presents three criteria in their order of priority as mentioned in the national network statement in Austria (ÖBB-Infrastruktur, 2020), Belgium (Infrabel, 2020), Czech (SŽDC, 2020), Germany (DB-Netze, 2020), Italy (RFI, 2020) and Spain (Adif, 2020). These countries have an explicit list of criteria which is ordered in priority. Some countries list the criteria without any explicit order whereas others use models for total social costs as in Sweden (Trafikverket, 2020b) or robustness as in France (SNCF-Réseau, 2020).

As mentioned before, some countries with organizational separation (and mostly low competition) have a capacity allocator with links to the incumbent which may create various conflicts of interest, and potential new entrants may see this as a risk when considering entering the market. Such conflicts and risks can be avoided with more transparent capacity allocation rules, i.e., clearer conflict resolution procedures.

Table 8 indicates that there are various procedures to allocate capacity in case of conflicting train path requests.

Table 7. Comparative overview of priority criteria in selected countries.

	Priority 1	**Priority 2**	**Priority 3**
Austria	Clockface traffic	Framework agreement (if not already declared congested), Passenger public interest traffic (during peak times)	Cross borders
Belgium	No previous underutilization of allocated capacity	Speed for passenger (on high-speed lines), for freight (on freight lines), domestic passenger (on mixed lines)	Highest monthly access charge
Czech	Passenger public services (priority to interregional and international trains)	International freight	Framework agreement
Germany	Regular interval services	Cross-border services	Freight services
Italy	International train services	General public transport services (agreements with central or regional bodies)	Highspeed or freight services on their dedicated infrastructures
Spain	Regular interval services	Cross-border services	Freight services
France	Traffic on European freight corridors, distance covered, commercial and financial importance, timetable robustness		
Britain	Improvement of the network capability, reflection of demand, short journey time, commercial interest of Network Rail (no priority for PSO services)		
Netherlands	Statutory priority rules specifying the services (passenger or freight) to prioritize on each route		
Sweden	Total social costs		
Switzerland	Prioritization depending on the type of traffic or bidding mechanism (with Vickrey auction)		

Table 8. Summary of the different procedures to solve capacity conflicts.

Procedure	**Main components**
General principles	Highest societal benefits (SE), robustness and commercial importance (FR), improvement of network capability and demand reflection (GB)
Specific priority	Specific criteria (CZ, DE, ES), statutory rules (NL), criteria depending on infrastructure type (BE) or congestion (AT), framework agreements (IT)
Market-based (pricing)	Vickrey auction (CH), highest bidder (DE)

The allocation rules are mainly either based on general principles (e.g., SE, FR) or specific priority criteria (e.g., CZ). Specific criteria are clearer allocation rules that can depend on the speed (e.g., BE), the type of traffic (e.g., CZ) or the level of congestion of the infrastructure (e.g., AT). Although transparent, such rules do not always yield (socio)economically efficient allocation outcomes. Additional special rules may be applied in certain countries, e.g., CZ where the running time in any allocated train path in the country is never beyond 20 hours (SŽDC, 2020).

More general criteria require the development and use of a certain capacity allocation model. For instance, the French procedure for capacity allocation is generally based on models for improving the robustness of the final annual timetable (Perez Herrero, 2016). In Sweden, the infrastructure manager uses an efficiency-based model that aims at evaluating the total societal benefits (and costs) of different outcomes and choose the best alternative (Trafikverket, 2020a). These models may lead to efficient solutions but are often less transparent, i.e., unclear to the railway undertakings.

The more specific the criteria are, the more transparent the procedure becomes. However, it is not always easy to list specific and transparent priorities valid for all conflict situations since these may sometimes lead to inefficient outcomes. Market-oriented procedures exist as allocation rules in some countries in the form of auction. In addition to the track access charges, the winner pays either the second highest bid, called Vickrey auction (e.g., CH), or the highest bid (e.g., DE). Although allowed by the EU legislation, such procedures are rarely, if not never, used.

3.6. Track access charges

When the market was vertically integrated, i.e., the same railway company was responsible for both infrastructure and operations, there was no need for access charges. This is not the case in vertically separated markets where access charges are regulated (EC, 2001), and become important for the capacity allocation (Nash et al., 2018). For instance, too low charges can lead to capacity overutilization from railway undertakings and financial deficit for the infrastructure manager(s), whereas too high charges can lead to capacity underutilization. This relationship between network capacity utilization and track access charges is presented in **Figure 7**. The figure indicates that markets with low utilization often have lower charges than in networks with

scarcer capacity. Such lower charges can be intended to increase the demand for capacity in these markets. However, this decreases the revenues of the infrastructure manager, and can therefore increase its financial deficit.

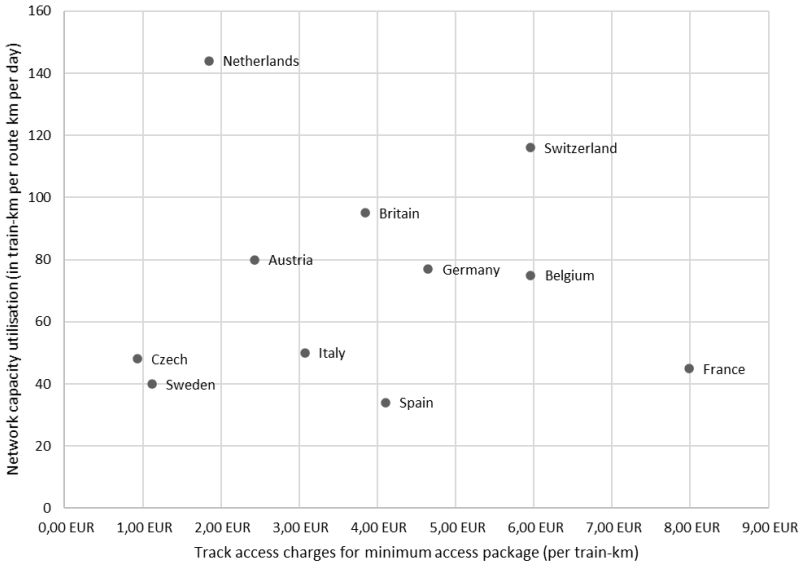


Figure 7. Network utilization and access charges in selected markets (IRG, 2019).

Although track access charging systems have such importance, EU policy only gives general guidelines on its principles, e.g., 1st package and its recast in the SERA directive (EC, 2001, EC, 2012). Although such access charges vary between the reviewed countries, this survey finds that many include similar components and converge to a similar charging system, just as predicted by Crozet (2004). This is also shown in **Table 9** which shows that all the reviewed markets include basic or minimum access/service charges covering the train path charges, and sometimes administrative and reservation costs.

The train path charges may also include direct, indirect, variable and fixed infrastructure costs. Access to certain infrastructure facilities (e.g., passenger stations, freight marshalling yards and terminals) and additional services (e.g., ticket sales, telecommunication and traffic information) are often also charged for. **Table 9** also indicates that most countries include performance regimes to encourage railway undertakings to use better rolling stock and ensure service punctuality. Moreover, financial bonus (or malus) is often used to incentivize (or penalize)

the use of capacity which decreases the number of unused (or cancelled) allocated train paths. Proper use of capacity is further ensured through capacity allocation-related charges, e.g., timetable planning, mark-up costs and congestion charges.

Table 9. The track access charges in the selected markets (* if the component exits).

Access charges	Components	AT	BE	CZ	FR	DE	GB	IT	NL	ES	SE	CH
Minimum access package	Administrative		*									
	Reservation				*				*	*		
	Train path	*	*	*	*	*	*	*	*	*	*	*
Service facilities	Station	*		*	*	*	*	*	*	*		*
	Marshalling yards	*			*	*	*	*	*		*	*
Additional services	Electric traction	*	*	*	*	*	*	*	*			*
	Traffic information	*	*		*	*		*	*		*	
	Ticketing		*		*			*				
Financial incentives	Non usage		*	*	*	*		*	*		*	*
	Cancellation		*	*	*	*		*	*		*	*
Performance regimes	Delay	*	*	*	*	*	*	*	*	*	*	
	Wear and tear	*			*	*	*	*			*	
	New traffic				*	*						
	Environment	*	*	*		*	*		*			*
Capacity allocation	Planning	*		*	*	*						
	Mark-ups	*	*		*	*	*		*			*
	Congestion	*			*	*			*		*	*

The values of the parameters that are used to calculate the costs are mostly set by the infrastructure manager itself and sometimes with approval from the regulator, e.g., ORR in GB (ORR, 2017). Even though the charging systems are similar, the values of the cost parameters can vary significantly from one market to the other as illustrated in **Figure 7**. For instance, countries such as SE and CZ have significantly lower values than others such as FR (Crozet, 2018). Such differences do not only depend on the capacity utilization as presented in **Figure 7**, but also on whether the pricing is based on marginal cost or average cost (for cost recovery) for infrastructure operations and maintenance. It can also depend on whether the infrastructure manager collecting the charges is a governmental agency, nonprofit or for-profit company. In some cases, costs or benefits are simply difficult to estimate or not well estimated, e.g., noise and environmental effects (Lan and Lin, 2005). The latter environmental effects are sometimes controlled beforehand when providing the license for the undertakings to operate in the national railways, e.g., as in GB (Network-Rail, 2020).

With a few minor exceptions (e.g., DE and CH), it is uncommon that access charges are used as a conflict resolution procedure. This appears to be a severely underused opportunity as it is allowed by EU legislation; it is difficult to understand why this is not more common. One hypothesis is that it is because most railway markets were vertically integrated until recently, and it simply takes time to develop the access charges principles to solve capacity conflicts necessary in a vertically separated market. However, the survey indicates that access charges are more commonly used to incentivize the railway operators to efficiently use the allocated capacity. For instance, through differentiated track access charges such as congestion charges, i.e., charging a higher price where capacity is scarcer, and also through performance regimes, i.e., delay compensation.

Note that track access charges may also change from one year to the other. Major revisions can be brought to the components as well as the cost values of these components to account for the recent developments in the national railway infrastructure and operations as well as the European legislation.

4. Conclusions

Several countries aim to introduce or increase competition among operators, both for passenger and freight services. For this to succeed, the capacity allocation process needs to be transparent and to some extent predictable, allowing prospective operators to foresee what capacity they will be allocated. It also needs to yield efficient outcomes by ensuring that certain operators (providing the best value for money for their customers) get capacity. Few if any countries have capacity allocation processes that satisfy all these criteria.

As to transparency and predictability, most countries have processes where it is difficult, especially for an outsider, to understand which path requests get priority when a conflict occurs, and it is even more difficult for a potential new operator to understand how it should act in order to get the capacity it needs to provide its services. There are a few exceptions where it is relatively clear how priority is given, and even fewer with market-based procedures (e.g., auction). But there are many more cases where capacity conflicts are resolved through various kinds of general priority criteria, where it is often difficult for an outsider to understand how they are applied. For example, several countries have priority criteria or decision rules which are not necessarily

consistent or mutually exclusive, or where it is not clear in what order they take precedence.

An additional concern is that the agency responsible for capacity allocation (usually the infrastructure manager) has sometimes organizational links to the incumbent, often dominating operator. A new operator considering whether to enter the market may have reasonable concerns that this may bias the judgment of priorities in a capacity conflict in favor of the incumbent operators – especially if the capacity allocation process is informal and non-transparent (e.g., using general principles as allocation rules). As noted in this survey, markets where the capacity allocator appears to have conflicts of interest tend to generally have less competition, and incumbents can often have larger market shares of the passenger and freight markets.

The capacity allocation process is crucial for a multi-operator railway market to function efficiently. The purpose of operator competition is to ensure, in the long run, that operators provide the services which give the best value for money to customers. For this to work, it is essential that the most efficient operator, i.e., the one providing the most attractive services from the market's point of view, also gets priority in a capacity conflict. From our review, we can conclude that such considerations are surprisingly absent. With a few exceptions, priority criteria have at best a vague relation to consumer demand and market efficiency. A vast majority of priority criteria and decision rules instead relates to simple administrative or technical criteria, for example, that longer train paths have higher priority than short ones, or that passenger services (or high-speed trains) get priority over freight services (or slower trains). There appears to be few explicit arguments based on market efficiency for how such criteria have been formulated.

Opening the market for railway services to competition can in principle yield substantial social benefits, partly because operators get more incentives to become more cost-efficient and more responsive to consumer demand, partly because evolutionary selection will ensure that services are weeded out whenever production costs exceed the market's willingness to pay. But for this to work, it is necessary that the process for resolving capacity conflicts between different operators is efficient and transparent. Our survey indicates that most countries still have some way to go in this respect. Thus, the need to develop and experiment with more efficient and transparent procedures within the legislation such as market-based capacity allocation.

Acknowledgments

The authors are grateful to Jan-Eric Nilsson and Yves Crozet for reference recommendations as well as Russell Pittman, Steven Harrod and several anonymous reviewers for the valuable discussions and comments.

References

Abbott, M. and B. Cohen (2017). "Vertical integration, separation in the rail industry: a survey of empirical studies on efficiency." European Journal of Transport and Infrastructure Research 17(2): 207-224.

Adif (2020). "Network Statement 2021."

Alexandersson, G. and K. Rigas (2013). "Rail liberalisation in Sweden. Policy development in a European context." Research in Transportation Business & Management 6: 88-98.

Asmild, M., T. Holvad, J. L. Hougaard and D. Kronborg (2009). "Railway reforms: do they influence operating efficiency?" Transportation 36(5): 617-638.

Bouf, D., Y. Crozet and J. Lévêque (2005). Vertical separation, disputes resolution and competition in railway industry. Thredbo 9, 9th conference on competition and ownership in land transport, 5-9 september 2005, Lisbonne., Lisbon Lisbon Technical University.

Crozet, Y. (2004). "European railway infrastructure: towards a convergence of infrastructure charging?" International Journal of Transport Management 2(1): 5-15.

Crozet, Y. (2016). Introducing competition in the European rail sector. Discussion Paper prepared for the Roundtable on Assessing regulatory changes in the transport sector.

Crozet, Y. (2016). Liberalisatin of passenger rail services - France.

Crozet, Y. (2018). Case Study – France: logic and limits of full cost coverage. Track access charges: reconciling conflicting objectives. CERRE, CERRE & University of Lyon (LAET).

Crozet, Y., C. Nash and J. Preston (2012). "Beyond the quiet life of a natural monopoly: Regulatory challenges ahead for Europe's rail sector." Policy paper, CERRE, Brussels, December 24.

DB-Netze (2020). Network statement 2021.

EC (1991). Council Directive 91/440/EEC of 29 July 1991 on the development of the Community's railways, European Commission.

EC (2001). Directive 2001/14/EC on the allocation of railway infrastructure capacity and the levying of charges for the use of railway infrastructure and safety certification, EU Parliament.

EC (2012). Directive 2012/34/EU on establishing a single European railway area, EU Parliament.

EC (2016). Fourth railway package of 2016, European Commission.

Friebel, G., M. Ivaldi and C. Vibes (2010). "Railway (De)regulation: A European efficiency comparison." Economica 77(305): 77-91.

Gibson, S. (2003). "Allocation of capacity in the rail industry." Utilities Policy 11(1): 39-42.

Gilbo, E. P. (1993). "Airport capacity: representation, estimation, optimization." IEEE Transactions on Control Systems Technology 1(3): 144-154.

Hansson, L. and J. E. Nilsson (1991). "A new Swedish railroad policy: Separation of infrastructure and traffic production." Transportation Research Part a-Policy and Practice 25(4): 153-159.
Infrabel (2020). Network Statement 2021.
IRG (2019). Seventh Annual Market Monitoring Working Document, Independent regulators' group rail.
Jensen, A. and P. Stelling (2007). "Economic impacts of Swedish railway deregulation: A longitudinal study." Transportation Research Part E-Logistics and Transportation Review 43(5): 516-534.
Klein, M. (1999). Competition in Network Industries.
Lan, L. W. and E. T. J. Lin (2005). "Measuring railway performance with adjustment of environmental effects, data noise and slacks." Transportmetrica 1(2): 161-189.
Laurino, A., F. Ramella and P. Beria (2015). "The economic regulation of railway networks: A worldwide survey." Transportation Research Part A: Policy and Practice 77: 202-212.
Link, H. (2004). "Rail infrastructure charging and on-track competition in Germany." International Journal of Transport Management 2(1): 17-27.
Link, H. (2016). Liberalisatin of passenger rail services - Germany.
Link, H. (2018). Case Study – Germany. Track access charges: reconciling conflicting objectives. CERRE, German Institute for Economic Research (DIW Berlin).
Ludvigsen, J. (2009). "Liberalisation of Rail Freight Markets in Central and South-Eastern Europe: What the European Commission Can Do to Facilitate Rail Market Opening." European Journal of Transport and Infrastructure Research 9(1).
Merkert, R. (2012). "An empirical study on the transaction sector within rail firms." Transportmetrica 8(1): 1-16.
Monami, E. (2000). "European passenger rail reforms: A comparative assessment of the emerging models." Transport Reviews 20(1): 91-112.
Murillo-Hoyos, J., M. Volovski and S. Labi (2016). "Rolling stock purchase cost for rail and road public transportation: random-parameter modelling and marginal effect analysis." Transportmetrica A: Transport Science 12(5): 436-457.
Nash, C. (2008). "Passenger railway reform in the last 20 years - European experience reconsidered." Reforms in Public Transport 22: 61-70.
Nash, C., Y. Crozet, H. Link, J.-E. Nilsson and A. Smith (2016). Liberalisatin of passenger rail services - project report.
Nash, C., Y. Crozet, H. Link, J.-E. Nilsson and A. Smith (2018). Track access charges: reconciling conflicting objectives - project report. CERRE, CERRE.
Nash, C., J. E. Nilsson and H. Link (2013). "Comparing Three Models for Introduction of Competition into Railways." Journal of Transport Economics and Policy 47: 191-206.
Nash, C. A., A. S. J. Smith, D. van de Velde, F. Mizutani and S. Uranishi (2014). "Structural reforms in the railways: Incentive misalignment and cost implications." Research in Transportation Economics 48: 16-23.
Network-Rail (2020). Network Statement 2021.
Nilsson, J.-E. (2016). Liberalisation of passenger rail services - Sweden.
Nilsson, J. E. (2018). Case Study – Sweden: Track access charges and the implementation of the SERA directive - promoting efficient use of railway infrastructure or not? Track access charges: reconciling conflicting objectives. CERRE, VTI Swedish National Road and Transport Research Institute.

OECD (2005). Structural Reform in the Rail Industry. Competition Policy Roundtables.
OECD (2013). Recent Developments in Rail Transportation Services. Competition Policy Roundtables.
ORR. (2017). "Track access guidance | Office of Rail and Road." from http://www.orr.gov.uk/rail/access-to-the-network/track-access/guidance.
Perez Herrero, M. (2016). Rail capacity constraints : an economic approach. PhD, Université Lumière Lyon 2.
ProRail (2020). Network Statement 2021.
RFI (2020). Network statement 2021.
SBB (2020). Network Statement 2021.
Smith, A. (2016). Liberalisatin of passenger rail services - Britain.
Smith, A. and C. Nash (2018). Case Study – Great Britain. Track access charges: reconciling conflicting objectives. CERRE, CERRE & University of Leeds.
SNCF-Réseau (2020). Network statement 2021.
SŽDC (2020). "Network statement 2021."
Talebian, A., B. Zou and A. Peivandi (2018). "Capacity allocation in vertically integrated rail systems: A bargaining approach." Transportation Research Part B: Methodological 107: 167-191.
Tomeš, Z., M. Kvizda, M. Jandová and V. Rederer (2016). "Open access passenger rail competition in the Czech Republic." Transport Policy 47: 203-211.
Trafikanalys (2014). Railway in Sweden and Japan - a comparative study.
Trafikverket (2020). Järnvägsnätsbeskrivning 2020 - Prioriteringskriterier, Swedish Transport Administration.
Trafikverket (2020). Network Statement 2021, Swedish Transport Administration.
UIC (2009). Noise Differentiated Track Access Charges, International Union of Railways.
Van de Velde, D., C. Nash, A. Smith, F. Mizutani, S. Uranishi, M. Lijesen and F. Zschoche (2012). "Economic effects of vertical separation in the railway sector." Report CER and Inno-V Amsterdam.
Yeung, R. (2008). Moving Millions: The Commercial Success and Political Controversies of Hong Kong's Railway, Hong Kong University Press.
ÖBB-Infrastruktur (2020). Network statement 2021.

Paper P2

Pricing Commercial Train Path Requests Based on Societal Costs.

Ait-Ali A.[1,3], Warg J.[2] and Eliasson J.[3] (2020)
[1] VTI Swedish National Road and Transport Research Institute, Transport Economics (TEK), Stockholm
[2] KTH Royal Institute of Technology, Division of Transport Planning, Stockholm
[3] Linköping University, Department of Science and Technology (KTS), Norrköping

Published in Transportation Research Part A: Policy and Practice, Volume 132, February 2020, Pages 452-464

https://doi.org/10.1016/j.tra.2019.12.005

Abstract

On deregulated railway markets, efficient capacity allocation is important. We study the case where commercial trains and publicly controlled traffic ("commuter trains") use the same railway infrastructure and hence compete for capacity. We develop a method that can be used by an infrastructure manager trying to allocate capacity in a socially efficient way. The method calculates the loss of societal benefits incurred by changing the commuter train timetable to accommodate a commercial train path request, and based on this calculates a reservation price for the train path request. If the commercial operator's willingness-to-pay for the train path exceeds the loss of societal benefits, its request is approved. The calculation of these benefits takes into account changes in commuter train passengers' travel times, waiting times, transfers and crowding, and changes in operating costs for the commuter train operator(s). The method is implemented in a microscopic simulation program, which makes it possible to test the robustness and feasibility of timetable alternatives.

We show that the method is possible to apply in practice by demonstrating it in a case study from Stockholm, illustrating the magnitudes of the resulting commercial train path prices. We conclude that marginal societal costs of railway capacity in Stockholm are considerably higher than the current track access charges.

Keywords: railway markets; vertical separation; competition; capacity allocation; access charges.

1. Introduction

The deregulation of railway markets in many countries have meant that it has become increasingly common that several operators run trains on the same track and hence compete for the same capacity. Inevitably, conflicts arise between different operators´ capacity requests, and in such cases the infrastructure manager needs to prioritize between conflicting path requests. Depending on the ownership of the railway infrastructure, the infrastructure manager can have different objectives. In this paper, we will consider the case where infrastructure is publicly owned, and the infrastructure manager wants to allocate capacity in a way that maximizes total societal benefits.

There is a clear need for methodological development to overcome these challenges. Very large social and commercial values are at stake, so making the capacity allocation process more efficient can potentially generate substantial benefits. Moreover, the process and decision criteria need to be transparent and consistent to ensure fair and efficient competition between operators. In current practice, however, capacity allocation processes are often opaque, and it is common that simplified decision criteria or rules of thumb are used, such as faster trains having priority over slower ones, or passenger services having priority over freight.

Highly simplified, there are two principal ways for an infrastructure manager to resolve conflicting path requests, either based on benefit judgments or on willingness-to-pay (WTP). In processes based on benefit judgments, the infrastructure manager compares some criteria or calculations intended to reflect the societal benefits generated by alternative capacity allocations. The allocation which scores highest, according to these criteria, is chosen. In processes based on WTP, the operator that is willing to pay the highest price gets priority. WTP-based processes include slot auctions and demand-differentiated track charges. Both methods have their respective strengths and weaknesses. The problem of processed based on benefit judgment is that the data necessary for judging the benefit of a service is not always available. There exist well-developed methods to calculate societal benefits and costs of transport services, but such calculations need detailed data about demand, ticket prices and operating costs. For commercial traffic, such data is almost never available to the infrastructure manager, either because it is sensitive business information known only to the operator(s), or because this data is unknown at the time of the decision. The problem in WTP-based processes is that the societal benefits of publicly controlled traffic – for example, the subsidized commuter trains serving many large urban regions – do not necessarily correspond to the responsible public agency's willingness (or ability) to pay. Since the societal benefits of urban public transport

cannot easily be "observed" in the same way as a commercial operator can observe its profits, it is much more difficult for public transport agencies to correctly assess its "willingness to pay" for capacity; both over- and underestimations may occur. Moreover, the societal benefits generated by urban public transport accrue to society at large, not to the agency, so there is no obvious link between the societal benefits generated by commuter train services and the responsible agency's financial resources or hence its ability to pay for capacity. This is illustrated by the observation that public transport agencies are often under great financial strain, despite generating substantial societal benefits, not least because urban commuter trains are usually heavily subsidized (often for good reasons).

In this paper, we propose a hybrid method to resolve capacity conflicts between publicly controlled traffic (i.e., train services where supply and fares are determined by a public transport agency striving to maximize social welfare) and commercial traffic (i.e., passenger or freight train services run by companies striving to maximize profits). For brevity, we will call the publicly controlled services "commuter trains" in the following. Such services are often run by one or more operators contracted by the agency, but this is inessential in this context.

The main idea of the proposed method is to calculate a reservation price for a commercial operator's path request by estimating the societal costs (i.e., loss of benefits) of the changes needed in a baseline commuter train timetable to accommodate this path request. If the commercial operator is willing to pay this reservation price, it is awarded the path and the commuter train timetable is adjusted; if not, the request is declined. The process can be extended to handle multiple commercial path requests which also allows an infrastructure manager to prioritize between several commercial operators competing for the same capacity, i.e., on-rail competition.

This circumvents the problems explained above – on the one hand that an infrastructure manager does not have access to information necessary to assess the societal benefits of a commercial train service, and on the other hand that there is no clear correspondence between public agencies' WTP:s and the societal benefits generated by the services they run. Instead of only comparing benefit calculations or only comparing WTP:s, the process proposed in this paper compares benefit calculations for commuter train services to the WTP of commercial train services. The advantages of such a process is that it utilizes the respective characteristics of commuter trains and commercial traffic: for commuter trains, the information needed for benefit calculations (fares, passenger volumes, operating costs etc.) is not secret and can be relatively obtained easily;

for commercial traffic, commercial operators have a relatively good perception of the profit they would make on a given train path, and hence of the price they are willing to pay for a train path request.

The economic logic of the idea is most easily explained as a way to price the negative externality the commercial train service incurs on the commuter trains. When the commuter train timetable is adjusted to make room for the commercial train request, this causes a loss for the commuter train passengers and the operator(s), which is an external effect from the point of view of the commercial operator. It is well known that pricing an externality will increase total societal benefits, if there are no other market imperfections. But there are at least two types of market imperfection that are worth considering: unpriced externalities of road traffic, and monopoly power of commercial train services. If there are unpriced road traffic externalities, then it may be motivated to subsidize either commuter train services or commercial train services, or both (so-called second-best pricing). If commercial train operators have some degree of monopoly power, then they will supply less services than socially optimal, which may warrant various kinds of regulations. The assumption in this paper is that such market imperfections are handled outside of the capacity allocation process. There are a variety of ways to do this, and many are already in use, for example subsidizing commuter train fares, setting track charges lower than marginal costs, regulating commercial train monopolies, and pricing road traffic externalities through fuel taxes, parking charges and congestion pricing. All of these methods are both simpler and more efficient ways to handle market imperfections than trying to handle them in the capacity allocation process. Hence, we assume that we can ignore these market imperfections in this context, and then it follows that pricing the externality caused by commercial train services interfering with commuter trains will increase the overall social welfare.

The purpose of the paper is to explain and develop the proposed method, and to apply it in a real-world application. In a case study from Stockholm, we calculate reservations prices for a commercial train service by adjusting a baseline commuter train timetable in three different ways: removing a commuter train service, shifting its departure time, and increasing its running time by letting it wait along the line while the commercial train passes by. The corresponding reservation prices are calculated for three different time periods, morning and afternoon peak hours and midday off-peak, and compared to the current track charges (Trafikverket, 2020b). The adjustments to the commuter train timetable are done in the microscopic simulation tool *RailSys* (Radtke and Bendfeldt, 2001) to guaranty conflict-free timetable solutions.

The paper contributes to the literature on capacity conflict resolution in several ways. First, it describes a concrete method to resolve conflicts between commuter and commercial trains. This, in turn, is an important component in an overall process of capacity allocation, which also needs to include resolving other kinds of conflicts such as conflicts between different commercial operators, and between train services with different planning horizons (Broman et al., 2018). Second, the method can also be used as a decision tool for traffic planners to evaluate societal effects of alternative rescheduling scenarios. Third, the implementation and application of the method shows its practical feasibility, and shows how cost-benefit analysis can be combined with a large-scale commercial microscopic simulation tool such as *RailSys* for real-world train timetabling and planning applications. Fourth, results from the case study indicate a realistic range of reservation prices that commercial trains should pay if capacity should be reallocated from commuter to commercial trains. Our results indicate that the reservation prices in our case study are considerably higher than current track charges in Stockholm.

The paper has the following structure: Section 2 gives a brief literature review of recent research on train timetable assessment. Section 3 describes the method for calculating the societal costs of changes in a baseline commuter train timetable. Section 4 presents the case study, with input data and results. Section 5 concludes.

2. Previous related research

The idea proposed in this paper is similar to the one proposed by Johnson and Nash (2008). They examine the feasibility of using opportunity costs for each train slot to price scarce rail capacity on congested franchised lines in Britain. Opportunity costs are calculated based on consumer surplus, externalities and operation costs, and based on the additional traffic attracted by the slot, the additional quality for the users due to the slot, the savings of external costs for users that switch from road to rail, as well as the costs for not running that train. A demand model is used to evaluate the effects of changes in the transport system that are larger than the ones considered in this paper. The basic idea is similar to the current paper, but there are a number of differences between the suggested models. For example, we avoid having to develop and calibrate a demand model by working directly with origin-destination matrices. The advantage is that one can then accurately evaluate even minor changes in the timetables; on the other hand, keeping demand constant is a disadvantage when evaluating large timetable changes (we return to this issue below).

Using pricing to allocate scarce rail capacity is still rather uncommon in practice, even if several countries use it to some minor extent (Nash, 2005). Nash et al. (2004) point out that the lack of a price for scarce capacity is the major defect in the current British system for capacity allocation. Several authors such as Affuso (2003), Nilsson (2002) and Newbery (2003) discuss and propose various auction-based methods to allocate capacity, focusing on competing commercial services. The current paper, similar to the one by Johnson and Nash (2008), complements these approaches by proposing a method to resolve conflicts between publicly controlled services and commercial ones.

There are still not many papers that use cost-benefit analysis methods to evaluate train timetables, although there a few examples such as Adler et al. (2010) who study how competition between airlines and high-speed rail may affect service levels, and how this affects the benefits of high-speed rail investments. Eliasson and Börjesson (2014) also apply cost-benefit analysis on timetable construction, highlighting the impact of timetable assumptions on appraisal of railway investments. (Brännlund et al., 1998) is also an example of capacity allocation with an economic objective; they present an algorithm that schedules a set of trains to obtain a profit-maximizing timetable without violating track capacity constraints. More generally, however, there is a vast literature on valuing elements of timetables, such as waiting times, transfer and in-vehicle time, e.g., (Hensher, 1997), (Hensher and Ton, 2002) and (Balcombe et al., 2004).

There is also of course a vast literature evaluating timetables from various perspectives. A classic and widely studied question is how timetabling affects total capacity. This can be studied with a combination of analytical methods, simulation and optimization methods (Abril et al., 2008). For reproducing railway operations in the best way, dynamic, synchronous, microscopic, stochastic simulation tend to work best (Borndörfer et al., 2018). In addition to capacity effects, operational characteristics of timetables such as punctuality, stability and robustness have been extensively studied. For example, Delorme et al. (2009) focus on the stability aspect of the train timetables by developing a timetable stability module that is used to build an optimization model for railway operation planning.

There are also studies evaluating the performance of timetables from a passenger perspective, for example Kunimatsu et al. (2012), who use microsimulation of both passengers and trains in the railway network. Passenger behavioral aspects such as avoiding transfers and choosing best routes are also accounted for. The timetable is evaluated based on the disutility of passengers due to delays and crowding. This evaluation

model is useful to capture detailed aspects of single line train timetables with a passenger perspective. The use of big data helps study even larger networks such as in the work by Jiang et al. (2016). Other studies have attempted to aggregate multiple passenger-related aspects of train timetables such as the sum of weighted waiting times, average of unit waiting time and maximum ratio of waiting time to travel time into a fuzzy analytic hierarchy process (Isaai et al., 2011).

3. Evaluation Model

The general idea of the method is to calculate the loss of benefits in commuter train traffic incurred when adjusting their timetable to accommodate an additional commercial train path. This loss of benefits, measured in monetary terms, yields the price that a commercial train operator requesting the path has to pay. If the operator is willing to pay the requested amount (the reservation price), the operator is awarded the requested path; otherwise the request is denied. As discussed in the introduction, this can be viewed as a way to price the externality caused by the commercial train service when it interferes with the commuter train services. It can also be seen as a way to correctly price the "input good" (the track capacity) need to produce the "good" that is the commercial train service. Both interpretations are well-known economic principles that lead to a socially efficient market equilibrium, provided that other market imperfections can be ignored. There are obviously several market imperfections in transport markets, such as road traffic externalities and some degree of monopoly power for commercial train operators, but there are both simpler and more efficient ways to handle these than in the capacity conflict resolution process, and many of them are also used in practice, such as subsidized fares and track charges, fuel taxes and congestion charges. Hence, it is reasonable to expect that introducing a price for track capacity that more accurately reflects its societal cost should lead to an overall increase in total societal benefits.

The method starts with an initial situation, where commuter trains are run according to a baseline timetable, and a train path request from a commercial operator which conflicts with the baseline timetable. Usually, commuter trains aim to be operated with regular headways (4-8 trains per hour) and frequent stopping pattern, while the commercial train operates longer distances less regularly with fewer stops and higher average speeds. Both the difference in average speed and the trains´ inflexibility to change their departure times due to marketing and capacity reasons makes it difficult to allocate capacity. The method calculates the value of the loss of societal benefits resulting from this modified

commuter train timetable compared to the baseline timetable, and this loss of benefits (measured in monetary terms) constitutes the reservation price.

The loss of benefits ΔB consists of two parts: the change in consumer surplus ΔC and the change in producer surplus ΔP. The change in consumer surplus reflects all changes affecting passengers, for example changes in waiting times, crowding, transfers and travel time. The change in producer surplus reflects all changes affecting the operator, for example changes in operations costs and revenues. It is quite possible that a change may result in, for example, a loss of consumer surplus but a gain in producer surplus. Cancelling a train service, for example, would normally produce a gain in producer surplus by decreasing the costs of operations, but a loss of consumer surplus, since some passengers may experience increased waiting times and increased crowding in the remaining trains. Other types of societal benefits, e.g., changes in road traffic emissions and tax revenues, are ignored since they are either relatively small or are internalized in other ways, e.g., corrective taxes.

3.1. Calculating the change in consumer surplus

Let T_{ijr} be the number of passengers arriving at time r to station i to travel to station j. The matrix $\{T_{ijr}\}$ is called the (dynamic or time-dependent) origin-destination (OD) matrix. Let c_{ijr} be the generalized cost of travelling from station i to station j starting at time r, and define it as the sum of the fare f_{ijr}, waiting time(s) w_{ijr} and the in-vehicle time t_{ijr}, converting the time components into money by multiplying with the value of waiting time β and the value of in-vehicle travel time α, see equation (1). As explained further below, α depends non-linearly on the crowding level in the train. The waiting time penalty β is in general a non-linear function of the headway, since when headways are long, passengers will spend part of the "waiting time" elsewhere by adjusting their schedule to the train's departure time. In this study, however, the service frequency is high, so the waiting time penalty can be assumed to be constant. Moreover, we do not distinguish between waiting at the first station and waiting at transfer stations in this study. Introducing this distinction, and allowing for non-linear headway valuations, is conceptually straightforward.

$$c_{ijr} = f_{ijr} + \alpha t_{ijr} + \beta w_{ijr} \tag{1}$$

A change in the baseline timetable will cause generalized travel costs to change, possibly inducing a change in passenger volumes as well. Let

exponents 0 and 1 denote variables before and after the change, respectively. Using the rule-of-a-half approximation, the change in consumer surplus[4] is defined as in equation (2).

$$\Delta C = \sum_{ijr} T^0_{ijr}\left(c^0_{ijr} - c^1_{ijr}\right) + \frac{1}{2}\sum_{ijr}\left(T^1_{ijr} - T^0_{ijr}\right)\left(c^0_{ijr} - c^1_{ijr}\right) \qquad (2)$$

In the following, we ignore the second term in this expression, effectively assuming that $T^0_{ijr} = T^1_{ijr}$. This is a reasonable assumption as we are considering small timetable changes. This simplifies the exposition and the following calculations, and the approximation error is small[5] for moderate changes in the generalized travel cost.

The timetable consists of a number of train services indexed by k. Given the timetable and the OD matrix, the number of passengers boarding and alighting train service k at station i, called B_{ik} and A_{ik} respectively, can be calculated (as explained in more detail further below). The number of passengers on train service k between station i and the subsequent station along the service line is $N_{ik} = \sum_{s=1}^{i} B_{sk} - A_{sk}$, where the summation is taken over the stations served by train service k. This allows us to rewrite the change in consumer surplus from equation (2) as the difference in the relevant part of the aggregate generalized cost, denoted by C, before and after the change, see equation (3).

$$\Delta C = C^0 - C^1\,, \qquad \text{where } C = \sum_k \sum_{i \in k} \alpha(N_{ik}) t_{ik} N_{ik} + \beta_{ik} w_{ik} N_{ik} \qquad (3)$$

The notation $i \in k$ means that the summation should be taken over stations served by train service k; t_{ik} denotes the travel time from station i to the next station with train service k; w_{ik} denotes the average waiting time for train service k at station i.

A number of things should be noted. First, we assume that the timetable change does not affect fares, so the fare terms in the generalized cost cancel out. Second, the value of in-vehicle time α depends on the number of

[4] This definition rests on some conventional assumptions such as negligible income effects and a locally linear demand curve.

[5] If the demand elasticity with respect to generalized cost is ε, the relative error of a relative change in generalized costs p is $\frac{\varepsilon p}{2}$. For example, a 10% change of the generalized cost with a demand elasticity of -0.5 gives a relative approximation error of $\frac{\varepsilon p}{2} = 2.5\%$. As we only consider small changes in the timetable, p will stay moderate.

passengers in the train between each pair of stations – the more passengers in the train, the higher is the weight (i.e., the disutility) of in-vehicle travel time. A change in the timetable will typically change N_{ik} and hence α, since the crowding levels change. Third, the value of waiting time β_{ik} depends on i and k. This is because the marginal valuation of waiting time is falling since when waiting times (or rather, headways) are long, passengers can adapt their schedule to avoid waiting at the platform, see (Fosgerau, 2009) for a theoretical analysis of this. This matters mainly for long headways, however; in our case study, headways are so short that this distinction does not matter much. Finally, and most importantly, expressing ΔC in this way simplifies the implementation substantially, since there is no need to calculate and represent the entire generalized cost matrix $\{c_{ijr}\}$, or even have precise data on the full OD matrix $\{T_{ijr}\}$. It is enough to have data on the number of passengers in each train on each link (N_{ik}) and the arrival times of passengers to their departure stations (which determines their waiting times w_{ik} based on their destination station and the next train that is serving the latter). This data is much easier to measure than the full OD matrix or generalized cost matrix; station arrival data is often directly available from entry measurements (e.g., smartcard gantries), and passenger loads can be measured either by on-board counts, automatic counting or weighing systems. Of course, if the OD matrix is available, passenger loads, travel times and average waiting times can be calculated with the corresponding timetable. The latter method simplifies the calculation of ΔB. Further, using a simulation software also allows for introducing delays in the model.

The assumption that passengers take the next train that is serving their destination station is made possible since the commuter network in the case study has two lines forming an X without any branches and the studied train timetable does not include skip-stops or overtaking. In other possible case studies, networks and/or timetables can be more complex (with branches, skip-stops or overtaking). In this case, the passengers will take the quickest path to their destinations which may not necessarily correspond to the next train or to a direct trip (without transfer).

3.2. Calculating the change in producer surplus

The producer surplus is the difference between total fare revenues and operating costs. In the following, we assume that the timetable change does not affect fare revenues (although allowing for this is straightforward). This means that we only consider the change in operating costs. The total relevant operating cost can be separated into three parts: costs

proportional to trains' running distances (e.g., maintenance), costs proportional to trains' running times (e.g., staff), and costs proportional to the number of wagons necessary to run the timetable (capital costs for vehicles). Hence, we can write the change in producer surplus ΔP as $\Delta P = P^0 - P^1$ where P is calculated as in equation (4).

$$P = (1 + K_{overhead})\left(K_{time}\sum_{k}\sum_{i \in k} t_{ik} + K_{distance}\sum_{k}\sum_{i \in k} d_{ik} + K_{wagon}N_{wagon}\right) \quad (4)$$

The number of wagons N_{wagon} necessary to run the timetable is estimated based on the total number of train services. Since the studied timetables include slight changes compared to the baseline timetable, it is possible to manually compute the changes in vehicle allocation compared to the reported number of wagons for the baseline timetable. For more substantial changes to the train timetable, an existing vehicle allocation model can be used to compute the changes in N_{wagon}. The parameters $K_{time}, K_{distance}, K_{wagon}$ and $K_{overhead}$ are calculated by analyzing the public transport agency's total operating costs by type (e.g., train staff, maintenance, capital costs etc.), allocating them to most relevant proxy for variable costs (distance, time or number of wagons) and dividing by the corresponding total (total distance, total time, total number of wagons).

3.3. Parameters

The parameters that are used in the benefit calculations in the case study are taken from various sources: partly from the research literature, partly from the Swedish cost-benefit guidelines (Trafikverket, 2016), and partly from the calculation manual of the Stockholm Public Transport Agency (SLL, 2017a).

The notations, values and sources of the parameters that are used for the computation of the consumer surplus are provided in **Table 1** whereas **Table 2** provides the ones for the producer surplus (10 SEK is around 1€).

To compute the consumer surplus, we use the daily average for the distribution of trip types, i.e., 50% leisure, 50% commuting (SLL, 2017a). Given that the valuations (e.g., for in-vehicle travel and waiting times) are quite similar for leisure and commuter trips (see **Table 1**) that is considered to be a reasonable assumption in the absence of detailed data on the temporal distribution of trip types over the studied periods of the day,

However, the model allows to easily include the temporal distribution for trip types given the corresponding data.

Table 1. Parameters for the computation of the consumer surplus.

Parameter	Formula	Values	Source
In-vehicle travel time	$\alpha = K_{crowding}\alpha_0$	Leisure: $\alpha_0 = 57$ SEK/h Commuting: $\alpha_0 = 74$ SEK/h	(Eliasson and Börjesson, 2014)
Crowding factor	Piecewise linear function: $K_{crowding} = \begin{cases} K_{0-0.5}, 0 \leq \frac{n}{C} < 0{,}5 \\ \vdots \\ K_{1.75-2}, 1{,}75 \leq \frac{n}{C} < 2 \end{cases}$ $\frac{n}{C}$ = load factor n = number of passengers onboard C = number of seats in the trains	- Leisure trips, passengers sitting: 0-50 % is 1.04 50-100 % is 1.14 100-125 % is 1.26 125-150 % is 1.39 150-175 % is 1.53 175-200 % is 1.69 - Leisure trips, passengers standing: 100-125 % is 1.94 125-150 % is 2.15 150-175 % is 2.39 175-200 % is 2.64 - Commuting trips, passengers sitting: 0-50 % is 0.86 50-100 % is 0.95 100-125 % is 1.05 125-150 % is 1.16 150-175 % is 1.27 175-200 % is 1.4 - Commuting trips, passengers standing: 100-125 % is 1.62 125-150 % is 1.79 150-175 % is 1.99 175-200 % is 2.2	(Wardman and Whelan, 2011)
Waiting time	Piecewise linear function: $\beta = \begin{cases} \beta_{0-10}, 0 \leq w_{ijr} < 10 \\ \vdots \\ \beta_{30-60}, 30 \leq w_{ijr} < 60 \end{cases}$ w_{ijr} = waiting time	- Leisure trips: 0-10 min is 86 SEK/h, 11-30 min is 70 SEK/h and 31-60 min is 34 SEK/h - Commuting trips: 0-10 min is 74 SEK/h, 11-30 min is 53 SEK/h and 31-60 min is 26 SEK/h	(Algers et al., 2010)

Table 2. Parameters for the computation of the producer surplus (SLL, 2017a).

Parameter	Notation	Value
Time-related production costs	K_{time}	2 000 SEK/h
Distance-related production costs	$K_{distance}$	30 SEK/(wagon km)
Fixed vehicle costs	K_{wagon}	5 000 000 SEK/wagon/year
Production overhead	$K_{overhead}$	9% on top of total production costs

In the producer surplus, vehicle operation costs are estimated based on the total yearly travel distance and operation time for the trains and cost values according to the cost-benefit analysis recommendation from the local operator. Operating costs include capital costs and maintenance of the rolling stock, fuel and staff. Further, indirect costs including capital costs, overhead and administration are included based on the total number of passenger and rolling stock kilometers. The total costs are converted to daily costs by dividing yearly costs by the factor 320.

In Sweden, most of these valuations are compiled and explained in the guide for cost-benefit analysis, i.e., ASEK guidelines (Trafikverket, 2016). Several other countries have similar documents and handbooks. The British practical guide by Balcombe et al. (2004) provides extensive evidence on travel time, crowding, interchanges or transfers. These values can be used in various applications, e.g., public transport demand model, rail franchise specification. The guidelines that are used in these different countries are mostly based on a common body of research results.

3.4. Estimating passenger flow data

An important input to the calculation of the consumer surplus is the time-OD-matrix $\{T_{ijr}\}$. We will briefly describe how it can be estimated from station entry data (i.e., smartcards), and how it then can be used together with a given timetable to estimate passenger loads in the trains for calculating in-vehicle travel and waiting time costs. Handling trips with transfers is explained last, since this requires additional calculations.

Most public transport operators have access to data on the number of passengers entering each station during a certain interval of time, since this can usually be registered by most kind of entry gates. If some kind of smartcards are used, this becomes even easier. Let O_{ir} be the number of passengers entering station i at time r. In our case study, we only have data on the total number of passengers during each 15-minutes time

period, so we assume that the arrival rate is constant over each such time period.

A common limitation (which is also the case in our case study) is that there is no station exit data, i.e., $\{D_{jr}\}$, the number of passengers exiting station j at time r, is unknown. To overcome this, we assume that trips are symmetric, meaning that the total number of passengers entering a station j during a day is equal to the number of passengers alighting at that station, i.e., $\sum_r O_{jr} = \sum_{ir} T_{ijr} = \sum_r D_{jr}$. Together, this allows us to estimate the time-OD-matrix with a standard entropy maximization problem. See (Ait-Ali and Eliasson, 2019a) for more details.

In our formulas so far, we have tacitly assumed that the discretization of time, indexed by r, is sufficiently fine-grained compared to the discretization of the timetable. In a practical implementation, however, it is obviously inefficient to actually calculate and store the number of departures per station for each minute, or even second; instead, a constant rate of passenger departures per unit time is assumed to be known for each time interval. The number of passengers boarding a train is then easily calculated by multiplying the corresponding arrival rate(s) with the time elapsed since the previous train departure. The number of passengers alighting a train can be calculated in a similar way.

Let $T_{ij}(r)$ be the continuous rate of passengers entering station i to travel to station j at time r, and let r_{ik} be the departure time of train service k from station i. The number of passengers B_{ik} boarding train k at station i is then given by equation (5) where $j \in k(i)$ denotes that the summation is taken over all stations j served by train service k from station i.

$$B_{ik} = \int_{r_{i,k-1}}^{r_{ik}} \sum_{j \in k(i)} T_{ij}(r)\, dr \tag{5}$$

The time resolution of the rate $T_{ij}(r)$ depends on available data; in our case study, we have data on arrival rates per 15 minutes time period. **Figure 1** provides an illustration of the calculation of the number of boarding passengers B_{ik}. A similar method is used to calculate the numbers of alighting passengers A_{jk}.

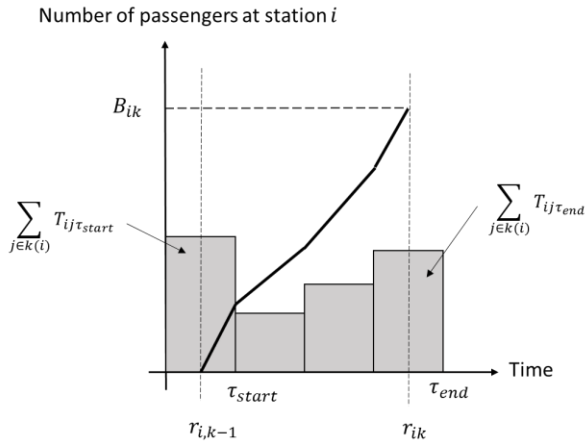


Figure 1. Calculation of the number of passengers B_{ik} boarding train k at station i using OD matrix $\{T_{ijr}\}$.

The model is intended to be used for commuter services with high frequency leading to short headways between departures and therefore short waiting times for the passengers at the origin stations. This makes it possible to assume that passengers enter the origin stations at a uniform rate. However, for longer waiting times between two consecutive train departures, passengers may adapt their time of entrance to the stations. As previously explained in section 3.1, this adaptation is accounted for using a piecewise linear function which reflects the decreasing value of longer waiting times, see the last row in **Table 1**.

When dealing with passenger flows in complex train networks, it is necessary to handle trips that require transfers between train services. Given the OD matrix, transfer trips are split into two; one from the origin station to the transfer station and another one from the latter to the destination station. The number of passengers in each of the two trips is calculated using the same method as for any direct trip. These are added to the direct trips which gives the overall train loads including transfers. When there are several possible transfer stations, passengers are assumed to transfer at the first possible station. An illustration of the passenger flow per link segment including transfer trips is given in **Figure 2**.

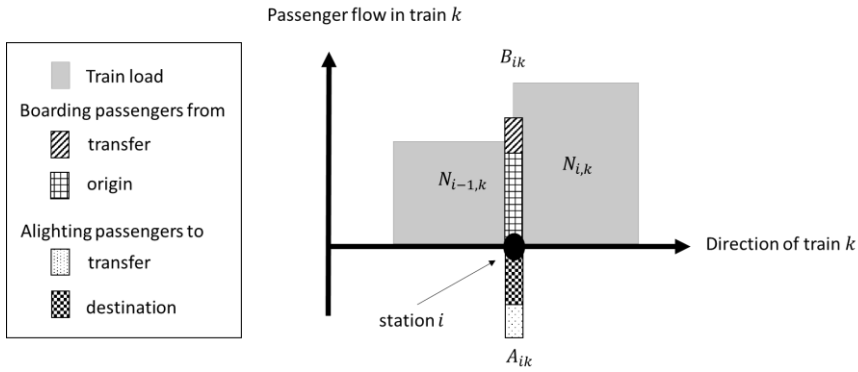


Figure 2. Passenger flow in train service k at station i including transfers.

3.5. Implementation

The network and the train timetables are modelled using the microscopic simulation tool *Railsys* which allows for instance to easily create, handle and simulate train timetables. There are many reasons for this choice, but the main motives for the purpose of this paper are:

- The Swedish Transport Administration provides a network, train and timetable models in *RailSys* that can be easily adjusted.
- Manipulation and visualization of networks and train timetables is facilitated.
- Conflict detection (for timetable alternatives).
- Timetables can be simulated to allow evaluation of feasibility and robustness (e.g., delay propagation). This is out of the scope of the current study, but a major advantage for practical applications.
- Data can be exported easily (e.g., in XML/CSV format) for integration with other software environments.

The network is microscopically modelled in the simulation tool down to individual tracks, switches and signals. This gives more flexibility in investigating the effect of various aspects of the railway infrastructure on passengers and operators and the opportunity to check if a timetable solution is feasible. Due to the complexity of the network model, the network is always handled within the simulation tool and never exported outside.

The train timetables are however exported and used outside the simulation tool. This data is processed in order to extract information about arrival and departure times at every station, travel times and distances between stations. It is easy to create, modify and delete train paths for different train timetables scenarios. It is also possible to add potential

delay distributions in the train timetable and simulate a certain number of days in order to estimate how the timetable will perform in operation.

4. Case study

This section applies the method on a case study from Stockholm, calculating reservation prices for a commercial train path request during three different time periods (morning peak, mid-day, afternoon peak). The case study illustrates the practical feasibility of the method, and also shows magnitudes of reservation prices, i.e., the loss of societal benefits in the commuter train system from allocating capacity to an additional commercial train.

4.1. Description of the Stockholm commuter trains

The Stockholm commuter trains (locally called *pendeltåg*) share parts of their tracks with commercial passenger[6] and freight trains (Froidh and Nelldal, 2015). **Figure 3** shows the network in 2016. For the central section, departure times and tracks are strictly regulated in order to coordinate the high frequency of services passing the double track line south of Stockholm Central. The signaling system in combination with the requirement to offer regular commuter traffic allows for 28 trains per hour (i.e., 24 commuter and 24 long distance), but four of these paths are not allocated in order to ensure some buffer (Trafikverket, 2013). While commuter and long-distance services are separated on most of the other lines, they share the tracks northwest of Stockholm between Karlberg and Bålsta (**Figure 3**) which is the focus of the conflicts treated in this study. During peak hours, six commuter trains and two longer-distance trains per hour and direction have to be coordinated. However, in principle the whole commuter network is included in this study in order to cover all trips than can be affected by the adjustments. For the studied working day, the data includes around 346 commuter train departures allowing more than 230,000 individual trips.

[6] For regional passenger trains, frequent commuters can get their fares partly subsidized through a cap on their total fare expenses, where the excess is covered by a combination of regional public transport authorities. However, this complication is irrelevant for the present case study.

Figure 3. Stockholm commuter network (Frohne, 2016).

The commuter network has an X-shape with around 50 stations along 240 km. All the trains pass a central section between *Karlberg* and *Älvsjö*. Commuter services are mainly operated in two channels, *Södertälje-Märsta* and *Nynäshamn-Bålsta*, complemented with additional services on shorter distances during peak hours. One of these complementary lines is prolonged to *Uppsala* via Sweden´s largest airport *Arlanda*. In addition, there is a branch connecting *Södertälje* to *Gnesta* (excluded in the case study for simplicity). Moreover, there are no skip-stop services or overtaking, the two daily departures that change branches are excluded in this study.

Between *Karlberg* and *Älvsjö*, the major lines share the infrastructure. *Stockholm central* is the hub of the network allowing connections to services such as long-distance passenger trains, local and regional buses, airport shuttles and metro.

In 2016, the Stockholm region had 2.3 million inhabitants (SCB, 2016). **Figure 4** presents the number of departing passengers from each station across the day. There are major travelling peaks in the morning and afternoon peak hours. Note the peak from Stockholm central station, especially in the afternoon peak hours.

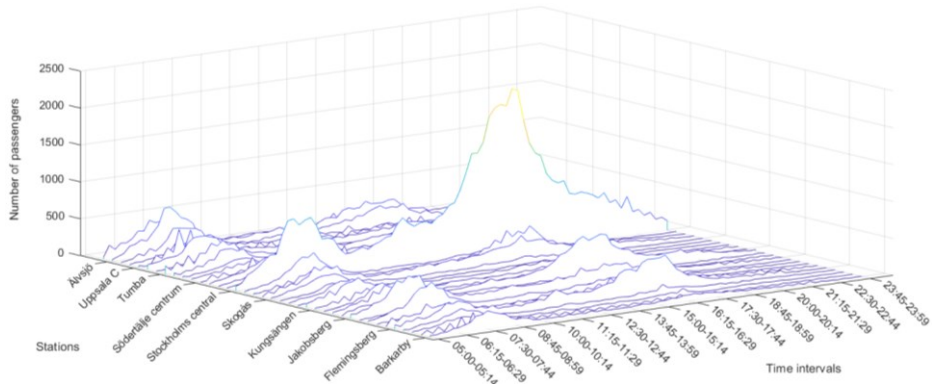


Figure 4. Number of passengers (from smartcard data) travelling from different stations per 15 minutes time intervals.

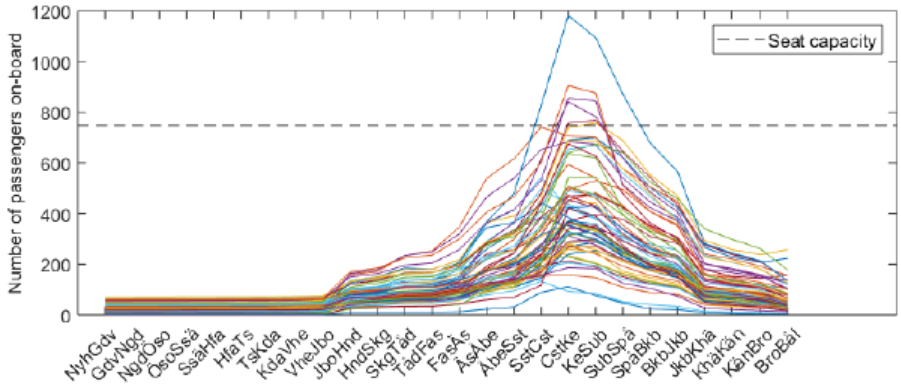


Figure 5. Typical load of passengers of all the train paths during a weekday between Bålsta and Nynäshamn.

Typical passenger loads for the different trains in a weekday in 2016 are presented in **Figure 5**. All trains running between Bålsta and Nynäshamn in both directions are included. The x-axis shows the links between two consecutive stations on the line, the passenger load is given on the y-axis. The horizontal dashed line represents the train (i.e., two wagon units) seat capacity. In the case study, each train consists of two wagons of the model Cordia X60 with a seating capacity of 374 passengers for each wagon unit (ALSTOM, 2004). Although there are extra train departures, crowding in the trains is common. The figure shows that passenger load on the line is well above the seating capacity for certain trains around the central station during peak hours.

4.2. Allocating capacity for a commercial train path

In this case study, a working day in September 2016 on the line *Bålsta-Nynäshamn* is used as baseline timetable. **Figure 6** shows a graphical timetable, with a train path request in blue conflicting with a commuter train path in yellow. The conflict area is colored red, and the commuter trains not involved in the conflict are in purple. In a graphical timetable, each line represents the scheduled time (horizontal axis) and location (vertical axis) in a train run surrounded by squares showing how long each train blocks the belonging signal sequence (block section). Each block section can only be occupied by one train at a time.

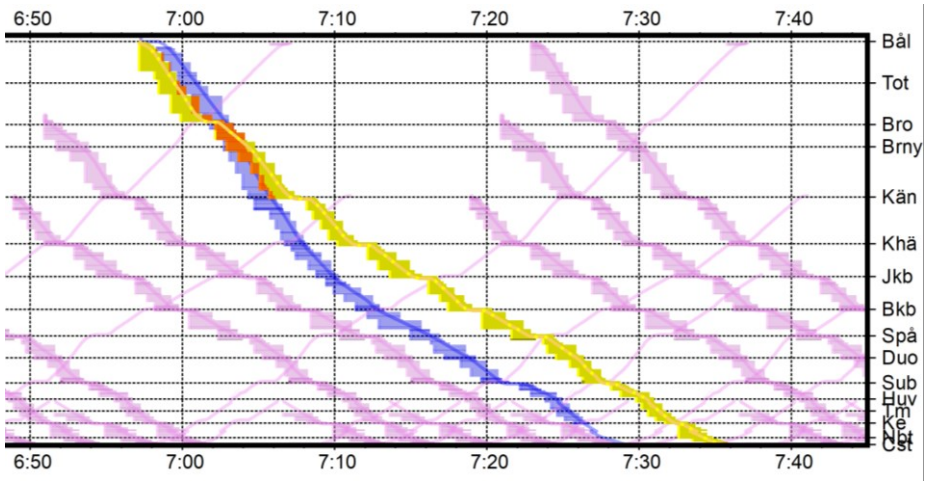


Figure 6. Graphical timetable showing a conflict (red) between a commuter train (yellow) and a commercial train (blue) during the morning peak (around 7:00 AM) on the line between Bålsta and Stockholm central station. Commuter trains not involved in the conflict are purple.

In order to resolve the conflict, the commuter train timetable can be changed in either of three ways:

- Scenario S1 – Remove: the conflicting commuter train service is cancelled. Note that this means that the whole roundtrip train service needs to be cancelled.
- Scenario S2 – Delay: the commuter train´s dwell time at a certain stopping station is prolonged, in order to let the commercial train overtake.
- Scenario S3 – Shift: the commuter train´s departure time at the first station is shifted, so that the commercial train can pass first.

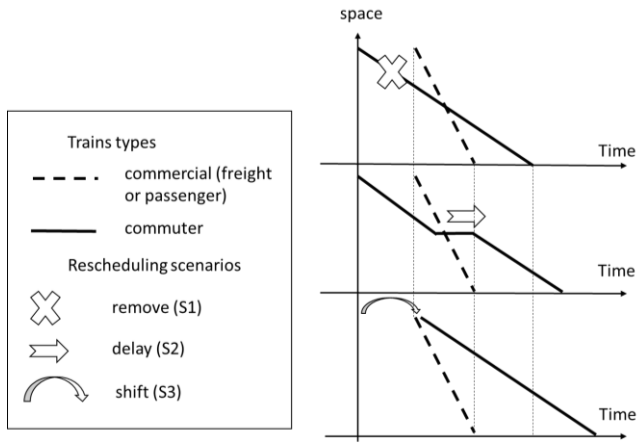


Figure 7. Different experiment scenarios for conflicts between subsidized and commercial services.

Changes were therefore made to the baseline timetable by rescheduling a certain train path using the different rescheduling scenarios, i.e., remove, shift and delay. These scenarios are illustrated in **Figure 7** with the graphical timetable showing the conflicts between the subsidized and the commercial trains and the different rescheduling scenarios. In order to avoid large costs, the adjusted timetables are designed not to largely affect vehicle and staff allocation. While the adjustments in S2 and S3 are within the available timetable margins, the whole roundtrip service is removed in S1.

The last two scenarios (i.e., delay (S2) and shift (S3)) differ from a passenger point of view mainly in the way delays are experienced. In both scenarios, average waiting and total crowding penalties will tend to increase, since passengers get less evenly spread across train services when departures are not at even intervals. In S2, passengers onboard also get a longer travel time.

These rescheduling scenarios, applied to the baseline timetable, are tested for conflicts in three different time periods, i.e., morning peak (6:00-9:00), mid-day off-peak (11:00-14:00) and afternoon peak (15:00-18:00). To illustrate this, **Figure 8** shows two graphical timetables presenting two rescheduling scenarios, i.e., delay (left) and shift (right). Lines show train paths; squares the infrastructure each path is blocking. Note that commercial trains that are not included in this study are not visualized here. The figure on the right shows the commuter timetable where the departure time of the conflicting commuter train path (yellow) has been shifted by +3 minutes whereas the one on the left shows that of delaying it in *Kungsängen* station for 4 minutes to let the commercial

train pass. The third type of conflict solution is not presented in the figure but is about completely removing the conflicting commuter train path.

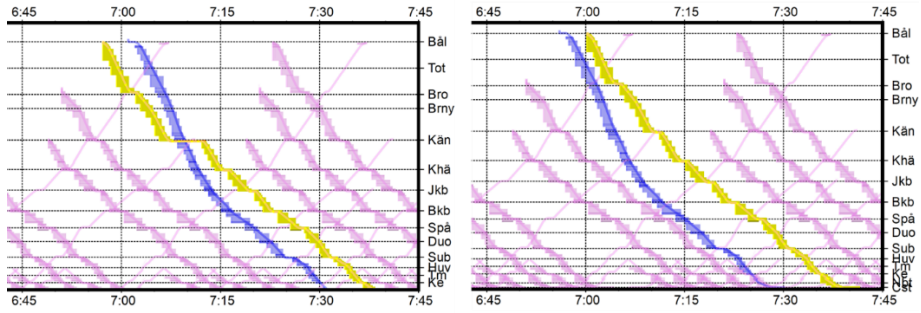


Figure 8. Illustration of two different commuter timetable changes for conflict solution: S2-delay (left) and S3-shift (right). Adjusted commuter train path in yellow, other commuter trains in purple, commercial train in blue.

4.3. Results

The total societal costs of the different rescheduling scenarios are presented in **Figure 9**. Changing the baseline timetable to accommodate the commercial train path can cause more than 100 kSEK (10 SEK is around 1€) in total societal losses in the commuter train system. Scheduling a commercial train leads to higher societal costs in the peak hours, especially in the morning. This can be explained by the variation in the trip distribution during the day, see **Figure 4**. The minimal cost is incurred by shifting the departure time of the conflicting commuter train, resulting in a societal loss of around 9-16 kSEK. Removing the conflicting commuter train path (i.e., scenario S1) always lead to the highest societal costs, especially during peak hours. A less costly rescheduling strategy is to shift the departure time of the conflicting commuter train (i.e., scenario S3). Delaying the conflicting train at a certain station (S2) is also less costly but is more expensive than shifting the departures, especially in peak hours.

In order to give more details about the results in **Figure 9**, we present the different elements forming the total societal costs in **Table 3** by showing the different consumer or passenger costs (in-vehicle travel time, waiting and transfer costs) and producer or operator costs (distance and time dependent costs and fixed costs). Negative costs in **Table 3** indicate savings, such as decreased operating or shorter transfer times. A dash means no difference in the costs relative to the baseline timetable.

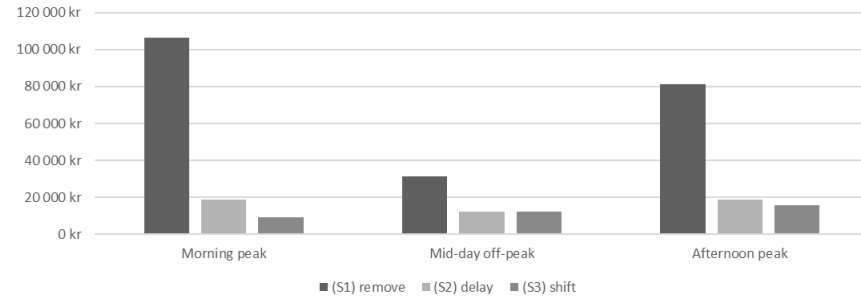


Figure 9. Total societal costs of the changes to the baseline commuter train timetable (in SEK, 1€ is around 10 SEK).

Removing the conflicting train always causes larger societal losses than delaying or shifting the commuter train. Indeed, cancelling the commuter train leads to higher consumer costs compared to the gains in producer costs, which in turns lead to a high total cost especially in peak hours. However, delaying the departure of the conflicting commuter train in a stopping station leads to higher costs in both consumer and producer costs. The latter is due to the increase in operation time due to the additional stopping time in the middle of the line. Shifting the departure times of the conflicting commuter train causes no additional producer costs and higher consumer costs. These results clearly show that shifting the trains is the societally optimal strategy to accommodate an additional commercial train path. Of course, this may not always be possible if the timetable is very dense.

Table 3. Components of the total societal costs of changes relative to the baseline timetable (in SEK, 1€ is around 10 SEK).

Case	Time	Consumer costs				Producer costs				Total
		travel	waiting	transfer	**total**	distance	time	fixed	**total**	**costs**
S1 – remove	Morning	97 499	18 835	-476	115 859	-6 370	-3 633	-900	-10 904	104 955
	Mid-day	32 433	9 482	-457	41 458	-6 370	-3 633	-900	-10 904	30 554
	Afternoon	70 817	21 298	-451	91 664	-6 370	-3 633	-900	-10 904	80 761
S2 – delay	Morning	14 059	2 038	-	16 096	-	133	12	145	16 242
	Mid-day	9 756	760	-	10 516	-	133	12	145	10 661
	Afternoon	12 138	6 920	-	19 058	-	133	12	145	19 203
S3 – shift	Morning	7 894	1 353	-	9 247	-	-	-	-	9 247
	Mid-day	9 571	1 059	-	10 630	-	-	-	-	10 630
	Afternoon	15 070	974	-	16 044	-	-	-	-	16 044

Based on these societal costs, one can compute the minimum cost that the commercial operator(s) should pay to compensate the resulting losses in societal benefits. The last column of **Table 3**, i.e., total costs, shows this minimal price for the different rescheduling scenarios and time periods of the day. Compared to the current track charge for the train path, these total societal costs are substantially higher. Swedish track charges are made up of several components, one of which is set to partly reflect congestion (lack of capacity) on the tracks. For a train passing Stockholm, this "congestion charge" is 433 SEK (Trafikverket,

2020b), i.e., less than 5% of the lowest societal loss caused by accommodating an additional commercial train path. For instance, the total track charge for a commercial passenger train Västerås - Stockholm C sharing the track with the subsidized trains between Bålsta and Stockholm (45 km of the 105 km in total) is 1 454.46 SEK in total, including electricity and so on. Since the tracks are fully used for most parts of the day, the calculated costs can also be interpreted as the marginal value of increased capacity.

It is possible that certain adjustments of the commuter train timetable may result in positive societal benefits rather than losses, since there is no absolute guarantee that the baseline commuter train timetable is strictly optimal. In such cases, the reservation price of the accommodated train path is of course zero (except for the usual charges for wear-and-tear etc.). In some cases, however, such results may be a sign that the parameters of the benefit-cost calculation, or its input data, may need to be adjusted.

It is important to note that reliability is not included in this study. When disturbances are added and operation simulated, a timetable's robustness to delays can be studied and might change results (e.g., optimal strategy). Including reliability aspects, part of the future work, shows the importance and advantage of using a microscopic timetable simulation tool such as *Railsys*. Non-optimal timetables may also sometimes be due to political considerations; there is anecdotal evidence of over-supply from other commuting train systems. Exploring the optimality of timetables from a benefit-cost perspective is an interesting research area of its own but is left for future research.

5. Conclusions and future work

This paper describes an approach to resolve conflicting capacity requests between commercial trains and publicly controlled traffic ("commuter trains"), by calculating the loss of societal benefits caused by accommodating the commercial train path and using this to price the commercial train path. The calculation of benefits takes into account in-vehicle times, waiting times, transfers, crowding and operating costs. The economic logic of this approach can be seen as a way to internalize the externality the commercial train causes on the commuter trains by having them to adjust their schedule. This means that if the commercial operator finds the price worth paying, overall social welfare increases by granting capacity to the commercial train service, and vice versa.

The case study of the commuter train services in Stockholm shows that the evaluation model can be used in different situations to help planners

evaluate the impact of their timetable choices. Results also provide insights into how the model can be used to price commercial train paths in conflict with the commuter train services. The results show that accommodating additional train paths in the busy commuter train timetable comes at a high societal cost – much higher than the current track charge intended to partly reflect scarce capacity in Stockholm. We also show that it is possible to substantially reduce the costs of changes in commuter train timetables by choosing the right rescheduling alternative. This best alternative might not be evident to the planners without the help of a cost assessment model as the one presented in this paper.

The evaluation model that is presented in this paper is not limited to the experiment in the case study. It can be used in a number of other real-world situations to help railway planners make efficient changes to the timetables with minimal societal costs. Ideas for possible applications include the assessment of the societal effects of using a train timetable with skip-stops instead of all-stops. It is also possible to link the model to an optimization model in order to find optimal train timetables given a certain trip distribution. Moreover, the results of the evaluation model can be used to assess project proposals for railway capacity expansion by comparing the costs of such projects with the societal benefits of having a new train timetable (with expanded capacity) in the long run. This point is further discussed by Eliasson and Börjesson (2014).

Additionally, one may also improve the model by including extra costs and accounting for negative and positive externalities. Many possible future works can help improve the quality of the assessment model and the case study that is presented in this paper. As mentioned in the limitations, the OD-data was incomplete, hence only an estimate was used. Applying the model in a case study with complete input data may provide more accurate results and insights. Moreover, socio-economic data on the passengers, if available, can be used with an improved version of the model where disaggregated results can be computed. In this way, one can get insights as to which (how much each) socio-economic group is winning or losing. This gives further insights on the equity of the train service operations.

Acknowledgements

This research is part of the project Socio-economically efficient allocation of railway capacity, SamEff (*Samhällsekonomiskt effektiv tilldelning av kapacitet på järnvägar*). The project is funded by a grant from the Swedish Transport Administration (*Trafikverket*). The authors are grateful to three anonymous reviewers for their helpful comments and suggestions.

References

ABRIL, M., BARBER, F., INGOLOTTI, L., SALIDO, M. A., TORMOS, P. & LOVA, A. 2008. An assessment of railway capacity. Transportation Research Part E-Logistics and Transportation Review, 44, 774-806.

ADLER, N., PELS, E. & NASH, C. 2010. High-speed rail and air transport competition: Game engineering as tool for cost-benefit analysis. Transportation Research Part B: Methodological, 44, 812-833.

AFFUSO, L. 2003. Auctions of rail capacity? Utilities Policy, 11, 43-46.

AIT-ALI, A. & ELIASSON, J. 2019. Dynamic Origin-Destination-Matrix Estimation Using Smart card Data: An entropy maximization approach. RailNorrköping2019. Norrköping, Sweden.

ALGERS, S., BÖRJESSON, M., SUNDBERGH, P., BYSTRÖM, C. & ALMSTRÖM, P. 2010. Valuation of Time in Transport – The National Studies 2007/08 in Sweden. WSP report.

ALSTOM 2004. CORDIA 60X Stockholm Transport Renews its Commuter Fleet.

BALCOMBE, R., MACKETT, R., PAULLEY, N., PRESTON, J., SHIRES, J., TITHERIDGE, H., WARDMAN, M. & WHITE, P. 2004. The demand for public transport: a practical guide.

BORNDÖRFER, R., KLUG, T., LAMORGESE, L., MANNINO, C., REUTHER, M. & SCHLECHTE, T. 2018. Handbook of Optimization in the Railway Industry, Springer International Publishing.

BROMAN, E., ELIASSON, J. & ARONSSON, M. 2018. A Mixed Method for Railway Capacity Allocation. 21st Meeting of the Euro Working Group on Transportation 2018. Braunschweig.

BRÄNNLUND, U., LINDBERG, P. O., NÕU, A. & NILSSON, J.-E. 1998. Railway Timetabling using Lagrangian Relaxation. Transportation Science, 32, 358-369.

DELORME, X., GANDIBLEUX, X. & RODRIGUEZ, J. 2009. Stability evaluation of a railway timetable at station level. European Journal of Operational Research, 195, 780-790.

ELIASSON, J. & BÖRJESSON, M. 2014. On timetable assumptions in railway investment appraisal. Transport Policy, 36, 118-126.

FOSGERAU, M. 2019. The marginal social cost of headway for a scheduled service. Transportation Research. Part B: Methodological, 43, 813-820.

FROHNE, E. 2016. Stockholm commuter train system map. Wikimedia Commons.

FROIDH, O. & NELLDAL, B. L. 2015. The impact of market opening on the supply of interregional train services. Journal of Transport Geography, 46, 189-200.

HENSHER, D. A. 1997. A practical approach to identifying the market potential for high speed rail: A case study in the Sydney-Canberra corridor. Transportation Research Part A: Policy and Practice, 31, 431-446.

HENSHER, D. A. & TON, T. 2002. TRESIS: A transportation, land use and environmental strategy impact simulator for urban areas. Transportation, 29, 439-457.

ISAAI, M. T., KANANI, A., TOOTOONCHI, M. & AFZALI, H. R. 2011. Intelligent timetable evaluation using fuzzy AHP. Expert Systems with Applications, 38, 3718-3723.

JIANG, Z., HSU, C.-H., ZHANG, D. & ZOU, X. 2016. Evaluating rail transit timetable using big passengers' data. Journal of Computer and System Sciences, 82, 144-155.

JOHNSON, D. & NASH, C. 2008. Charging for scarce rail capacity in Britain: a case study. Review of Network Economics, 7.

KUNIMATSU, T., HIRAI, C. & TOMII, N. 2012. Train timetable evaluation from the viewpoint of passengers by microsimulation of train operation and passenger flow. Electrical Engineering in Japan, 181, 51-62.
NASH, C. 2005. Rail Infrastructure Charges in Europe. Journal of Transport Economics and Policy (JTEP), 39, 259-278.
NASH, C., COULTHARD, S. & MATTHEWS, B. 2004. Rail track charges in Great Britain—the issue of charging for capacity. Transport Policy, 11, 315-327.
NEWBERY, D. M. 2003. Network capacity auctions: promise and problems. Utilities Policy, 11, 27-32.
NILSSON, J.-E. 2002. Towards a welfare enhancing process to manage railway infrastructure access. Transportation Research Part A, 36, 419–436.
RADTKE, A. & BENDFELDT, J.-P. 2001. Handling of railway operation problems with RailSys. Proceedings of the 5th world congress on rail research, 2001 Cologne. World congress on rail research.
SCB 2016. Population in the country, counties and municipalities on 31 December 2016 and Population Change in 2016. Statistics Sweden.
SLL 2017. Dokumentation av SAMS 3.0. Stockholm.
TRAFIKVERKET 2013. Planning conditions for Stockholm sector T15.
TRAFIKVERKET 2016. English summary of ASEK recommendations.
TRAFIKVERKET 2017. Network Statement 2018. Swedish Transport Administration.
WARDMAN, M. & WHELAN, G. 2011. Twenty Years of Rail Crowding Valuation Studies: Evidence and Lessons from British Experience. Transport Reviews, 31, 379-398.

Paper P3

Are commuter train timetables consistent with passengers' valuations of waiting times and in-vehicle crowding?

Ait-Ali A.[1,2], Eliasson J.[2] and Warg J.[3] (2020)
[1] VTI Swedish National Road and Transport Research Institute, Transport Economics (TEK), Stockholm
[2] Linköping University, Department of Science and Technology (KTS), Norrköping
[3] KTH Royal Institute of Technology, Division of Transport Planning, Stockholm

Submitted for journal publication

Abstract

Social cost-benefit analysis (CBA) is often used to analyze transport investments, and can also be used for transport operation planning and capacity allocation. If it is to be used for resolving capacity conflicts, however, it is important to know whether transit agencies' timetable requests are consistent with the valuations in the CBA framework. In this study we compare passengers' valuations of waiting times and in-vehicle crowding with the valuations implied by a transit agency's choice of commuter train timetables. Comparing the optimal and the actual frequencies of Stockholm commuter trains allows us to estimate the agency's implicit valuations of waiting time and crowding. The results suggest that the agency adopted service frequencies slightly higher than socially optimal, implying a higher implicit valuation of waiting time and crowding than the ones used in CBA guidelines, which are estimated from passenger preferences. We also find that the optimal frequencies are more sensitive to the waiting time valuation than to that of crowding.

Keywords: waiting time; crowding; cost benefit analysis; implicit preference; commuter train

1. Introduction

Public transport is a central part of most urban transport systems, and the decisions of public transport agencies about which services to run are hence very important. In this paper, we explore to what extent timetables determined by a local public transport agency (PTA) are consistent with passengers' valuations of waiting times and in-vehicle crowding, as codified in guidelines for social cost-benefit analysis (CBA). We present a way to do this based on analyzing the timetables' trade-offs between service frequency and operations costs, and apply the method on commuter trains in Stockholm.

Just as passengers' valuations can be inferred by observing their choices between travel options, analogous implicit valuations can be derived from the choices which the agency makes when determining timetables. These implied valuations can then be interpreted as the agency's "revealed preferences" regarding, e.g., waiting times and crowding. The question is then whether these coincide with passengers' valuations. Similar studies of the implicit preferences of public agencies have been conducted by, e.g., McFadden (1975), McFadden (1976), Nellthorp and Mackie (2000) and Eliasson and Lundberg (2012). These papers have studied various administrations' decisions about infrastructure investments, estimating implicit valuations of different kinds of benefits and costs. To our knowledge, this is the first similar study based on an agency's train timetable decisions.

There are several reasons why the question of consistency between an agency's timetable decisions and passenger preferences is important. First and most obvious, possible differences between agencies' and passengers' implicit valuations raise several interesting questions. Is the agency simply failing to construct cost-efficient timetables? In that case, the methods developed in this paper can indicate how timetables can be improved. Or is there something missing from the conventional CBA framework, that the agency correctly considers? In that case such studies can identify improvements of the framework. Or are the agency's decisions affected by other considerations – perhaps (hidden) political pressure? Or are there some constraints on the agency's decisions which prevent it from making optimal decisions from the passengers' point of view? Exploring whether an agency's decisions are consistent with passenger valuations forms a starting point for interesting and deep discussions about an agency's objectives and efficiency, as well as the ability of the CBA framework to capture the relevant aspects of public transport service provision.

A second motivation for our interest in this issue is that it has been suggested by Johnson and Nash (2008) and Ait-Ali et al. (2020) that CBA of timetable adjustments can be used to resolve capacity conflicts between commercial train services and publicly controlled commuter trains. The idea is to calculate the net loss of social benefits incurred when the commuter train timetable is adjusted to make room for a commercial train service, and use this loss as a reservation price for the commercial train slot: capacity is allocated to the commercial train only if the commercial operator is willing to pay an access charge equal to this reservation price. However, this idea rests on the assumption that the CBA framework used to calculate the social loss from the timetable adjustment is consistent with the PTA's timetable preferences. If not, the idea might be difficult to accept for the PTA– for example, if the results of the cost-benefit analysis would indicate that it would yield a positive net social benefit to reduce the number of commuter train services compared to what the agency wants. At least, one would need to investigate the reasons for possible differences between the valuations in the CBA framework and the preferences of the PTA, and determine whether or how they can be reconciled.

A third reason for investigating the principles underlying agencies' timetable choices is that knowledge of these principles is necessary for evaluating infrastructure investments. As discussed in Eliasson and Börjesson (2014), the benefits of a railway capacity improvement are determined by the difference in timetables with and without the investment. To conduct a CBA of a railway investment, these timetables must hence be constructed, and the analyst needs a guiding principle to determine them. Eliasson and Börjesson (2014) suggest that, in the lack of better evidence, an analyst could assume that the PTA strives to maximize net social benefits – but empirical evidence of the principles implicit from agencies' actual timetable choices is obviously better.

Section 2 provides an overview of the relevant research literature. Section 3 describes the analytic model. Data for the numerical analysis is presented in section 4 and results in section 5. We conclude the paper with section 6.

2. Literature Review

There is a vast literature on passengers' valuations of trip dimensions, such as in-vehicle travel time, in-vehicle crowding, waiting time, walking time and delays. Abrantes and Wardman (2011) provided an overview and meta-analysis for the valuation of the in-vehicle travel time based on British evidence. Wardman and Whelan (2011) also performed a meta-analysis to evaluate the British value of crowding in rail trips. The two

authors collected data on crowding valuations from the last 20 years from 15 different studies. The meta-analysis quantified the variations in the large set of time multipliers. The study aggregated these values into implied multipliers for seated and standing travelers for commuter and leisure trips. Most such valuation studies are based on stated choice experiments, but there are also studies based on observed behavior (i.e., revealed preference), for example Tirachini et al. (2016) who estimated crowding valuations (standing and sitting multipliers) based on smart card data from the metro system in Singapore.

The valuations of in-vehicle travel time and waiting time used in this study are based on the Swedish value of time study, reported by Algers et al. (2010) and Börjesson and Eliasson (2012), and subsequently updated and included in the Swedish CBA guidelines (Trafikverket, 2016). The valuation of crowding is based on the study by Björklund and Swärdh (2017), who estimated crowding multipliers for different modes and areas from Swedish data, reaching similar results as the Wardman and Whelan (2011) meta-study in the UK.

Just as passengers' implicit valuations can be inferred by analyzing their choices between options with different benefits and costs, agencies' "implicit preferences" can be inferred by analyzing their decisions. One of the earliest studies is by McFadden (1975) who looked at the implicit valuations of benefits and costs of road infrastructure investments implied by the decisions of a transport agency. The author developed a theoretical framework based on discrete choice modelling to infer the implicit choice criteria and benefit valuations used by the agency when selecting infrastructure investments. The inference relies on *ex-post* evaluation of the consequences and outcomes of the selection decisions. Similar studies have been presented by Nellthorp and Mackie (2000) and Eliasson and Lundberg (2012).

Compared to the many studies of travelers' valuations, there are only a few studies of the implicit preferences of bureaucracies. Even fewer have compared the two sets of valuations, and as far as we know none in the context of public transport planning (e.g., commuter train systems). A number of studies looked at the optimal supply of public transport, such as Qin and Jia (2013) who studied a crowded rail transit line in China, Börjesson et al. (2017) who analyzed optimal bus fares and frequencies in Stockholm, and Asplund and Pyddoke (2019) who did a similar study but in a medium-sized Swedish city. These are examples of studies comparing actual supply with the optimal one (according the CBA framework used), although they do not estimate the valuations implied by the actual transit supply.

Basu (1980), in a book about the revealed preference of governments, formalized a model by Weisbrod and Chase (1966) which studied income redistribution weights in CBA studies. This formalization is based on a standard model of a social welfare function and the distinction between local and global welfare. Such a model allows to estimate the weights of a welfare function based on information about the projects chosen by the government. Another formalization by the same author used fuzzification for analysis of revealed binary preferences (Basu, 1984). Brent (1991) discussed the previous techniques for revealing a government's distributional weights. The author contrasted stochastic methods, e.g., McFadden (1975), with deterministic ones, e.g. Basu (1980), indicating a preference for stochastic methods. The latter is applied and discussed in the case of the UK railway closure at that time. The stochastic approach is also used in the more recent work by Scarborough and Bennett (2012). They applied choice modelling techniques to estimate distributional weights in CBA models for environmental policy analysis.

A somewhat related literature considers consumer (i.e., personal and self-interested) versus citizen (i.e., social and moral) preferences. Im et al. (2014) looked at the extent to which citizen preferences are reflected in the resource allocations from the budget of the city of Seoul both at city and district level. The authors found that there is no perfect reflection of such preferences, meaning that resource or budget allocation in the city seem to be non-participatory. The authors highlighted and discussed the potential of a participatory budgeting which reflects the citizen preferences. Similar studies were also conducted in the US (Franklin and Carberry-George, 1999), the Netherlands (Michels and De Graaf, 2010) and Malaysia (Manaf et al., 2016). Most of these studies claim a positive impact of citizen involvement in decision making, and that participatory decision making is desirable in representative democracies. However, Bossert and Weymark (2004) found it difficult to include all the citizen groups and show that the social welfare maximizing function can be dictatorial. Therefore and according to Arrovian social choice theory, certain individual preferences must be considered over others, see Arrow's impossibility theorem or paradox (Arrow, 1963).

Lewinsohn-Zamir (1998) criticized the distinction between consumer and citizen preferences in the context of the provision of public goods, e.g., public transport services. The author claimed that such a distinction is unrealistic, and no quantitative difference can be made, arguing that both preferences are driven by other trade-offs that are less manifested in daily life. Moreover, since such preferences are successfully considered, in many cases, in the political arena, citizen preferences should be given more weight and be carefully used in tools such as cost-benefit analysis.

3. Analytic model

Our evaluation framework includes passenger benefits and train operating costs. Passenger benefits include travel times, in-vehicle crowding and waiting times, while operating costs include fixed costs, time- and distance-dependent costs and overhead costs. Since we are studying relatively minor changes in the timetables, we use a fixed origin-destination matrix, which means that there are no changes in external effects due to modal shifts from road transport, and no changes in fare revenues or tax revenues. Adding such effects is straightforward, provided that demand effects can be estimated.

In the present study, we also ignore unexpected delays. This is an important issue for future research, since robustness towards incidents, minimizing knock-on delays, may be an important consideration when constructing optimal timetables. As we shall see, however, this omission does not seem to affect our conclusions in the present paper.

Consider a commuter train line with a given time-dependent origin-destination (OD) matrix, specifying the number of passengers traveling from station i to station j for each time interval. Train services with different departure times are indexed by k. We assume that all services have the same travel time between station pairs, so passengers will simply take the first arriving train from their origin station. Given the time-dependent OD matrix and a schedule of train services, the number of boarding and alighting passengers on each station can be calculated for each train service. Let B_{ik} and A_{ik} be the number of boarding and alighting passengers, respectively, at station i and departure k. We distinguish train directions by letting stations have different indices depending on which direction passengers are travelling in, so each physical station will have two indices, one for each direction of the line. Let $F_{ik} = \sum_{l \le i}(B_{lk} - A_{lk})$ be the number of passengers onboard train service k at the link segment following station i.

The total social (or societal) cost for the services is the sum of passengers' generalized travel costs and the operating costs for the train services. Generalized travel costs consist of two parts: waiting times and in-vehicle travel times (including in-vehicle crowding). In the present study, we assume that the valuation β of waiting time (i.e., headway between services) is constant, since we study high-frequency services. Analyses for low-frequency services need to take into account that the marginal valuation of headway decreases with the length of the headway, since travelers can adjust their schedule to avoid waiting at the platform (Fosgerau, 2009). The valuation of in-vehicle travel time increases with the

crowding in the vehicle, since traveling in crowded conditions incurs a higher disutility per minute on travelers (i.e., longer perceived travel time). Let the valuation of in-vehicle time be $\alpha\left(1+\gamma\left(\frac{F_{ik}}{S}\right)^{\theta}\right)$, where α is the value of in-vehicle travel time, γ and θ are the studied parameters for the valuation of in-vehicle crowding, S is the number of seats in the train and F_{ik} is the number of passengers onboard train service k on link i. This function fits the results of the crowding studies by Wardman and Whelan (2011) and Björklund and Swärdh (2017), the two studies also yield similar parameter values.

Operating costs include staff, maintenance and other operation-related costs. We ignore fixed costs since they do not vary with the service frequency.

If we assume that the timetable is regular for a certain time period, and the service frequency is N trains per hour, we can write the total social costs $\mathrm{TC}(N)$ in equation (1) as the sum of operating costs KN, waiting time costs $\Sigma_{ik}\beta\frac{B_{ik}}{2N}$ and in-vehicle time costs $\Sigma_{ik}\alpha\left(1+\gamma\left(\frac{F_{ik}}{S}\right)^{\theta}\right)t_iF_{ik}$.

$$\mathrm{TC}(N) = KN + \sum_{\substack{\text{station } i \\ \text{train } k}} \frac{\beta B_{ik}}{2N} + \sum_{\substack{\text{link } i \\ \text{train } k}} \alpha\left(1+\gamma\left(\frac{F_{ik}}{S}\right)^{\theta}\right)t_iF_{ik} \qquad (1)$$

The formulation in (1) can be rewritten by rearranging the terms that include the frequency as in equation (2).

$$\mathrm{TC}(N) = KN + \alpha\sum_{i,k} t_iF_{ik} + \frac{\alpha\gamma}{S^{\theta}}\sum_{i,k} t_i(F_{ik})^{\theta+1} + \frac{\beta}{2N}\sum_{i,k} B_{ik} \qquad (2)$$

Note that the term $\sum_{i,k} t_iF_{ik}$ is constant since it is the total passenger travel time which does not depend on the level of supply (i.e., frequency). Another constant term is $\mathcal{B} := \sum_{i,k} B_{ik}$ which is the total number of passengers boarding the trains. The optimal service frequency N^* is the one that minimizes the total social cost as defined in (3).

$$N^* = N^*(\gamma,\beta,\theta) = \underset{N\in\mathbb{N}^*}{\operatorname{argmin}}\,\mathrm{TC}(N) \qquad (3)$$

If crowding is ignored ($\theta = 0$), the optimal frequency is the well-known square root principle stated by Mohring (1972) as in (4).

$$\mathrm{TC}'(N) = K - \frac{\beta \mathcal{B}}{N^2} = 0 \Rightarrow N^* = \sqrt{\frac{\beta \mathcal{B}}{K}} \quad (4)$$

If the crowding penalty is linear in seating occupancy ($\theta = 1$), the optimal frequency is given in (5) where $\mathcal{F}_{\theta=1} = \frac{2\alpha}{S} \sum_i t_i \sum_k F_{ik} \frac{dF_{ik}}{dN}$ is the in-vehicle crowding term. Note that in the absence of an analytic formulation of the passenger load F_{ik}, its variation can only be calculated using a numerical differentiation method.

$$\mathrm{TC}'(\mathrm{N}) = 0 \Rightarrow N^* = \sqrt{\frac{\beta \mathcal{B} + \gamma \mathcal{F}_{\theta=1}(N^*)}{K}} \quad (5)$$

The study of the analytic expression of the optimal frequency for general values of θ can yield very complex analytic formulations. In some cases (e.g., non-integer values or integer values $\theta \geq 4$), there is no possible closed form (Abel, 1824). Therefore, we explore the optimal frequency using a numerical approach, and also use it to study the effects of the valuation parameters on the optimal train frequency.

Table 1. Values that are used for the parameter valuation in producer and consumer costs (10 SEK ≈ 1 EUR).

Parameters	Values	References
Travel time	$\alpha_0 = 65.5$ SEK/h	Weighted travel time valuation (57 SEK/h) for leisure trips and (74 SEK/h) for commuting trips (Eliasson and Börjesson, 2014)
Waiting time	$\beta_0 = 80$ SEK/h	Valuation for average waiting times less than 10 min (i.e., headway ≤ 20 min), average between (86 SEK/h) leisure and (74 SEK/h) commuting (Algers et al., 2010).
Crowding	$\gamma_0 = 0.085$ $\theta_0 = 3$	Using results by Björklund and Swärdh (2017) from stated preference crowding valuation study in Stockholm, a curve fitting of the function $1 + \gamma \left(\frac{F_{ik}}{S}\right)^{\theta}$ was performed to find γ and θ.
Operation	$K_{distance} = 30$ SEK/wagon-km $K_{time} = 5\ 205$ SEK/wagon-h $K_{overhead} = 9\%$	All parameter values for the operating costs are from the Stockholm transit agency (SLL, 2017b). Vehicle costs are 5 000 000 SEK per wagon-year; we calculate wagon costs per hour assuming each wagon is operated 6 hours per day and 260 days per year.

The valuations in the CBA framework are meant to reflect passenger preferences. The relevant values for our case study have been estimated in previous studies, and are presented in **Table 1** together with source references.

Table 1 presents the adopted values (column 2) for the different cost parameters (column 1) as well as the references to the original studies

(column 3). The travel parameters $(\alpha, \beta, \gamma, \theta)$ are all indexed by 0 to distinguish them from the parameter estimates obtained from the timetable analysis presented below. The seating capacity of a commuter train (with two coupled trainsets) is $S = 748$ seats in total (ALSTOM, 2004). Most of the trains during the studied time intervals are operated with two wagons even during midday (off-peak with lower passenger loads) due to other considerations such as infrastructure restrictions and punctuality. For comparison, we also analyze optimal off-peak frequency assuming operations with short trains (single trainset, $S = 374$). Valuations estimated in different years are scaled to a common price level.

In order to estimate the agency's implicit valuation, we assume that it strives to minimize total social costs, i.e., service frequencies N_r for each line and time period combination r should fulfill $\mathrm{TC}'(N) = 0$. Based on this assumption, we can estimate the agency's implicit valuations such that the observed service frequencies N_r are indeed the optimal choices, or as close to optimal as possible. In our case study, we have frequencies for two lines and three time periods (we assume that frequencies have to be the same in each direction of a line), but only three valuation parameters γ, θ and β (the baseline value of travel time α cannot be identified separately, since it is confounded with γ). We hence estimate the valuation parameters by minimizing the squared deviations from the optimality conditions $\mathrm{TC}'(N) = 0$, summed over line/time period combinations r as in (6) where $(+)$ and $(-)$ denote the two directions of each line.

$$\begin{aligned}\{\gamma, \theta, \beta\} &= \underset{\gamma,\theta,\beta}{\operatorname{argmin}} \sum_r \left(\frac{\partial \mathrm{TC}(\gamma, \theta, \beta)}{\partial N_r}\right)^2 \\ &= \underset{\gamma,\theta,\beta}{\operatorname{argmin}} \sum_r \left(2K_r + \gamma\left(\frac{d\mathcal{F}_r^{(+)}}{dN}(\theta; N_r) + \frac{d\mathcal{F}_r^{(-)}}{dN}(\theta; N_r)\right) - \beta\left(\frac{\mathcal{B}_r^{(+)}}{N_r^2} + \frac{\mathcal{B}_r^{(-)}}{N_r^2}\right)\right)^2\end{aligned} \tag{6}$$

The term K_r is the marginal operating cost, $\mathcal{F}_r^{(.)} = \frac{\alpha}{S^\theta} \sum_{i,k} t_i \left(F_{ik,r}^{(.)}\right)^{\theta+1}$ denotes the general crowding term and $\mathcal{B}_r$ is the total number of passengers boarding on line and time period combination r.

An analytic estimate of the effect of heterogeneous demand and crowding variation is helpful to understand the consequences of our assumption to base our social cost benefit analysis on the average demand and that passengers are spread evenly in the train. This analytic estimate can be obtained as follows. We consider crowding variation as an example: day-to-day demand variation is analogous. Let us focus on a train service k and station i, and let $F = F_{ik}$ be the number of passengers on the train.

Normalize the train length to 1, so there are on average F passengers per distance unit in the train. Passengers are spread out unevenly: there are more passenger in the ends of the train than in the middle. Assume that in the ends of the train there are $(1+\delta)F$ passengers per distance unit, and $(1-\delta)F$ in the middle. Hence, the average passenger-to-seat ratio is $\frac{F}{S}$, see **Figure 1** for an illustration. In other words, the passenger density f at a distance from the train's end x in the interval $x \in [0,0.5]$ is $f(x) = F(1+\delta-4\delta x)$.

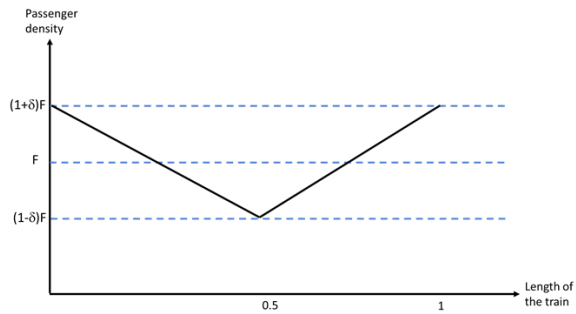


Figure 1. Model for uneven in-vehicle crowding and spreading of passengers in the train.

Given passenger density f, assume that the total social cost of in-vehicle time per train hour is $\alpha\left(1+\gamma\left(\frac{f}{S}\right)^{\theta}\right)f$, i.e., last term in equation (1). Since the passenger density f is not constant, we integrate this cost function from one train end to the other to get the total social cost of in-vehicle time per train hour. Straightforward integration yields the total in-vehicle time cost per hour given in (7).

$$\alpha\left(1+\gamma\left(\frac{F}{S}\right)^{\theta}\frac{(1+\delta)^{\theta+2}-(1-\delta)^{\theta+2}}{2\delta(\theta+2)}\right)F \qquad (7)$$

Note that when δ tends to zero and using a Taylor series, the cost in (7) will tend to the crowding penalty function used in equation (1). If $\delta = 0.5$, which is a realistic value for crowding variation along a train, and with $\theta = 3$, the crowding penalty factor $\gamma\left(\frac{F}{S}\right)^{\theta}$ increases by a factor of 3. As an illustration, this means that if the passenger-to-seat ratio is $\frac{F}{S} = 1.5$ and $\gamma = 0.085$, the value of in-vehicle time increases by around 90%, compared to the around 30% it would have increased if passengers had been

evenly spread across the train. Further sensitivity analyses on this can be found later in the case study.

4. Data

In this section, we present the input data for the numerical analysis performed on a commuter train line in Stockholm. **Figure 2** presents the commuter network (as of 2015). We will first concentrate on one line and direction, i.e., the J35 line in **Figure 2** filled in black from Kungsängen (Kän) to Västerhaninge (Vhe). We then present summary results for the other main lines and directions (i.e., between Upplands Väsby and Tumba). The J35 line includes 17 stations (from a network total of around 50) with Stockholm central station as the largest passenger station. Part of the studied line (i.e., between Karlberg and Älvsjö) are shared with other lines.

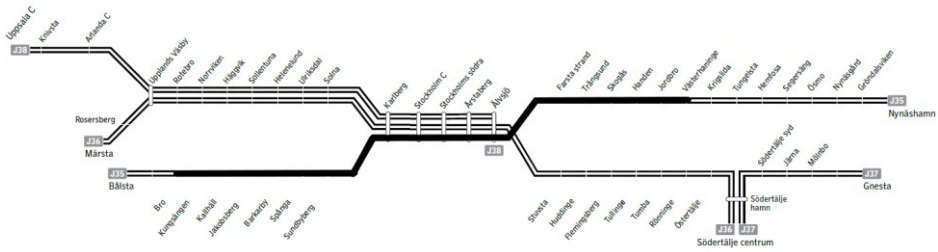


Figure 2. Studied (filled in black) line of the commuter train network in Stockholm, adapted from (SLL, 2015).

For each pair of stations, we know the number of trips for every 15 minutes over a normal day in September 2015, i.e., the time-dependent OD matrix. This matrix is estimated from smart card data (Ait-Ali and Eliasson, 2019b). It also includes passengers transferring to and from other lines. Some services start or continue outside the studied line, but we assume that those passenger flows start (terminate) at the first (last) station.

Figure 3 illustrates the number of passengers entering each station per 15 minutes interval over the day. Given the OD-matrix and the train timetable, the number of passengers boarding B_{ik} and alighting A_{ik} each train service k and station i are calculated. Travel times between stations t_i are known and constant for all trains. We study three main time intervals: morning peak (6:00 – 9:00), afternoon peak (15:00 – 18:00) and midday off-peak (10:00 – 13:00).

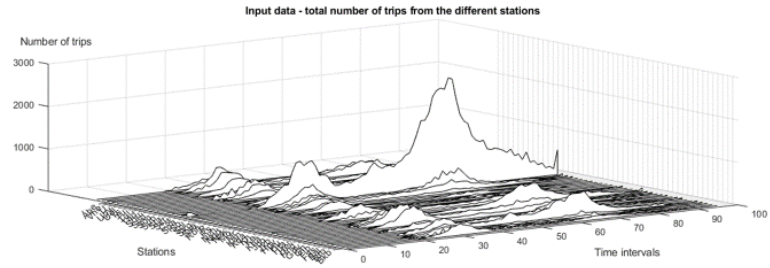


Figure 3. Number of passengers entering each station per 15min time interval over the day.

SL, the public transport agency in Stockholm, adopted in 2015 the commuter train timetable summarized in **Table 2**. The table shows service frequencies from Kän to Vhe, including extra departures on parts of the line during peak hours. There is a regular service frequency during all the studied time intervals, see column 2. However, peak hours include additional (or extra) train departures which are not all regular, see column 3. Those extra departures do not operate the whole line but for the sake of simplicity, we assume they do which leads to the total frequency presented in column 4. Thus, there are 7 departures in total during the morning peak (i.e., train every 8.6 min), 6 departures in the afternoon (i.e., train every 10 min) and 4 departures during midday (i.e., train every 15 min).

Table 2. SL's service frequency from Kän to Vhe for different time intervals during a working day (winter 2015).

Time interval	No. regular departures (per hour)	No. extra departures (per hour)	Total frequency (per hour)
Morning peak (6:00 - 9:00)	4[7]	3[8]	7.0
Midday off-peak (10:00 - 13:00)	4	0	4.0
Afternoon peak (15:00 - 18:00)	4	2	6.0

Using the total frequency and trip distribution (A_{ik} and B_{ik}), it is possible to estimate the total number of passengers onboard (F_{ik}) each train from Kän to Vhe for the studied time intervals as illustrated in **Figure 4**. The horizontal dashed line shows the total number of seats per train, i.e., $S = 748$ (ALSTOM, 2004).

[7] Certain trains are running parts or beyond the studied line, e.g., to Älvsjö or Nynäshamn, from Jakobsberg.

[8] The provided frequency for extra departures is an average since not all are regularly running every X minute.

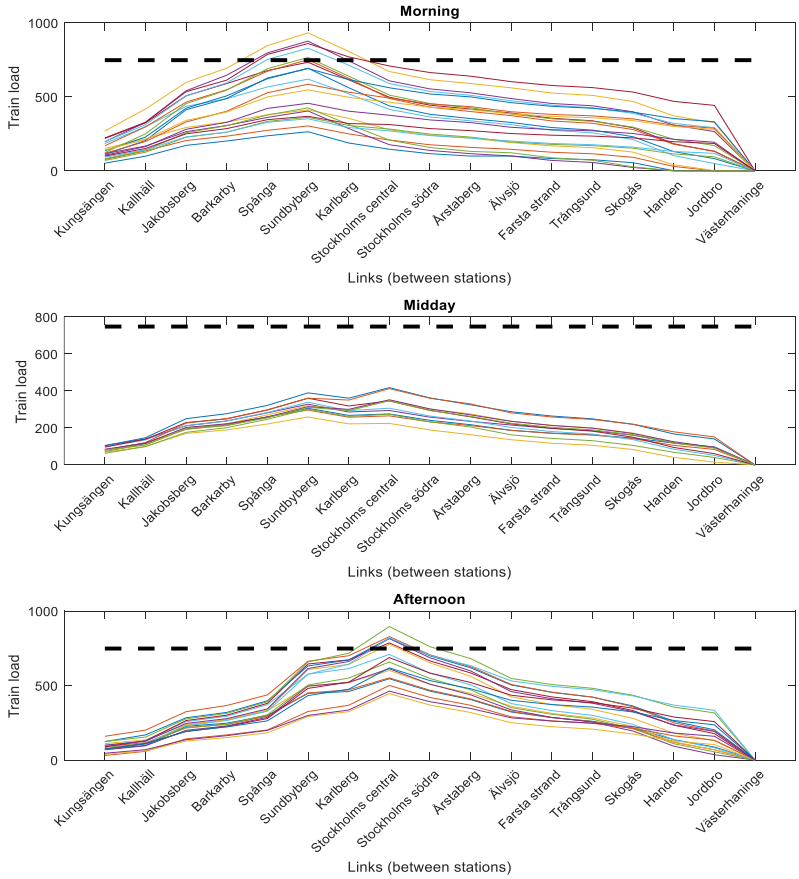


Figure 4. Estimated passenger load per train from Kän to Vhe during different time intervals. Each color refers to a train service. Dashed horizontal line shows the total seating capacity of the train.

5. Results

Given the input data and the analytic model presented above, we can study the total social cost as a function of service frequency (i.e., trains per hour) for different time intervals: morning, midday and afternoon. The results are presented in **Figure 5** where the total costs (in SEK) are presented in the y-axis as a function of the service frequency in the x-axis. Each color refers to one of the studied time intervals. Note that for the midday time interval, both long and short trains are presented with the corresponding total costs with a scale on the right y-axis.

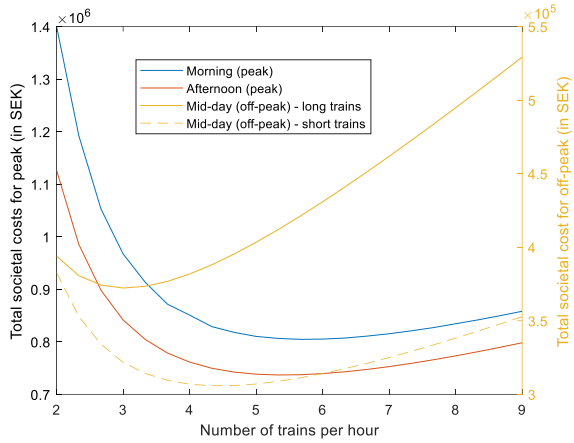


Figure 5. Total social costs as a function of the frequency for different time intervals.

The total social cost has a minimum which corresponds to the optimal frequency, given the parameters assumed above. A higher frequency than the optimum leads to operating costs larger than passenger benefits, whereas a lower frequency than the optimum leads to passenger costs increasing faster than savings in operating costs. The numerical values for the optimal frequencies in the different time intervals are reported in column 2 of **Table 3**, and are compared to the public transport agency's actual frequencies.

Table 3. Optimal service frequencies (given valuation parameters from **Table 1**), compared to SL's actual frequency.

Time interval	Optimal frequency (trains/h)	SL's frequency (trains/h)
Morning	5.7	7.0
Midday (long)	3.0	4.0
Midday (short)	4.3	-
Afternoon	5.3	6.0

SL's actual frequencies are slightly higher than the optimal ones, given the valuation parameters in **Table 1**. Based on these parameters, the public transport agency should reduce the service frequency on this line and direction. But this result can also be interpreted as the agency putting a higher implicit valuation on either waiting time or crowding (or both) than the valuations in **Table 1**. We thus turn to the question of identifying the agency's implicit valuations, as implied by the timetable choice (or the frequency).

For the sake of comparison, we also study the optimal frequency on the opposite direction of the studied line (i.e., from Vhe to Kän). In addition,

we also look at the second main line between Upplands Väsby (Upv) and Tumba (Tu) in both directions (i.e., southwards to Tu and northwards to Upv), see **Figure 2**. The results are presented in **Table 4**.

Table 4. Optimal frequencies (in trains/h) on the studied main lines and directions, compared to SL's actual frequency.

Time interval	SL	Kän → Vhe (Southwards)	Vhe → Kän (Northwards)	Upv → Tu (Southwards)	Tu → Upv (Northwards)
Morning	7.0	5.7	5.7	6.3	7.3
Midday (long)	4.0	3.0	3.3	3.3	3.3
Midday (short)	-	4.3	4.7	5.3	5.3
Afternoon	6.0	5.3	5.7	6.7	6.3

Just as for the Kän-Vhe line studied above, optimal frequencies on the other lines and directions are generally slightly lower than the actual frequencies chosen by SL. The exceptions are in the afternoon peak hour on the line between Upv and Tu in both directions as well as in the morning peak hour from Tu on the same line. These are the only cases where SL is running fewer trains than optimum. Such higher optimal frequencies on these lines and directions is mainly due to the higher ridership which requires running more trains to reduce waiting time and crowding costs.

Note that using short trains during off peak hours always leads to a higher optimal frequency than SL's, especially on the Upv-Tu line. The latter has a generally higher optimal frequency which is mostly due to the cheaper production costs (for short trains) justifying the increase in service frequency.

From an operational point of view (e.g., efficient rolling stock circulation), the frequency for the two directions on the same line need to be similar. This constraint can be satisfied in different possible ways. For instance, the line frequency could be set as the average or maximum of the two optima. Another alternative is to modify the model in order to include all the line (i.e., both directions) with a single variable (line frequency).

In what follows, we focus on the Kän-Vhe line to illustrate the implicit valuation of waiting time and crowding. The same method can then be used to study other directions and lines.

5.1. Valuation of waiting time

The higher the waiting time valuation β is, the higher the optimal service frequency will be. The fact that the actual frequency is higher than the optimum (using the parameter β_0 from **Table 1**) suggests that SL's

implicit valuation is higher than β_0. **Figure 6** shows the variation of the optimal frequency when varying the waiting time valuation β. The horizontal dashed lines show SL's frequencies for morning, midday and afternoon whereas the vertical dashed line is the baseline valuation β_0.

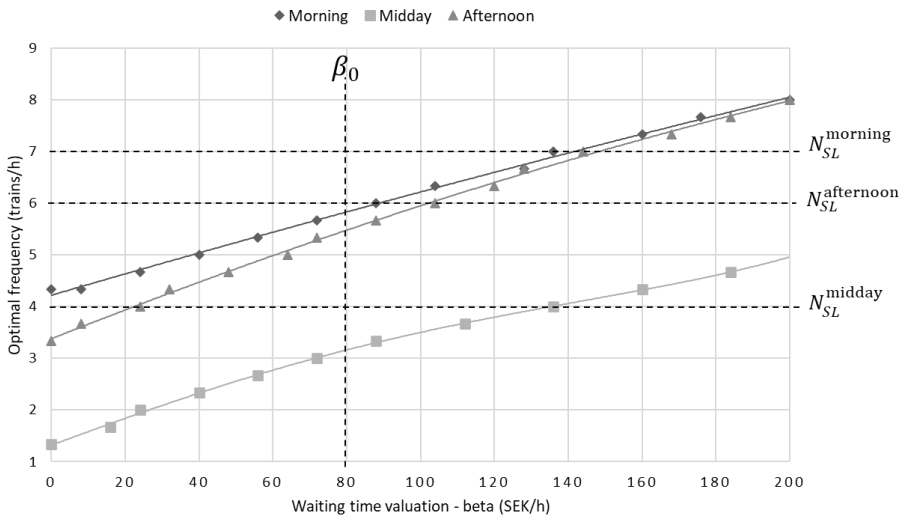


Figure 6. The optimal frequency as a function of the waiting time valuation for different time intervals.

As expected, we find in **Figure 6** that a higher waiting time valuation leads to a higher optimal frequency. If the actual frequencies for the different time intervals are projected on the x-axis for each curve, the corresponding x-axis projection values can be used for the implicit valuation of waiting time (if other parameters are kept fixed). **Table 5** presents the numerical values for the agency's implicit valuations of waiting time for different time periods.

Table 5. Implicit waiting time valuations (in SEK/h) for train services from Kän to Vhe, other valuations fixed, $\beta_0 = 80$ SEK/h.

Time interval	Implicit waiting time valuation (in SEK/h)	Deviation from β_0
Morning	144	+80%
Midday (long)	144	+80%
Afternoon	156	+95%

The agency's implicit valuations in **Table 5** are almost twice as high as the valuation β_0 according to the studies that the CBA guidelines are based on, see **Table 1**. Note, though, that these are the implicit valuations obtained if only the waiting time valuation is changed, while the

crowding valuation is kept constant. Next, we study the crowding valuation parameters.

5.2. Valuation of in-vehicle crowding

The crowding valuation depends on two variables, the factor γ and the exponent θ. **Figure 7** presents optimal frequencies as functions of γ (left) and θ (right), *ceteris paribus*. The meaning of the horizontal and vertical dashed lines is similar to that in **Figure 6**.

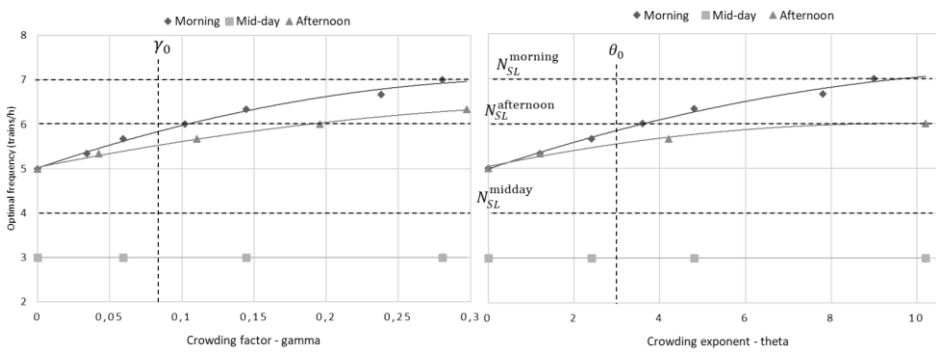


Figure 7. Optimal train frequency as a function of the crowding parameter (left – factor, right - exponent).

As expected, higher crowding valuations generally yield higher optimal frequencies. For midday (with long trains), there is no effect since crowding is negligible. The numerical results showing the implicit valuation of crowding are given in **Table 6** for peak hours. Off peak hours are not shown due to the negligible crowding levels during this time interval.

Table 6. Implicit valuation of crowding during peak hours for train services from Kän to Vhe, ceteris paribus, $\theta_0 = 3$ and $\gamma_0 = 0.085$.

Time interval	Implicit crowding factor	Implicit crowding exponent
Morning	0.281	9
Afternoon	0.196	10

The implicit in-vehicle crowding valuation is more than twice as high as the one based on passenger valuations in **Table 1**. Note, though, that our calculation is based on average seating occupancy, i.e., that passengers are spread evenly across the train. In fact, there are usually more passengers towards the ends of the train due to the layout of the stations (Peftitsi et al., 2020). Since the crowding valuation is a nonlinear function of the seating occupancy, heterogenous occupancy along the train will increase the total crowding penalty, even more so if one also considers that more

passengers will by definition experience high crowding than will experience low crowding (since there are by definition more passengers where there is higher crowding). The losses for high-crowded parts of the train will hence outweigh the benefits of the low-crowded parts. Below, we perform a sensitivity analysis to study the effect of taking heterogeneous occupancy into account.

Moreover, demand varies across days, and since the crowding valuation is nonlinear, the higher crowding penalties during high-demand days will outweigh the corresponding lower crowding penalties during low-demand days. Another sensitivity analysis is also presented later, it looks at the effect of varying the demand (i.e., OD matrix) on the optimal frequency.

5.3. Joint valuations

In order to see the joint effect of varying crowding and waiting valuations, we study the variation of the optimal frequency when varying these two parameters jointly. **Figure 8** plots the results for (from left to right) morning and afternoon peak hours. Each line shows combinations of valuations (crowding γ and waiting time β), for which the corresponding frequency is optimal. Remember that the actual frequencies are 7 trains/hour in the morning peak and 6 trains/hour in the afternoon peak.

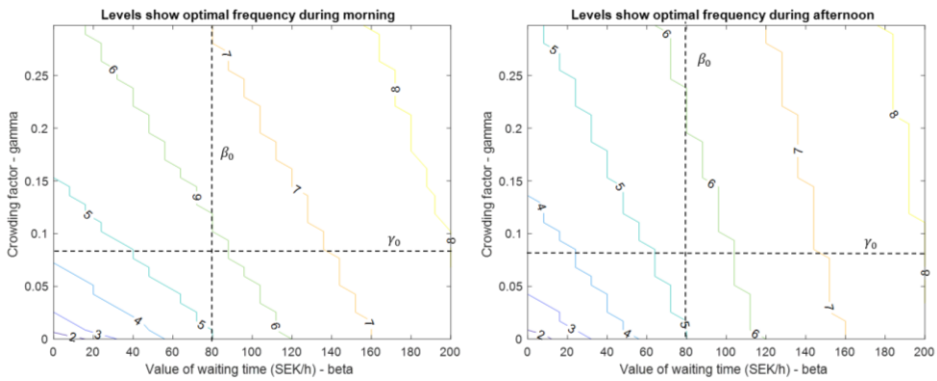


Figure 8. Contour lines showing the optimal frequency when varying both crowding and waiting time valuations during morning and afternoon time intervals.

Figure 8 shows that the sensitivity of the optimal frequency to the waiting time is much higher than to the crowding factor, i.e., level curves or contours are almost vertical, especially in the afternoon. This is explained once again by the fairly low crowding in the trains.

5.4. The agency's average implicit valuations

In order to estimate SL's implicit valuation averaged over all the lines and time periods, we use equation (6) that was previously presented. An equivalent formulation is given in (8) where r is over 6 different combinations, i.e., two lines (Kän-Vhe and Upv-Tu) and three time periods (morning, midday and afternoon).

$$\underset{\gamma,\theta,\beta}{\operatorname{argmin}} \sum_{r} \left(2K + \gamma \frac{\alpha}{S^{\theta}} \left(\sum_{i} t_i \frac{d}{dN} \left(\sum_{k} \left(F_{ik,r}^{(+)} \right)^{\theta+1} + \sum_{k} \left(F_{ik,r}^{(-)} \right)^{\theta+1} \right) \right) - \frac{\beta}{N_r^2} \sum_{ik} \left(B_{ik,r}^{(+)} + B_{ik,r}^{(-)} \right) \right)^2 \quad (8)$$

The derivatives over N of the parameterized function of the train loads $F_{ik,r}$ can be calculated using a numerical differentiation method; we use the central difference method. **Table 7** compares the implicit and baseline valuations of waiting time.

Table 7. Comparison between SL's estimated implicit valuations of waiting time and baseline CBA guidelines.

Valuations	Waiting time	Crowding factor	Crowding exponent
CBA guidelines	$\beta_0 = 80$ SEK/h	$\gamma_0 = 0.085$	$\theta_0 = 3$
SL's implicit (fixed θ, γ)	$\beta_{SL} = 144$ SEK/h	$\gamma_{SL} = \gamma_0$	$\theta_{SL} = \theta_0$
Deviation	+80%	fixed	fixed-

Table 7 indicates that SL's implicit valuation for waiting time given their choice of frequency (over all the lines and the time intervals) is different than the baseline passengers' stated valuation from the Swedish CBA guidelines. For instance, SL's implicit valuation of waiting time is around twice as high as the values from the guidelines. When performing the estimation for crowding valuation, we find implicit valuations that are substantially different from the baseline values. These differences are due to the low crowding in the case study.

The results for the study of all lines and time intervals as well as those from specific lines and directions are based on several assumptions. First, the OD demand is assumed to be fixed. Second, the in-vehicle crowding is assumed to be equally distributed between train wagons. These two assumptions are studied in the following sensitivity analysis.

5.5. Heterogeneous crowding and demand

The results presented above are based on average passenger demand and average crowding. However, this demand varies randomly between days, and crowding vary across the train. Since the crowding penalty function is nonlinear, these variations will not necessarily average out in the total

social cost. To study the effect of these variations, we perform sensitivity analyses.

Based on the already used OD matrix, we analyze the optimal service frequencies using varying passenger demand, i.e., 11 variants of OD matrix from −50% to +50% of the average demand for passengers boarding and alighting. The results are presented in the bar chart in **Figure 9** for commuting services from Kän to Vhe. For each time interval, the bar to the right shows the optimal frequency for the variable demand where the total social cost function is taken as the average of the costs for each OD matrix variant. The figure also shows the previously reported optimal frequencies, i.e., baseline or SL (bars to the left) and fixed (bars in the middle).

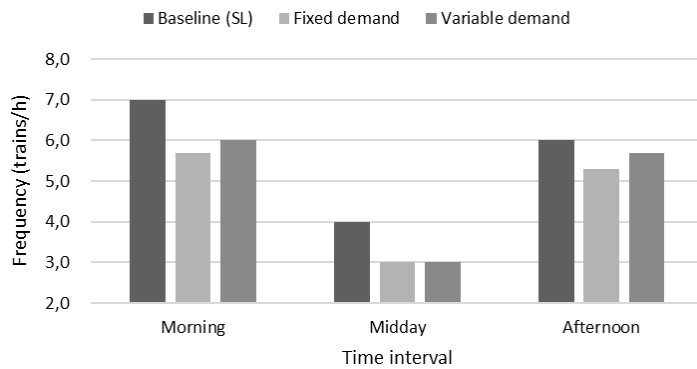


Figure 9. Optimal frequency from Kän to Vhe when varying the OD-matrix for different time intervals, compared to SL's chosen frequency and the previously presented optimum for fixed passenger demand.

The optimal frequency increases when studying variable passenger demand. The increase is almost similar for the different time intervals except midday where there is no change in optimal frequency. The magnitude of the increase in optimal frequency for varying demand is relatively small if compared to that of the fixed demand, e.g., up to 8% increase in the afternoon. This can be due to the low crowding levels, meaning that waiting times dominate the passenger costs. Moreover, this increase in optimal frequency (due to variable demand) is still below SL's frequency. Therefore, demand variation (alone) do not explain SL's choice and other factors should also be considered.

Previously, we also considered that passengers are assumed to enter train wagons in a uniform way. However, passengers tend to choose wagons that are near the entrance of the train station (for boarding) or exit (for alighting) which leads to increased crowding in these wagons and

decreased in others (Fang et al., 2019). Since the crowding penalty function is nonlinear, these variations do not necessarily cancel each other out. One way to study this difference in crowding levels between wagons is to vary the available seating capacity compared to the total number of seats in the trains. For instance, decreasing seating capacity in the analysis could reflect the unbalanced loads between train wagons.

We perform a sensitivity analysis on the available seating capacity (i.e., total number of seats in the trains) in a way that is similar to the analysis which was previously done for passenger demand, i.e., 11 variants with vehicle capacity ranging from -50% to +50% of a standard long train with two coupled trainsets. Similar to **Figure 9**, **Figure 10** shows the results of the analysis for seat supply which also indicate that the magnitude of the increase in optimal frequency (for variable seat supply) is small and still below SL's choice of frequency.

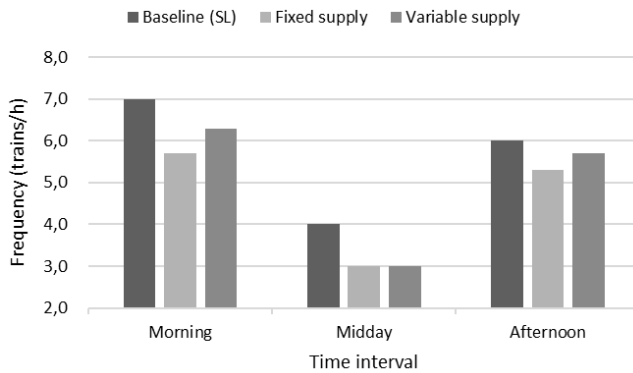


Figure 10. Optimal frequency from Kän to Vhe when changing the seating capacity for different time intervals, compared to SL's chosen frequency and the previously presented optimum for a standard long train.

The optimal frequency is not sensitive enough to variable demand or seat supply in order to justify SL's choice of frequency. **Figure 9** indicates at most 8% increase in optimal frequency when varying demand and at most 11% in **Figure 10** for variable seat capacity. No impact is found during midday off-peak hours. Thus, the two sensitivity analyses reveal that variations in day-to-day ridership (demand) and disparities in crowding between wagons (seating capacity) both lead to an increase in the optimal frequency. However, they are relatively small and (only) partially explain SL's higher frequency and valuations of waiting times (and crowding).

In order to reduce total costs, it is important to smoothen the demand. Moreover, seat supply should be fully exploited and unbalances between

train wagons should be avoided in order to further decrease costs. Traffic information on the platforms can be used to guide passengers waiting on platforms to the least crowded wagons.

5.6. Summary and discussions

Our case study shows that the frequencies are not quite consistent with a CBA based on passenger valuations. Actual frequencies are generally higher than what is optimal according to the CBA framework. Estimating the agency's implicit valuations of waiting time and crowding leads to higher values than the ones based on passenger stated preference studies.

The results of estimating implicit valuations (on separate lines, directions and time intervals) indicate waiting time valuations around twice as high as in the passenger valuation studies. Analyzing waiting time and crowding jointly shows that the former influences the optimal frequency more than the latter, due to the low crowding in the case study.

However, these results are based on the assumption that trains actually run according to the schedule, i.e., no delays or cancellations. Delays and other sources of variability lead to heterogeneity in crowding across services, days and wagons. Since the crowding penalty is nonlinear, these variations do not cancel out on average, but tend to increase the optimal frequency. A sensitivity analysis indicates that variations in OD demand and seating capacity have an impact on the optimal frequency but are not the main determinants. Hence, they (only) partially give an answer to why the PTA's frequencies are not consistent with passengers' stated preferences.

6. Conclusions and Future Works

This study attempts to check if commuter train timetables are consistent with passengers' stated valuations of trip parameters such as waiting times and in-vehicle crowding. Using Stockholm's commuter services as a case study, the results of a CBA calculation of the optimal frequencies indicate that the PTA's timetables are not quite consistent with passengers' valuations. In order to explain these inconsistencies, this work estimates the PTA's implicit valuation of waiting time and crowding.

From an analytic CBA model, we studied numerically the socially optimal frequency on one of the highly frequented lines in Stockholm commuter train system 2015. The results show that the PTA (i.e., SL) generally adopts a slightly higher frequency than the optimal one on most of the commuter lines and directions. Conversely, estimating the corresponding implicit valuations indicate that SL has higher implicit valuations

than the ones based on passenger studies, for instance, 95% higher for waiting times on the studied line and direction (and 80% for all lines). Apart from the studied valuations of waiting time and crowding, there might be some other more subtle factors (e.g., political lobbying, labor laws, infrastructure limitations, etc.) which can also lead to nonoptimal frequencies (or inconsistent timetables with passengers' valuations).

Such results have several policy implications on different stakeholders, e.g., PTA, infrastructure manager and passengers. On the one hand, the PTA should justify the choice of the timetables (or service frequencies), and should explain any inconsistencies with passengers' preferences or from optimum (suggested by CBA models). On the other hand, if passengers' preferences (e.g., waiting time valuation) are one of the main driving powers in the decision process, their participation in stated or revealed preference studies is important for more consistent and better timetables by the PTA. Moreover, when using CBA models to solve train capacity conflicts (Ait-Ali et al., 2020), infrastructure managers often assume that the actual timetable for commuter trains is optimal (i.e., consistent with passengers' valuations). This study shows that such assumptions may not always hold, e.g., removing a certain commuter train path (or reduce service frequencies) may lead to timetables yielding higher social benefits, according to the CBA framework. This assumption is also important when the infrastructure manager or the government is considering new infrastructure investments where actual existing capacity is not efficiently used.

The model in this study is not only applicable to commuter train services, it can also be used to study other transport modes (e.g., bus and metro) or other systems such as public utilities. Another possible use of the model is in train timetabling for local commuter services. Combined with an optimization model, it allows to study different timetabling strategies such as skip-stop, periodic or cyclic timetables. Line planning may also be studied where central parts of the network have more frequent trains as opposed to outer parts, and passengers have to make transfers to travel between the two parts. Further extensions to the study are also possible, such as additional performance measures for punctuality and interchanges.

Acknowledgment

This research is part of the SamEff project about socio-economically efficient allocation of railway capacity (*samhällsekonomiskt effektiv tilldelning av kapacitet på järnvägar)*. The project is funded by a grant from the Swedish Transport Administration (*Trafikverket*). The authors

are grateful to Roger Pyddoke and John Nellthorp for the valuable discussions and comments.

References

ABEL, N. H. 1824. Mémoire sur les équations algébrique: où on démontre l'impossiblité de la résolution de l'equation générale du cinquième dégré. Faculty of Science - University of Oslo.

ABRANTES, P. A. L. & WARDMAN, M. R. 2011. Meta-analysis of UK values of travel time: An update. Transportation Research Part A: Policy and Practice, 45, 1-17.

AIT-ALI, A. & ELIASSON, J. 2019. Dynamic Origin-Destination Estimation Using Smart Card Data: An Entropy Maximisation Approach. arXiv preprint arXiv:1909.02826.

AIT-ALI, A., WARG, J. & ELIASSON, J. 2020. Pricing commercial train path requests based on societal costs. Transportation Research Part A: Policy and Practice, 132, 452-464.

ALGERS, S., BÖRJESSON, M., SUNDBERGH, P., BYSTRÖM, C. & ALMSTRÖM, P. 2010. Valuation of Time in Transport – The National Studies 2007/08 in Sweden. WSP report.

ALSTOM 2004. CORDIA 60X Stockholm Transport Renews its Commuter Fleet.

ARROW, K. J. 1963. Social choice and individual values.

ASPLUND, D. & PYDDOKE, R. 2019. Optimal fares and frequencies for bus services in a small city. Research in Transportation Economics, 100796.

BASU, K. 1980. Revealed Preference of Government, Cambridge University Press.

BASU, K. 1984. Fuzzy revealed preference theory. Journal of Economic Theory, 32, 212-227.

BJÖRKLUND, G. & SWÄRDH, J.-E. 2017. Estimating policy values for in-vehicle comfort and crowding reduction in local public transport. Transportation Research Part A: Policy and Practice, 106, 453-472.

BOSSERT, W. & WEYMARK, J. A. 2004. Utility in Social Choice. In: BARBERÀ, S., HAMMOND, P. J. & SEIDL, C. (eds.) Handbook of Utility Theory: Volume 2 Extensions. Boston, MA: Springer US.

BRENT, R. J. 1991. On the estimation technique to reveal government distributional weights. Applied Economics, 23, 985-992.

BÖRJESSON, M. & ELIASSON, J. 2012. The value of time and external benefits in bicycle appraisal. Transportation Research Part A: Policy and Practice, 46, 673-683.

BÖRJESSON, M., FUNG, C. M. & PROOST, S. 2017. Optimal prices and frequencies for buses in Stockholm. Economics of Transportation, 9, 20-36.

ELIASSON, J. & BÖRJESSON, M. 2014. On timetable assumptions in railway investment appraisal. Transport Policy, 36, 118-126.

ELIASSON, J. & LUNDBERG, M. 2012. Do Cost–Benefit Analyses Influence Transport Investment Decisions? Experiences from the Swedish Transport Investment Plan 2010–21. Transport Reviews, 32, 29-48.

FANG, J., FUJIYAMA, T. & WONG, H. 2019. Modelling passenger distribution on metro platforms based on passengers' choices for boarding cars. Transportation Planning and Technology, 42, 442-458.

FOSGERAU, M. 2009. The marginal social cost of headway for a scheduled service. Transportation Research. Part B: Methodological, 43, 813-820.

FRANKLIN, A. L. & CARBERRY-GEORGE, B. 1999. Analyzing How Local Governments Establish Service Priorities. Public Budgeting & Finance, 19, 31-46.

IM, T., LEE, H., CHO, W. & CAMPBELL, J. W. 2014. Citizen Preference and Resource Allocation: The Case for Participatory Budgeting in Seoul. Local Government Studies, 40, 102-120.
JOHNSON, D. & NASH, C. 2008. Charging for scarce rail capacity in Britain: a case study. Review of Network Economics, 7.
LEWINSOHN-ZAMIR, D. 1998. Consumer Preferences, Citizen Preferences, and the Provision of Public Goods. The Yale Law Journal, 108, 377-406.
MANAF, H. A., MOHAMED, A. M. & LAWTON, A. 2016. Assessing Public Participation Initiatives in Local Government Decision-Making in Malaysia. International Journal of Public Administration, 39, 812-820.
MCFADDEN, D. 1975. The Revealed Preferences of a Government Bureaucracy: Theory. The Bell Journal of Economics, 6, 401-416.
MCFADDEN, D. 1976. The Revealed Preferences of a Government Bureaucracy: Empirical Evidence. The Bell Journal of Economics, 7, 55-72.
MICHELS, A. & DE GRAAF, L. 2010. Examining Citizen Participation: Local Participatory Policy Making and Democracy. Local Government Studies, 36, 477-491.
MOHRING, H. 1972. Optimization and Scale Economies in Urban Bus Transportation. The American Economic Review, 62, 591-604.
NELLTHORP, J. & MACKIE, P. J. 2000. The UK Roads Review—a hedonic model of decision making. Transport Policy, 7, 127-138.
PEFTITSI, S., JENELIUS, E. & CATS, O. 2020. Determinants of passengers' metro car choice revealed through automated data sources: a Stockholm case study. Transportmetrica A: Transport Science, 16, 529-549.
QIN, F. & JIA, H. 2013. Modeling Optimal Fare and Service Provisions for a Crowded Rail Transit Line. Journal of Transportation Systems Engineering and Information Technology, 13, 69-80.
SCARBOROUGH, H. & BENNETT, J. 2012. Cost Benefit Analysis and Distributional Preferences, Edward Elgar Publishing, Incorporated.
SLL 2015. Fakta om SL och länet 2015.
SLL 2017. SAMS 3.0 Documentation In: ADMINISTRATION, T. (ed.) Strategic Development of Socio-economic Analysis. 2017-03-03 ed. Stockholm: Transport Administration.
TIRACHINI, A., SUN, L., ERATH, A. & CHAKIROV, A. 2016. Valuation of sitting and standing in metro trains using revealed preferences. Transport Policy, 47, 94-104.
TRAFIKVERKET 2016. English summary of ASEK recommendations.
WARDMAN, M. & WHELAN, G. 2011. Twenty Years of Rail Crowding Valuation Studies: Evidence and Lessons from British Experience. Transport Reviews, 31, 379-398.
WEISBROD, B. A. & CHASE, S. B. 1966. Problems in Public Expenditure Analysis, Washington, DC, The Brookings Institution.

Paper P4

Disaggregation in Bundle Methods: Application to the Train Timetabling Problem.

Ait-Ali, A.[1,2], Lindberg, P. O.[1], Eliasson, J.[2], Nilsson, J.[1] and Peterson, A.[2] (2020)
[1] VTI Swedish National Road and Transport Research Institute, Transport Economics (TEK), Stockholm
[2] Linköping University, Department of Science and Technology (KTS), Norrköping

Published in Journal of Rail Transport Planning & Management, 100200

https://doi.org/10.1016/j.jrtpm.2020.100200

Abstract

The train timetabling problem (TTP) consists of finding a feasible timetable for a number of trains which minimises some objective function, e.g., sum of running times or deviations from ideal departure times. One solution approach is to solve the dual problem of the TTP using so-called bundle methods. This paper presents a new bundle method that uses disaggregate data, as opposed to the standard bundle method which in a certain sense relies on aggregate data. We compare the disaggregate and aggregate methods on realistic train timetabling scenarios from the Iron Ore line in Northern Sweden. Numerical results indicate that the proposed disaggregate method reaches better solutions faster than the standard aggregate approach.

Keywords: Train timetabling; disaggregation; bundle methods; lagrangian relaxation; mathematical programming

1. Introduction

The train timetabling problem (TTP) refers to finding a feasible train timetable that minimises some objective functions. Such a timetable specifies where each train is located at given times over a certain period and is often presented as a graphical space-time diagram. That the timetable is feasible means that it should be free of conflicts between trains and satisfy certain functional constraints given by the railway system, such as the track capacity resulting from the physical infrastructure and the signalling system.

Train path requests (e.g., ideal departure time, latest arrival time and stopping stations) are received from the train operator(s). An infrastructure manager is tasked to produce a feasible train timetable that maximises a certain total objective function (e.g., total utility) based on the received train path requests. Due to network capacity restrictions, certain path requests are sometimes adjusted or rejected (i.e., not included in the final timetable).

TTPs are often formulated as mathematical programs, e.g., Integer Linear Programs (ILPs) or Mixed Integer Programs (MIPs). Solving such models, i.e., finding an optimal (or good quality) solution, is not always easy. Solution methods that make use of simplifications or heuristics are often needed to make the computational solution time realistic and tractable.

Relaxation methods, such as *lagrangian relaxation*, have been widely used as solution methods for solving the TTP models. In these solution methods, the (easier) dual program resulting from the relaxation becomes the focus rather than the (harder) TTP program (called *primal*). *Bundle methods* are often used to solve the dual programs resulting from lagrangian relaxation of TTP models. For instance, Brännlund et al. (1998) adopted a standard bundle method where aggregate information from all the train requests are used to solve the dual problem arising from lagrangian relaxation of a discrete-time and space formulation of TTP.

In the present paper, we propose an improved variant of bundle method using a disaggregate approach where the optimisation is performed with separate dual information for each train request. The aim of the paper is to derive the novel approach (called *disaggregate*) based on the same TTP model by Brännlund et al. (1998), and study its performances on some real-world timetabling scenarios. We show that the proposed

approach results in substantial reductions in computation times, up to 45% (excluding the initialisation phase), compared to the standard bundle method.

In the following section, several related works are reviewed and compared to this paper. The mathematical program of the TTP is formulated in section 3. We derive in section 4 the two solution methods, i.e., aggregate and disaggregate approach. In section 5, we test the two approaches and compare their performances on realistic train timetabling scenarios from the single-track Iron Ore railway line (*Malmbanan*) in Northern Sweden. Section 6 concludes the paper.

2. Related Work

The research literature related to the topic is rich. Many research papers focused on modelling TTP or on solving it, while others treated both. In this section, we present some of the related work from the research literature.

TTPs can be modelled using alternative basic formulations leading to various mathematical programs. The main variables, i.e., space and time, can be discretised and lead to ILPs (Yue et al., 2016) or continuous and lead to MILPs such as Bach et al. (2018) and Forsgren et al. (2013). Most TTP models are linear but they can also be nonlinear if the constraints or objective function include a nonlinear term (Xu et al., 2014). Some TTP models focused on single track lines (Brännlund et al., 1998) whereas others on more general railway networks (Meng and Zhou, 2014). In addition to standard off-line TTPs, certain models are also used for real time (online) timetabling under disturbances, i.e., operational planning such as the models by Törnquist and Persson (2007), Törnquist (2012) or Quaglietta et al. (2016). The final timetables can have a specific format, e.g., cyclic as in Zhang et al. (2019b) or periodic as in Jamili et al. (2012).

Different TTP models include various types of constraints and variables. Track occupancy constraints and blocking rules are particularly important for single track. Traditionally, these constraints are included in the ILP using big-M techniques. Alternatively, Meng and Zhou (2014) introduced cumulative flow variables. Instead of using standard space and time variables (i.e., departure or arrival time at specific stations as in time–space network modelling framework), Cacchiani et al. (2008) used variables where each variable corresponds to the timetable of a train, i.e., all the train path from origin to destination. Additional

constraints such as maintenance can also be included (Caprara et al., 2006, Forsgren et al., 2013, Zhang et al., 2019a).

The objective function in the TTP models is often based on an estimation of the value of train paths. Such value depends on several parameters such as total travel time, departure or arrival time. Brännlund et al. (1998) assumed a simple linear profit function which decreases when departing before or after the ideal departure time. In a model that combines train timetabling and stopping patterns, Yue et al. (2016) used a profit function that includes the number of stops, stopping time and the passenger demand, i.e., origin-destination (OD) matrix. On the importance of the choice of the objective function, Törnquist (2015) found that it has a significant impact on the computational time.

To choose between the alternative TTP models often means making trade-offs between the (dis-)advantages of each variant. Discrete formulations produce timetable information at certain points in space-time whereas continuous variants allow to produce more detailed timetables. However, both include integer variables (i.e., combinatorial) and thus difficult to solve for realistic train timetabling scenarios. Most TTP models attempt to find train timetables for single track lines which can be often also used for networks. Similarly, off-line TTPs can be used for real time (online) timetabling where the objective function is to reduce the disruptions but such online models have more requirements for the computational times. Another example is the choice of variables, Cacchiani et al. (2008) chose the whole train path as a variable and therefore reduces the model complexity (i.e., decreases the number of variables or unknowns) but the approach requires generating a set of good alternative train paths to choose from.

Lagrangian relaxation is commonly used as a solution methodology to solve different variants of TTP models. In an early study by Brännlund et al. (1998), the track capacity constraints are assigned prices, i.e., lagrangian multipliers. A dual iterative method is used together with a heuristic to find feasible timetables for small to medium-sized realistic scenarios. The authors show that lagrangian relaxation can be used to solve TTPs and indicate that bundle method (to solve the dual) generally performs better than alternative methods such as (modified) subgradient. In a related study, Caprara et al. (2002) introduced a graph theoretic model (i.e., multigraph) based on the ILP. Using lagrangian relaxation, the authors develop a heuristic that is based on a lagrangian profit function which relates train paths in the multigraph with their profits. Such profits are used to rank alternative paths that are included in the final

timetable solution. In a follow up study, Caprara et al. (2006) presented a basic discrete ILP model for TTP and the corresponding graph representation. The authors applied the lagrangian profit together with a subgradient iterative heuristic. In addition to the basic model, they test their methods on extended TTP models including manual block signalling (as opposed to automatic in the basic), station capacity, prescribed timetable (for a subset of trains) and maintenance.

Several more recent studies also make use of lagrangian relaxation to solve TTP models. For instance, Meng and Zhou (2014) developed a TTP model, on an N-track network, by simultaneously rerouting and rescheduling trains using a time–space network model. They show that their new approach provides more efficient solutions when the capacity constraints are dualized using lagrangian relaxation. Another study by Yue et al. (2016) developed a new TTP model that considers the stopping patterns (passenger service demand) and train timetable at the same time. The authors use lagrangian relaxation to formulate a simpler linear program which is solved using a column-generation-based heuristic. The final solution is found by iteratively updating the restricted master problem and the sub-problems. The authors show an improvement in the profit function and capacity utilisation with their algorithm. In a recent study, Zhang et al. (2019b) developed a new ILP by extending the time–space network and periodic event scheduling problem (PESP). They transform the PESP into multi-commodity network flow model with two coupled schedules and capacity constraints. These constraints are dualized using lagrangian relaxation and Alternating Direction Method of Multipliers (ADMM). For each train request, the cheapest master schedule is found in the time-space network. An iterative primal dual framework allows to find the optimal solution.

Other alternative solution methodologies that are also applied to solve TTPs include Linear Programming (LP) relaxation (mostly applied to MILP models) by Cacchiani et al. (2008), simulated annealing and particle swarm by Jamili et al. (2012) or genetic algorithms by Xu et al. (2014).

The current paper focuses on solving the basic single track TTP, similar to the early formulation by Brännlund et al. (1998), and later by Caprara et al. (2002) using a multigraph model. Both studies include capacity constraints using binary variables for block or arc occupation which allows to model their occupation by at most one train. Such constraints link the different trains and tracks and increase exponentially with the number of train requests and multi-tracks (e.g., at stations). An

alternative model that limits this increase in complexity is proposed by Meng and Zhou (2014). They introduced a cumulative flow for modelling temporal and spatial occupancy as well as safety headways suitable for multi-tracks. Hence, the authors use integer variables that captures the sum of capacity consumption (or cumulative flow) that is constrained by the total number of tracks (or station capacity). We adopt this modelling approach for the capacity constraints (and safety blocking rules), and do not consider any additional constraints which can be added to the basic TTP model as in Caprara et al. (2006).

Lagrangian relaxation is also used in the current paper to dualize the capacity constraints. Adopting a multigraph approach with a lagrangian profit function as in Caprara et al. (2006), the dual problem is solved iteratively using bundle methods. Brännlund et al. (1998) show that such methods indicate better solution performances compared to sub-gradients used by Caprara et al. (2002) and Caprara et al. (2006). Based on the dual solution, the final feasible timetable can be obtained using one of the existing combinatorial heuristics such as rapid branching (Borndörfer et al., 2013).

Several TTP studies use *Malmbanan* as a case study, mostly for rescheduling scenarios. For instance, Törnquist (2015) studied different scenarios where part of the train timetable is fixed and shows that optimal solutions for a 4 h time window can be found within 1 min or less. In the European ON-TIME project, a proof-of-concept is presented by Quaglietta et al. (2016) who look at the use, in real world scenarios and realistic simulation environments, of two mathematical algorithms for solving TTPs during traffic perturbations. The authors demonstrate and compare these algorithms and their results indicate reductions in total delays by 35%. In a more recent study, Bach et al. (2018) showed how they have successfully used MILPs models for practical real-time train scheduling. They report finding solutions within 2 seconds for planning horizons covering 2h.

A number of cited works use other case studies and report varying results. With limited computational power, Brännlund et al. (1998) studied a medium size single track line (17 stations) in Sweden with 30 trains (freight and passenger) to be scheduled over a day. The authors report good quality solutions (within a few percent of optimality) but rather modest computational times. A similar but more recent study by Caprara et al. (2006) looked at various scenarios from Italy (17 to 49 stations, 54 to 221 trains) with a time step in minutes. The authors report varying

computational times between few minutes to around 2 hours with quality solutions reaching between 1% to 16%.

Table 1 presents an overview of some of the most related references comparing the adopted TTP models, solution methods and reported results (largest instance, computational time and/or solution quality). There are, however, many other related studies which are not cited here and interested readers are referred to, e.g., the review papers by Lusby et al. (2011) and Harrod (2012).

Table 1. Literature overview of the most related ILP models for TTPs, the adopted lagrangian relaxation (LR) solution method(s) and the reported results (largest instance, computational results and/or solution quality).

Reference	ILP model(s)	LR solution method(s)	Instance, results
Brännlund et al. (1998)	Single track	Dual iterative methods (bundle method, subgradient), heuristic feasible solution	17 stations and 30 trains, 3.8%
Caprara et al. (2002)	Single track, multigraph	Lagrangian profit function, sub-gradient, heuristic feasible solution	39 stations and 500 trains, 1.5h and 14%
Caprara et al. (2006)	Single track, multigraph, additional constraints (e.g., maintenance)	Lagrangian profit function, sub-gradient, heuristic feasible solution	17 stations and 221 trains, 1.7h and 13%
Meng and Zhou (2014)	Multi-track networks, cumulative flow (for track occupancy)	Priority rules	85 stations (network) and 40 trains, 5min and 34%
Yue et al. (2016)	Profit function (with stopping pattern and passenger demand)	Column generation-based heuristic (master and subproblems)	23 stations and 280 trains, 2.5min and 9%
Zhang et al. (2019b)	Cyclic TTP using time-space network and PESP	ADMM	23 stations (double-track) and 34 trains, 3min
This paper	Multigraph, cumulative flow	Lagrangian profit function, two bundle method variants (aggregate, disaggregate), rapid branching (suggested)	14 stations (51 blocks) and 32 trains, 2.8h (dual disaggregate)

The contribution of the paper is to develop a new TTP model (column 2 in **Table 1**) based on the basic single track model by Brännlund et al. (1998), and makes use of modelling improvements such as multigraph (Caprara et al., 2002) and cumulative flow for track occupancy (Meng and Zhou, 2014). Besides, it contributes with a new lagrangian relaxation-based solution method (column 3 in **Table 1**) which is an improved variant of the standard bundle method to solve TTP models. The new solution method is suitable when there are several train path requests which are concurrent and from different train operating companies (i.e., on-track competition such as in open access lines).

3. Mathematical Model

In this section, we present some background information about the considered TTP. Thereafter, we introduce and describe the main notations. Finally, we state and explain the mathematical model.

3.1. Background

The network that is considered in the train timetabling problem is a single-track line. The line is discretised into different blocks which are of two types:

a) *Station blocks* are crossings with sidings where the trains can wait for a scheduled stop or for another train to overtake or pass in the opposite direction. The block capacity, i.e., number of parallel track sidings, often allows for more than one train to stop and wait.
b) *Signalling blocks* are line sections where only one train can pass at a time. These are often between traffic signal points and thus the name, i.e., signalling blocks.

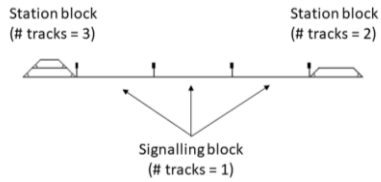


Figure 1. A stretch of a single-track line with station and signalling blocks, adapted from (Gurdan and Kaeslin, 2015)

Station blocks have therefore a capacity higher or equal to 1 whereas signalling blocks have a capacity equal to 1. **Figure 1** illustrates the two types of blocks by representing a stretch of a single-track line with two station blocks with a capacity of 3 and 2 (i.e., number of parallel tracks) and three signalling blocks between the two station blocks. Note that it is possible to have more than one train between two consecutive station blocks if the safety headway blocking rules allow it. These rules will be explained later in the paper.

Figure 2 shows that there are different speed scenarios for any train passing the blocks. The train can pass at full speed in a certain block if there is no stop before entering or after leaving the block, as shown in scenario (1) in **Figure 2**. If there is a stop before or after the block (or both), the speed is lower, as shown in scenarios (2), (3) and (4). Trains can also wait in a side track at the station block, as shown in scenario (5) to let other trains pass and in (6) for scheduled stop.

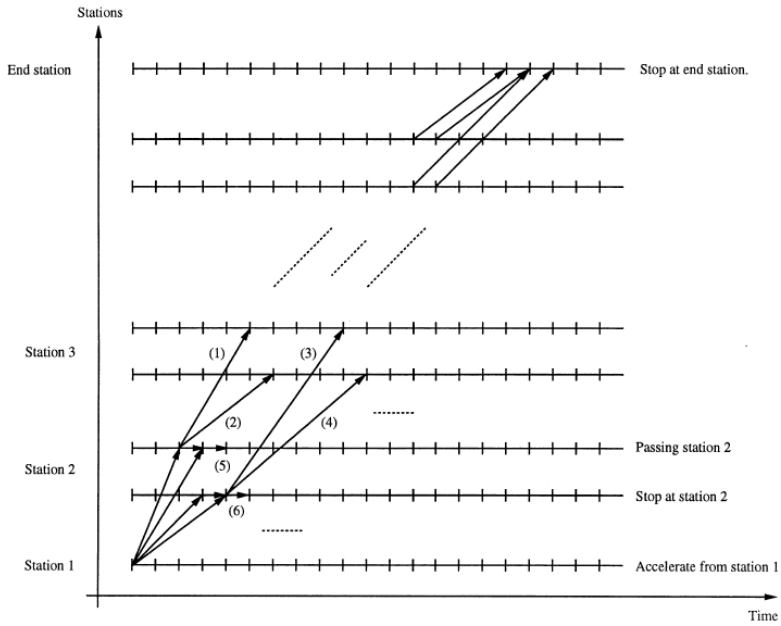


Figure 2. Speed scenarios between train stations (Brännlund et al., 1998).

Hence, there are two different possible movement states in each end of a block: *passing at full speed* (noted F) or *stopping* (noted S). These lead to four different speeds scenarios depending on the state at the start and the end of the block: FF, FS, SF and SS. These different speed scenarios can be seen in **Figure 2**, FF is the fastest scenario (1) and SS is the slowest (4). SF and FS are slower than FF but faster than SS. These different speeds reflect the acceleration and deceleration (i.e., braking) properties of the trains which can play an important role in determining the travel time for the different scenarios.

3.2. Model

A train request $r \in \mathcal{R}$ specifies a set of possible paths $\mathcal{P}_r$ and assigns a utility value v_p to each of them (depending on the total travel time and the deviation from the ideal departure time). The sum of utilities v_p for the selected train paths p is the objective function in the TTP model. The criteria, for whether a path is possible or not, are determined by the stopping stations, the departure time window and the latest arrival time. The "null path", i.e., not scheduling or removing the requested train, is always a possible path. The set of all possible paths is denoted $\mathcal{P} = \bigcup_{r \in \mathcal{R}} \mathcal{P}_r$. Note that we do not store all the possible paths for a train request $r \in \mathcal{R}$.

Time is discretised into time intervals $t \in \mathcal{T}$ and the single-track line is separated into blocks $b \in \mathcal{B}$. The combinations (t, b) make up nodes (or vertices) in a time-space directed graph. Each arc (denoted a) represents a possible train movement (i.e., FF, SF, SS or FS). For instance, train leaving block b_1 at time t_n accelerating from standstill, going to block b_2 at time t_m. This arc (denoted a_1) is illustrated in the time-space graph in **Figure 3**.

Each train path $p_r \in \mathcal{P}_r$ is an ordered subset of train movements $a_r \in \mathcal{A}_r$ (set of arcs of request r) that describes the trajectory of the train in time and space from origin to destination station. We therefore have $a_r \in p_r \subseteq \mathcal{A}_r$ and we construct and store the block-time graph (as shown in **Figure 2**) from origin to destination for each train request. The train movement represented by an arc in the graph leads to the (not necessarily physical) occupation of a certain block-times, given by the binary matrix $\delta_{bt}^{a} \in \{0,1\}$ indicating (for a certain train movement arc a) whether the time-blocks (t, b) are occupied or not. For example, arc a_1 (of path p_1) leads to the physical occupation of block b at times $t_n^1, \dots, t_{m-1}^1$, i.e., $\delta_{bt}^{a_1} = 1$ if $t_n^1 \leq t \leq t_{m-1}^1$ (grey area in **Figure 3**). The parameters δ_{bt}^{a} are used in the capacity constraints of the TTP program to account for the capacity consumption of a certain train movement arc a (part of train path p) on time-block (t, b). Which paths (i.e., arcs and nodes) are possible for a specific train request is determined by the requested train departure windows, latest arrival time as well as the acceleration and deceleration (speed properties) of the train. These are provided in the input data, the speed properties are given as the travel time in each block for the different speed scenarios (i.e., FF, SF, SS or FS).

To ensure a safety headway or distance between trains, several blocking rules have been adopted. For instance, if two trains are moving (as in **Figure 3**), the first train (path p_1) occupies the block b before certain minutes (S_{before}^1) of physically entering the block (at t_n^1). Thus, the block-time occupation parameter $\delta_{bt}^{a_1}$ for train movement a_1 (including the blocking rules) is $\delta_{bt}^{a_1} = 1$ for all t such that $t_n^1 - S_{before}^1 \leq t \leq t_{m-1}^1$ and $\delta_{bt}^{a_1} = 0$, otherwise. The same blocking rules (with S_{before}^2) apply to the second train (path p_2 in **Figure 3**). Since this train stops at the next block (for certain minutes S_{stop}^2), it can keep occupying block b for certain minutes (S_{after}^2) after physically leaving the block (at t_m^2). Thus, the parameter $\delta_{bt}^{a_2}$ for train movement a_2 (including the blocking rules) is $\delta_{bt}^{a_2} = 1$ for all t such that $t_n^2 - S_{before}^2 \leq t \leq t_{m-1}^2 + S_{after}^2$ and $\delta_{bt}^{a_2} = 0$, otherwise.

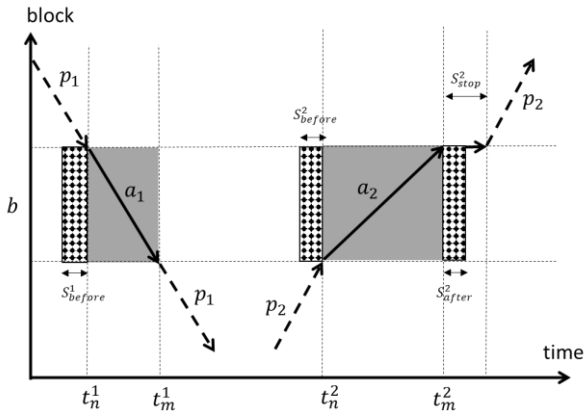


Figure 3. Graphical representation of block occupancy. Black arrows (or arcs) show train movements, grey areas the corresponding physical block occupancy and dotted areas the additional safety block occupancy.

The number of minutes S_{before} and S_{after} depends on the speed scenario and direction of the meeting trains before and after the block. **Figure 3** illustrates the adopted blocking rules and the Swedish timetabling guidelines recommend using $S_{before} = S_{after} = 3$ minutes. (Note that we do not make any difference between headway and clearance time).

These rules also guarantee that two trains in opposite directions cannot instantly swap their respective physically occupied consecutive blocks. Readers who are unfamiliar with blocking rules are invited to refer to the work by Hansen and Pachl (2014) on blocking time theory. A related work by Harrod and Schlechte (2013) compares physical and timed block occupancy.

In **Figure 3**, the total capacity consumption $\sum_{p\in\mathcal{P}} d^p_{bt} = \sum_{a_1\in p_1} \delta^{a_1}_{bt} + \sum_{a_2\in p_2} \delta^{a_2}_{bt}$ (for both p_1 and p_2 if scheduled) at block b is 1 in the grey and dotted area and 0 elsewhere. Note that the total capacity consumption can be greater than 1 (e.g., 2 or double occupation) if several paths (e.g., p_1 and p_2) occupy the same block at the same time. The constraints (1.i) enforce that the total capacity consumption (or cumulative flow for track occupancy as in Meng and Zhou (2014) at any block b (and any time t) is at most equal to the capacity limit c_b.

We summarise the notations in **Table 2** for the main sets, parameters and variables in the mathematical model.

Table 2. Summary of the adopted notations.

Type	Notation	Description
Sets	$\mathcal{T}$	Time intervals $\{1, 2, \dots, t, \dots\}$
	$\mathcal{B}$	Blocks $\{1, 2, \dots, b, \dots\}$
	$\mathcal{R}$	Train requests $\{1, 2, \dots, r, \dots\}$
	$\mathcal{A}_r$	Train movement arcs for request $r \in \mathcal{R}$
	$\mathcal{P}_r$	Possible train paths for request $r \in \mathcal{R}$
	$\mathcal{P} = \bigcup_{r \in \mathcal{R}} P_r$	All possible train paths
Parameters	c_b	Capacity of block $b \in \mathcal{B}$
	v_{p_r}	Utility value of train path $p_r \in \mathcal{P}_r$
	$d_{bt}^{p} = \sum_{a \in p} \delta_{bt}^{a}$	Block occupation of time t of block b by train path $p_r \in \mathcal{P}_r$
Variables	$x_p \in \{0,1\}$	Allocation state of path $p \in \mathcal{P}$

The TTP is now to select exactly a path $p_r \in \mathcal{P}_r$ for each train request $r \in \mathcal{R}$ such that the total path value $\sum_r v_{p_r}$ is maximised and the timetable is feasible (capacity constraints are met). Let $x_p \in \{0,1\}$ be an indicator variable denoting whether path p is selected or not. With this, we can formulate the TTP as the mathematical program (1).

Constraints (1.i) are the capacity constraints of the blocks which do not allow the presence of more trains than the capacity limit c_b in the block-time (b, t). The constraints (1.ii) specify that exactly one path is chosen for each train request. The constraints (1.iii) are the binary constraints on the path selection variable x_p.

Each train request $r \in \mathcal{R}$ thus has a finite set of possible paths $p \in \mathcal{P}_r$ (from the starting station through the network to the ending station) to perform its requested duties (e.g., scheduled stops, departure window and latest arrival time). Each path is an ordered subset of movement arcs a that describes the trajectory of the train in time-space graph.

A utility value v_p can be computed for each possible path p. Summing up over all the selected paths that are in the final timetable gives the total utility value of the train timetable, i.e., the objective value to maximise $\sum_{p \in \mathcal{P}} v_p x_p$ in program (1). In this study, we use a basic model for the utility function which reflects the deviation penalty from an ideal departure time. Moreover, we assume that the utility values of the different trains are mutually independent which is not the case in many realistic scenarios.

The TTP is now to select one path $p \in \mathcal{P}_r$ for each train request $r \in \mathcal{R}$, such that capacity limits c_b are not violated, and such that the total timetable utility (based on v_p) is maximised. The term d_{bt}^p (equal to $\sum_{a \in p} \delta_{bt}^a$) in constraints (1.i) corresponds to the capacity consumption of path p on block-time (b, t). Note that these capacity consumption parameters are also used for enforcing the blocking rules (i.e., safety occupations).

$$\text{(TTP)} \quad \begin{cases} \max\limits_{x_p} \sum\limits_{p \in \mathcal{P}} v_p x_p \\ \text{s.t.} \begin{cases} \sum\limits_{p \in \mathcal{P}} d_{bt}^p x_p \leq c_b, & \forall (b,t) \in \mathcal{B} \times \mathcal{T} \quad (i) \\ \sum\limits_{p \in \mathcal{P}_r} x_p = 1, & \forall r \in \mathcal{R} \quad (ii) \\ x_p \in \{0,1\}, & \forall p \in \mathcal{P} \quad (iii) \end{cases} \end{cases} . \tag{1}$$

The stated TTP model is an ILP with very large number of binary variables for real world instances. The combinatorial nature of the problem makes it difficult to solve for these instances using the-state-of-the-art ILP solvers. In order to get around the computational complexity of the problem, we use the classical lagrangian relaxation technique as the starting point for the solution method.

4. Solution Methods

This section focuses on deriving and describing aggregate and disaggregate solution methods for the stated TTP model. Both are based on the lagrangian relaxation and use a bundle method to solve the lagrangian dual problem.

4.1. Dual problem

In the lagrangian relaxation of (TTP), we allow the constraints (1.i) to be violated, thus allowing the presence of more trains in a time-block than the capacity limit. Moreover, this relaxation also allows the violation of the blocking rules. This violation of the capacity limit is done however at a certain price given by the lagrangian multipliers $\mu = \{\mu_{bt}\} \geq 0$. The relaxed version of (TTP) is noted $(\text{TTP})_\mu$ and is formulated in problem (2). The optimal value of the relaxed problem is noted $\varphi(\mu)$.

$$(\mathrm{TTP})_{\mu}\begin{cases}\varphi(\mu) \coloneqq \max\limits_{x_p} \sum\limits_{p\in\mathcal{P}} v_p\, x_p + \sum\limits_{(b,t)\in\mathcal{B}\times\mathcal{T}} \mu_{bt}(c_b - \sum\limits_{p\in\mathcal{P}} d_{bt}^{p}\, x_p) \\ \text{s.t.}\begin{cases}\sum\limits_{p\in\mathcal{P}_r} x_p = 1, & \forall r \in \mathcal{R} \\ x_p \in \{0,1\}, & \forall p \in \mathcal{P}\end{cases}\end{cases} . \qquad (2)$$

$(\mathrm{TTP})_{\mu}$ is a relaxation of (TTP) for two reasons. First, every feasible solution to (TTP) is also a feasible solution to $(\mathrm{TTP})_{\mu}$. Second, the objective value of any feasible solution in (TTP) is not greater than that in $(\mathrm{TTP})_{\mu}$. Hence, for each $\mu \geq 0$, the value of $\varphi(\mu)$ in $(\mathrm{TTP})_{\mu}$ is larger than or equal (i.e., an upper bound) to the optimal value of (TTP).

It is possible to further simplify the objective value of $(\mathrm{TTP})_{\mu}$. For a given $\mu \geq 0$ and under the same constraints as in (2), $\varphi(\mu)$ can be rewritten as in (3).

$$\varphi(\mu) = \sum_{(b,t)\in\mathcal{B}\times\mathcal{T}} c_b \mu_{bt} + \max_{x_p} \sum_{p\in\mathcal{P}} \left(v_p - \sum_{(b,t)\in\mathcal{B}\times\mathcal{T}} \mu_{bt} d_{bt}^{p} \right) x_p. \qquad (3)$$

$v_p - \sum_{(b,t)\in\mathcal{B}\times\mathcal{T}} \mu_{bt} d_{bt}^{p}$ can be interpreted as a reduced utility revenue for choosing path $p \in \mathcal{P}$ (i.e., $x_p = 1$) given the multipliers μ. This means that $(\mathrm{TTP})_{\mu}$ is equivalent to finding the shortest path in the time-space graph for each train request. Shortest in the sense that the path yields the maximum reduced utility revenue where the multipliers represent the cost of traversing the arcs in that path. Shortest path problems have well-established solution algorithms and are relatively easy to solve. In this model, we developed a shortest path algorithm based on topological sorting. This is justified by the fact that the time-space graph is a weighted directed acyclic graph (Cormen et al., 2009).

As just noted, $\varphi(\mu)$ is an upper bound to the optimal value of (TTP). Thus, the dual problem (D) is to find the optimal solution μ^* that gives the smallest bound as in (4). Since there are only a finite number of shortest path combinations, φ is piecewise linear. It is therefore a convex function since it is the maximum of a set of linear functions. Moreover, φ has a lower bound, i.e., any feasible solution to the original problem (TTP). Therefore, (D) has a global minimum φ^* at the optimal multipliers μ^*.

$$(\mathrm{D}) \begin{cases} \mu^* := \operatorname{argmin} \varphi(\mu) \\ \quad \text{s.t.} \quad \mu \geq 0 \end{cases}. \tag{4}$$

Let us assume that for an arbitrary value $\bar{\mu} \geq 0$, the maximum in $(\mathrm{TTP})_{\bar{\mu}}$ is achieved at $\tilde{x}(\bar{\mu}) = (\tilde{x}_p)_{p \in \mathcal{P}}$. Inserting the corresponding $\tilde{x}(\bar{\mu})$ in the objective of $(\mathrm{TTP})_{\mu}$ gives a linear (in μ) function $\tilde{\varphi}(\mu) = \sum_{p \in \mathcal{P}} v_p \tilde{x}_p + \sum_{bt} \mu_{bt} (c_b - \sum_{p \in \mathcal{P}} d_{bt}^p \tilde{x}_p)$, that is equal to $\varphi(\mu)$ at $\bar{\mu}$. This linear function corresponds to a supporting plane to the graph of φ. The slope of this function is given by the matrix $g(\bar{\mu}) = (\bar{g}_{bt}) := (c_b - \sum_{p \in \mathcal{P}} d_{bt}^p \tilde{x}_p)$ which is a subgradient of φ at $\bar{\mu}$. Thus, the supporting linear function to φ at $\bar{\mu}$ can be written as in equation (5) where * denotes the inner product between two matrices (i.e., component wise).

$$\varphi(\bar{\mu}) + g(\bar{\mu}) * (\mu - \bar{\mu}). \tag{5}$$

In order to solve (D), we use the (aggregate) bundle method by Kiwiel (1990), described in the next section. Based on this method, the subsequent section derives the disaggregate approach to solve (D) for the train timetabling problem TTP.

4.2. Aggregate bundle method

For $\mu = \bar{\mu}$, the (possibly many) $\tilde{x}(\bar{\mu})$, i.e., maximum in $(\mathrm{TTP})_{\mu}$, give the subgradients to φ at $\bar{\mu}$. Suppose that we currently are at $\mu = \mu_k$, and that we have chosen to approximate φ by the supporting planes computed in iterations $l \in \mathcal{L}_k$, where $\mathcal{L}_k$ is the *bundle* of active supporting planes of φ at iteration k. Let the corresponding subgradients be $\{g_l\}_{l \in \mathcal{L}_k}$. Then in the standard aggregate bundle method, we compute a new tentative solution as the solution to the subproblem formulated in (6) where $|\cdot|$ denotes the Euclidean norm (2-norm) of a matrix reshaped into a vector, and $\bar{\varphi}^k(\mu) := \max_{l \in \mathcal{L}_k} \{\varphi(\mu_l) + g_l * (\mu - \mu_l)\}$ is the maximum of the supporting linear functions at μ_l, for $l \in \mathcal{L}_k$, giving an outer linearization of φ. The quadratic second term helps to avoid taking large steps and the step size is adjusted using the control parameter u_k at each iteration.

$$(\bar{\mathrm{D}}_k^{\mathrm{agg}}) \begin{cases} \min \bar{\varphi}^k(\mu) + \dfrac{u_k}{2} |\mu - \mu_k|^2 \\ \quad \text{s.t.} \quad \mu \geq 0, \end{cases}. \tag{6}$$

In order to get around the inner maximisation, $(\bar{\mathrm{D}}_k^{\mathrm{agg}})$ can be formulated as a single minimisation problem by adding an additional variable as well

as new constraints for the supporting linear functions. This leads to the equivalent problem (7).

$$(\overline{\mathrm{D}}_k^{\mathrm{agg}}) \begin{cases} \min\limits_{v,\mu} v + \dfrac{u_k}{2}|\mu - \mu_k|^2 \\ \text{s.t.} \begin{cases} v \geq \varphi(\mu_l) + g_l * (\mu - \mu_l), & \forall l \in \mathcal{L}_k \quad (i) \\ \mu \geq 0 & (ii) \end{cases} \end{cases} \tag{7}$$

The matrices μ_l for $l \in \mathcal{L}_k$ can be extremely large and lead to an excessive memory usage. Therefore, we suggest an equivalent formulation of the supporting linear functions in which scalars, $\Psi_{kl} \coloneqq \varphi(\mu_l) + g_l * (\mu_k - \mu_l)$ at k and for all $l \in \mathcal{L}_k$, are stored instead of the matrices. The right-hand side of the constraint (7.i) can now be rewritten according to equation (8).

$$\varphi(\mu_l) + g_l * (\mu - \mu_k) + g_l * (\mu_k - \mu_l) = \Psi_{kl} + g_l * (\mu - \mu_k). \tag{8}$$

Hence, problem (7) (or equivalently (6)) can be reformulated as in (9).

$$(\overline{\mathrm{D}}_k^{\mathrm{agg}}) \begin{cases} \min\limits_{v,\mu} v + \dfrac{u_k}{2}|\mu - \mu_k|^2 \\ \text{s.t.} \begin{cases} v \geq \Psi_{kl} + g_l * (\mu - \mu_k), & \forall l \in \mathcal{L}_k \quad (i) \\ \mu \geq 0 & (ii) \end{cases} \end{cases} \tag{9}$$

The advantage of the formulation in (9) is that it allows to save in the memory usage. Instead of storing the bundle of matrices μ_l, as in (7), we only store the scalars Ψ_{kl} and use the matrix of multipliers μ_k at the current iteration k. Moreover, Ψ_{kl} can be updated recursively, without the need for retrieving the matrices μ_l, whenever the multipliers are updated (i.e., $\mu_{k+1} \neq \mu_k$) as

$$\Psi_{k+1,l} = \Psi_{kl} + g_l * (\mu_{k+1} - \mu_k)\,, \tag{10}$$

so that the supporting linear functions always have the current μ_k as "foot point".

Let y_{k+1} be the optimal solution to $(\overline{\mathrm{D}}_k^{\mathrm{agg}})$. At y_{k+1} we evaluate the dual objective φ by solving $(\mathrm{TTP})_\mu$ for $\mu = y_{k+1}$. We might then get a new supporting plane, including a new subgradient g_{k+1}. We define the achieved descent as $\bar{\varphi}^k(\mu_k) - \varphi(y_{k+1})$ and the forecasted one as $\bar{\varphi}^k(\mu_k) - \bar{\varphi}^k(y_{k+1})$. If the ratio of the achieved descent by the forecasted one is larger than a certain step quality threshold $m_L \in [0,1]$ (e.g., 10%) then we

set $\mu_{k+1} = y_{k+1}$, and the new $\mathcal{L}_{k+1}$ will incorporate the active supporting planes from $\mathcal{L}_k$ as well as the newly generated supporting plane. If otherwise the ratio is not large enough, we set $\mu_{k+1} = \mu_k$ and $\mathcal{L}_{k+1}$ will only add the newly generated supporting plane to $\mathcal{L}_k$. Thus, the polyhedral approximation of φ is always improved at each iteration.

The step control parameter u_{k+1} is adjusted in both cases (i.e., ratio is large enough or not). It is set so that the curvature of the objective in $(\overline{\mathrm{D}}_k^{\mathrm{agg}})$ between μ_k and y_{k+1} fits that of φ. We use the same step control update strategy as the one adopted by Kiwiel (1995).

4.3. Disaggregate bundle method

In the disaggregated bundle method, the main idea is to linearly approximate (with supporting planes) the function φ for each train request $r \in \mathcal{R}$ instead of the approximation of the aggregated sum of all the requests. Thus, we extract more information from the solutions to the subproblems $(\mathrm{TTP})_\mu$ by considering a disaggregate function for each train request that we note φ_r for all $r \in \mathcal{R}$.

$$\varphi(\mu) = \sum_{r \in \mathcal{R}} \varphi_r(\mu) + \sum_{(b,t) \in \mathcal{B} \times \mathcal{T}} c_b \mu_{bt} \ . \tag{11}$$

The formulation in (11) shows how the dual objective function is separated into request-dependent functions φ_r which are defined in program (12) as a maximisation problem for a given value of the multipliers μ. These μ_{bt} can be interpreted as prices for using the block b at time t, an interpretation that we will return to in the concluding section.

$$\begin{aligned} \varphi_r(\mu) \coloneqq \max_{x_p} & \sum_{p \in \mathcal{P}_r} (v_p - \sum_{(b,t) \in \mathcal{B} \times \mathcal{T}} \mu_{bt} d_{bt}^p) x_p \\ \text{s.t.} & \begin{cases} \sum_{p \in \mathcal{P}_r} x_p = 1 & (i) \\ x_p \in \{0,1\}, \quad \forall p \in \mathcal{P}_r & (ii) \end{cases} \end{aligned} \ . \tag{12}$$

The disaggregate dual problem (13), noted $(\mathrm{D}^{\mathrm{dis}})$, will be slightly different from (D) that is used in the aggregate approach.

$$\left(\mathrm{D}^{\mathrm{dis}}\right)\begin{cases}\min\limits_{\mu}\sum\limits_{r\in\mathcal{R}}\varphi_r(\mu)+\sum\limits_{(b,t)\in\mathcal{B}\times\mathcal{T}}\mu_{bt}\,c_b\\ \text{s.t.}\quad \mu\geq 0\end{cases}. \tag{13}$$

At iteration k in the disaggregate approach, each objective component φ_r has its own bundle $\mathcal{L}_k^r$, with the subgradients defined as $g_{rl}:=-d^{\hat{p}_{rl}}$, where $\hat{p}_{rl}\in\mathcal{P}_r$ is the shortest path in the sense that it leads to the maximal revenue for $\mu=\mu_l$ in $(\mathrm{TTP})_\mu$.

As with the aggregate approach, we use the subgradients to build supporting linear functions that are used as an outer approximation. In the disaggregate approach the outer approximation is computed for each objective component φ_r. Thus, the disaggregate bundle method problem is written as follows

$$(\overline{\mathrm{D}}_k^{\mathrm{dis}})\begin{cases}\min\sum\limits_{r\in\mathcal{R}}v_r+\sum\limits_{(b,t)\in\mathcal{B}\times\mathcal{T}}c_b\mu_{bt}+\dfrac{u_k}{2}|\mu-\mu_k|^2\\ \text{s.t.}\quad\begin{cases}v_r\geq\varphi_r(\mu_l)+g_{rl}*(\mu-\mu_l), & \forall l\in\mathcal{L}_k^r\ \forall r\in\mathcal{R}\quad (i)\\ \mu\geq 0 & (ii)\end{cases}\end{cases}. \tag{14}$$

In order to minimise the memory storage in the implementation, as with the aggregate approach in equation (8), instead of storing all the previous matrices of multipliers μ_l, we only store the corresponding scalar parameters Ψ_{kl}^r that we define similarly as follows

$$\Psi_{kl}^r:=\varphi^r(\mu_l)+g_{rl}*(\mu_k-\mu_l). \tag{15}$$

The parameters are updated in a similar way to the aggregate case that is previously described in (10). Thus, the formulation of $(\overline{\mathrm{D}}_k^{\mathrm{dis}})$ can be rewritten using the scalar parameters. So, at iteration k, we have

$$(\overline{\mathrm{D}}_k^{\mathrm{dis}})\begin{cases}\min\sum\limits_{r\in\mathcal{R}}v_r+\sum\limits_{(b,t)\in\mathcal{B}\times\mathcal{T}}c_b\mu_{bt}+\dfrac{u_k}{2}|\mu-\mu_k|^2\\ \text{s.t.}\quad\begin{cases}v_r\geq\Psi_{kl}^r+g_{rl}*(\mu-\mu_k),\forall\, l\in\mathcal{L}_k^r,\forall r\in\mathcal{R} & (i).\\ \mu\geq 0 & (ii)\end{cases}\end{cases} \tag{16}$$

Note that formulation (16), unlike (9), includes the term $\sum_{(b,t)\in\mathcal{B}\times\mathcal{T}}c_b\mu_{bt}$ since this term is shared between the different train requests $r\in\mathcal{R}$ and is therefore not separable.

We can further simplify (16) by introducing $s = \mu - \mu_k$ to get the following formulation

$$(\overline{\mathrm{D}}_{\mathrm{k}}^{\mathrm{dis}}) \begin{cases} \min \sum_{r \in \mathcal{R}} v_r + \sum_{(b,t) \in \mathcal{B} \times \mathcal{T}} (\mu_{bt}^k + s_{bt})\, c_b + \frac{u_k}{2} |s|^2 \\ \text{s.t.} \begin{cases} v_r \geq \Psi_{kl}^r + g_{rl} * s, & \forall\, l \in \mathcal{L}_k^r, \forall r \in \mathcal{R} \quad (i) \\ s \geq -\mu^k & (ii) \end{cases} \end{cases}. \tag{17}$$

In theory, the disaggregate bundle formulation improves the aggregate one in that it allows to make a more accurate linear approximation of the dual objective function φ by taking advantage of the individual train requests (Ψ_{kl}^r in contrast with Ψ_{kl}). This allows to extract and use more dual information per iteration from the shortest path results. Another advantage is that the disaggregate formulation can be implemented as a parallel program for each train request to improve the computational time (Gurdan and Kaeslin, 2015).

5. Experimental Setup and Results

This section provides details about the case study. We first give implementation-related information before presenting the input data and the experiment scenarios. Finally, the results are presented and discussed before discussing further notes.

5.1. Implementation

Both solution methods, i.e., the aggregate and the disaggregate bundle method, are developed in MATLAB. The methods call a C++ function that computes the shortest path given the prices $\mu = \{\mu_{bt}\}$ of occupying the block-times. The information between the two programming environments is exchanged using *mex functions* which are subroutines that allow MATLAB programs to call C++ functions which have faster access to memory (MathWorks, 2016). **Figure 4** gives an overview of the software architecture that was implemented.

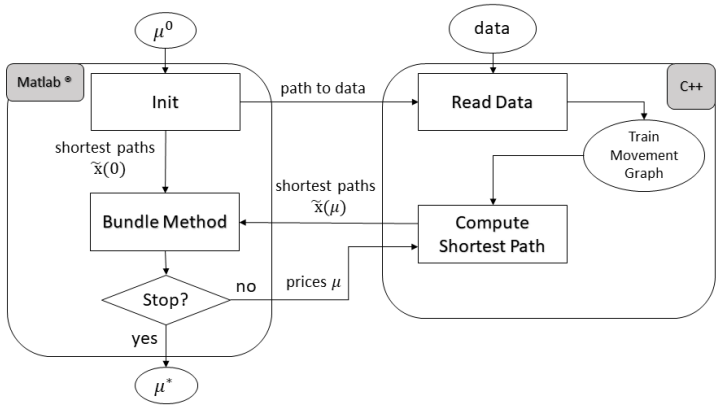


Figure 4. Software architecture of the model implementation.

In order to speed up the computation of the shortest path algorithm, the path networks (the graph of possible train movements in time-space) are constructed once and are stored in the C++ environment memory for use in all the iterations of the bundle method. Dijkstra algorithm is used for the shortest path computation (Dijkstra, 1959). The MATLAB program (i.e., bundle method) calls at first a C++ function that allocates memory and constructs the path networks from the input data before performing the bundle iteration which is implemented using interior point method.

5.2. Input data

The input data that is used to test the two algorithms is based on train operations on the Iron Ore line (*Malmbanan*) in northern Sweden. The stretch that is considered is between *Kiruna* (Sweden) and *Narvik* (Norway) as in **Figure 5**.

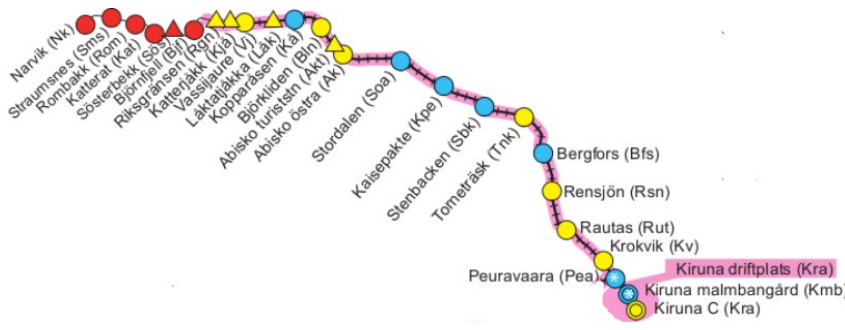


Figure 5. The Iron Ore line between Narvik (in Norway) and Kiruna (in Sweden).

The input data was provided by the Swedish National Transport Administration (*Trafikverket*). It consists of the following:

- Signalling blocks (with waiting stations) on the line
- Travel time to traverse the associated block for different speed scenarios (SF, SS, FS, FF)
- Number of tracks (i.e., capacity) of each waiting station

In this study we consider, for the sake of simplicity, only one type of trains having the same speed properties for the different requested train paths. This means that all the trains have similar travel times between blocks on the line.

We consider 32 train requests during a weekday with 6 (SJ AB) passenger trains and 26 freight trains, operated by the freight operators (e.g., LKAB, Green Cargo and Hector Rai). Each request includes the following information:

- Departure and arrival stations
- Ideal departure time
- Latest arrival time

Each train request specifies an ideal departure time. We consider in this study that trains have a departure windows around this ideal departure time of around 30 minutes, i.e., trains are can depart up to 30 minutes before and up to 30 minutes after the ideal departure time.

The value of a path is the (weighted) sum of the deviation from ideal departure time (as in **Figure 6**) and the path's total running time from departure to arrival station. We distinguish between freight and passenger services in the peak (v_{max} in **Figure 6**) of the deviation function. In this experiment, we set it to 500 for all the 26 freight trains and to 1000 for the remaining 6 passenger trains in order to reflect the assumption that the departure time for passenger trains is generally more valuable. The choice of values for train requests getting fulfilled at all, i.e., its value compared to the null path, has been arbitrarily set at different levels for freight (500) and passenger (1000) services. It has been suggested that this input to the optimisation exercise emanates from an explicit bidding process (on-track competition) where different (competing) operators define the paths they request and the value function of being allocated a path. The utility value v_p associated with the allocated path $p \in \mathcal{P}_r$ then specifies the operator's benefit of being able to run each service. Solving the track allocation problem more generally comprises two components; the optimisation problem which is addressed by the present paper; and the valuation problem which can be handled by operators submitting

bids for each path. The latter problem is out of the scope of this paper but was addressed by Nilsson (2002).

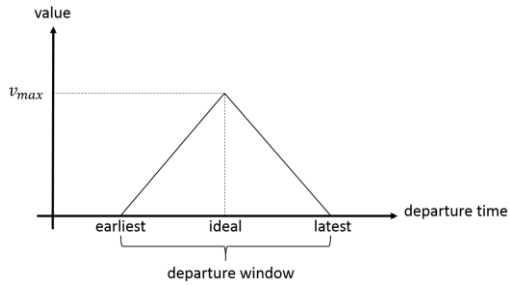


Figure 6. Simplified valuation function of train requests (Brännlund et al., 1998).

Moreover, we consider that each (passenger) train has a minimal compulsory waiting time of 2 min at every scheduled waiting. We also consider the blocking rules to ensure a certain safety distance between trains.

In order to check the models on different problem instances, we constructed three different test cases from the given data. **Table 3** lists the different experiment scenarios and their characteristics.

Table 3. Test cases and their characteristics.

Test case	Terminal stations	# of stations	# of blocks
S1	Bjørnfjell - Narvik	5	14
S2	Kiruna – Torneträsk	7	23
S3	Kiruna – Vassijaure	14	51

The test case scenarios correspond to an increasingly longer stretch of the Iron Ore line, i.e., increasing number of stations and blocks. The 32 requested train paths are as previously described and are adjusted to run between the terminal station of the considered line stretch of each scenario.

5.3. Results

The experimental tests were executed on a remote computer with two processors Intel(R) Xeon(R) CPU E5645. Each processor has a clock frequency of 2.40GHz and 12MB cache memory; the RAM memory is 80GB.

The models have several parameters and initialisations to be set before starting the execution. The parameter values are given in **Table 4**.

Table 4. Algorithm parameters and their values

Parameter	Value
Time discretisation step	30 seconds
Step quality threshold	$m_L = 0.1$ (= 10%)
Initial step control value	$u_0 = 1$
Minimal step control	$u_{min} = 10^{-10}$
Maximal number of iterations	$k_{max} = 200$
Initial prices (multipliers)	$\mu_0 = 0$
Tolerance (stopping condition)	$\epsilon = 10^{-13}$

Both models were tested under the same conditions, i.e., same machine, parameters and input data. **Figure 7** shows the comparison between the dual objectives in the aggregated and disaggregate approaches for the test cases S1 – S3.

In the three test cases, the optimisation of the dual objective function has a similar behaviour for the two approaches in the first iterations. However, after a certain number of iterations, the minimisation in the disaggregate approach becomes faster as more information is collected in the iteration bundle. This leads to a faster convergence using the disaggregate approach.

Table 5. Initialisation and execution time in the test cases S1 – S3

Case	Initialisation time (in min)	Execution time – aggregate (in min)	Execution time – disaggregate (in min)	Execution time improvement (*)
S1	26.44	40.36	24.11	40.3%
S2	39.17	49.71	27.13	45.4%
S3	213.19 (≈ 3.5h)	209.09 (≈ 3.5h)	169.72 (≈ 2.8h)	18.8%

(*) The improvement is relative to the aggregate approach.

Table 5 presents the computation times for the initialisation and execution time of the test cases. The initialisation time (column 2) is the time needed to read data and construct the train movement graph (see **Figure 4**). This step is similar for both variants and is performed only once for each test case. The execution times (columns 3 and 4) relate to the time to execute the bundle method (with iterative calls to the shortest path algorithm). Both variants call the same algorithm for computing the shortest path on the same train movement graph, so the improvement in execution time is mainly due to the fewer number of iterations performed by the disaggregate variant (see **Figure 7**). Moreover, if the solution algorithm stops after reaching a maximum number of iterations (e.g.,

$k_{max} \geq 7$ for **Figure 7**), the disaggregate solution is generally more accurate, i.e., better dual objective value, which is mainly due to the faster convergence of the approach after a number of iterations.

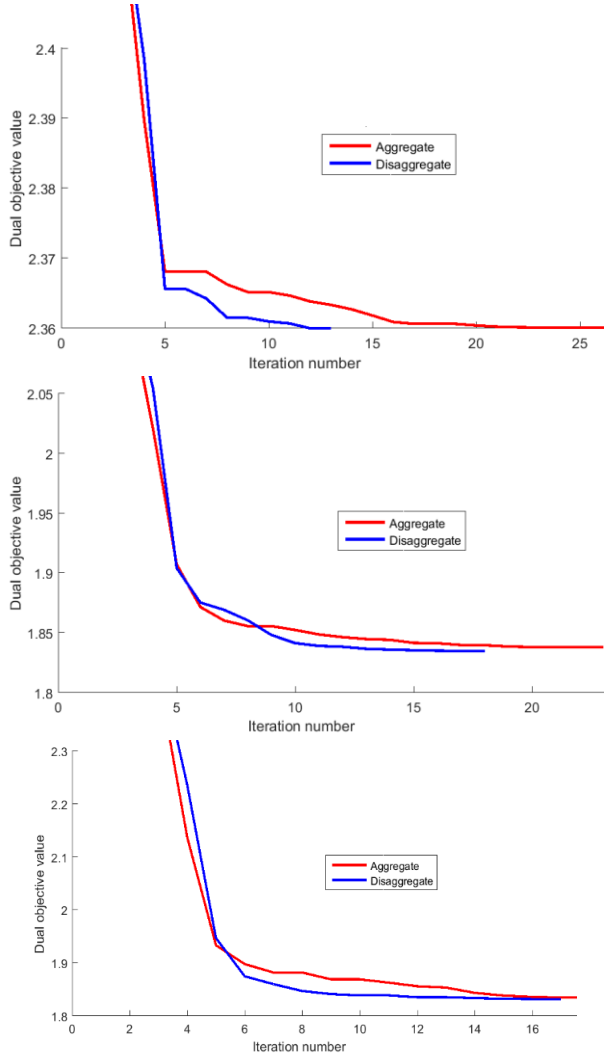
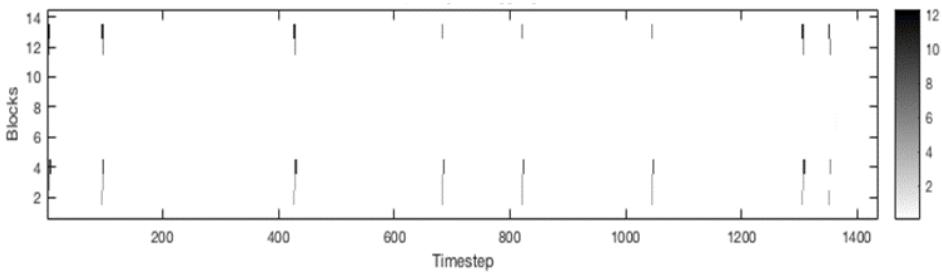


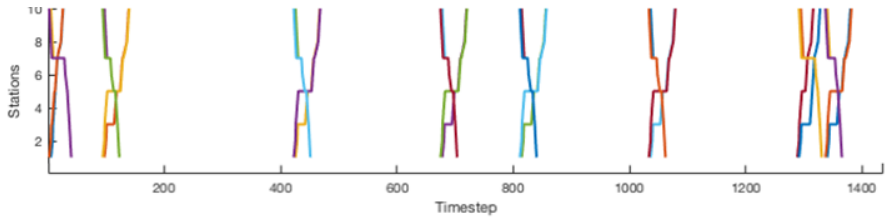
Figure 7. Dual objective for the two approaches in the test cases S1 to S3 (from top).

Note that the computation times in **Table 5** increase substantially for S3 (larger instance). Moreover, the savings in the execution times, when using disaggregation, decrease. Therefore, disaggregation may give limited savings when used to solve larger instances of the problem, but more tests are needed to reveal this variant's convergence properties.

Testing the two approaches also provides additional useful information. For instance, the solution to the dual problem provides values for the multipliers μ_{bt} which can be used for pricing the infrastructure capacity in time-space. **Figure 8(a)** presents the resulting optimal infrastructure pricing (or multipliers) for test case S1 using the disaggregate approach. The vertical axis refers to the blocks and the horizontal one to the time steps (30 seconds for each timestep), higher prices are shown in black. The corresponding optimal (not necessarily feasible) dual solution timetable (i.e., selected train path for each request) can be visualised in the graphical (time-space) timetable in **Figure 8(b)** where colours are used to differentiate between the train requests. Each selected train path is shown in a different colour. Since this is the optimal solution of the relaxed TTP, plotting all the selected train paths at the final iteration of the algorithm will not necessarily lead to a feasible timetable. Such a timetable can be obtained from the relaxed solution (together with all the generated train paths) as explained in section 5.4.



(a) Optimal mulipliers (or capacity prices) for each block and time, higher values in black.



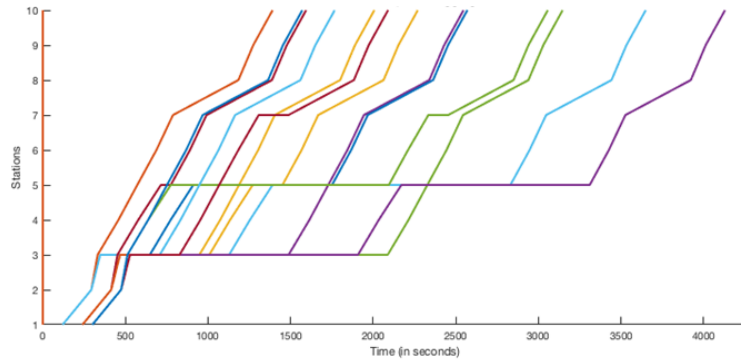
(b) Selected train paths at optimal dual solution, colours to distinguish between train paths.

Figure 8. Disaggregate optimal solution (i.e., multipliers and train paths) from test case S1.

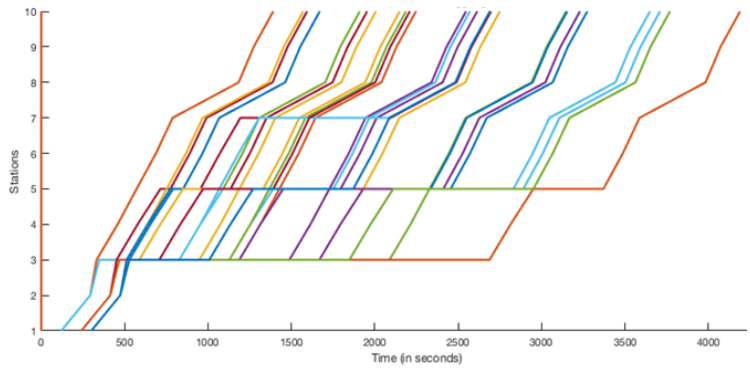
Similar figures can be produced for the aggregate case. However, the disaggregate approach yields higher multiplier values (prices), train paths with more stops and more generated train path alternatives. The results in **Figure 8** indicate that having train path requests with close (or

similar) ideal departure times (i.e., competition for the slot) yields higher pricing of capacity around the block-times with most conflicts, e.g., mainly around the departure stations and ideal departure times. If these conflicting train paths are requested by different concurrent operators (e.g., open access), such prices can be used to sell departure slots or as congestion fees in the access charges.

In each iteration, the shortest path algorithm is called to find the best paths for a given pricing. This leads to an additional output of the algorithm, i.e., a set of generated (possible) train paths for each request. Therefore, each request has a set of alternative paths which contains "the" optimal path (including the null path, i.e., cancelling the train) that will be selected in the final optimal (feasible) timetable.



(a) Standard aggregate approach.



(b) Proposed disaggregate approach.

Figure 9. Generated train path alternatives for the first request in test case S1, colours to distinguish between alternatives.

Figure 9 illustrates an example of a set of generated paths for the first train request in test case S1. The first graph (a) shows the aggregate approach and the second one (b) the disaggregate approach. The disaggregate approach generates a larger set of paths compared to the aggregate one, e.g., 24 paths instead of 14 as in **Figure 9**. The null path is always part of such a set to allow cancel train requested train (i.e., vertical line in the origin). The distribution of the generated paths is affected by several parameters, e.g., the optimisation approach (aggregate or disaggregate), the utility value of the objective function (freight or passenger) and other conflicting train requests.

The generated set of paths from the disaggregate methods has therefore more potentially optimal train paths per request that can be selected in the final feasible timetable. This larger set of generated paths is particularly important in finding better quality feasible solutions. Branch-and-bound is a technique that can be used to find such a solution based on the resulting optimal solution of the relaxation problem.

In practice, both variants can be used to solve small to medium size instance of the problem. In particular, train timetabling instances with highly congested traffic (i.e., several conflicting train requests and/or limited railway capacity). The case study has shown that disaggregation allows savings in computation time. Moreover, the construction of the train movement graph is done once and can be performed ahead of executing the bundle solution methods which can allow substantial savings in the total solution time. Another advantage, specific to the disaggregate method, is the potential of a parallel implementation which would lead to further savings in computation times. However, even with the advantages of the disaggregate variant (including the parallelisation), the method seems to be unsuitable for real time rescheduling, especially for larger instances. Moreover, the formulation of the objective function assumes that each train request has a certain utility function which is not always available and may be difficult to estimate. Alternative recent approaches exist such as auction (Pena-Alcaraz, 2015) or societal costs (Ait-Ali et al., 2020).

5.4. Fractional solution and final feasible timetable

The relaxed fractional solution from the solution methods will not necessarily lead to a feasible timetable, as in **Figure 8(b)**, unless the studied instance correspond to an unimodular constraint matrix, the TTP is solved in this case.

However, the relaxed optimal solution often yields fractional values for the integer variables. In this case, the multipliers and the generated train paths combined with a branch-and-bound method can be used to find a good quality feasible timetable. A unique path (including "null-path") is selected for each request from the generated train paths to form the feasible timetable.

In order to compute the fractional solution (i.e., $x = (x_p)$ not necessarily integer) of the relaxed (TTP), we use the dual solution of the dual disaggregate problem ($\mathrm{D}^{\mathrm{dis}}$). For this, we note $\lambda = (\lambda_{rl})$ as the dual solution of ($\overline{\mathrm{D}}_K^{\mathrm{dis}}$) in (14) at the last iteration K of the iterative disaggregate bundle algorithm. The dual multipliers λ_{rl} are associated with the problem constraint for request $r \in \mathcal{R}$ and bundle iteration $l \in \mathcal{L}_K^r$. The fractional solution for a path $p_r \in \mathcal{P}_r$ is formulated in (18) where SP_l^r denotes the shortest path for request r at iteration l in the iterative disaggregate bundle algorithm.

$$x_{p_r} = \sum_{\substack{l \in \mathcal{L}_K^r \\ p_r = SP_l^r}} \lambda_{rl} \geq 0 \,. \tag{18}$$

It can be shown that the dual solution satisfies $\sum_{l \in \mathcal{L}_K^r} \lambda_{rl} = 1$ at the optimum. Thus, the fractional solution is $x_{p_r} \in [0,1]$, $\forall p_r \in \mathcal{P}_r, \forall r \in \mathcal{R}$ and can be interpreted as the likelihood of choosing path p_r for request r in the final feasible timetable. For instance, if a certain path p_r is the shortest path during all the K iterations of the algorithm ($\forall l \in \mathcal{L}_K^r$), we have $x_{p_r} = \sum_{\substack{l \in \mathcal{L}_K^r \\ p_r = SP_l^r}} \lambda_{rl} = \sum_{l \in \mathcal{L}_K^r} \lambda_{rl} = 1$ and hence this path (i.e., p_r) is chosen in the final feasible timetable.

Adapted variants of branch-and-bound methods, e.g., rapid branching (Borndörfer et al., 2013), can be used to find an optimal feasible combination of the generated paths using their respective fractional solution value from the disaggregate approach.

6. Conclusions

Finding a faster solution to the train timetabling problem is useful for many situations in applied railway planning, both for long-term and short-term planning, and when trying to construct capacity allocation processes with several stakeholders. This paper contributes to this by presenting an improved, parallelisable and easily implementable bundle

method that can be used when solving the dual problem arising from the relaxation of a train timetabling problem. Compared to the standard bundle method, the new method uses disaggregate information and yields more accurate solutions faster. When solving the problem, lagrangian multipliers are obtained for each train request that can be interpreted as the price incurred by the network for using the train path. This is useful as it can give insights to the infrastructure manager on different aspects such as track access charges or lack of capacity.

Numerical results suggest that the new, disaggregate method tends to give shorter execution times and better accuracy, compared to the standard, aggregate bundle method. Moreover, the disaggregate method generates larger sets of possible train paths, which is a useful feature for branch-and-bound algorithms. The disaggregate approach has therefore the potential to improve lagrangian-based solution methods for discrete time and space formulations of TTPs. Possible future works include investigating the scalability of this new approach with further tests and analysis using other case studies. The proposed disaggregation approach can also be used with solution methods (other than bundle methods) and compare its performances with the corresponding standard variant.

Acknowledgment

This research is part of the project Socio-economically efficient allocation of railway capacity, SamEff (*Samhällsekonomiskt effektiv tilldelning av kapacitet på järnvägar*) which is funded by a grant from the Swedish Transport Administration (*Trafikverket*). The authors are grateful to Martin Aronsson, Tobias Gurdan and Alain Kaeslin for the valuable discussions and comments as well as to three anonymous reviewers and journal editors for helpful suggestions.

References

AIT-ALI, A., WARG, J. & ELIASSON, J. 2020. Pricing commercial train path requests based on societal costs. Transportation Research Part A: Policy and Practice, 132, 452-464.

BACH, L., MANNINO, C. & SARTOR, G. MILP approaches to practical real-time train scheduling: the Iron Ore Line case. International Network Optimization Conference (INOC), 2018.

BORNDÖRFER, R., LÖBEL, A., REUTHER, M., SCHLECHTE, T. & WEIDER, S. 2013. Rapid branching. Public Transport, 5, 3-23.

BRÄNNLUND, U., LINDBERG, P. O., NÕU, A. & NILSSON, J.-E. 1998. Railway Timetabling using Lagrangian Relaxation. Transportation Science, 32, 358-369.

CACCHIANI, V., CAPRARA, A. & TOTH, P. 2008. A column generation approach to train timetabling on a corridor. 4OR, 6, 125-142.

CAPRARA, A., FISCHETTI, M. & TOTH, P. 2002. Modeling and solving the train timetabling problem. Operations Research, 50, 851-861.
CAPRARA, A., MONACI, M., TOTH, P. & GUIDA, P. L. 2006. A Lagrangian heuristic algorithm for a real-world train timetabling problem. Discrete Applied Mathematics, 154, 738-753.
CORMEN, T. H., LEISERSON, C. E. & RIVEST, R. L. 2009. Introduction to Algorithms, Cambridge, US, The MIT Press.
DIJKSTRA, E. W. 1959. A note on two problems in connexion with graphs. Numerische mathematik, 1, 269-271.
FORSGREN, M., ARONSSON, M. & GESTRELIUS, S. 2013. Maintaining tracks and traffic flow at the same time. Journal of Rail Transport Planning & Management, 3, 111-123.
GURDAN, T. & KAESLIN, A. 2015. Parrallelised Shortest Path Algorithm for Railway Timetable Construction. MA Master thesis, KTH.
HANSEN, I. A. & PACHL, J. 2014. Railway timetabling & operations. Eurailpress, Hamburg.
HARROD, S. & SCHLECHTE, T. 2013. A direct comparison of physical block occupancy versus timed block occupancy in train timetabling formulations. Transportation Research Part E: Logistics and Transportation Review, 54, 50-66.
HARROD, S. S. 2012. A tutorial on fundamental model structures for railway timetable optimization. Surveys in Operations Research and Management Science, 17, 85-96.
JAMILI, A., SHAFIA, M. A., SADJADI, S. J. & TAVAKKOLI-MOGHADDAM, R. 2012. Solving a periodic single-track train timetabling problem by an efficient hybrid algorithm. Engineering Applications of Artificial Intelligence, 25, 793-800.
KIWIEL, K. C. 1990. Proximity control in bundle methods for convex nondifferentiable minimization. Mathematical Programming, 46, 105-122.
KIWIEL, K. C. 1995. Approximations in proximal bundle methods and decomposition of convex programs. Journal of Optimization Theory and Application, 84, 529-548.
LUSBY, R. M., LARSEN, J., EHRGOTT, M. & RYAN, D. 2011. Railway track allocation: models and methods. OR Spectrum, 33, 843-883.
MATHWORKS. 2016. Introducing MEX Files [Online]. source MEX fileC, C++, or Fortran source code file. Available: http://se.mathworks.com/help/matlab/matlab_external/introducing-mex-files.html [Accessed October 2015].
MENG, L. & ZHOU, X. 2014. Simultaneous train rerouting and rescheduling on an N-track network: A model reformulation with network-based cumulative flow variables. Transportation Research Part B: Methodological, 67, 208-234.
NILSSON, J.-E. 2002. Towards a welfare enhancing process to manage railway infrastructure access. Transportation Research Part A, 36, 419–436.
PENA-ALCARAZ, M. M. T. 2015. Analysis of Capacity Pricing and Allocation Mechanisms in Shared Railway Systems. PhD, MIT - Massachusetts Institute of Technology.
QUAGLIETTA, E., PELLEGRINI, P., GOVERDE, R. M. P., ALBRECHT, T., JAEKEL, B., MARLIÈRE, G., RODRIGUEZ, J., DOLLEVOET, T., AMBROGIO, B., CARCASOLE, D., GIAROLI, M. & NICHOLSON, G. 2016. The ON-TIME real-time railway traffic management framework: A proof-of-concept using a scalable standardised data communication architecture. Transportation Research Part C: Emerging Technologies, 63, 23-50.
TÖRNQUIST, J. 2012. Design of an effective algorithm for fast response to the rescheduling of railway traffic during disturbances. Transportation Research Part C: Emerging Technologies, 20, 62-78.

TÖRNQUIST, J. 2015. Computational decision-support for railway traffic management and associated configuration challenges: An experimental study. Journal of Rail Transport Planning & Management, 5, 95-109.
TÖRNQUIST, J. & PERSSON, J. A. 2007. N-tracked railway traffic re-scheduling during disturbances. Transportation Research Part B: Methodological, 41, 342-362.
XU, X., LI, K., YANG, L. & YE, J. 2014. Balanced train timetabling on a single-line railway with optimized velocity. Applied Mathematical Modelling, 38, 894-909.
YUE, Y., WANG, S., ZHOU, L., TONG, L. & SAAT, M. R. 2016. Optimizing train stopping patterns and schedules for high-speed passenger rail corridors. Transportation Research Part C: Emerging Technologies, 63, 126-146.
ZHANG, Y., D'ARIANO, A., HE, B. & PENG, Q. 2019a. Microscopic optimization model and algorithm for integrating train timetabling and track maintenance task scheduling. Transportation Research Part B: Methodological, 127, 237-278.
ZHANG, Y., PENG, Q., YAO, Y., ZHANG, X. & ZHOU, X. 2019b. Solving cyclic train timetabling problem through model reformulation: Extended time-space network construct and Alternating Direction Method of Multipliers methods. Transportation Research Part B: Methodological, 128, 344-379.

Paper P5

The Value of Data for Demand Estimation in Public Transport Systems.

Ait-Ali A.[1,2] and Eliasson J.[2] (2020)
[1] VTI Swedish National Road and Transport Research Institute, Transport Economics (TEK), Stockholm
[2] Linköping University, Department of Science and Technology (KTS), Norrköping

Submitted for journal publication

Abstract

Passenger origin-destination data is an important input for public transport planning. In recent years, new data sources have become increasingly common through the use of automatic collection of entry counts, exit counts and link flows. However, collecting such data can be sometimes costly. The value of additional data collection hence has to be weighed against its costs. We study the value of additional data for time-dependent origin-destination matrix estimation, using a case study from the London Piccadilly underground line. Our focus is on how the precision of the estimated matrix increases when additional data on link flow, destination count and/or average travel distance is added, starting from only origin counts. We concentrate on the precision of the most policy-relevant estimation outputs, namely link flows and station exit flows. Our results suggest that link flows are harder to estimate than exit flows, and only entry and exit data is far from enough to estimate link flows with any precision. Information about average trip distance adds greatly to the estimation precision. The marginal value of additional destination counts decreases only slowly, so a relatively large number of exit station measurement points seem warranted. Link flow data for a subset of links hardly add to the precision, especially if other data has already been added.

Keywords: Dynamic origin destination, OD estimation; entropy maximisation; lagrangian relaxation; smart cards; public transport

1. Introduction

Passenger origin-destination data is an important input for public transport (PT) planning. PT demand is summarised in time-dependent origin destination (OD) matrices, which state the number of trips between pairs of stations, i.e., number of passengers from an origin to a destination station per time interval, such as 15-minute intervals. The knowledge of such matrices may improve the efficiency of PT supply (Pelletier, Trépanier, and Morency 2011), e.g., cost-effective timetable designs (Sun et al. 2014), or for studying passenger costs from timetable changes to solve track capacity conflicts (Ait-Ali, Warg, and Eliasson 2020).

In recent years, many PT systems have adopted new technological solutions such as automated fare collection (AFC), automated vehicle location (AVL) and vehicle weighing systems that measure passenger link flows. These solutions generate useful data, e.g., smart card data or automatic vehicle weighs, which can be used for OD matrix estimation. However, acquiring such data can sometimes be costly, since it often requires installation and maintenance of measurement equipment on stations, tracks and vehicles. Having measurements on all stations and links can be prohibitively costly, so a PT agency needs to weigh these costs against the benefits of a more precisely estimated OD matrix.

In this study, we investigate how much the precision of an estimated dynamic OD matrix for a single train line increases when additional data becomes available. We use a case study from the London Piccadilly underground line. Starting with origin counts only, we incrementally add data about exit counts, link flows and average trip distance, and measure how the precision of the estimated matrix increases with additional data.

We concentrate on the precision of the most policy-relevant variables, namely time-dependent link flows and station arrival rates, since these determine policy decisions such as service frequency (Ait-Ali, Eliasson, and Warg 2020) and capacity of stations and trains. They are also the key variables when analysing passenger costs and benefits when adjusting timetables to solve capacity conflicts with other trains, as explained by Ait-Ali, Warg, and Eliasson (2020).

Our results suggest that only entry and exit data is far from enough to estimate link flows with any precision. Information about average trip distance adds greatly to estimation precision. Moreover, extrapolating from a limited number of destination counts or link flow measurements

to the rest of the network results in lower added value, especially if prior data such as average travel distance is already included. Measuring a relatively large number of link flows and exit stations thus seem warranted.

Section 2 briefly summarises the large literature on OD estimation. Section 3 describes the methodology and the case study. Results are presented in section 4, and section 5 concludes.

2. Literature Review

The OD estimation-related research literature is rich, and has a long history which can be traced back to the early 20th century, e.g., with gravity models (Reilly 1931), entropy maximisation (Cesario 1973) and Furness methods (Morphet 1975). There are various origin destination problems which appear in many fields of transportation research (Doblas and Benitez 2005). Most studies treat the time-independent (or static) problem (Wang, Gentili, and Mirchandani 2012), but there has been an increasing interest in the (harder) time-dependent (or dynamic) version (Cho, Jou, and Lan 2009) which is the focus in this paper. This is partly due to increased data availability through AFC data, which is valuable for more precise estimation of (dynamic) OD matrices (Gordillo 2006). Better OD estimates can be used to improve PT services in various ways, for instance by inferring the purpose of the trips (Alsger et al. 2018), pricing and allocating railway capacity (Ait-Ali, Warg, and Eliasson 2020), or by estimating in-vehicle crowding costs (Hörcher, Graham, and Anderson 2017; Yap, Cats, and van Arem 2018).

The OD estimation problems also differ in terms of the considered zones and the studied type of transport traffic. Some studies looked at the flow of road vehicles (Wang, Gentili, and Mirchandani 2012) whereas fewer considered passenger flow in PT systems (Alsger et al. 2018) such as buses (Wang, Attanucci, and Wilson 2011), freight (Shen and Aydin 2014) or passenger rail (Gordillo 2006). Similar to the study by Wong and Tong (1998), this paper focuses on the passenger flow in a commuter rail system.

The formulation of the problem also depends on whether prior (target) matrices exist. Many authors assume the existence of such a matrix (Wang and Zhang 2016). However, this is not the case in our study and many others (Cho, Jou, and Lan 2009).

Generally speaking, the OD estimation problem consists of finding the most probable matrix that is consistent with observations, or minimising

the deviation from observations. The definition of "most probable" (and "deviation") leads to different formulations of the objective function and functional constraints, and thus to a number of OD estimation models. For instance, deviation functions can be modelled in various ways, e.g., using discrete choice models (Ben-Akiva 1985), generalised least square (Cascetta and Nguyen 1988), Kalman filters (Cho, Jou, and Lan 2009), mean least square with entropy (Xie, Kockelman, and Waller 2011), gravity models (Shen and Aydin 2014). Other modelling approaches also exit such as genetic algorithms with entropy (Fu 2012), principal component analysis (Djukic et al. 2012), Bayesian inference (Carvalho 2014), trip chaining (Alsger et al. 2016; Hora et al. 2017) or Markov chain models (Abareshi, Zaferanieh, and Safi 2019). Due to the continuous development of new approaches, several authors summarise and compare many of the different OD estimation models, for example Cascetta and Nguyen (1988), Abrahamsson (1998), Peterson (2007), Bera and Rao (2011), Deng and Cheng (2013), and more recently Alsger (2017) and Li et al. (2018).

In the absence of a target matrix, this paper adopts the entropy maximisation (EM) principle which is also equivalent to a number of models such as gravity (Wilson 1967), minimum information (Van Zuylen and Willumsen 1980) and discrete choice models (Mishra et al. 2013). Entropy maximisation (EM) originates from the statistical theory of probability. In the context of OD estimation, the EM principle relies on the idea that there are many possible trip distributions (or system states) and that the most probable OD estimate (or state) is the one that maximises the total entropy (or randomness). Variants of such formulation have been adopted in a number of OD estimation studies. Fisk (1988) used a similar (time-independent) formulation and considered that the choice of the path depends on the total travel time (or congestion). Similarly, Brenninger-Göthe, Jörnsten, and Lundgren (1989) used it in a multi-objective program for OD estimation using traffic counts. The same formulation was also more recently used by Xie, Kockelman, and Waller (2011) and Fu (2012).

Different types of data have been used to estimate OD matrices, e.g., cell phone network (Wang et al. 2013), tolling (Wang and Zhang 2016), GPS data and travel surveys (Ge and Fukuda 2016). In PT systems, the increasing adoption of AFC and AVL, and thus the availability of the corresponding smart card data, has led to the emergence of new applications based on such data (Nassir et al. 2011; Alsger et al. 2015).

Such studies focus on different aspects in order to improve the PT planning and its efficiency. For instance, the estimation (Mosallanejad, Somenahalli, and Mills 2019) or validation (Alsger et al. 2016) of OD matrices in different PT systems, or more particularly trip destinations (Trépanier, Tranchant, and Chapleau 2007). Moreover, different case studies exist from various PT networks around the world. These include entry-only and/or entry-exit systems from New York (Barry et al. 2002), Santiago (Munizaga and Palma 2012), China (Chen and Liu 2016) and London (Wang, Attanucci, and Wilson 2011). London is also the case study in this paper. Readers interested in a summary of the different OD estimation studies using smart card PT data are referred to the review paper by Li et al. (2018).

Wang, Gentili, and Mirchandani (2012), one of the only studies on the value of data, looked at the additional value of well-located sensors for improving road traffic OD estimates. However, although rich, the research literature on PT data does not include, to our knowledge, studies that look at the value of knowing such additional data for dynamic OD estimation. Hence, the purpose of this paper is to fill this gap in the literature by studying the value of smart card and additional PT data.

3. Methodology and Data

In this section, we describe and formulate the main problem, i.e., the OD matrix estimation, the solution method (details in the appendix) and the case study.

Let n_{ij}^t be the number of passengers starting from station i in time interval t, going to station j. The (dynamic) OD matrix estimation consists of finding a time-dependent origin-destination matrix $\{n_{ij}^t\}$ that is consistent with observations. This is done by estimating the entropy-maximising matrix (the "most probable" matrix) that is consistent with observations of origin counts O_i^t, destination counts D_j^t, link flows F_l^t and the average trip distance $\bar{d}$.

In the following, we assume that origin counts are always available, since virtually all PT systems collect such data at entry gates. Destination counts, however, is not always collected, since equipping exit gates with data collection equipment is costly. Link flow measurements require specialised equipment, such as automated vehicle weighing. The average trip distance is usually estimated using travel surveys.

From the network and timetable, the travel time matrix τ_{ij} can be calculated. Given these, the estimated number of arriving passengers at station j in time interval t can be calculated as $\Sigma_i n_{ij}^{t-\tau_{ij}}$. The estimated flow on link $l = (s, e)$ at time t can be calculated as $\sum_{\substack{i<l \\ j>l}} n_{ij}^{t-\tau_{is}}$, where $\{i < l\}$ denote all stations i preceding link l (including s), and $\{j > l\}$ denote all stations succeeding it (including e). Given distances between stations d_{ij}, the estimated average trip distance is calculated as $\frac{\Sigma_{ijt} n_{ij}^t d_{ij}}{\Sigma_{ijt} n_{ij}^t}$.

The core question of the paper is how the precision of the estimated matrix improves when more and more data becomes available. Let L be the subset of links where link flows are known, and Δ the subset of stations where exit counts are known. The dynamic OD estimation problem can then be formulated as

$$\begin{cases} \max_{n \geq 0} -\sum_{ijt} \left(n_{ij}^t \log\left(n_{ij}^t\right) - n_{ij}^t \right) \\ \sum_j n_{ij}^t = O_i^t \; ; \;\; \forall i, \forall t \quad (1.1) \\ \sum_i n_{ij}^{t-\tau_{ij}} = D_j^t \; ; \; \forall t, \forall j \in \Delta \;\; (1.2) \\ \sum_{\substack{i<l \\ j>l}} n_{ij}^{t-\tau_{is}} = F_l^t \; ; \forall t, \forall l = (s,e) \in L \quad (1.3) \\ \frac{\sum_{ijt} n_{ij}^t d_{ij}}{\Sigma_{ijt} n_{ij}^t} = \bar{d} \quad (1.4) \\ n_{ii}^t = 0 \; ; \;\; \forall i \end{cases} . \qquad (1)$$

The central question can now be stated as: How much is the precision of the estimated OD matrix n_{ij}^t improved when additional data becomes available, i.e., when the sets Δ and L become larger?

We must thus define what kind of "precision" we are interested in. In applied policy making, for example timetable design and investments in links or stations, the exact cells of the OD matrix is actually less important. What matters most are station flows and link flows, since this determines the crowding levels in vehicles and stations. This is used for decisions about link and station capacity upgrades, for station staffing

planning, and for timetable design (timetable optimisation depends mainly on passenger departure and arrival rates per line segment, and on crowding levels on different links). Hence, we will concentrate on how close to reality the estimated OD matrix is in terms of link flows and arrival rates per station (origin rates are assumed to be known). We thus measure the relative root mean square error (or deviation) for link flows ($\mathrm{RMSE}_{\mathrm{link}}$) and arrival rates at destination stations ($\mathrm{RMSE}_{\mathrm{dest}}$), and study how these vary with more available information, i.e., when the sets of available link flows and destination counts, L and Δ, becomes larger.

Let $\widehat{D}_j^t = \sum_i n_{ij}^{t-\tau_{ij}}$ be the estimated number of passengers arriving at station j and time interval t, and $\hat{F}_l^t = \sum_{\substack{i<l \\ j>l}} n_{ij}^{t-\tau_{is}}$ the estimated link flow on link l and time interval t. The relative errors are then defined as in equation (2) and (3).

$$\mathrm{RMSE}_{\mathrm{dest}} = \frac{\sqrt{\Sigma_{jt}\left(\widehat{D}_j^t - D_j^t\right)^2}}{\sqrt{\Sigma_{jt}\left(D_j^t\right)^2}} \tag{2}$$

$$\mathrm{RMSE}_{\mathrm{link}} = \frac{\sqrt{\Sigma_{lt}\left(\hat{F}_l^t - F_l^t\right)^2}}{\sqrt{\Sigma_{lt}(F_l^t)^2}} \tag{3}$$

For those stations and links where data is available ($j \in \Delta$ and $l \in L$) the errors will of course be zero (assuming that the optimisation problem is feasible). The errors hence measure the deviations for the unobserved stations and links – in other words, how well the available station and link data can be extrapolated to unobserved stations and links.

Table 2 lists the different combinations of destination counts, link flows and average travel distance that we will explore.

3.1. Solution method

The EM estimation model is a concave (nonlinear) maximisation program with a nonlinear objective function (the total entropy) and linear constraints. Finding a solution for time-dependent real-world instances (e.g., large networks and/or longer time periods) is generally hard. Thus, instead of using state-of-the-art solvers, we derive the iterative solution methods using Lagrangian relaxation.

We first relax the constraints and associate corresponding Lagrangian multipliers as presented in **Table 1**. This leads to the formulation of a Lagrange function (or relaxed dual objective function). More details can be found in the appendix.

Table 1. Lagrangian multipliers and the corresponding relaxed constraints.

Constraint(s)	Description	Lagrangian multiplier(s)
(1.1)	Origin counts	$\lambda_{it};\ \forall i, \forall t$
(1.2)	Destination (or exit) counts	$\mu_{jt};\ , \forall t, \forall j \in \Delta$
(1.3)	Link flow counts	$\varphi_{lt};\ \forall t, \forall l \in L$
(1.4)	Average travel distance	θ

Using first order optimality conditions on the Lagrange function, we can formulate the (primal) solution, i.e., OD estimate as a function of the (dual) Lagrangian multipliers. Depending on the studied data, we find different solution formulations of the dynamic OD estimate n_{ij}^t. **Table 2** presents the formulations for the different studied variants. A more detailed derivation of these solution formulations is described in the appendix.

Table 2. Solution models and formulations of different variants.

Variant	Model	Formulation
O	Origin counts only, for all stations (basic model)	$e^{\lambda_{it}}$
O-D	Origin counts (for all stations) and destination counts for a subset of stations Δ	$\begin{cases} e^{\lambda_{it}+\mu_{j,t+\tau_{ij}}}; j \in \Delta \\ e^{\lambda_{it}}; j \notin \Delta \end{cases}$
O-d-D	As O-D plus average travel distance	$\begin{cases} e^{\lambda_{it}+\theta d_{ij}+\mu_{j,t+\tau_{ij}}}; j \in \Delta \\ e^{\lambda_{it}+\theta d_{ij}}; j \notin \Delta \end{cases}$
O-F	Origin counts (for all stations) and link flows for a subset of links L	$\begin{cases} e^{\lambda_{it}+\varphi_{lt}};\ l=(i,j) \in L \\ e^{\lambda_{it}};\ l=(i,j) \notin L \end{cases}$
O-D-F	As O-F but with destination counts for all stations	$\begin{cases} e^{\lambda_{it}+\varphi_{lt}+\mu_{j,t+\tau_{ij}}}\ l=(i,j) \in L \\ e^{\lambda_{it}+\mu_{j,t+\tau_{ij}}};\ l=(i,j) \notin L \end{cases}$
O-d-D-F	As O-D-F plus average travel distance	$\begin{cases} e^{\lambda_{it}+\theta d_{ij}+\varphi_{lt}+\mu_{j,t+\tau_{ij}}}; l=(i,j) \in L \\ e^{\lambda_{it}+\theta d_{ij}+\mu_{j,t+\tau_{ij}}}; l=(i,j) \notin L \end{cases}$

In order to estimate the multipliers, we use the problem constraints. In some trivial cases, it is possible to find a closed-form expression such as in the basic O model where $\sum_j n_{ij}^t = O_i^t \Rightarrow n_{ij}^t = e^{\lambda_{it}} = \frac{O_i^t}{|S|-1}$, i.e., all destinations have similar attractivity. In other (more interesting) cases, this is often difficult (sometimes impossible). Thus, we attempt to find

numerical solutions by iteratively balancing the relaxed constraints corresponding to additional studied data. **Figure 1** is an example of an iterative algorithm to find the numerical solution of the multipliers for the O-d model. More details about the iterative algorithms can be found in the appendix.

The iterative solution algorithm stops when the constraints are satisfied, up to a certain tolerance ϵ. Note that the use of the (hard) constraint of origin counts (from smart cards) to derive an analytic expression of the dynamic OD estimate yields the formulation in (4).

$$n_{ij}^t = O_i^t \; p(j|i,t)$$

$$\text{where } p(j|i,t) = \frac{e^{u_{ij}^t}}{\sum_j e^{u_{ij}^t}} \text{ and } u_{ij}^t = K_j^t + \theta_1 k_{ijt}^{(1)} + \cdots + \theta_m k_{ijt}^{(m)} \tag{4}$$

The term $p(j|i,t)$ can be seen as the probability of choosing destination j when departing from origin i at time interval t. In this case, the exponent u_{ij}^t can be interpreted as the total utility for travelling to j from i during time interval t. Such utility may include parameters $k_{ijt}^{(1)}, \ldots, k_{ijt}^{(m)}$ corresponding to m types of additional data, if available. The coefficients (or multipliers) $\theta_1, \ldots, \theta_m$ are estimated to reflect utilities (if $\theta \geq 0$) or disutilities (if not). The constant K_j^t is specific to the destination station j and time interval t.

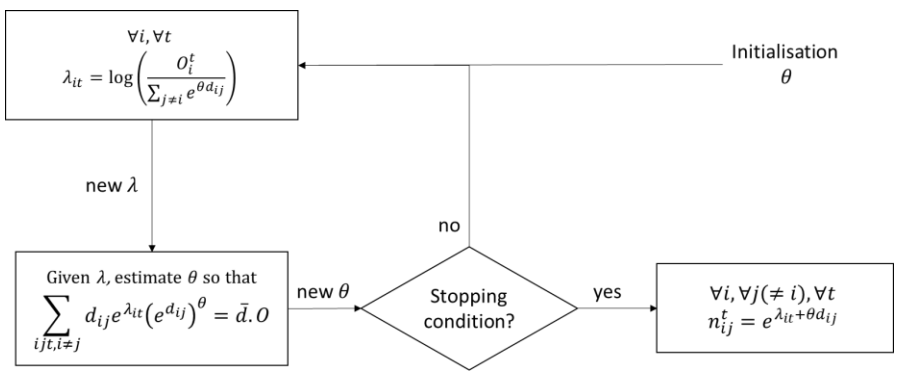


Figure 1. Iterative algorithm for the O-d variant.

Such interpretation can also be found in discrete choice models (without the random error term) where the discrete choices are between the different destination stations j given an origin i. The (alternative-specific)

constants K_j^t and parameters $\theta_1, \ldots, \theta_m$ are specific to the PT system where the OD estimation is performed, and need to be estimated to reflect the (dis-)utilities explaining the choice of the passengers. It is possible to estimate the values of these parameters using additional data, e.g., from smart cards, stated (or revealed) preference surveys, old OD or target matrices from the same PT system.

3.2. Case study data

To explore the question formulated above we use a case study based on the London Piccadilly underground line. Transport for London (or TfL) provides open access to a comprehensive multi-rail demand dataset as part of the NUMBAT project (TfL 2018). Based on the use of smart cards at entry/exit station gates during a typical 2018 autumn weekday, the dataset provides information about the number of passengers boarding and alighting at each station (per 15 min), and link flows (per 15 min) for a subset of the links (data for around 100 links are available, but we study 12 of the most crowded links). The data also contains an estimated OD matrix for longer time periods which is used in this case study to calculate the average travel distance.

The Piccadilly line (**Figure 2**) is more than 70 km long and consists of 53 stations with two different western branches at Acton Town station. Note the one-way trajectory around the Heathrow airport from/to Hatton Cross through terminal 4 then 2 & 3.

In **Figure 3**, stations are sorted according to their location on the studied line in order to make it easy to visualise the symmetry of the distance matrix. However, the matrix, as shown in the figure, is not completely symmetric, see around the airport due to the previously mentioned one-way trajectory.

In order to calculate the average travel times τ_{ij}, we use the travel distances between each pair of stations which is illustrated in **Figure 3**. We assume that all trains are running according to the train timetable (headways) presented in **Table 3**, and that their average speed is 33 kmph (TfL 2018).

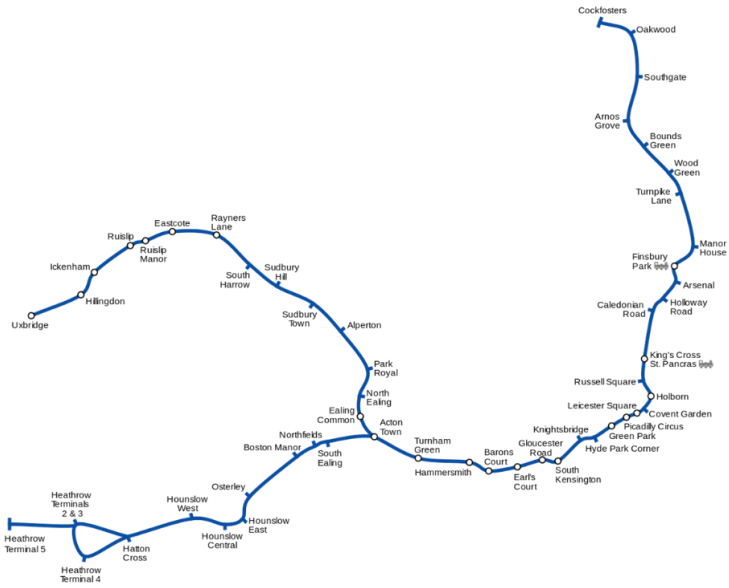


Figure 2. Piccadilly line of the London commuter network.

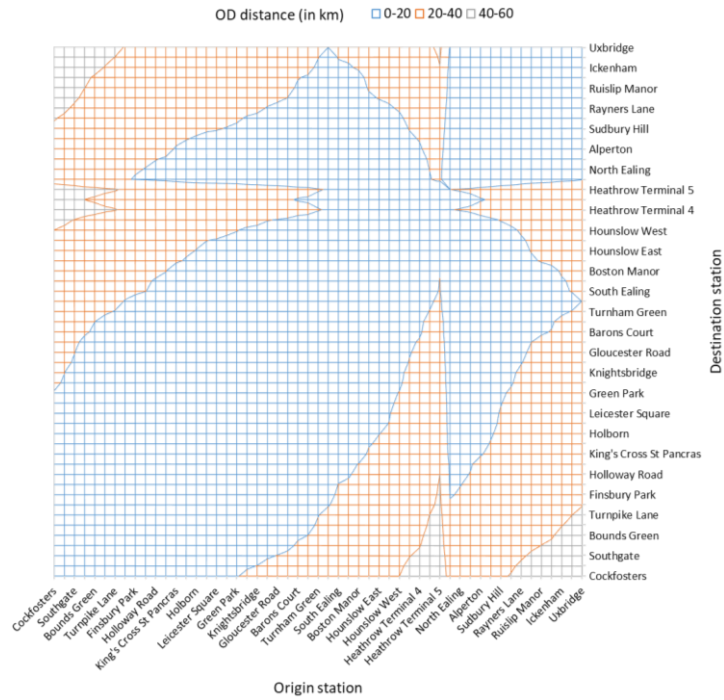


Figure 3. Travel distances (in km) between the different pairs of stations.

Table 3. Train headways in the studied time periods for both directions (TfL 2018).

	Peak **i.e., morning (7.00 – 10.00) and afternoon (16.00 – 19.00)**	**Off-peak** **e.g., midday (10.00 – 16.00)**
Main	5/2 min	5 min
Branches	5 min	10 min

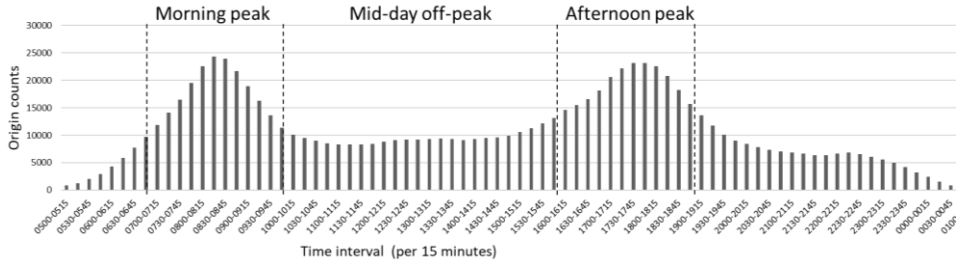


Figure 4. Temporal variation of the total number of boarders and alighters.

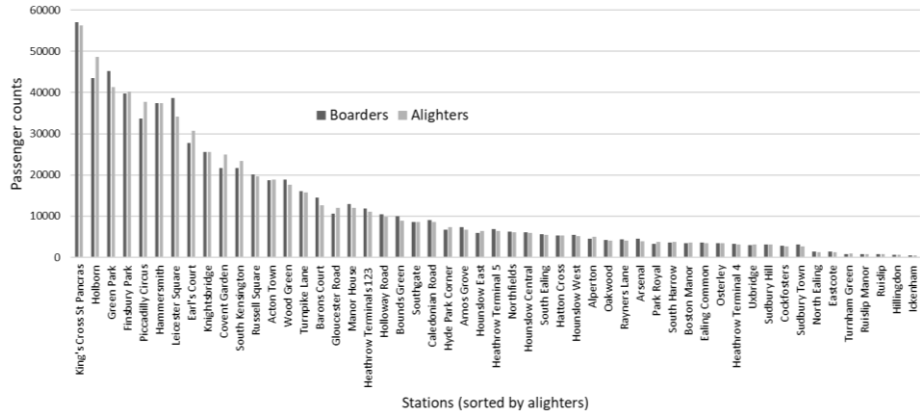


Figure 5. Spatial variation of the total number of boarder and alighters.

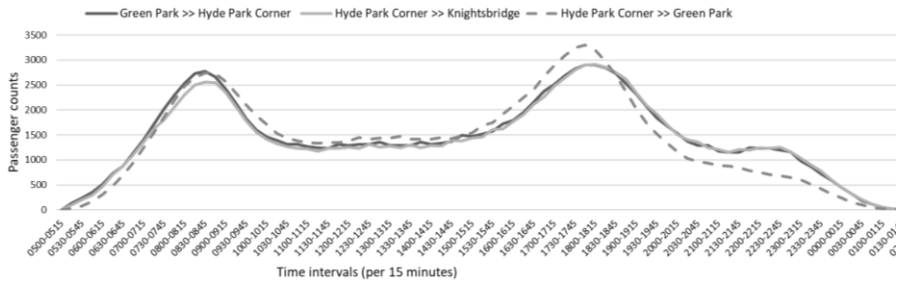


Figure 6. Temporal variation of the passenger flow in three of the most crowded links.

As presented in **Table 3**, we focus in this study on three different time periods, i.e., morning and afternoon (peak) as well as midday (off-peak). These time periods are also illustrated in **Figure 4** which also shows the variation of both the origin (boarders) and the destination or exit (alighters) counts per 15 minutes time interval over the day. The studied time periods are separated by dashed vertical lines in the figure.

In addition to the temporal variation (per time interval) of the number of boarders and alighters as shown in **Figure 4**, we present the spatial variation (per station) in **Figure 5** over the day. The stations on the horizontal axis are sorted by the number of alighters (from highest). **Figure 6** presents the link flows during the day for three of the largest links.

The average travel distance per passenger $\bar{d}$ is usually estimated from demand travel surveys. For our case study, we calculate it based on the available OD matrices (per time period). **Table 4** shows the average travel distance in km per passenger for the different studied time periods of the day.

Table 4. Average travel distance (in km per pax) for the different time periods.

Time period	Average travel distance (km per pax)
Morning (peak)	9.7
Midday (off-peak)	8.6
Afternoon (peak)	9.1

4. Results

In this section, we present results on how the precision of the estimated matrix improves when more data is included in the estimation. We focus on the precision of arrival rates at destination stations and on link flows. Several scenarios with different types of additional data are tested. **Table 5** presents an overview of the reported results, i.e., tested models and the corresponding presented estimation errors.

Table 5. Overview of the presented results and the tested models.

Incrementally added	Destination data	Link flow data
$\mathrm{RMSE}_{\mathrm{dest}}$	O-D, O-d-D	O-F
$\mathrm{RMSE}_{\mathrm{link}}$	O-D, O-d-D	O-F, O-D-F, O-d-D-F

Two types of data are incrementally added, i.e., destination and link data. The value of such additional data is studied by testing different estimation models. The average travel distance is also studied in certain models.

Note that when incrementally including link flows in O-D-F and O-d-D-F, exit counts from all stations are considered unlike other models (i.e., O-D and O-d-D variants) where these are also incrementally included.

4.1. Estimating arrivals rates per station

We first focus on the destination estimation, i.e., number of alighters in the system per 15 minutes time interval and station. **Figure 7** show how the relative error ($\mathrm{RMSE}_{\mathrm{dest}}$) varies when data for more and more stations is added. Stations are sorted according to their total number of alighters, and for each step along the x-axis, data for one additional station is added. The $\mathrm{RMSE}_{\mathrm{dest}}$ error is presented separately for three parts of the day (morning, midday and afternoon). When data for all destinations has been added, the error is of course zero.

Surprisingly, including a relatively small number of destinations actually increases the error for both the midday and afternoon periods. Only after a majority of destinations have been added does the error decrease. Adding the average travel distance to the estimation improves the estimation quite substantially. For the midday and afternoon periods, just adding the average travel distance decreases the error by as much as adding data for almost all destinations but without the average distance. These results suggest that only having data for a subset of destinations is not enough – in fact, it may even increase the overall error. Having enough exit counts is apparently important to get good precision in the estimate. Information about average trip distances, on the other hand, is highly valuable.

With focus on the O-F model, **Figure 8** shows how $\mathrm{RMSE}_{\mathrm{dest}}$ varies when incrementally adding link flow data from 12 of the most crowded links. The error is presented for the three time periods of the day, and links are sorted and added according to their passenger flows. The order of added links may differ between the different time periods, see later in **Figure 10**, for the specific added links.

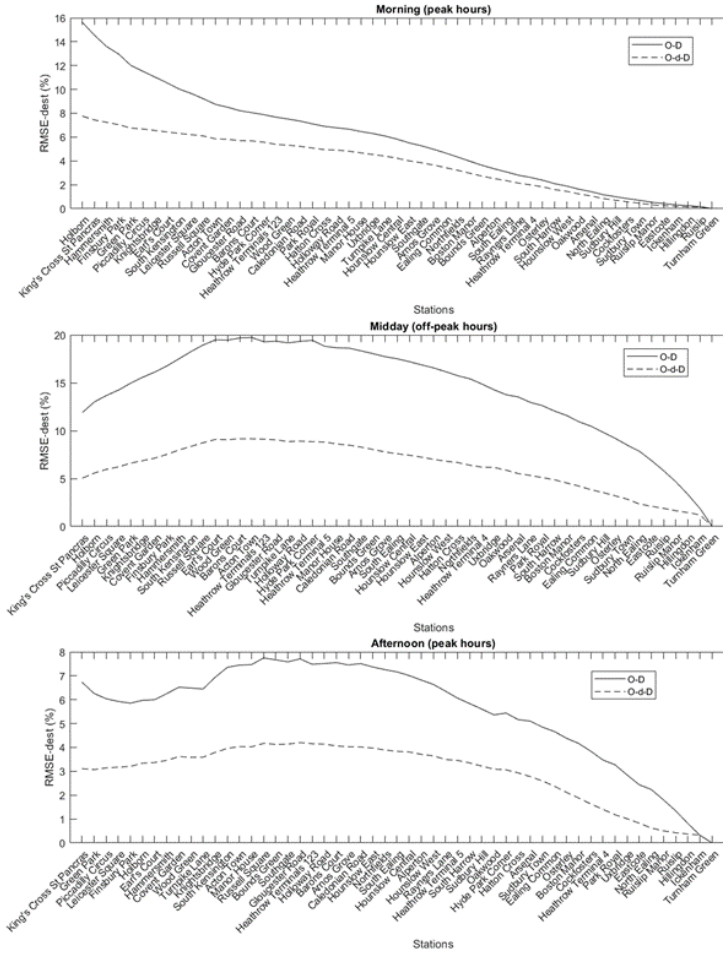


Figure 7. Variation of $\mathrm{RMSE}_{\mathrm{dest}}$ as destination data is incrementally added.

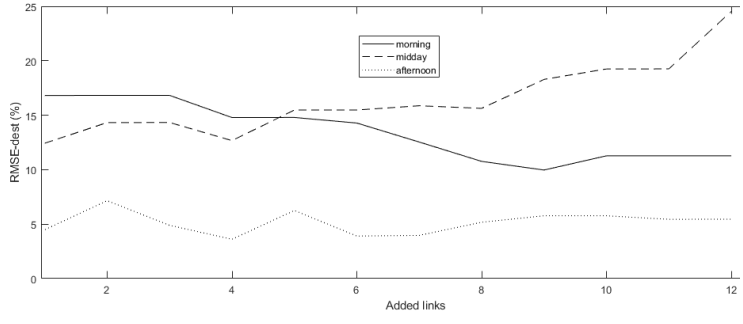


Figure 8. Variation of $\mathrm{RMSE}_{\mathrm{dest}}$ as link data is incrementally added to the O-F model.

Including link flow data does not seem to reduce the destination error. It increases during midday peak hours whereas it remains almost constant in the afternoon. The exception is the morning time period as the error is slightly reduced when more links are added. Later in **Figure 10,** we study the combination with other additional data, i.e., average travel distance and exit counts.

4.2. Estimating passenger flows per link

Figure 9 shows how the link flow error $RMSE_{link}$ varies when additional destination data is added, i.e., O-D and O-d-D models. Even after all exit counts have been included, the link flow error is far from zero. In fact, adding destination data hardly improves the link flow estimation, apart for the first few destinations.

As above, information about the average trip distance greatly decreases the error. Surprisingly, however, adding more destination data tends to increase the error in this case.

To get decent precision in the link flow estimation, link flow data is apparently necessary. **Figure 10** shows the change in $RMSE_{link}$ when incrementally including flow data for 12 of the most crowded links (as in **Figure 8**). The figures show estimations without destination data (O-F), with all destination data (O-D-F) and with destination data and average travel distance (O-D-F-d).

In the O-F model, the error decreases with more data, as expected. However, although the relative error is lower (than O-F), in models with destination data (O-D-F) and average travel distance (O-d-D-F), it remains almost constant when link flow data is added. This is the case for all the time periods except for the morning peak hours where the relative error decreases after adding the 4^{th} link, but remains constant after. These results indicate that the marginal value of additional data decreases when prior data is already added and, in some cases, e.g., link data when destination (and average travel distance) is already included, it may be insignificant. Including more link data can provide further insights, however, the estimation models tend to become more computationally expensive.

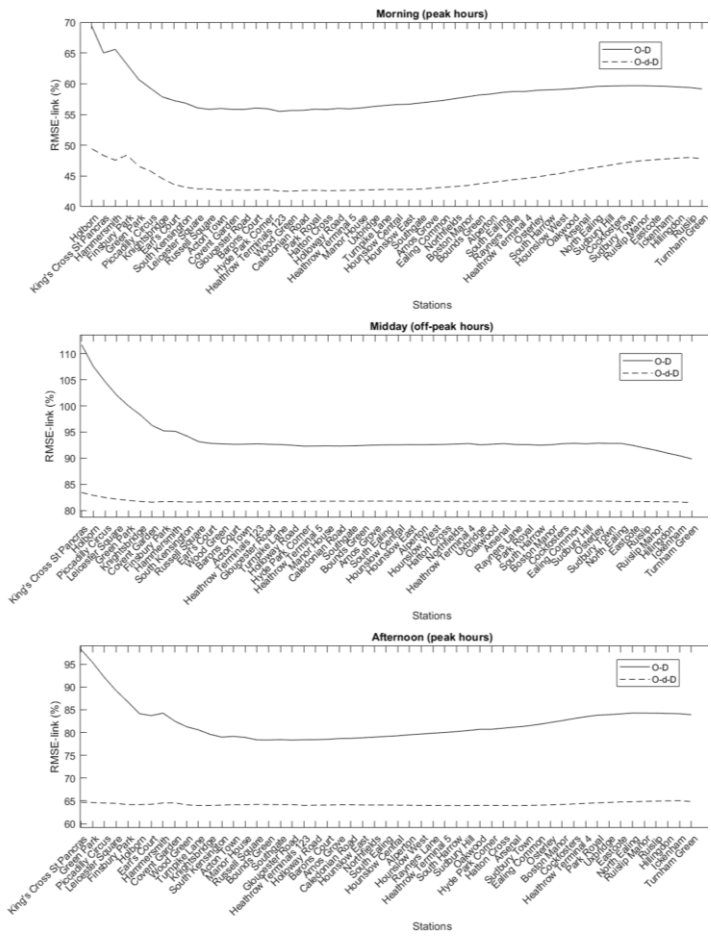


Figure 9. Variation of $RMSE_{link}$ as destination data is incrementally added.

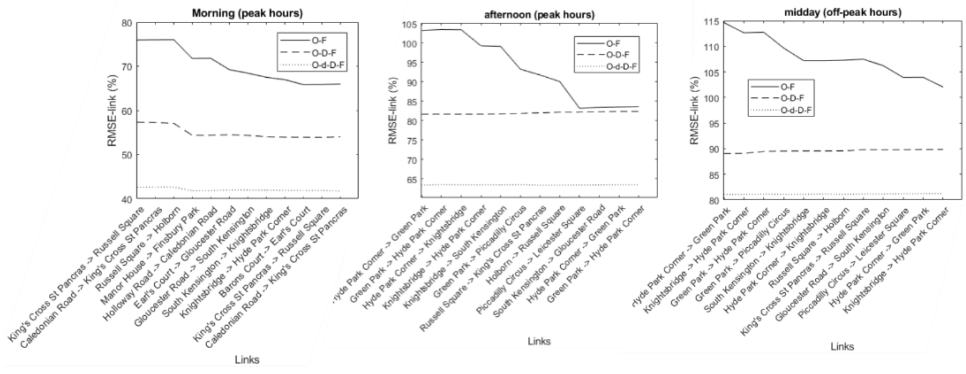


Figure 10. Variation of $RMSE_{link}$ as link flow data is added incrementally.

5. Conclusions and Future Works

Even if the literature includes several studies on (dynamic) OD-matrix estimation, this work attempts to assess the marginal value of additional data in terms of estimation precision. The additional data we have studied is arrival rates per station (which may be collected through AFC systems or specialised equipment), link flows (which may be collected by vehicle weighing systems) and average trip distances (from travel surveys). We explore this through a case study based on the London Piccadilly line in 2018, separating three time periods of the day (i.e., morning and afternoon peak hours, midday). We focus on the precision of the estimated time-dependent arrival rates and link flows, rather than on individual cells in the time-dependent OD matrix.

The results indicate that arrival rates per destination station (if enough) may improve the estimation, but in two cases of three, including data for a subset of destinations actually made the estimation worse. Perhaps contrary to expectations, it turns out to be valuable to have data for a very large share of destinations: the marginal value of acquiring more data, even for the last stations, is surprisingly high. Arrival rates (exit counts) can be collected easily for AFC systems which are based on "tap-in/tap-out" (such as London), but for entry-only AFC systems (such as Stockholm), special data collection equipment needs to be installed at exit gates. Our results suggest that installing such equipment may only lead to marginal improvements of the estimated OD-data, unless a large (enough) share of stations are equipped. If such equipment is costly, it might be more cost-efficient to consider other forms of data collection, and to study the value of the collected data.

Similarly, the study of a subset of added link flows indicates that link estimation may improve but only if no prior additional data is already added. Otherwise, the estimates are better (than with no prior data) but do not improve with added link data. Thus, the marginal value of such detailed data may be insignificant if specific prior data is already included. These results show that detailed (often expensive) data may have a lower marginal value for the demand estimation and can therefore lead to less accurate demand-sensitive policy decisions, e.g., setting welfare-optimal line frequencies.

There are a number of possible future works that can further validate these results, e.g., using other estimation models, metrics for the valuation of the estimation quality, and by studying additional data sources in other case studies. For instance, we used the relative RMSE to quantify

the precision of these estimates, but other metrics can be tested in future work, such as the implied optimal service frequencies (Ait-Ali, Eliasson, and Warg 2020), or levels of in-vehicle crowding (Çelebi and İmre 2020). Full-day estimation instead of per time period can also be tested when additional data is lacking. However, assuming that the time-aggregated OD matrix is symmetric is a strong assumption, and is for example violated in our data set. Furthermore, such full day estimation also requires additional computational power and can be intractable for large networks.

Overall, information about average trip distances gives by far the greatest improvement of the estimation. Acquiring such estimates, from travel surveys, link flow measurements or by other means, is hence a priority. In this study we have only used one average distance (per time period) for the whole line, but obviously, getting more detailed data (for parts of the line) would be highly valuable. Furthermore, instead of gradually including data based on the magnitude of the counts or flow, other orders can also be tested, for instance based on job or home locations during peak hours. Closely related, the model can be adapted to find the optimal data types and their spatiotemporal locations that yield the most valuable data collection strategy.

Acknowledgments

This research is part of the project Socio-economically efficient allocation of railway capacity, SamEff (*Samhällsekonomiskt effektiv tilldelning av kapacitet på järnvägar*) which is funded by a grant from the Swedish Transport Administration (*Trafikverket*). The authors are grateful to Jan Lundgren for improvement suggestions on an earlier version.

References

Abareshi, M., M. Zaferanieh, and M. R. Safi. 2019. 'Origin-Destination Matrix Estimation Problem in a Markov Chain Approach', *Networks and Spatial Economics*.

Abrahamsson, T. 1998. *Estimation of Origin-Destination Matrices Using Traffic Counts - A Literature Survey* (IR-98-021: Austria, Europe).

Ait-Ali, Abderrahman, Jonas Eliasson, and Jennifer Warg. 2020. "Are commuter train timetables consistent with passengers' valuations of waiting times and in-vehicle crowding?" In *VTI Working Papers*. Swedish National Road & Transport Research Institute.

Ait-Ali, Abderrahman, Jennifer Warg, and Jonas Eliasson. 2020. 'Pricing commercial train path requests based on societal costs', *Transportation Research Part A: Policy and Practice*, 132: 452-64.

Alsger, Azalden A., Mahmoud Mesbah, Luis Ferreira, and Hamid Safi. 2015. 'Use of Smart Card Fare Data to Estimate Public Transport Origin–Destination Matrix', *Transportation Research Record*, 2535: 88-96.
Alsger, Azalden AM. 2017. 'Estimation of transit origin destination matrices using smart card fare data'.
Alsger, Azalden, Behrang Assemi, Mahmoud Mesbah, and Luis Ferreira. 2016. 'Validating and improving public transport origin–destination estimation algorithm using smart card fare data', *Transportation Research Part C: Emerging Technologies*, 68: 490-506.
Alsger, Azalden, Ahmad Tavassoli, Mahmoud Mesbah, Luis Ferreira, and Mark Hickman. 2018. 'Public transport trip purpose inference using smart card fare data', *Transportation Research Part C: Emerging Technologies*, 87: 123-37.
Barry, James J., Robert Newhouser, Adam Rahbee, and Shermeen Sayeda. 2002. 'Origin and Destination Estimation in New York City with Automated Fare System Data', *Transportation Research Record*, 1817: 183-87.
Ben-Akiva, Moshe E. 1985. *Discrete choice analysis : theory and application to travel demand* (Cambridge, Mass. : MIT Press: Cambridge, Mass.).
Bera, Sharminda, and KV Rao. 2011. 'Estimation of origin-destination matrix from traffic counts: the state of the art'.
Brenninger-Göthe, Maud, Kurt O. Jörnsten, and Jan T. Lundgren. 1989. 'Estimation of origin-destination matrices from traffic counts using multiobjective programming formulations', *Transportation Research Part B: Methodological*, 23: 257-69.
Carvalho, Luis. 2014. 'A Bayesian Statistical Approach for Inference on Static Origin–Destination Matrices in Transportation Studies', *Technometrics*, 56: 225-37.
Cascetta, Ennio, and Sang Nguyen. 1988. 'A unified framework for estimating or updating origin/destination matrices from traffic counts', *Transportation Research Part B: Methodological*, 22: 437-55.
Çelebi, Dilay, and Şükrü İmre. 2020. 'Measuring crowding-related comfort in public transport', *Transportation Planning and Technology*, 43: 735-50.
Cesario, Frank J. 1973. 'A note on the entropy model of trip distribution', *Transportation Research*, 7: 331-33.
Chen, Su Ping, and Dai Zong Liu. 2016. "Bus Passenger Origin-Destination Matrix Estimation Using Available Information from Automatic Data Collection Systems in Chongqing, China." In *Advanced Materials Research*, 878-89. Trans Tech Publications.
Cho, H. J., Y. J. Jou, and C. L. Lan. 2009. 'Time dependent origin-destination estimation from traffic count without prior information', *Networks and Spatial Economics*, 9: 145-70.
Deng, Qionghua, and Lin Cheng. 2013. 'Research Review of Origin-destination Trip Demand Estimation for Subnetwork Analysis', *Procedia - Social and Behavioral Sciences*, 96: 1485-93.
Djukic, Tamara, CiTG Department of Transport and Planning, Delft University of Technology, 2628CN, Netherlands, Gunnar Flötteröd, Hans van Lint, and Serge Hoogendoorn. 2012. "Efficient real time OD matrix estimation based on Principal Component Analysis." In *2012 15th International IEEE Conference on Intelligent Transportation Systems*, 115-21. IEEE.
Doblas, Javier, and Francisco G. Benitez. 2005. 'An approach to estimating and updating origin–destination matrices based upon traffic counts preserving the prior structure of a survey matrix', *Transportation Research Part B: Methodological*, 39: 565-91.
Fisk, CS. 1988. 'On combining maximum entropy trip matrix estimation with user optimal assignment', *Transportation Research Part B: Methodological*, 22: 69-73.

Fu, Guo Jiang. 2012. 'Study of Solving Crossing Origin-Destination Matrix Based on Entropy Maximizing Model', *Applied Mechanics and Materials*, 182-183: 970-74.
Ge, Qian, and Daisuke Fukuda. 2016. 'Updating origin–destination matrices with aggregated data of GPS traces', *Transportation Research Part C: Emerging Technologies*, 69: 291-312.
Gordillo, Fabio. 2006. 'The value of automated fare collection data for transit planning: an example of rail transit od matrix estimation', Massachusetts Institute of Technology.
Hora, Joana, Teresa Galvão Dias, Ana Camanho, and Thiago Sobral. 2017. 'Estimation of Origin-Destination matrices under Automatic Fare Collection: the case study of Porto transportation system', *Transportation Research Procedia*, 27: 664-71.
Hörcher, Daniel, Daniel J. Graham, and Richard J. Anderson. 2017. 'Crowding cost estimation with large scale smart card and vehicle location data', *Transportation Research Part B: Methodological*, 95: 105-25.
Li, Tian, Dazhi Sun, Peng Jing, and Kaixi Yang. 2018. 'Smart card data mining of public transport destination: A literature review', *Information*, 9: 18.
Mishra, Sabyasachee, Yanli Wang, Xiaoyu Zhu, Rolf Moeckel, and Subrat Mahapatra. 2013. "Comparison between gravity and destination choice models for trip distribution in Maryland." In.
Morphet, Robin. 1975. 'A note on the calculation and calibration of doubly constrained trip distribution models', *Transportation*, 4: 43-53.
Mosallanejad, Mona, Sekhar Somenahalli, and David Mills. 2019. 'Origin-Destination Estimation of Bus Users by Smart Card Data.' in Stan Geertman, Qingming Zhan, Andrew Allan and Christopher Pettit (eds.), *Computational Urban Planning and Management for Smart Cities* (Springer International Publishing: Cham).
Munizaga, Marcela A., and Carolina Palma. 2012. 'Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile', *Transportation Research Part C: Emerging Technologies*, 24: 9-18.
Nassir, Neema, Alireza Khani, Sang Gu Lee, Hyunsoo Noh, and Mark Hickman. 2011. 'Transit Stop-Level Origin–Destination Estimation through Use of Transit Schedule and Automated Data Collection System', *Transportation Research Record*, 2263: 140-50.
Pelletier, Marie-Pier, Martin Trépanier, and Catherine Morency. 2011. 'Smart card data use in public transit: A literature review', *Transportation Research Part C: Emerging Technologies*, 19: 557-68.
Peterson, Anders. 2007. 'The Origin–Destination Matrix Estimation Problem - Analysis and Computations', Dissertations, University of Linköping.
Reilly, William J. 1931. *The law of retail gravitation* (W.J. Reilly: New York).
Shen, Guoqiang, and Saniye Gizem Aydin. 2014. 'Origin–destination missing data estimation for freight transportation planning: a gravity model-based regression approach', *Transportation Planning and Technology*, 37: 505-24.
Sun, Lijun, Jian Gang Jin, Der-Horng Lee, Kay W. Axhausen, and Alexander Erath. 2014. 'Demand-driven timetable design for metro services', *Transportation Research Part C: Emerging Technologies*, 46: 284-99.
TfL. 2018. "A comprehensive multi-rail demand data set for London." In *Project NUMBAT*.
Trépanier, Martin, Nicolas Tranchant, and Robert Chapleau. 2007. 'Individual trip destination estimation in a transit smart card automated fare collection system', *Journal of Intelligent Transportation Systems*, 11: 1-14.

Van Zuylen, Henk J., and Luis G. Willumsen. 1980. 'The most likely trip matrix estimated from traffic counts', *Transportation Research Part B: Methodological*, 14: 281-93.
Wang, Hua, and Xiao Ning Zhang. 2016. "Estimation of Origin-Destination Matrix with Tolling Data." In *Applied Mechanics and Materials*, 239-44. Trans Tech Publications.
Wang, Ming-Heng, Steven D. Schrock, Nate Vander Broek, and Thomas Mulinazzi. 2013. 'Estimating Dynamic Origin-Destination Data and Travel Demand Using Cell Phone Network Data', *International Journal of Intelligent Transportation Systems Research*, 11: 76-86.
Wang, Ning, Monica Gentili, and Pitu Mirchandani. 2012. 'Model to Locate Sensors for Estimation of Static Origin–Destination Volumes Given Prior Flow Information', *Transportation Research Record*, 2283: 67-73.
Wang, Wei, John Attanucci, and Nigel Wilson. 2011. 'Bus passenger origin-destination estimation and related analyses using automated data collection systems'.
Wilson. 1967. 'A statistical theory of spatial distribution models', *Transportation Research*, 1: 5.
Wong, S. C., and C. O. Tong. 1998. 'Estimation of time-dependent origin–destination matrices for transit networks', *Transportation Research Part B: Methodological*, 32: 35-48.
Xie, Chi, Kara M. Kockelman, and S. Travis Waller. 2011. 'A maximum entropy-least squares estimator for elastic origin-destination trip matrix estimation', *Procedia - Social and Behavioral Sciences*, 17: 189-212.
Yap, Menno, Oded Cats, and Bart van Arem. 2018. 'Crowding valuation in urban tram and bus transportation based on smart card data', *Transportmetrica A: Transport Science*: 1-20.

Appendix 1: Solution formulation

The lagrangian relaxation of the different constraints with the corresponding multipliers leads to the following Lagrange function

$$\mathcal{L}(n,\lambda,\mu,\theta,\varphi) = E(n) + \sum_{it}\lambda_{it}\left(\sum_{j} n_{ij}^{t} - O_i^t\right) + \sum_{j\in\Delta,t}\mu_{jt}\left(\sum_{i} n_{ij}^{t-\tau_{ij}} - D_j^t\right) + \theta\left(\sum_{ijt} d_{ij}n_{ij}^{t} - \bar{d}.O\right) + \sum_{l=(s,e)\in L,t}\varphi_{lt}\left(\sum_{\substack{i<l\\ j>l}} n_{ij}^{t-\tau_{is}} - F_l^t\right)$$

The first order optimality condition for the function $\mathcal{L}$ in terms of the variable n_{ij}^{t} is as follows

$$\frac{\partial\mathcal{L}}{\partial n_{ij}^{t}} = 0 \Rightarrow \mathcal{L}(\lambda,\mu,\theta,\varphi) = -log\left(n_{ij}^{t}\right) + \lambda_{it} + \mu_{j,t+\tau_{ij}} + d_{ij}\theta + \varphi_{lt}$$

Note that the multiplier $\mu_{j,t+\tau_{ij}}$ is only included if $j \in \Delta$, i.e., known data on the exit counts at station j. Similarly, φ_{lt} is also only included if $l \in L$, i.e., known flow at link l. Thus, we have the following general solution formulation

$$n_{ij}^{t} = e^{\lambda_{it}+\theta d_{ij}+\varphi_{lt}+\mu_{j,t+\tau_{ij}}} = e^{\lambda_{it}}e^{\theta d_{ij}}e^{\varphi_{lt}}e^{\mu_{j,t+\tau_{ij}}}$$

To sum up, depending on the studied additional data, we have different variants of the solution formulation as presented in **Table 2**.

Appendix 2: Iterative algorithm

The iterative algorithm aims at estimating the multipliers, i.e., $\lambda, \mu, \theta, \varphi$ and ν (full day). For that, each iteration of the algorithm attempts to balance the different constraints until these are satisfied (up to a certain error tolerance ϵ). To derive the algorithms, we first use the (hard) constraints for the origin counts ($O_i^t \neq 0$) to estimate λ_{it} as follows

$$\sum_j n_{ij}^t = O_i^t \Rightarrow \sum_j e^{\lambda_{it}+\theta d_{ij}+\varphi_{lt}+\mu_{j,t+\tau_{ij}}} = O_i^t \Rightarrow e^{\lambda_{it}} = \frac{O_i^t}{\sum_j e^{\theta d_{ij}+\varphi_{lt}+\mu_{j,t+\tau_{ij}}}}$$

With smart card data on counts ($D_j^t \neq 0$) at large destination stations, we use the corresponding constraints to estimate μ_{jt} as follows

$$\sum_i n_{ij}^{t-\tau_{ij}} = D_j^t \Rightarrow \sum_i e^{\lambda_{i,t-\tau_{ij}}+\theta d_{ij}+\varphi_{l,t-\tau_{ij}}+\mu_{jt}} = D_j^t \Rightarrow e^{\mu_{jt}} = \frac{D_j^t}{\sum_i e^{\lambda_{i,t-\tau_{ij}}+\theta d_{ij}+\varphi_{l,t-\tau_{ij}}}}$$

Similarly, the constraints for additional data on the average travel distance $\bar{d}$ can be used to estimate θ by finding the solution (root) of the following equation

$$\sum_{ijt} d_{ij} n_{ij}^t = \bar{d}O \Rightarrow \sum_{ijt} d_{ij} e^{\lambda_{it}+\varphi_{lt}+\mu_{j,t+\tau_{ij}}} \left(e^{d_{ij}}\right)^{\theta} = \bar{d}O$$

When we include additional data on flows ($F_{l=(s,e)}^t \neq 0$) at crowded links, we estimate φ_{lt} by solving the following system of linear (in $e^{\varphi_{lt}}$) equations

$$\sum_{\substack{i<l \\ j>l}} n_{ij}^{t-\tau_{is}} = F_l^t \Rightarrow e^{\varphi_{lt}} e^{\lambda_{st}+\theta d_{se}+\mu_{e,t+\tau_{se}}} + \sum_{\substack{i<l \\ j>l \\ l^*=(i,j)\in L \\ l^*\neq l}} e^{\varphi_{l^*,t-\tau_{is}}} e^{\lambda_{i,t-\tau_{is}}+\theta d_{ij}+\mu_{j,t+\tau_{ij}-\tau_{is}}} = F_l^t$$

The iterative algorithm stops either after a certain number of iterations or when all the constraints are satisfied, e.g., using RMSE and an error tolerance ϵ.