TU Delft

Delft University of Technology

Incorporating Congestion Phenomena into Large Scale Strategic Transport Model Systems

Brederode, L.J.N.

DOI
10.4233/uuid:9363fddf-aeed-4fcc-82bd-23bcced5cc6d

Publication date
2023

Document Version
Final published version

Citation (APA)
Brederode, L. J. N. (2023). *Incorporating Congestion Phenomena into Large Scale Strategic Transport Model Systems*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:9363fddf-aeed-4fcc-82bd-23bcced5cc6d

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

TRAIL

Summary

Strategic traffic assignment (TA) models assess long-term effects of policies on route choices of travelers. To meet stability requirements, current strategic TA models lack modelling of queues. This thesis develops two TA models that include queue modelling whilst satisfying stability requirements along with a method to fuse observed link flows, congestion patterns and -delays. All methods are shown to be applicable in the large-scale strategic application context.

About the Author

Luuk Brederode has worked as a transport model innovator at Goudappel and DAT. Mobility since 2005. In parallel, he carried out his PhD research from 2013 to 2023 at the transport and planning department of Delft University of Technology.

TRAIL Research School ISBN 978-90-5584-330-5

Erasmus University Rotterdam · Radboud University · rijksuniversiteit groningen · TU Delft Delft University of Technology · University of Twente. · TU/e Technische Universiteit Eindhoven University of Technology

TRAIL THESIS SERIES T2023/15

Luuk Brederode Incorporating Congestion Phenomena into Large Scale Strategic Transport Model Systems

Incorporating Congestion Phenomena into Large Scale Strategic Transport Model Systems

Luuk Brederode

Incorporating Congestion Phenomena into Large Scale Strategic Transport Model Systems

Luuk Brederode

Delft University of Technology

Cover illustration by Karien Brederode – Luiting Maten

Incorporating Congestion Phenomena into Large Scale Strategic Transport Model Systems

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus, Prof.dr.ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Friday 27 October 2023 at 12:30 o'clock

by

Lucas Johannes Nicolaas BREDERODE
Master of Science in Civil Engineering, University of Twente, the Netherlands
born in Tilburg, the Netherlands

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	chairperson
Dr.ir. A.J. Pel	Delft University of Technology, Promotor
Prof.dr.ir. S.P. Hoogendoorn	Delft University of Technology, Promotor

Independent members:

Prof.dr. M.C.J. Bliemer	University of Sydney, Australia
Prof.dr.ir. C.M.J. Tampère	Katholieke Universiteit Leuven, België
Prof.dr. C. Osorio	HEC Montréal, Canada
Prof.dr.ir. T. van Vuren	University of Leeds, UK
Prof.dr.ir. B. van Arem	Delft University of Technology
Prof.dr.ir. J.W.C. van Lint	Delft University of Technology, reserve member

Goudappel
MOBILITEIT BEWEEGT ONS

Dat.mobility
ADVANCING ANALYTICS

TRAIL Thesis Series no. T2023/15, the Netherlands Research School TRAIL

TRAIL
P.O. Box 5017
2600 GA Delft
The Netherlands
E-mail: info@rsTRAIL.nl

ISBN: 978-90-5584-330-5

Copyright © 2023 by Luuk Brederode

All rights reserved. No part of the material protected by this copyright notice may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without written permission from the author.

Printed in the Netherlands

“Adding car lanes to deal with traffic congestion is like loosening your belt to cure obesity."

Lewis Mumford (1956)

Preface

Even with seven years of professional experience at Goudappel, little did I know when I started this PhD research back in 2013. Inspired by Michiel Bliemer (who was my direct colleague at that time), and immediately backed up by the company (through Luc Wismans and Wim Korver) closely followed by support from Delft University (through Serge Hoogendoorn and Adam Pel), I started this endeavor. The whole research was set out to last only three (part-time) years, as my first prototypical implementation of STAQ was 'almost' done (yes, its application on the tactical model of Amsterdam indeed produced output that -at first sight- looked somewhat plausible) and I would only need to 'wrap up' the implementation, conduct some convincing case studies and write 'a little booklet' about it. Whether my fellow colleagues knew to what extent I underestimated the challenge ahead remains unknown, but what I do know is that I wouldn't have missed it for the world!

Why? There are many reasons, but the most important one was that during the last decade, I could create, maintain and realize my own roadmap for some of the most important parts of strategic transport model systems in use in the Netherlands, whilst covering the other parts as part of my job at Goudappel and DAT.Mobility. It was also a decade in which I could get a taste of what opportunities academic life has to offer and a decade in which I learned that both academic writing and perseverance are very useful skills in general.

I would like to thank the people that kept believing in me throughout this journey. First and foremost Luc Wismans: you shielded me with a very solid 'management umbrella' on top of your active contributions to the research. Adam and Serge, as promotors and daily supervisor you also never lost faith and provided me with many insights and in-depth discussions, as well as practical suggestions that greatly contributed to the research. From a slightly greater distance, I would also like to thank Wim Korver, Joost de Bruijn, Eric Pijnappels and Jos van Kleef for their continued support.

I would also like to thank the many other people that contributed to the research itself. Firstly, Michiel: thank you for giving me the space to join you after the first conception of STAQ and bring it to the state it is in now. Erik de Romph for your efforts to prioritize the implementation of STAQ in OmniTRANS transport planning software and Mark Raadsen and Jeroen van Oorspronk for actually implementing it in production code. Furthermore, I'm grateful to a bunch of clever interns that I supervised throughout this time period: Arjan van Leeuwen, Robbert van der Kleij, Anton Dijkstra, Berend Steenhuisen, Tanja Hardt, Jelle Neeft, Marije Siemann, Bernike Rijksen, Aswin Nandakumar, Lotte Gerards, Matthew Maat and Glenn Lang: thank you for the fun times and your internship contributions to the development and valorization of the methods developed in this thesis.
Although you contributed in an indirect way, I would also like to thank my uni roommate from the first hour: Erik-Sander Smits. I hope we may continue our periodic discussions on node modelling, the link transmission model, route choice models and any transport-data related topic while enjoying some beers.

I'd also like to thank all members of the doctoral committee for examining my thesis. I am honored to have you all on my defence ceremony and I'm looking forward to your questions and the debate.

Dan de mensen die voor wat sociale afleiding zorgden tijdens mijn PhD-periode. Aan mijn goede vriend en paranimf Ties Brands: bedankt voor onze spitsvondige gesprekken en analyses over hoe de menselijke samenleving (en het openbaar vervoer daarin) werkt en zou moeten werken tijdens onze frequente stedentrips, kroegentochten en andere onverwachte gebeurtenissen. Aan mijn zus en paranimf Liesbeth Brederode: je hebt meestal maar één woord nodig om te weten wat ik bedoel en om ons aan het lachen te maken, terwijl we tegelijkertijd zeer zinvolle gesprekken voeren over het leven op deze planeet (ja, inderdaad, we kunnen het best een hovercraft gebruiken om de afstand tot de maan te bepalen…). En ook dank aan mijn andere vrienden met wie ik hockey speel(de), weekendtrips maakte, zeilde en nog veel meer leuke ervaringen deelde in de afgelopen 10 jaar. Ik weet zeker dat we samen nog veel meer herinneringen zullen creëren!

Voor mijn ouders Ank en Niek: dank voor jullie onvoorwaardelijke en nog steeds voortdurende steun vanaf mijn geboorte, en vooral bedankt dat jullie me alle ruimte en vrijheid gegeven hebben om mijn eigen keuzes te maken in het leven. Ook dank voor jullie tomeloze inzet, creativiteit, liefde en gezelligheid wanneer we met ons gezin bij jullie op bezoek in Frankrijk zijn of omgekeerd. Het is altijd weer vertederend om jullie bezig te zien met Sam en Tijmen. Ik ken geen andere opa en oma van jullie leeftijd die daar zoveel energie in weten te stoppen!

En dan nog de belangrijkste supporters van dit onderzoek: mijn gezin! Sam en Tijmen, ik heb jullie tijdens dit PhD onderzoek een stormachtige ontwikkeling zien meemaken van de schattige baby's (waar je verder 'niets' mee kon) tot de kleine mensjes die jullie vandaag de dag zijn. Jullie zien opgroeien naast mijn baan en een part time promotietraject heeft me geleerd dat alles relatief is in het leven en ik ben ontzettend trots op jullie. Sam, jouw interesse in alles wat staat, zit, beweegt of juist stil zit in combinatie met je tomeloze verlangen om iedereen het naar de zin te brengen is hartverwarmend, terwijl de door mij op jou geprojecteerde wiskunde knobbel mijn verkeerde studiekeuze (ja, het had wiskunde moeten zijn) volledig wegpoetst. Tijmen, jouw duidelijke hang naar zelfstandigheid (je kunt op verdacht veel vlakken al evenveel of meer dan je grote broer) en je vastberadenheid (nee is nee, maar ja is ook echt ja) zijn een mooi contrast tot jouw aanstekelijke lach wanneer we bijvoorbeeld een 'glote belg' doen.

Maar de meeste dank ben ik verschuldigd aan jou, Karien, mijn ongelooflijk mooie vrouw. Bedankt voor al je steun, zorg en liefde, en dat je me hebt weten te overtuigen dat we toch niet 'even' mijn promotietraject gingen afwachten voor we aan kinderen zouden beginnen. Je was altijd bereid om de gaten te vullen die ik achterliet als ik weer eens de drukke wetenschapper wilde uithangen, bood altijd een luisterend oor wanneer ik moest 'wennen' aan de laatste commentaren van de peer-reviewers en ja, je bent inderdaad 'volgens mij wel leuk'…

Luuk
Deventer, juli 2023

Contents

Preface v
Contents vi
Chapter 1 Introduction 1
1.1 Background and motivation 2
1.2 Research objectives 3
1.3 Scope 4
1.4 Methodological contribution 8
1.5 Practical contribution 9
1.6 Outline 10
Chapter 2 Genetics of traffic assignment models for strategic transport planning 11
2.1 Introduction 12
2.2 DNA of traffic assignment models 14
2.3 Gene 1: Spatial assumptions 19
2.4 Gene 2: Temporal assumptions 22
2.5 Gene 3: behavioural assumptions 25
2.6 Classification of existing traffic assignment models 26
2.7 Discussion and conclusions 29
Chapter 3 Static Traffic Assignment With Queuing: model properties and applications 31
3.1 Introduction 32
3.2 Concept and methodology of STAQ 34
3.3 Model implementation 40
3.4 Demonstration of model properties using case study examples 45
3.5 Conclusions and discussion 58
Chapter 4 Extension of a static into a semi-dynamic traffic assignment model with strict capacity constraints 63
4.1 Introduction 64
4.2 From static to semi-dynamic: relaxing the empty network assumption 66
4.3 Solution algorithm 67
4.4 Applications 76
4.5 Discussion 83
4.6 Conclusions and recommendations 85
A. Appendix: two derivations for determination of queue size at link level 87
Chapter 5 Travel demand matrix estimation methods integrating the full richness of observed traffic flow data from congested networks 89
5.1 Introduction 90
5.2 Methodologies 91
5.3 Software 96
5.4 Practical Insights from applications 98

5.5 Conclusion and recommendations 101
Chapter 6 Travel demand matrix estimation for strategic road traffic assignment models with strict capacity constraints and residual queues 103
6.1 Introduction 104
6.2 The Matrix estimation problem for SCCTA models 106
6.3 Solution algorithm 117
6.4 Application on a small network 126
6.5 Application on a large network 135
6.6 Conclusions, discussion and further research 139
A. Appendix: convexity 142
B. Appendix: Discontinuities in flow acceptance factor function 142
C. Appendix: approximated sensitivities on turn level breaking convergence 143
Chapter 7 Conclusions, implications and discussion 147
7.1 Conclusions on the SCCTA model STAQ 147
7.2 Conclusions on the travel demand estimation method 148
7.3 Implications 148
7.4 Discussion 149
7.5 Conclusions on the semi-dynamic version of STAQ 152
Acknowledgements 153
References 154
Summary 166
Samenvatting 171

Chapter 1

Introduction

Since the late 1950s, strategic transport model systems are used to support decision making by assessing the long-term impact of transport policies and land-use scenarios (Castiglione, 2015; Ortuzár and Willumsen, 2011). Strategic transport model systems provide (forecasted) travel patterns by mathematically modelling the underlying behavioral decisions made by travelers before and during their travels. Subsequently, expected long-term impact of scenarios are derived by comparing forecasted travel patterns of the scenarios with a reference (do-minimum) scenario.

Strategic transport model systems consist of different submodels, one for each type of behavioral decision that is included (black boxes in Figure 1.1, left). Only the most important decision types influencing the key indicators used in strategic applications are included. Most model systems only include models for trip or tour frequency, destination, mode and route choices, whereas some of them are extended to include models on e.g. departure time choice, public transport and mobility service subscriptions, car and driver's license possession and relocation.

This thesis focuses on traffic assignment (TA) models that describe route choices of car traffic. TA models confront the travel demand with network supply (digitized networks), determining the routes that travelers choose as well as the resulting traffic state on the network (i.e., traffic conditions, including congestion). A strategic TA model usually imposes user equilibrium (UE, Wardrop, 1952) conditions on its outcomes to facilitate fair comparison of model outcomes for different scenario's. The TA model is often the most computationally expensive component of a strategic model system because an iterative approach is required to impose UE conditions.

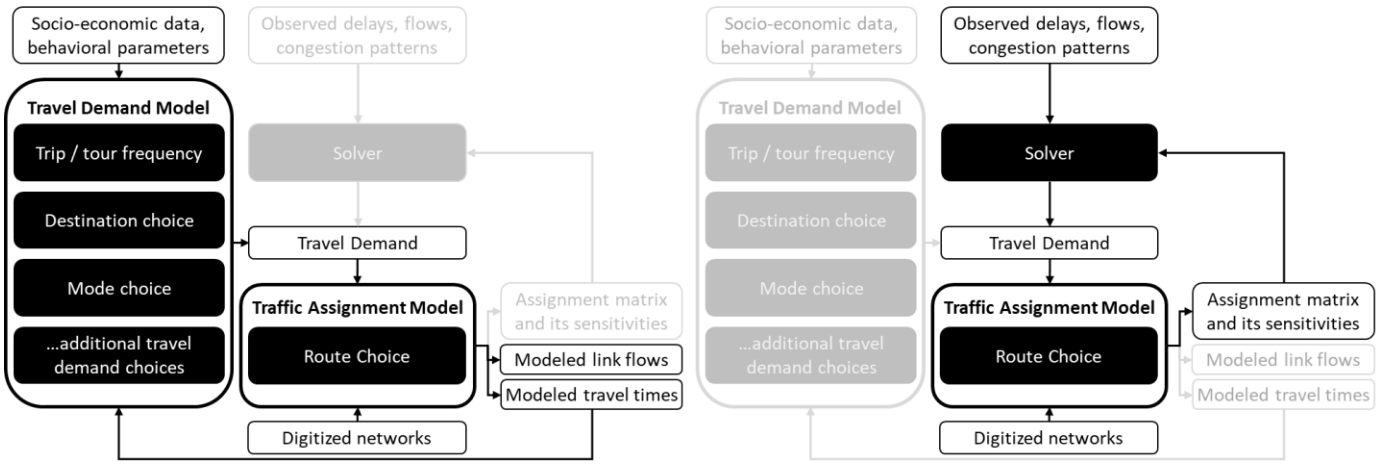


Figure 1.1: the role of a TA model in strategic transport model application (left) and travel demand estimation (right)

There are two use cases for TA models in strategic transport models systems, both of which are subject of this thesis. The primary use case (Figure 1.1, left) is the application of a strategic transport model system to evaluate (policy) scenarios. The secondary use case (Figure 1.1, right) is the estimation of travel demand from observed network data, which is only conducted for a base year (reference scenario) when constructing a new (version of a) strategic transport model system.

In the application context of strategic transport model systems (Figure 1.1, left), TA models are used to determine traffic states (e.g. link flows, congestion patterns, travel times) under UE conditions, given a travel demand profile that is determined by all other choice models in the system, together referred to as 'the travel demand model'. Although in practice often omitted or simplified for computational reasons, travel times from the TA model are fed back into the travel demand models in an iterative fashion, such that (changes in future) network delays are also considered in the choice models within the travel demand model.
In the travel demand estimation context (Figure 1.1, right), TA models provide information to a solver that alters the travel demand, such that it better fits observed flows. In this context, the TA models are used to determine the relationships between the travel demand and link flows (referred to as 'assignment matrices') under UE conditions. The travel demand estimation method proposed in this thesis extends the application range to also include observed (route) delays and observed congestion patterns in the estimation. This requires that also the response of the assignment matrices to changes in demand need to be determined by the TA model.

1.1 On the creation of this thesis

This thesis is a result of part-time research conducted by the author between 2013 and 2023 as part of the innovation program of Goudappel mobility consultants, the Netherlands. Before the start of this research in 2013, the concept behind STAQ was already developed by Michiel Bliemer (also a Goudappel employee at that time) and the author. Furthermore, the first prototypical implementation of STAQ had already been implemented by the author (Brederode et al., 2010). After Michiel Bliemer moved to the University of Sydney, the work with respect to the positioning of the model (ultimately resulting in Chapter 2) was continued in close collaboration, whereas the research described in the other chapters of this thesis was conducted by the author, apart from minor research contributions by Goudappel interns on the semi-dynamic version of STAQ (Chapter 4) and the matrix estimation method (Chapter 6). To further clarify the contributions of the (co-)authors to this thesis, a CRediT author statement is added to the title page of chapters 2 through 6.
This thesis represents only part of the research output, the software implementations of STAQ (Chapter 3), its semi-dynamic counterpart (Chapter 4) and the demand estimation method (Chapter 6) are an equally, if not more important result. At the time of writing, STAQ is included in OmniTRANS transport planning software and is used in eight different Dutch strategic transport model systems, while its semi-dynamic counterpart and the travel demand estimation method have already successfully been applied in pilots on full scale Dutch strategic transport model systems.
To increase value of this thesis for practitioners that are more interested in the capabilities of the software than the methods behind it, subsection 1.4.3 contains a provisional positioning of STAQ compared to TA models with similar functionalities, as a side note to Chapter 2, which methodologically positions STAQ in the field of strategic TA models.

1.2 Background and motivation

For TA models, there is a clear trade-off between model accuracy and model complexity, whereas for application in the strategic context, model stability (i.e.: the satisfaction of UE conditions) is conditional (Table 1.1). This trade-off is discussed below; a more thorough description of the underlying model properties (which are in line with Bliemer et al., 2013; Flötteröd and Flügel, 2015; Flügel et al., 2014) is given in sections 3.1 and 4.1, where these properties form the argumentation for the development of the static capacity constrained traffic

assignment (SCCTA) model and semi-dynamic capacity constrained traffic assignment (S-DCCTA) model proposed in this thesis.
In general, TA model accuracy should be maximized to increase model applicability, whilst at the same time, complexity should be minimized to attain low calculation times, input requirements and high model accountability. Specific to the strategic application context, model stability is a requirement, because when comparing model outputs of different scenarios, we aim to single-out differences only caused by or related to the different scenario inputs. In the context of TA models, this means that (scenario specific) differences due to a lack of UE conditions and/or differences caused by random variables should be negligible or non-existent.

Property	Role in the trade-off	Derivative model properties
Accuracy	Should be maximized	Model applicability
Complexity	Should be minimized	Computational- and input requirements, model tractability and accountability
Stability	Required in strategic context	Comparability of model outputs

Table 1.1 trade-off between desired model properties for strategic TA models.

Driven by e.g., cost-benefit analysis and spatial accessibility studies, the field of use of strategic TA models has shifted from solely forecasting traffic volumes on networks with relatively little detail towards forecasting both traffic volumes as well as travel times on networks with much more detail. Most strategic transport models in use to date use static capacity restrained traffic assignment (SCRTA) models. These models have a low complexity and satisfy the stability requirement but are insufficiently accurate on congested networks as they cannot model queues, leading to inaccurate travel time estimates and flow patterns on congested networks.
Given the structural occurrence of congestion on transport networks around the world, the 'abuse' of SCRTA models as a source for flow and travel time estimates on highly detailed networks has contributed to (legitimate) doubt about the use and even right of existence of transport models and model systems as a whole (e.g., Erhardt et al., 2020; Gordon and Lalanne-Tauzia, 2020; Kager, 2007; L, 2020; NM Magazine, 2015).
With the rise of activity based travel demand models (Castiglione, 2015) some strategic transport model systems have switched to microscopic simulation or macroscopic dynamic capacity and storage constrained traffic assignment (DCSTA) models (Tajaddini et al., 2020) that are sufficiently accurate on congested networks, but have a high level of complexity and do not satisfy the stability requirement (Chiu et al., 2011; Peeta and Ziliaskopoulos, 2001; Szeto and Lo, 2006).
To conclude, the level of accuracy of SCRTA models is no longer sufficient, whereas the complexity and instability of DCSTA models renders them unsuitable for use in strategic applications. At the same time, although the paradigm behind policy making is changing from a predict and provide towards a vision and validate philosophy (e.g., Filippi, 2022; Givoni and Perl, 2020; Lyons and Davidson, 2016; Soria-Lara and Banister, 2018), practitioners argue that the need for strategic TA models with sufficient accuracy in congested conditions, sufficient stability and low complexity will remain (Clerx, 2022; de Graaf, 2021; Hofman, 2018; van Vuren, n.d.).

1.3 Research objectives

The first research objective is to develop a TA model that provides better accuracy in congested conditions compared to SCRTA models, whilst maintaining a relatively low complexity and satisfying the stability requirement for strategic applications. This objective affects both the

model application as well as the travel demand estimation context in (Figure 1.1) Figure 1.1.
The second research objective is to embed the developed TA model from the first objective in a travel demand estimation methodology that exploits the improved TA model accuracy in congested conditions, such that it allows to include observed travel times and congestion patterns. Inclusion of travel times and congestion patterns is relevant for three reasons. Firstly, it allows to identify the conditions in which flows are observed (i.e.: it distinguishes cases where low flow is caused by low demand from cases where flow is reduced due to active bottlenecks), removing the need to impose (possibly incorrect) assumptions with respect to these conditions (Brederode and Verlinden, 2019). Secondly, adding more datapoints reduces the under-specification of the mathematical problem that the travel demand estimator solves (Frederix, 2012); especially observed route queuing delays may reduce the under-specification as it relates different links (as opposed to observed link flows). Thirdly, additional datapoints may improve spatiotemporal coverage and/or density of the observations.

1.4 Scope

This research is scoped by the requirement that both the TA model and the travel demand estimation method should be applicable on real world strategic transport model systems. Further specification of the research scope is derived from the adverb 'real world' (in subsection 1.4.1) and the adjective 'strategic' (in subsection 1.4.2) in this single sentence scope formulation, summarized in Table 1.2.

(derived) model property		#	Criterion
Accuracy	Applicability	R1	TA model is accurate on congested networks containing both highways and urban roads and for different road user classes.
		R4	Demand estimator can handle different data sources and aggregation levels
		S5	Demand estimator is an off-line model
Complexity	Computational requirements	R2	TA model and demand estimator are scalable up to 1.5 million links and 13000 zones
		R3	Computation time for TA model below 16h, Computation time for demand estimator below 3 days
	Input requirements	S6	May be slightly higher than for static capacity restrained TA models, but much lower than macroscopic dynamic TA models
	Accountability	S7	Each model (component) solves an explicitly formulated mathematical problem with known assumptions
	Tractability	S8	TA model consistent with (simplified) kinematic wave theory
		R5	Demand estimator is suitable to identify data inconsistencies
Stability		S1	TA model should adhere to UE conditions
		S2	TA model should not contain randomness
		S3	TA model should be macroscopic
		S4	Demand estimator has a unique solution

Table 1.2: criteria per model property. Subsections 1.4.1 and 1.4.2 discuss criteria with references numbers starting with 'R' and 'S' respectively

Criteria R1 and R4 presented in the table serve, within the model accuracy property, as the objectives to be maximized, whereas the other criteria need to (at least) be met. All criteria values for the (maximum) model complexity and (minimum) stability are specified such that they are at least still acceptable for users that currently employ static capacity restrained TA models.

1.4.1 Real world application criteria

Below, the criteria (R1-R5) that are required for real world applicability are briefly discussed:

- **R1** - Strategic transport model users today are mostly interested in measures targeting different road user classes on networks in or around urban regions where congestion occurs. This means that the TA model needs to be **accurate on congested networks for both highways and urban roads and for different road user classes**. This requires that the TA model explicitly models queues, uses an accurate link model and an explicit junction modelling component (see e.g.: Bezembinder, 2021 and references herein) and allows for user class specific route choice parameters, free flow speeds and network restrictions.
- **R2** - Today, real world strategic transport model systems cover large regions or even countries, whereas their spatial granularity (the level of detail of the digitized networks and zoning system) is relatively high (in the Netherlands a typical strategic transport model contains anywhere between 100K and 1.5million links and between 1500 and 13000 zones). This means that the TA model and demand estimator should be **scalable** to these dimensions.
- **R3** - At the same time, there is a limited amount of **computation time** available. Based upon experience as a practitioner, the main author argues that regular application of strategic transport demand models (including the TA model) has a desired calculation time of sixteen hours at most, such that an overnight run in between two working days is possible, whereas for travel demand estimators, calculation times up to three days are considered acceptable. This means that calculation times put a restriction on the methodologies and algorithms to consider.
- **R4** – Ever more data sources on observed transport network conditions become available. Currently available data includes flows, speeds, densities, and travel times on various levels of aggregation (e.g.: link, node, turning movement and route). The travel demand estimator should therefore be able to use **different data sources on different aggregation levels**. This means that methods that can only use a single type of data source and/or aggregation level are not considered.
- **R5** - Because multiple data sources on multiple aggregation levels are considered, data inconsistencies will occur in practice. This requires a tractable travel demand estimation method that can be used to **identify (clusters of) datapoints that are inconsistent**, such that the model user can choose which data points to remove. This means that only methods that allow to identify inconsistencies (e.g., no heuristics that sequentially handle data points) are considered.

1.4.2 Strategic criteria

Below, the criteria (S1-S8) that are required for applicability in the strategic context are briefly discussed:

- **S1** – The desired stability property that guarantees comparability of model outputs (Table 1.1) requires that differences due to a lack of UE conditions should be negligible or non-existent. This means that (possibly more accurate) TA models that not adhere to **UE conditions** (Wardrop, 1952) are not considered.
- **S2** - For the same stability criterion, TA models that contain **randomness** (due to e.g. random variables or stochastic processes) are not considered.

- **S3** - Although advances are being made with respect to 'frozen' or 'quenched' randomness in micro simulators for travel demand models (Brederode et al., 2020; Engelson et al., 2022; Horni et al., 2011; Zill and Veitch, 2022), application of such methods to TA models adhering to UE conditions are considered too complex. Therefore, only **macroscopic TA models** are considered.
- **S4** – Similar to the TA model, for comparability reasons, the travel demand estimator should solve a mathematical problem for which there's a **unique solution**.
- **S5** - In the strategic context, observed information is available for the entire time period, which means that an **off-line** travel demand estimator is considered (see e.g. Frederix, 2012 and references herein).
- **S6** – Strategic models are used to evaluate scenarios in distant futures, for which model inputs have a high level of uncertainty. Therefore, the **input requirements** (mainly with respect to the specification of network supply for the TA model) may be only slightly higher than for static capacity restrained TA models, but need to be much lower than macroscopic dynamic TA models. Further specification of this criterion will be provided in subsection 3.3.1)
- **S7** – To improve model accountability, the transport model should allow to analyze how differences between modelled scenarios are built up from different (behavioral) mechanisms within the model system (so called 'storytelling'). This requires that each model (component) solves an **explicitly formulated mathematical problem with known assumptions**.
- **S8** – For model tractability, queues and delays from the TA model need to be **consistent with (simplified) kinematic wave theory** (Lighthill and Whitham, 1955; Richards, 1956) or (Newell, 1993) such that outcomes on link level can be verified and compared with observed data.

1.4.3 Methodologies described only in non-peer reviewed publications

As this thesis focusses on methods applicable on real world strategic transport model systems, its methodological contributions are possible competitors for the methodologies that have been embedded in platforms from the major strategic transport modelling software vendors (e.g. Aimsun, Atkins (SATURN), Bentley (CUBE, EMME), Calliper (TransCAD), DAT.Mobility (OmniTRANS), PTV (Visum)). The initial versions of such methods are mostly described in (relatively old) peer-reviewed publications (e.g., Bakker et al., 1994; Bundschuh et al., 2006; the publications in "SATURN 11.6 Manual," n.d., app. C), but once a methodology has been embedded, its documentation mainly lives on in the form of (non-peer reviewed) manuals. Such documents generally only describe (incremental changes and improvements to) user options and sometimes the solution algorithm, without considering (or updating) the underlying mathematical problem formulation from the original publication(s).

Because evaluating the criteria from subsection 1.4.2 based upon such descriptions is hard if not impossible (especially with respect to uniqueness (S4), accountability (S7) and tractability (S8)), and because the credibility of commercial, non-peer reviewed publications is not guaranteed, only the peer-reviewed publications have been considered integrally in literature research conducted for this thesis. This holds for the literature scan on capacity constrained TA models presented in section 3.1 as well as the literature research on semi-dynamic TA models in section 4.2.2, on matrix estimation strategies in chapter Chapter 5 and on TA models and matrix estimation methods in section 6.1.1.

1.4.3.1 Provisional positioning of STAQ compared to methods from non-peer reviewed publications

To increase the practical contribution of this research, as a side note to this thesis, a non-exhaustive scan has been conducted through manuals of software packages containing static TA models that resemble (underlying goals of) STAQ or dynamic TA models being used in a strategic context. This has led to the provisional comparison displayed in Table 1.3, which is further discussed below.

Vendor / software	**Model class**	**Strategic criteria**					
		S1 Adheres to UE conditions	**S2** No Random-ness	**S3** Macro-scopic	**S6** Input require-ments	**S7** Accoun-table	**S8** Tract-able
Atkins / SATURN	'Enhanced' SCRTA	✓	✓	✓	✓		
PTV / assignm. w. ICA							
Rijkswaterst. / QBLOK							
DAT.Mobility / OmniTRANS	SCCTA (STAQ)	✓	✓	✓	✓	✓	✓
	S-DCCTA (S-DTAQ)						
	DCSTA (CTM)		✓	✓		✓	✓
	DCSTA (LTM)						
Bentley / Dynameq	DCSTA (micro/meso)						
PTV / SBA	DCSTA (micro/meso)					✓	✓
Calliper / TransModeller	DCSTA (micro/meso/macro)						

Table 1.3: provisional comparison of STAQ and S-DTAQ with models in considered software

With respect to static TA models resembling (underlying goals of) STAQ, the SatSim/SatEasy modules in SATURN (chapters 7 and 8 of "SATURN 11.6 Manual," n.d.), the 'assignment with ICA' in VISUM ("PTV Visum Help," n.d.) and the QBLOK module in the GM software of Rijkswaterstaat (Significance, 2021) have been considered[1]. From this analysis, the author concludes that to improve accuracy in congested conditions (the first research objective, subsection 1.3) STAQ integrally adopts the SCCTA model class, whereas the considered static TA models remain in the SCRTA model class, but include capacity constrained and/or semi-dynamic components into the underlying cost function. This means that these 'enhanced SCRTA models' dissatisfy the accountability (S7) and tractability (S8) criteria from subsection 1.4.2, the most pronounced example being that these models deliver ambiguous results in the form of flows associated with the cost function (that explicitly considers queues) coexisting with the flows in the SCRTA model component.

With respect to dynamic TA models being applied in the strategic context, only software from Bentley, PTV, Calliper and DAT.Mobility has been considered. Bentley's Dynameq uses microscopic and mesoscopic approaches (hence dissatisfying criteria S2 and S3) and its description explicitly makes no claims about existence or uniqueness of a solution (Mahut and Florian, 2010) thereby dissatisfying criteria S1 and S4. The same holds for PTV's SBA, as this uses the same mathematical model formulation as Dynameq. Similarly, based on ("Caliper Publications," n.d.), the author concludes that TransCAD contains (very efficient) algorithms to solve SCRTA models, but does not contain TA models resembling (underlying goals of) STAQ. Instead, Caliper shifts to microscopic, mesoscopic and macroscopic dynamic models in TransModeller (that dissatisfy criteria S1, S2, S3 and S4) when more accuracy in congested conditions is required. Next, based on extensive application experience with the macroscopic DCSTA models in DAT.Mobility's OmniTRANS the author concludes that the cell

[1]Note that (some version of) STAQ has also been implemented into Aimsun (Casas et al., 2015), but the author could not find any additional documentation on the specifics.

transmission model implementation (Raadsen et al., 2010) dissatisfies the stability (S1) and accountability (S7) criteria, whereas the link transmission model implementation (Raadsen et al., 2016) only dissatisfies the stability criterion. Given the underlying mathematical models (Daganzo, 1994; Yperman, 2007), the author believes that these conclusions will hold for all cell- and link transmission model based implementations.
The author concludes that for users that need improved congestion modelling, but do not require satisfaction of the accountability and tractability requirements and are comfortable interpreting two types of ambiguous model outputs, the considered 'enhanced' SCRTA models are good alternatives to STAQ. With respect to the DCSTA models, the author concludes that their basis in kinematic wave theory and underlying microscopic (car following) models makes them accountable and tractable, but their dynamic nature causes them to lack the low input requirements and stability and non- randomness properties that the author considers required for the strategic application context. From the above, it is concluded that in order to meet all strategic criteria from subsection 1.4.2, an S(-D)CC approach is needed.
Note again that the findings in this subsection are not an integral part of this research in this thesis as they are based on non-peer reviewed literature, the literature scan was not conducted exhaustively and not all manuals were available to the author.

1.5 Methodological contribution

The methodological contribution of this thesis is five-fold:

- **A framework for classification of all macroscopic first order strategic traffic assignment models** is developed. This classification provides a deeper understanding of the often implicit assumptions made in traffic assignment models described in the literature. It further allows for comparing different models in terms of functionality, and paves the way for developing novel traffic assignment models.
- **A complete description of the concept and implementation of the SCCTA model STAQ** is given, along with several model variations, one of which extends it by adding storage constraints. Because STAQ is derived from the dynamic generalized link transmission model (Gentile, 2010), the simplifying assumptions are explicit, providing a solid theoretical basis instead of merely providing a heuristic (Bliemer et al., 2012). Additionally, insights in how the model addresses the shortcomings of static capacity restrained traffic assignment (SCRTA) models and dynamic capacity and storage constrained traffic assignment (DCSTA) models in the strategic context for large congested networks are given using case study examples.
- **A semi-dynamic version of STAQ** is developed by relaxing the assumption in static TA models that the network is assumed to be empty at the start of the study period. This version further improves accuracy whilst maintaining stability and scalability properties required for application on large scale strategic transport model systems. To the best of the authors knowledge, this is the only semi-dynamic TA model that places vertical queues at the correct location (on the upstream node of the link affected by capacity constraint(s)) and removes flow downstream from bottlenecks as part of the assignment model.
- Considering that observed flow values should be interpreted differently depending on (four) different types of network conditions, **three new solution strategies for travel demand estimation on congested networks** are proposed. The new strategies require a capacity constrained TA model and are compared to the reference method from current practice which uses a static capacity restrained TA model and does not consider the effect of network conditions on the interpretation of observed flows.
- **A travel demand estimation method** is developed **using a SCCTA model** (such as STAQ) which combines the favorable properties of SCRTA and DCSTA models whilst allowing for

inclusion of route queuing delays and congestion patterns (from e.g. floating car data) besides the traditional link flows and prior demand matrix. The proposed solution method is robust, tractable and reliable because conditions under which a solution to the underlying optimization problem exist are known and because the problem is convex and has a smooth objective function.

1.6 Practical contribution

With respect to the development of STAQ, this thesis contains the following practical contributions:

- It **describes the methodology and solution algorithm of STAQ** in chapter Chapter 3, which replaces earlier attempts in (Bliemer et al., 2012; Brederode et al., 2010) and is a culmination of partial descriptions in (Bliemer et al., 2014, 2013).
- It **demonstrates the tractability and accuracy of STAQ on small size networks** by showing that all calculations can be done and understood using only the law of flow conservation, the shape of the fundamental diagram and the mathematical specification of the route choice model and by comparing outcomes to its static capacity restrained and dynamic counterparts.
- It **demonstrates the effect on societal benefits** in the context of a (social) cost-benefit analysis by replacing an SCRTA model with STAQ **in a case study on a large scale strategic transport model system:** the strategic transport model of the province of Noord-Brabant, the Netherlands (Heynickx et al., 2016).
- It **demonstrates the robustness and computational efficiency** on one small, two medium and four large scale strategic transport model systems **for all twelve model variations of STAQ**. Each variation is defined by the method used to average the route choice probabilities over iterations to improve convergence, the level of inclusion of junction modelling and whether spillback effects are included in the route choice model.

With respect to the semi-dynamic version of STAQ, this thesis contributes by:

- **Providing a description of the solution algorithm of the semi-dynamic version of STAQ** as an extension of the solution algorithm of its static ancestor, along with methods to derive collective losses and average delays from it outputs from both the networks operator's and traveler's perspective on link, route- and network level.
- **Demonstrating the absolute, temporal and spatial effects on collective loss when replacing an SCCTA model with a S-DCCTA model** using a comparative model application on the large scale strategic transport model system of Noord-Brabant. For this transport model system, this comparison quantifies the effect of the empty network assumption in static TA models.
- **Comparing accuracy and stability of SCCTA, S-DCCTA and DCSTA models on small networks** by looking at the location and length of queues and amount of collective loss and its temporal distribution, and by comparing the course of the adapted relative duality gap over iterations; and
- **Comparing stability and scalability of SCCTA and S-DCCTA models on the large scale strategic transport model system** of Noord-Brabant, by comparing calculation times and the required number of iterations to reach equilibrium, along with an outlook on expected calculation time reduction that could be gained when converting the prototypical implementation into production code.

With respect to travel demand estimation this thesis gives insights on the three proposed solution strategies compared to the strategy mostly employed in current practice based on practical applications. More specific:

- **for the reference strategies and two alternative strategies only brief insights are provided** based on earlier applications described in e.g. (Brederode et al., 2017; Lockwood, 2018; Verlinden and van Grol, 2022); whereas
- **for the (proposed) third strategy extensive findings on the accuracy, computational efficiency and scalability are provided** using applications on the Sioux Falls network (Transportation Networks for Research Core Team, 2019) as well as a large scale strategic transport model system of Noord-Brabant.

1.7 Outline

Figure 1.2 depicts the outline of the remainder of this thesis. Chapter 2 defines a framework to classify strategic macroscopic first order TA models, which contains STAQ (Chapter 3) as well as its semi-dynamic version (Chapter 4) that are part of this thesis. In Chapter 5, three solution strategies for travel demand estimation on congested networks using an SCCTA model are methodologically compared to the (SCRTA model based) method from current practice. An implementation of the third alternative strategy using STAQ is described and extensively tested in Chapter 6. The thesis concludes with Chapter 7 which contains conclusions along with implications for practice and recommendations for further research.
Together, Chapter 2, Chapter 3 and Chapter 4 focus on the first research objective (section 1.3) in the context of the primary use case for TA models: regular application in strategic transport model systems (corresponding to the left part of Figure 1.1), whereas Chapter 5 and Chapter 6 focus on the second research objective in the context of the secondary use case for TA models: the estimation of travel demand from observed network data (corresponding to the right part of Figure 1.1).

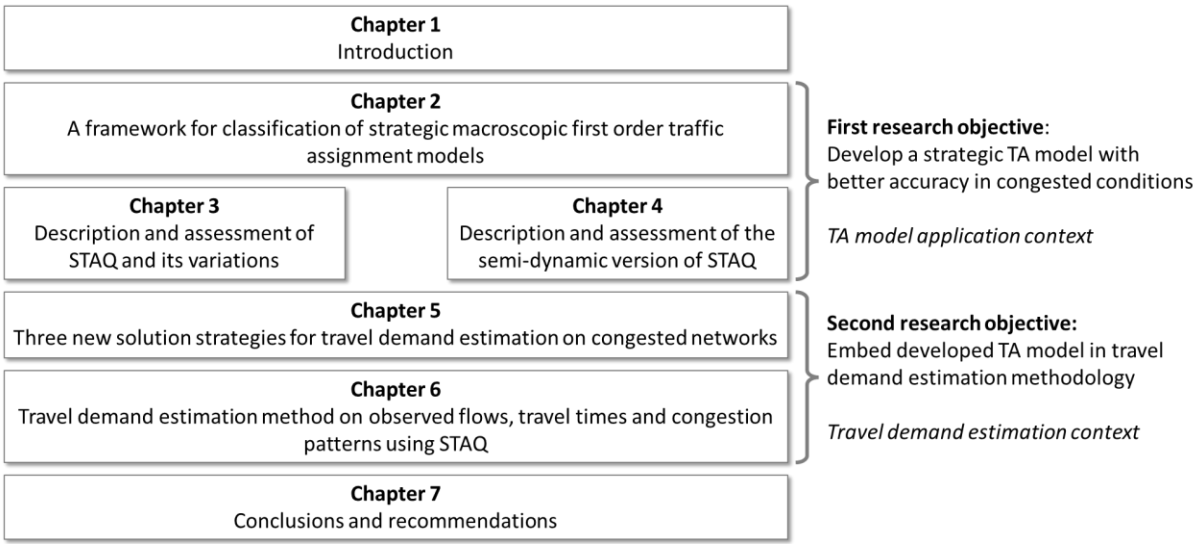


Figure 1.2: thesis outline

Chapter 2

Genetics of traffic assignment models for strategic transport planning

Abstract
This paper presents a review and classification of traffic assignment models for strategic transport planning purposes by using concepts analogous to genetics in biology. Traffic assignment models share the same theoretical framework (DNA), but differ in capability (genes). We argue that all traffic assignment models can be described by three genes. The first gene determines the spatial capability (unrestricted, capacity restrained, capacity constrained, and capacity and storage constrained) described by four spatial assumptions (shape of the fundamental diagram, capacity constraints, storage constraints, and turn flow restrictions). The second gene determines the temporal capability (static, semi-dynamic, and dynamic) described by three temporal assumptions (wave speeds, vehicle propagation speeds, and residual traffic transfer). The third gene determines the behavioral capability (all-or-nothing, one shot, and equilibrium) described by two behavioral assumptions (decision-making and travel time consideration). This classification provides a deeper understanding of the often implicit assumptions made in traffic assignment models described in the literature. It further allows for comparing different models in terms of functionality, and paves the way for developing novel traffic assignment models.

Keywords: Traffic assignment, strategic transport planning, spatial assumptions, temporal assumptions, behavioral assumptions, fundamental diagram, model capabilities

This chapter is a version of the following publication:
Bliemer, M.C.J., Raadsen, M.P.H., Brederode, L.J.N., Bell, M.G.H., Wismans, L.J.J., Smith, M.J., 2017. Genetics of traffic assignment models for strategic transport planning. Transport Reviews 37, 56–78. https://doi.org/10.1080/01441647.2016.1207211

CRediT author statement:
Michiel Bliemer: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Visualization, Project administration, Funding acquisition. **Mark Raadsen**: Conceptualization, Methodology, Investigation, Writing – review & editing. **Luuk Brederode**: Conceptualization, Methodology, Investigation, Writing – review & editing. **Michael Bell:** Writing – review & editing, Funding acquisition**. Luc Wismans**: Writing – review & editing. **Mike Smith**: Writing – review & editing, Resources.

2.1 Introduction

2.1.1 Background

Traffic assignment models are used all over the world in strategic (long term) transport planning and project appraisal to forecast future traffic flows and travel times. Road authorities typically apply traditional models on large scale road networks for this purpose. These models describe the interaction between road travel demand (in particular passenger cars) and road infrastructure supply and were initially developed in the 1950s. The overall structure as depictured in Figure 2.1 has not changed much since (although solution algorithms have become more efficient). Traffic assignment models consist of a route choice sub-model that determines path flows and a network loading sub-model that propagates these path flows through the network and yields travel times. The route choice sub-model has a (possibly time-varying) origin-destination travel demand matrix as input, while the network loading sub-model considers infrastructure characteristics including road segment length, number of lanes, maximum speed, and possibly intersection layout and average green times of traffic controls.

Over the past few decades, there have been many new developments (especially in dynamic network loading models) leading to more advanced traffic assignment models that describe flows and travel times more realistically and (in certain ways) enhance their applicability. Such advancements can be categorised as being spatial, temporal, or behavioural in nature. We will refer to models incorporating such advancements as more capable models that have a larger ability to incorporate phenomena observed in reality.

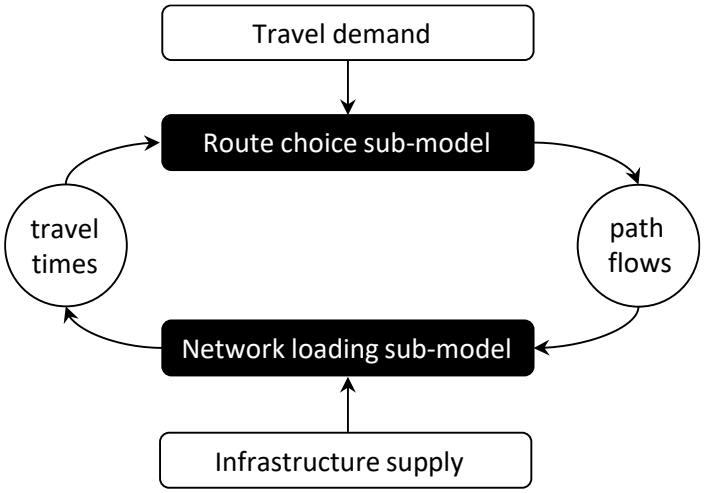


Figure 2.1: Equilibrium between travel demand and infrastructure supply

There exists a wide range of traffic assignment models proposed in the literature, ranging from static to dynamic models, ranging from models that consider only free-flow conditions to models that consider congestion with queuing and spillback, and ranging from all-or-nothing assignment to equilibrium models. These models differ in capabilities, each making their own underlying assumptions.

In this paper we aim to disentangle some of the characteristics of traffic assignment models and explicitly state the assumptions underlying these models. Deeper insights in these assumptions allows a better understanding of the capabilities of each model and the circumstances under which models may reasonably be applied, as well as develop new more capable models.

2.1.2 Scope

In this paper we focus on capabilities of traffic assignment models with a focus on motorised private transport. This means we do not consider public transport or active modes of transport

(such as walking and cycling). We would like to point out that "capability" is only one aspect when selecting suitable models for strategic transport planning. There are many other relevant aspects, such as ease of use (i.e., short run times, easy calibration, low input requirements), accountability (convergence of algorithms, existence and uniqueness of solutions, model complexity), and robustness (i.e., does the model generate stable outcomes). It is for example likely that a highly capable model has a higher computational complexity and less favourable solution properties, so a transport planning analyst should always balance these aspects when choosing a suitable model. We refer to (Bliemer et al., 2013) for a more general discussion on these requirements for traffic assignment models.

We narrow the scope of this paper further by making the following eight limiting assumptions: (i) macroscopic description of traffic flow, (ii) only first order effects are considered, (iii) only pre-trip route choice is considered, (iv) no day-to-day dynamics are considered, (v) individual travellers are guided by selfish (non-cooperative) behaviour, (vi) inelastic travel demand, (vii) only a single user class is considered, and (viii) only travel time is considered in route choice.

The first five assumptions are made because the focus is on traffic assignment models for strategic transport planning purposes, which in general do not consider mesoscopic or microscopic representations of traffic flows (with possible random components), ignore dynamical second order effects (such as capacity drop, stop-and-go waves, and hysteresis), do not consider en-route travel decisions (which are more relevant for short term traffic operations), do not consider learning processes and disequilibria (partly due to difficulties when comparing base and future scenarios), and does not consider system optimal conditions (which can be useful for network design).

The last three assumptions are made to restrict ourselves to core components of traffic assignment models in which we assume a given travel demand (and do not include departure time choice, mode choice, destination choice, or other travel choices influencing demand) for a single user class (passenger cars) considering only travel time (and do not include tolls, travel time reliability, parking costs, etc.). These last three assumptions can be relaxed and are not strictly necessary for our framework, but they allow a more focussed presentation of the concepts in this paper. For example, one can replace travel time with a generalised cost or (dis)utility function that includes travel times and travel costs. Further, multiple user classes can be taken into account by considering different sensitivities to time and cost in these generalised cost functions (e.g., people with a high or low willingness to pay for travel time savings). Taking different vehicle types into account in a macroscopic model is usually more challenging due to asymmetric interactions between for example cars and trucks (see e.g. Bliemer and Bovy, 2003), which is partly why modellers often choose to convert all vehicle types into passenger car units.

2.1.3 Genetics

In this paper we describe the 'genetics' of traffic assignment models, which allows us to describe and characterise models in a qualitative fashion. Although the various traffic assignment models proposed in the literature may seem very different and sometimes incompatible, they share the same DNA and can be seen as descendants of the same ancestors having different genes.

In biology, DNA is the blueprint of life that consists of instructions that control the functions of cells. Each species (e.g., humans) shares more or less the same DNA. The building blocks of DNA are called nucleotides, which store genetic information. Genes describe basic functions of living organisms and consist of a specific sequence of nucleotides. The genetic code therefore describes all characteristics of the organism. DNA is inherited from parents through recombination, and evolves through mutation (i.e., genetic variation).

Traffic assignment models can be thought of as being characterised by a genetic code containing model assumptions and genes that describe functionality. Each traffic assignment model for strategic transport planning shares the same theoretical framework (namely, DNA). We identify three different genes: (i) a gene that describes spatial interactions, (ii) a gene that describes temporal interactions, and (iii) a gene that describes behaviour. These genes are composed of nucleotides that delineate each individual assumption that impacts on the functional capability of the model. By combining different temporal, spatial, and behavioural assumptions, different traffic assignment models are created.

A very capable organism with many positive characteristics is sometimes said to have 'good genes'. Advanced traffic assignment models may be thought of as having 'better' genes than their simpler traditional counterparts regarding realism. An organism is defined by physical appearance and its behaviour, both defined by genes.[2] In strategic macroscopic models, the network loading sub-model can be seen as a physical process in which traffic flow is modelled as a fluid following hydrodynamic theories. While traffic flow is a result of underlying individual driving behaviour (e.g., speed choice and lane choice), this level of behaviour is not described by macroscopic models; instead it is aggregated to a physical relationship through a cost function or the fundamental diagram of traffic flow (see subsections 2.2.1 and 2.3.2). Thus, the network loading sub-model is physical in nature and described by a spatial and temporal gene. In contrast, the route choice sub-model describes a behavioural process and is described by a, third, behavioural gene.

Just like living organisms, traffic assignment models have evolved over time, often by small mutations in one of the underlying assumptions, sometimes by recombination of existing models into a new model. By discovering basic underlying assumptions of each model (genetic code), we can investigate model functionality and limitations, as well as propose improved models. It also allows genetic modifications of existing models to develop novel models.

2.1.4 Paper outline

In section 2.2 we describe the DNA of traffic assignment models, which allows us to classify each traffic assignment model. Section 2.3 describes the first gene using four nucleotides that represent the spatial assumptions. Section 2.4 describes the second gene, consisting of two nucleotides that represent the temporal assumptions. Section 2.5 discusses the third gene, consisting of two nucleotides representing behavioural assumptions. Section 2.6 establishes the genetic code for a selection of traffic assignment models proposed in the literature based on the spatial, temporal, and behavioural assumptions. Finally, we draw conclusions in section 2.7 and state some potential for new model development.

2.2 DNA of traffic assignment models

In the literature, the main distinction that is often made between models is with respect to temporal assumptions, i.e. whether a model is static or dynamic. Dynamic models are typically seen as superior over static models. However, in terms of spatial interactions, certain static models are capable of accounting for queues and even spillback while certain dynamic models may not. Also, regarding the underlying route choice behaviour, some simple static models may be more advanced than certain dynamic models. We therefore need a more elaborate classification of traffic assignment models that describes their characteristics and capabilities in greater detail.

[2] Although there is debate in the literature whether behaviour is determined by genes or by the environment (or both), in biology the field of study called behavioural genetics examines the origins of individual differences in behaviour.

In this section we propose a unified theoretical framework (DNA) for traffic assignment models. This classification leads to model types and capabilities that result from three different genes that describe spatial, temporal, and behavioural assumptions, see Figure 2.2. Details of these underlying assumptions will be discussed in sections 2.3, 2.4 and 2.5.
Gene 1 describes the assumptions regarding spatial interactions, resulting in four distinct model classes (see subsection 2.2.1). Gene 2 describes the assumptions regarding temporal interactions, resulting in three model classes (see subsection 2.2.2). Finally, Gene 3 describes the behavioural assumptions, leading to three model classes (see subsection 2.2.3). Combining the different model classes, the framework in Figure 2.2 describes in total 36 different model types, each with their own capabilities. The most capable model type according to this framework is a dynamic capacity and storage constrained equilibrium traffic assignment model, while the least capable model type is a static unrestrained all-or-nothing traffic assignment model. Each less capable model type is a special case of a more capable model type. In other words, less capable models can typically be derived from more capable models by making simplifying assumptions.

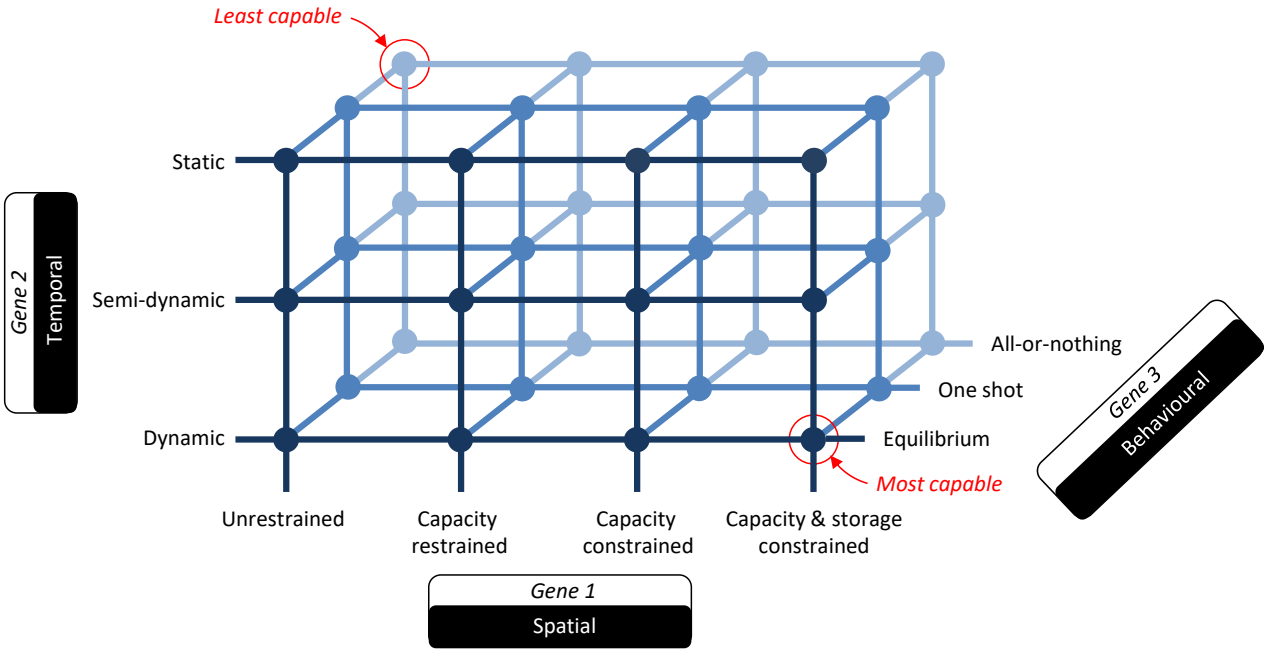


Figure 2.2: DNA of traffic assignment models

2.2.1 Model classes and capabilities resulting from spatial assumptions

As a result of spatial assumptions (Gene 1), the following model types are distinguished (in increasing order of capability):
Unrestrained models;

- Capacity restrained models;
- Capacity constrained models;
- Capacity and (queue) storage constrained models.

The most capable traffic assignment models are models that constrain both the capacity (of flow) and the storage (of queues) on road segments. These models ensure that flow does not exceed capacity by diverting traffic to routes with spare capacity or by buffering vehicles in a physical queue. If the length of the queue exceeds the length of the road segment, the queue will spillback to upstream road segments. A capacity constrained model is a special case in which there are no constraints on the (queue) storage and as such spillback does not occur. An even more simplified model class is the capacity restrained model. In this model class, flows can exceed the physical road capacity and therefore queues are not described explicitly. To

mimic the effect of queues (in these models) travel times simply increase with increasing levels of flow. Finally, the simplest and least capable model is an unrestrained model with fixed (usually free flow) travel conditions and travel times.
Capacity restrained models are the most common model class in strategic transport planning, although the use of capacity (and storage) constrained models is gaining in popularity. Unrestrained models are rarely used. Each model class has different capabilities and a particular model should ideally only be used in cases where the underlying spatial assumptions are valid; however, as remarked above, there are many other factors which may influence model choice.
Figure 3 indicates a fundamental diagram describing the theoretical relationship between flow and density that can be empirically observed from traffic counts, and depends, among other things, on the number of lanes, the maximum speed limit, and the road type. Such a fundamental diagram may be assumed to hold for each cross-section on a homogeneous road segment (and is independent of the length of the road segment). Each point on this diagram represents a specific steady-state traffic state.[3] While the diagram only shows flows (veh/h) and densities (veh/km), the speed of a vehicle (km/h) can be determined using the fundamental relationship that (space-mean) speed equals flow divided by density. For low densities (indicated by A and B in Figure 2.3) there is no congestion and no queues appear. Such traffic states are called hypocritical states (below the critical density) in which flow increases with density (i.e., throughput increases with more vehicles on the road). High densities (indicated by C and D) are a result of congestion and queues on the road. These traffic states are called hypercritical states in which flow decreases with density (i.e., throughput deteriorates with more vehicles on the road). The jam density provides an upper bound on the number of vehicles that can be stored on a certain road segment (assuming zero speed). For more information on the fundamentals of traffic flow theory and the fundamental diagram we refer to e.g. (Cascetta, 2009).

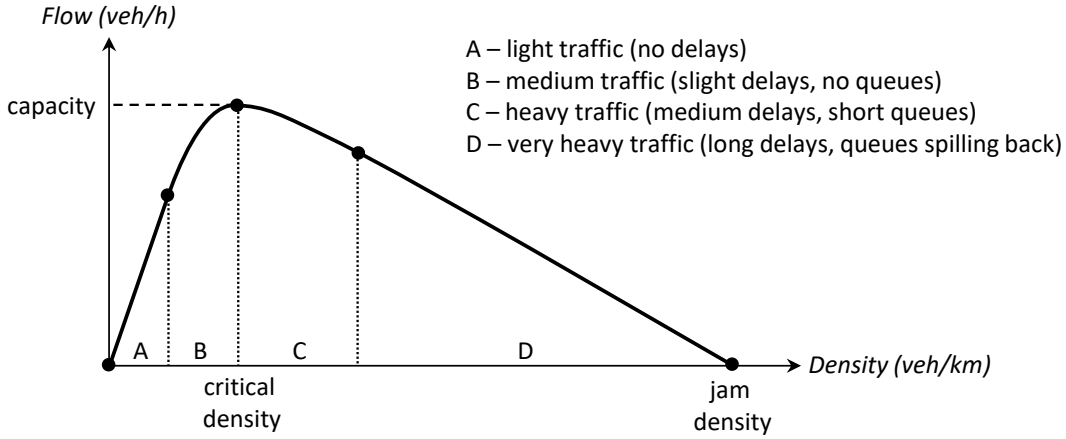


Figure 2.3: Spatial assumptions and model capabilities

Unrestrained models are only suitable for light traffic conditions (A) in which flow increases linearly with density, indicating that vehicles drive at maximum speed. Capacity restrained models are only suitable for light to medium traffic conditions (A and B) in which the flow does not exceed capacity, but some slight delays may occur due to increasing density. These models do not describe the hypercritical part of the fundamental diagram. Capacity constrained

[3] In other words, this relationship only describes first order effects and does not explicitly describe transitions between traffic states (which requires explicit modelling of braking and acceleration as second order effects). As mentioned in Section 1.2, second order effects are usually not considered in large scale strategic transport planning for tractability reasons, but also to avoid illogical behaviour such as negative flows and traffic going backwards as outlined by Daganzo (1995b).

models are suitable for light to heavy traffic conditions (A, B, and C) in which short queues can form.[4] These models cannot describe queues longer than the length of the road. Most capable is a capacity and storage constrained model, which can be applied to all traffic conditions (A, B, C, and D); including very heavy traffic when queues can grow longer than the road length and spillback to upstream road segments occurs.
Section 2.3 describes the underlying assumptions of these model classes in more detail.

2.2.2 Model classes and capabilities resulting from temporal assumptions

As a result of temporal assumptions (Gene 2), the following model types can be distinguished (in increasing order of capability):

- Static models;
- Semi-dynamic models;
- Dynamic models.

Dynamic models consider time-varying travel demand and multiple time periods for route choice and within each time period there exist (smaller) time steps for network loading in which flows are propagated through the network. These models explicitly account for variations over time in path flows, link flows, and travel times, and are the most capable models considered. Semi-dynamic models are special cases that only consider part of the dynamics. They often consider only a single time step for network loading within each route choice period, but may propagate traffic flows between route choice periods. Finally, static models are the simplest and least capable models that consider a stationary travel demand and only a single time period (with a specified or unspecified duration) for both route choice and network loading.
Some models are referred to as quasi-dynamic, which can be confusing. Quasi-dynamic models only consider a single time period and do not explicitly model time-varying flows. As such, these models are essentially static; they may be thought of as static models with certain dynamic elements (such as queues), see (Miller et al., 1975) and (Payne and Thompson, 1975). Due to lack of a formal definition, we define quasi-dynamic models as static models that impose capacity and/or storage constraints and thereby can explicitly account for queues (similar to more advanced dynamic models).
Static models are the most common model class adopted for strategic transport planning purposes, although semi-dynamic models are used in some countries. Dynamic models are increasing in popularity, but applications for strategic planning purposes remain in minority due to the much higher model complexity and related needed computation times. As before, model classes defined by temporal assumptions have different capabilities and should ideally only be used in cases where these assumptions are valid; however, as remarked above, there are many other factors that may influence model choice.
Figure 2.4 illustrates how static, semi-dynamic, and dynamic models represent travel demand. The solid red line indicates the actual travel demand for a single origin-destination pair, and the grey bars represent the average demand in the model during each period. The areas of the grey bars (indicating the number of vehicles) are equal to the areas underneath the demand curves.
A static model considers a single time period, typically consisting of an entire peak period (e.g., a three hour period from 6.30am till 9.30am), and assumes that traffic outside this time period does not influence flows or travel times in the considered period. In other words, traffic in different periods can be assigned separately. Route choice proportions are assumed stationary during this period and network loading also considers a single time period in which all traffic reaches the destination and link flows are interpreted as average flows during this period.

[4] Note that the line that separates traffic conditions C and D in Figure 2 is plotted somewhat arbitrary between the critical density and jam density since it is case specific, i.e. depends on the inflow rate and the link length.

In a semi-dynamic model multiple time periods are considered (e.g., one hour time slices, such as periods 6-7am, 7-8am, 8-9am, and 9-10am). It can be seen as a sequence of static models, however it takes the result from a previous period (such as vehicles in a queue) into account, for example by passing on residual traffic to the next period. As such, semi-dynamic models are more capable in describing travel demand variations as well as interactions of vehicles across time periods. Route choice proportions are assumed stationary during each time period, while network loading within each time period is usually done in a simple fashion similar to a static model. However, this typically does include the limitation that vehicles cannot be propagated for more than the duration of each period. In other words, vehicles that do not reach their destination within a single time period may be transferred to the next time period.
Dynamic models are capable of describing interactions between vehicles within and across each time period. They usually consider many smaller time periods (e.g., time slices of 15 minutes), which allows them to more accurately represent time-varying travel demand. Route choice proportions are typically assumed stationary during each time period. Network loading is much more sophisticated and similar to simulation models, i.e. they typically consider small time steps (e.g., 1 second) in which vehicles are propagated through the network.
Section 2.4 describes the underlying assumptions of these model classes in more detail.

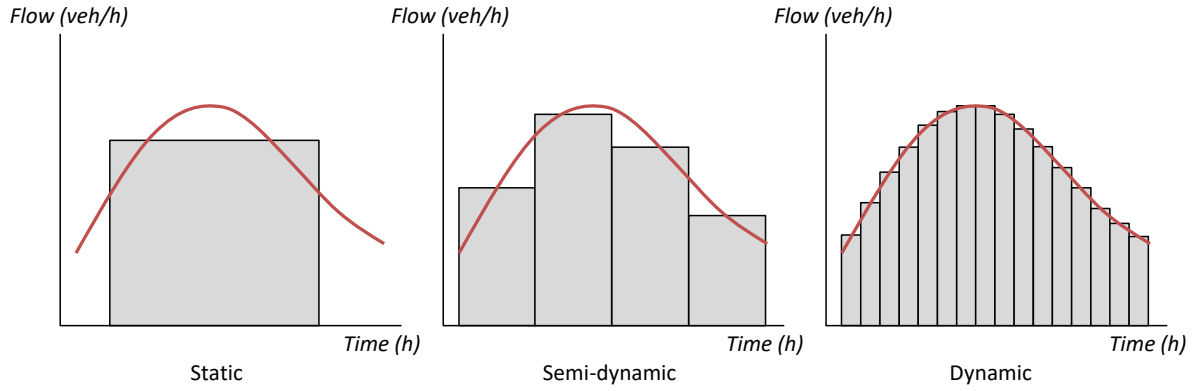


Figure 2.4: Temporal interaction assumptions and model capabilities

2.2.3 Model classes and capabilities resulting from behavioural assumptions

As a result of behavioural assumptions (Gene 3), the following model types can be distinguished (in increasing order of capability):

- All-or-nothing models;
- One shot models;
- Equilibrium models.

Equilibrium models are the most capable models in which travellers consider congested travel times when choosing their route. In an equilibrium state, often referred to as a user equilibrium in which travellers are assumed to be non-cooperative (i.e., exhibit selfish behaviour), no traveller can unilaterally change routes to improve his or her travel time (Wardrop, 1952). This is in contrast to system optimal models that assume travellers cooperate and minimise the total (or average) travel time in the system. In this context, in this paper, only user equilibrium models are considered. One shot models are simplified models in which there is no feedback from previous travel time experience but rather a single network loading is performed based on initial path flow proportions. Such path flow proportions are either pre-determined or based on instantaneous travel times considering current traffic conditions. Finally, the simplest and least capable is an all-or-nothing model that is a special case of a one shot model in which all travellers follow the fastest route based on given (typically free-flow) travel times.

Each of these model classes can be further differentiated into deterministic and stochastic models. Deterministic models usually assume perfect information, such that travellers base their decisions on actual travel times. In contrast, stochastic models assume imperfect information, such that travellers make decisions based on perceived travel times (Daganzo and Sheffi, 1977). Equilibrium models are the most widely used model class in strategic transport planning, while system optimal assignments are mainly used to provide a benchmark solution. One shot models are often applied to simulate traffic using a more advanced (dynamic) network loading model based on route choice proportions from a simpler (static) model. All-or-nothing assignments, static or time-dependent, are not that common (anymore), but are often sub-models in equilibrium models.
Section 2.5 describes the underlying assumptions of these model classes in more detail.

2.3 Gene 1: Spatial assumptions

The first gene represents the spatial assumptions, which describe how traffic flows in network loading spatially interact and directly impact on the realism of the model (see also Figure 2.2). These spatial interactions are a combination of assumptions on the link level (shape of the fundamental diagram, capacity and storage constraints), and the node level (turn flow restrictions yielding turn reduction factors). These spatial interactions have been analysed separately or jointly in the literature and can be calibrated empirically.
The four specific assumptions (nucleotides) within this gene are summarised in Table 2.1 and are discussed in more detail in the following subsections. The nucleotide level refers to the spatial level at which interactions are described. The spatial assumptions of a traffic assignment model can be indicated using a sequence of letters representing the genetic code. For example, the most widely used assignment model for strategic transport planning purposes is a static capacity restrained model with the following code for Gene 1: CN-UU-U-N. The most sophisticated and capable model according to this classification is defined by genetic code CC-CC-C-F.

Nucleotide	Level	Type	Code	Explanation
Shape of the fundamental diagram	Link	Hypocritical	L, P, Q, C	Linear / Piecewise linear / Quadratic / Concave
		Hypercritical	L, P, Q, C, H, V, N	Linear / Piecewise linear / Quadratic / Concave / Horizontal / Vertical / Not available
Capacity constraints	Link	Inflow	U, C	Unconstrained / Constrained
		Outflow	U, C	Unconstrained / Constrained
Storage constraints	Link		U, C	Unconstrained / Constrained
Turn flow restrictions	Node		F, O, N	First order / Other / No restrictions

Table 2.1: Genetic code for Gene 1 (spatial assumptions)

2.3.1 Nucleotide 1 – Shape of the fundamental diagram

All traffic assignment models explicitly or implicitly assume a fundamental diagram. The shape of the fundamental diagram plays an important role in traffic flow theory and different shapes lead to different traffic patterns on a link (some more realistic than others). We indicate the maximum flow through any part of a homogeneous road segment by physical road capacity C,

also referred to as the saturation flow, which depends among other things on the number of lanes and the speed limit. The inflow and outflow capacity, however, are at best equal to C and in many cases lower. For example, the outflow capacity may be restricted due to traffic controls and competing traffic (e.g., a merge) and the inflow capacity may be restricted due to spillback of a downstream bottleneck. This does not influence the fundamental diagram itself, but rather means that only specific traffic states on the diagram are observed in practice.

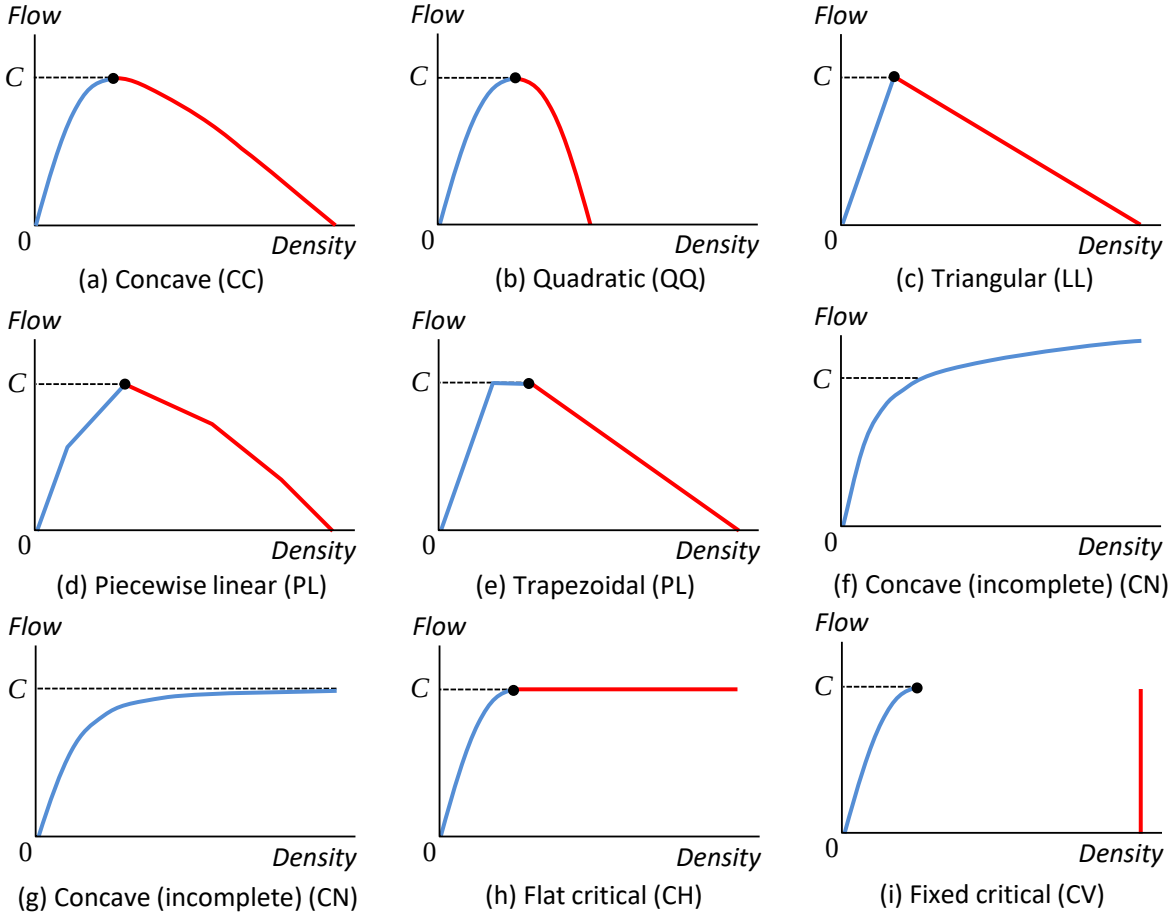


Figure 2.5 Shapes of the fundamental diagram

The fundamental diagram is generally defined by an increasing concave hypocritical branch (for densities lower than the critical density, indicated in blue in Figure 2.5, consistent with traffic conditions A and B in Figure 2.3) and a decreasing concave hypercritical branch (for densities higher than the critical density, indicated in red in Figure 2.5, consistent with traffic conditions C and D in Figure 2.3). The shape of such a general function can be indicated by CC using the coding from Table 2.1.

The first fundamental diagram was described by (Greenshields, 1935). He proposed a linear relationship between speed and density, which results in a quadratic fundamental diagram QQ, see Figure 2.5(b). Such a symmetric fundamental diagram may describe hypocritical traffic conditions quite accurately, but performs poorly for hypercritical states. A popular choice in traffic flow theory due to computational advantages has been an asymmetric triangular fundamental diagram LL (Newell, 1993) as shown in Figure 2.5(c). While a linear relationship in the hypercritical branch is often considered sufficiently realistic, a linear relationship in the hypocritical branch is less realistic (since it assumes that the speed at capacity is equal to the maximum speed). Therefore, piecewise linear fundamental diagrams PP as shown in Figure 2.5(d) have been proposed (e.g., Yperman, 2007), which maintain many of the computational

benefits. A special case of such a piecewise linear fundamental diagram is the trapezoidal fundamental diagram (Daganzo, 1994) shown in Figure 2.5(e).
Diagrams shown in Figure 2.5(a)-(e) result in models with physical queues since they have a downward sloping hypercritical branch, while the diagrams in Figure 2.5(f)-(g) do not result in any queues since the hypercritical branch is absent. Other shapes of the hypercritical branch of the fundamental diagram have been proposed that result in specific types of queues. A fundamental diagram with a horizontal hypercritical branch as shown in Figure 2.5(h) is consistent with a model with vertical (non-spatial) queues, while a vertical hypercritical branch as shown in Figure 2.5(i) yields a model with horizontal (spatial) queues in which all queues are assumed to have a fixed queuing density, either equal to the jam density (leading to very compact queues) or some other fixed queuing density (Bliemer, 2007).
Fundamental diagrams have been used extensively in more advanced capacity and storage constrained dynamic traffic assignment models; in contrast, static models have mainly relied on link performance functions (also called volume-delay functions or travel time functions or cost-flow functions), which describe the relationship between link travel time and link flow (volume) or between speed and flow ((Branston, 1976) reviews link performance functions). The most well-known link performance function is the BPR link performance function (Bureau of Public Roads, 1964) . The corresponding fundamental diagram that is implicitly assumed is plotted in Figure 2.5(f). Two things can be observed from this CN shape. First, the BPR function gives rise to only the hypocritical branch of the fundamental diagram and ignores the hypercritical branch. Secondly, the hypocritical branch increases beyond the physical road capacity *C*, making it suitable only for capacity restrained models. Another popular choice in capacity restrained models is the conical link performance function proposed by (Spiess, 1990), which exhibits less rapid increases in link travel times when flows exceed capacity.
(Davidson, 1966) proposed a specific function in which the travel time goes to infinity as the flow approaches capacity (as suggested by Beckmann et al., 1956). Such a function is called a barrier function and guarantees that flows do not exceed the road capacities, hence this function can be used in a capacity constrained model. The corresponding fundamental diagram is shown in Figure 2.5(g) in which the hypocritical branch has a horizontal asymptote at capacity. However, this model may give rise to computational problems and perhaps unrealistic travel times when flow approaches capacity. Several others have discussed modifications to eliminate these problems (e.g., Akcelik, 1991; Daganzo, 1977; Taylor, 1984).
Link performance functions have also been used in several dynamic models (e.g., Bliemer and Bovy, 2003; Friesz et al., 2013; Janson, 1991; Ran and Boyce, 1996) in which travel times are calculated for vehicles at the time of link entrance (based on the flow at link entrance or all flows that previously entered or exited the link). These computed travel times, also referred to as predictive travel times, are then used to calculate the link exit times for flow propagation. Such link performance functions cannot realistically describe flows and travel times under (very) heavy traffic conditions (at densities C and D in Figure 2.3) since these functions do not represent the hypercritical branch of the fundamental diagram and do not explicitly describe queues.

2.3.2 Nucleotide 2 – Capacity constraints

Some models consider capacity constraints, while others assume no upper bounds on traffic flows. In case no constraints on the link entrance and exit flows are assumed, i.e., UU in Table 1, no queues build up. This is consistent with fundamental diagrams of the shape shown in Figure 2.5(f). When considering both link entrance and exit capacity constraints, i.e. CC, these are typically set to the single physical link capacity *C*. In this case, residual queues will form upstream the bottleneck link. Some models consider UC, in which only link exit capacities are considered. In other words, flow is not restricted to flow in, but is restricted when flowing out.

Such an assumption leads in some situations to queues inside the bottleneck link. Finally, models can also consider CU with link entrance capacity constraints and no explicit outflow constraints.

2.3.3 Nucleotide 3 – Storage constraints

When the number of vehicles in a queue exceeds the available link storage, the queue will spill back to upstream links. The theoretical maximum number of vehicles that can be physically stored on a link should be equal to the jam density times the link length, although in moving queues (with a density lower than the jam density) the number of vehicles that can be present on the link is much lower. Some models do not consider spillback, thereby implicitly assuming no storage constraints (U). This essentially means an infinite jam density, which is consistent with the fundamental diagram presented in Figure 2.5(h). Models that take storage constraints into account (C) have a finite jam density, consistent with the fundamental diagrams in Figure 2.5(a)-(e) and Figure 2.5(i).

2.3.4 Nucleotide 4 – Turn flow restrictions

Given that queues and travel time delays mainly arise due to interactions at the node level (i.e., merges, intersections), it is perhaps surprising to see that many static traffic assignment models and some dynamic models completely lack a node model description. In case there are no capacity constraints on the link entrance or exit flows, queues will never occur and hence a node model can often be omitted (N). In addition to node models (or sometimes instead of node models), junction models can be used to calculate additional delays per turn and may also impose turn capacities as well (based on junction configurations and controls).

In the presence of capacity constraints, node models determine the turn flows at intersections, merges, and diverges. (Tampère et al., 2011) describe requirements for a first order node model for a node with any number of incoming and outgoing links. These requirements include holding free solutions (Jabari, 2016; inaccurately called flow maximisation in earlier papers on macroscopic node models), non-negativity, satisfying demand and supply constraints, satisfying the conservation of turn fractions (CTF) and the invariance principle (Lebacque and Khoshyaran, 2005). Merge constraints that follow the capacity based weighted queuing rule (Ni and Leonard, 2005) satisfy the invariance principle, in which the outflow rates are capacity proportional in case both in-links are congested. An often used merge constraint that does not satisfy the invariance principle is the fair merging rule in which inflow rates are demand proportional (Jin and Zhang, 2003)

(Bliemer, 2007) combines a first-in-first-out diverging rule and the fair merging rule into a closed form demand proportional model for general cross nodes. Several node models for general nodes have been proposed in the last decade (e.g., Jin, 2012a, 2012b; Jin and Zhang, 2004), none of them satisfy both CTF and the invariance principle and are therefore classified under other turn flow restrictions (O). More recently, models have been proposed that satisfy all requirements for first order node models (F), including CTF and the invariance principle, see e.g. (Flötteröd and Rohde, 2011; Gibb, 2011; Smits et al., 2015; Tampère et al., 2011).

2.4 Gene 2: Temporal assumptions

In this section we consider temporal assumptions in network loading identified in the second gene. Temporal assumptions determine whether a model is static, semi-dynamic, or dynamic. These assumptions consider interactions within time periods (wave speeds and vehicle propagation speeds) as well as across time periods (residual traffic transfer). They can be used to remove or simplify time dynamics within the model.

The three specific assumptions (nucleotides) within this gene are summarised in Table 2.2 and are discussed in more detail in the following subsections. Note that the level refers to the temporal level (within-period or across periods) at which the interactions are described. The temporal assumptions for traditional static models can be described by the following code for Gene 2: IN-IN-N. The most capable dynamic model is defined by genetic code KK-VV-T.

Nucleotide	Level	Type	Code	Explanation
Wave speeds	Within	Hypocritical	K, V, I	Kinematic / Vehicular / Infinite
		Hypercritical	K, I, Z, N	Kinematic / Infinite / Zero / Not available
Vehicle propagation speeds	Within	Hypocritical	V, I	Vehicular / Infinite
		Hypercritical	V, I	Vehicular / Infinite / Not available
Residual traffic transfer	Across		T, N	Transfer / No transfer

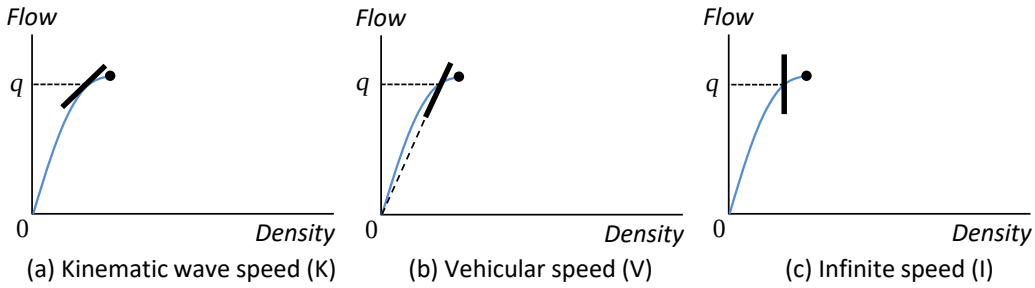
Table 2.2: Genetic code for Gene 2 (temporal interaction assumptions)

2.4.1 Nucleotide 5 – Wave speeds

Temporal interactions on a network are described by wave speeds as well as vehicle propagation speeds. Wave speeds are used to propagate traffic states through the network while vehicle propagation speeds describe how vehicles move through the network. Vehicle propagation speeds are discussed in the next nucleotide.

We first consider wave speeds in the hypocritical branch (i.e., forward waves). In the first order kinematic wave model proposed by LWR (Lighthill and Whitham, 1955; Richards, 1956), traffic conditions travel at the kinematic wave speed (K) equal to the slope of the hypocritical branch of the fundamental diagram as shown in Figure 2.6(a) for traffic flow rate q. It is important to realise that the speeds at which traffic states propagate and the speeds at which vehicles are propagated through the network are in general not the same. In case of a concave hypocritical branch, the kinematic wave speed is always smaller than (or equal to) the vehicular speed (V), which is equal to the flow divided by the density and hence equal to the slope of the line connecting the origin to the traffic state as shown in Figure 2.6(b). Only if the hypocritical branch is linear, these speeds are equal. More recent dynamic models consider kinematic wave speeds, but especially earlier dynamic models and semi-dynamic models consider vehicular speeds.

All static models simplify the within-period interactions by implicitly assuming infinite forward wave speeds (I) in which traffic states instantaneously propagate through the network and reach their destination within the single period. This situation is illustrated in Figure 2.6(c). This assumption effectively removes the necessity (and possibility) to track traffic states over time.



(a) Kinematic wave speed (K) (b) Vehicular speed (V) (c) Infinite speed (I)

Figure 2.6 Speeds in hypocritical branch

Backward waves track how traffic states in the hypercritical branch propagate backwards on a road segment, and are responsible for queue build up and possible spillback to upstream road segments. In the LWR model traffic conditions travel at the (negative) kinematic wave speed (K) equal to the slope of the hypercritical branch of the fundamental diagram as shown in Figure 2.7(a) for traffic state *q*. Similar to forward waves, it requires a dynamic model to explicitly deal with the effects of such backward kinematic waves over time.

An unconstrained static model gives rise to a fundamental diagram, which does not have a hypercritical branch; and so backward wave speeds are not available (N). A capacity constrained static model however does give rise to a hypercritical branch. In these fundamental diagrams two different temporal assumptions regarding backward waves can be made (since the time dimension does not exist in a static model). The most widely adopted assumption is that backward wave speeds are zero (Z) as shown in Figure 2.7(b). In this case, traffic conditions never move backwards, which usually means vertical non-spatial queues and no spillback. (Note that stationary physical queues are also consistent with zero backward wave speeds.) The zero speed assumption is consistent with fundamental diagrams of the shape shown in Figure 2.5(h). Another assumption is that there is a (negative) infinite speed (I) as depicted in Figure 2.7(c); this allows the model to describe spillback when the number of vehicles in the queue exceeds the available link storage. Note that an infinite backward wave speed does not mean that queues build up indefinitely, since the length of the queue is constrained by the number of vehicles in the queue. The fundamental diagram in Figure 2.5(i) is consistent with the infinite speed assumption.

2.4.2 Nucleotide 6 – Vehicle propagation speeds

Instead of looking at the speeds at which traffic states propagate, we now look at the assumption on the speed with which vehicles propagates on a road segment. As mentioned in the previous section, traffic states and vehicles in general do not move at the same speed.

In the hypocritical branch, traffic states and vehicles both move forward, but vehicles never move slower than traffic states (see Figure 2.6). In static models, the vehicle propagation speed is assumed to be infinite (I) such that vehicles move instantaneously through the network within a single time period. Note that although vehicles are propagated instantaneously in static models, this does not mean that the travel times are zero, since travel times are calculated separately from vehicular speeds. In contrast, dynamic models consider finite vehicular speeds (V), such that travel times are consistent with vehicle propagation speeds.

Traffic states in the hypercritical branch (if considered in the model) move upstream (i.e., have a negative speed), while vehicles move downstream (i.e., have a positive speed), see Figure 2.7.

In dynamic models the vehicle propagation speed is assumed to be equal to the finite vehicular speed (V). In static models that do not describe residual queues the vehicle propagation speed is implicitly assumed to be infinite (I), however, in static models that consider residual queues, the vehicle propagation speed is assumed to be finite and set to the vehicular speed (V). Note that this does not make the model dynamic since it only requires applying capacity and storage constraints to traffic flows instead of explicitly tracking vehicles over time.

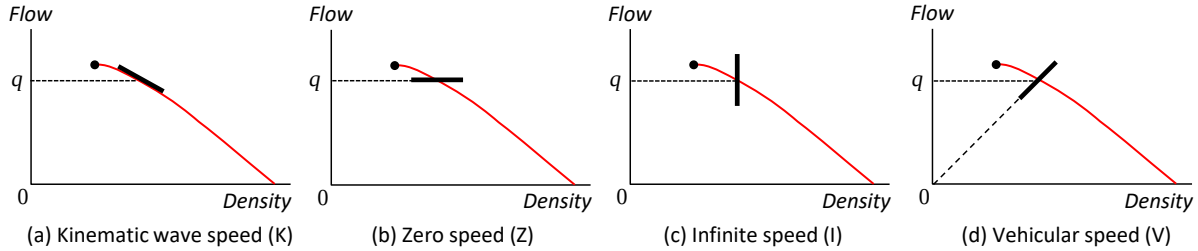


Figure 2.7 Speeds in hypercritical branch

2.4.3 Nucleotide 7 – Residual traffic transfer

Residual traffic at the end of a time period results when vehicles are not able to reach their final destination within the considered time period (or the smaller network loading time step). These residual vehicles are either (i) in a residual queue due to a bottleneck downstream, or (ii) simply are not able to reach their final destination because the travel time to reach the destination is longer than the considered time period. Residual traffic influences traffic flows and travel times in the next time period. This dependency of traffic across time periods can be eliminated by assuming that any residual traffic has no impact on the next time period, in other words, assuming that the network is empty at the beginning of each time period.

Dynamic models transfer all traffic (T), thereby describing the full temporal interactions within and across time periods. Static models have just one (fairly long) time interval and so do not consider residual traffic transfer (N). Thus static models are unsuitable for modelling short time periods in a congested network. The main difference between static and semi-dynamic models is that the latter does assume residual traffic transfer across time periods as discussed in subsection 2.2.2.

2.5 Gene 3: behavioural assumptions

The third and final gene represents the behavioural assumptions, which describe travellers' route choice. From biology we know that describing which genes affect behaviour is difficult, since behavioural characteristics are complex and polygenic (i.e., influenced by multiple genes). The same holds for describing route choice behaviour in traffic assignment models, and many types of behaviours have been described in the literature.

In this section we put route choice behaviour into a single gene with two nucleotides as summarised in Table 2.3 and discussed in more detail in the following subsections. We note that while we try to be as inclusive as possible, this list is not exhaustive and is limited by the scope set out in subsection 2.1.2 (for example, we do not consider day-to-day learning effects). The most capable model considered is a (equilibrium) model with the following code for Gene 3: BI-E, while the simplest model is a (all-or-nothing) model defined by genetic code FP-I.

Nucleotide	Type	Code	Explanation
Decision making	Rationality	F, B	Full / Bounded
	Information	P, I	Perfect / Imperfect
Travel time consideration		I, P, E	Instantaneous / Predictive / Experienced

Table 2.3: Genetic code for Gene 3 (behavioural assumptions)

2.5.1 Nucleotide 8 – Decision making

Decision making behaviour has many dimensions. We limit ourselves to the ones that have most often been used in the context of route choice, namely rationality, uncertainty, and motivation.

In terms of rationality, most traffic assignment models consider full rationality (F) which assumes that travellers consider all alternatives and eventually all travellers select their own best routes. In reality, travellers are unlikely to behave in such an optimal way due to resistance in change (inertia effects) and the fact that people often minimise effort and time in decision making. Bounded rationality (B) is a term that is often used to describe such decision making behaviour, which includes habitual route choice, or route choice in which travellers expose satisficing behaviour and consider routes with travel times sufficiently close to the fastest route travel time (Di et al., 2013; Han et al., 2015).

If travellers have perfect information (P), then decision making can be described by a deterministic process. In contrast, if travellers are considered to have imperfect information (I) with a given level of uncertainty, then decision making is referred to as probabilistic or stochastic. For example (Fisk, 1980) proposed a stochastic assignment model that adopts a logit model, (Zhou et al., 2012)adopt a C-logit model, and (Kitthamkesorn and Chen, 2013) adopt a path-size weibit model, where the latter two aim to correct the path choice probabilities for path overlap. Deterministic models can be seen as special cases of stochastic models where the level of uncertainty is equal to zero.
Although outside of the scope, we point out that travellers may be driven by different motivations for choosing a certain route. As stated in subsection 2.1.2, here we only consider selfish drivers who minimise their individual travel time leading to a user equilibrium based model. Other models exist in which drivers are guided by different motivations, yet these models are hardly ever used in the context of strategic transport planning.

2.5.2 Nucleotide 9 – Travel time consideration

In (semi-)dynamic models, different types of path travel times can be considered in route choice, see e.g., (Ran and Boyce, 1996) and (Buisson et al., 1999). Instantaneous path travel times (I) for a certain departure time consider only the traffic states at this time instant and the corresponding link travel times, and hence ignores any changes in traffic conditions while driving. Models that consider instantaneous travel times are often referred to as reactive. Predictive path travel times (P) consider the addition of link travel times based on the traffic conditions at the time of link entrance, hence time-varying traffic conditions along the path are taken into account. Such travel times can be considered as an estimate, since changing traffic conditions while traversing the link are ignored. More recent models calculate experienced travel times (E), which consider the actually experienced link travel times at the time of link exit (instead of link entrance). In static models (in which no such differences in path travel times exist) we assume that travel times are instantaneous.

2.6 Classification of existing traffic assignment models

Many traffic assignment models have been proposed in the literature that we can classify using the nine nucleotides in the three genes. Table 2.4 provides a list of some prototypical models described in the literature, which is by no means intended to be complete.
Looking at temporal assumptions, all static models assume infinite wave and vehicle propagation speeds in the hypocritical branch and no residual traffic transfer. In case a hypercritical branch of the fundamental diagram is considered, either zero or infinite backward wave speeds are assumed, and vehicle propagation speeds equal to vehicular speeds or infinity. On the other hand, all dynamic models assume forward wave speeds that are not infinite, i.e. either equal to the vehicular speed or kinematic wave speed. Backward wave speeds are equal to the kinematic wave speeds and follow the shape of the fundamental diagram (and can therefore be equal to zero or infinity if the hypercritical branch of the fundamental diagram is horizontal or vertical, respectively). Vehicle propagation speeds are equal to the vehicular speed in both the hypocritical and the hypercritical branch (if considered). Further, dynamic models assume residual traffic transfer.
Regarding behavioural assumptions, all models in Table 2.4 are (user) equilibrium models. Exceptions are (Bovy, 1990) who describes a one shot model for uncongested situations, while (Bliemer, 2007; Daganzo, 1995, 1994; Gentile, 2010; Yperman et al., 2005) mainly describe the network loading sub-model and omit behavioural route choice information.
Finally, looking at spatial assumptions, many models are capacity restrained using a strictly increasing link performance function, although more recently several capacity constrained

models have been proposed that can explicitly account for queues. Relatively few models are storage constrained in which spillback is described. A wide variety of shapes of fundamental diagrams has been used. More advanced models include turn flow restrictions through the incorporation of a node model, which allow more realistic queueing and spillback of traffic.
Semi-dynamic models are neither completely static nor completely dynamic. This means with respect to the temporal assumptions that they typically assume a sequence of connected static models as described in (Nakayama and Connors, 2014). In such a case, wave speeds and vehicle propagation speeds in the hypocritical branch are infinite. However, vehicle propagation speeds in the hypercritical branch are considered finite and vehicles that reside in a queue at the end of a time period are transferred to the next time period. We have omitted semi-dynamic models from the list in Table 4 because the papers are either in Japanese (Akamatsu et al., 1998; Fujita et al., 1989, 1988; Miyagi and Makimura, 1991; Nakayama, 2009) or have been described as operational procedures and algorithms rather than mathematical problems (e.g., Davidson et al., 2011; Van Vliet, 1982), which makes them difficult to classify accurately.

Table 2.4: Overview of assumptions made in different traffic assignment models proposed in the literature

	Gene 1: Spatial assumptions				*Gene 2: Temporal assumptions*			*Gene 3: Behavioural assumptions*	
	fundamental diagram	capacity constraints	storage constraints	turn flow restrictions	wave speeds	vehicle prop. speeds	residual traffic transfer	decision making	travel time consideration
Static models									
Bovy (1990)	LN	UU	U	N	IN	IN	N	FI	I
Beckmann et al. (1956)	CN	UC	U	N	IN	IN	N	FP	I
Irwin et al. (1961)	CN	UU	U	N	IN	IN	N	FP	I
Fisk (1980)	CN	UU	U	N	IN	IN	N	FI	I
Smith (1987)	LH	UC	U	N	IN	IN	N	FP	I
Bell (1995)	LH	UC	U	N	IZ	II	N	FI	I
Bifulco and Crisalli (1998)	CH	UC	U	N	IZ	IV	N	FI	I
Lam and Zhang (2000)	CH	UC	U	N	IZ	IV	N	FP	I
Zhou et al. (2012)	CN	UU	U	N	IN	IN	N	FI	I
Smith (2013)	LH	UC	U	N	IZ	II	N	FP	I
Smith et al. (2013)	CV	UC	C	N	II	IV	N	FP	I
Bliemer et al. (2014)	LC	CC	U	F	IZ	IV	N	FI	I
Dynamic models									
Janson (1991)	CN	UU	U	N	VN	VN	T	FP	I
Daganzo (1994, 1995a)	PL	CC	C	O	KK	VV	T	--	--
Chen and Hsueh (1998)	CN	UU	U	N	VN	VN	T	FP	P
Li et al. (2000)	LH	UC	U	N	KZ	VV	T	FP	I
Chabini (2001)	CN	UU	U	N	KN	VN	T	FP	P
Bliemer and Bovy (2003)	CN	UU	U	N	KN	VN	T	FP	P
Yperman et al. (2005)	LL	CC	C	O	KK	VV	T	--	--
Bliemer (2007)	CV	UC	C	O	KI	VV	T	--	--
Gentile (2010)	CC	CC	C	O	KK	VV	T	--	--
Friesz et al. (2013)	CH	UC	U	N	KZ	VV	T	FP	E
Han et al. (2015)	LL	CC	C	O	KK	VV	T	BP	E

2.7 Discussion and conclusions

In this paper we have presented a theoretical framework, which classifies traffic assignment models for strategic transport planning purposes. This framework is described in terms of a genetic code with three genes and nine nucleotides consisting of four spatial assumptions, three temporal assumptions, and two behavioural assumptions. This framework leads to in total 36 different model types, each with their own underlying assumptions and their own capabilities.
As a special case, the widely applied capacity restrained equilibrium static traffic assignment model can be derived by assuming (i) a concave hypocritical part and no hypercritical part of the fundamental diagram, (ii) no flow capacity constraints, (iii) no storage constraints, (iv) no turn flow restrictions, (v) infinite forward wave speeds and no backward waves, (vi) infinite vehicle propagation speeds, and (vii) no residual traffic transfer, (viii) perfectly rational travellers with full information, and (ix) instantaneous travel time consideration. Such strict assumptions limit the capability and hence realism of this particular model in certain instances. At the same time, we acknowledge that more capable models often have other less favourable characteristics, such as higher computational complexity and possible non-uniqueness of solutions. As a result, transport planners may decide to choose less capable models, but should be aware of model limitations when interpreting outputs.
Capacity constrained (and possibly also storage constrained) models are more capable and can explicitly describe queues (and possibly spillback). Several sophisticated dynamic models exist that are capable of describing flows and travel times under all traffic conditions. Such static models also exist, which extend the capability (realism) of static models in congested situations by sharing the same spatial assumptions made in advanced dynamic models. This opens up possibilities for static models that are derived from advanced dynamic models by simply using static temporal assumptions. Therefore, the framework described in this paper may not only be useful for classifying models, but also for developing new models with new genetic codes by combining different spatial, temporal, and behavioural assumptions (and hence inherit genetic properties from other models).

Chapter 3

Static Traffic Assignment With Queuing: model properties and applications

Abstract
This paper describes the road traffic assignment model Static Traffic Assignment with Queuing (STAQ) that was developed for situations where both static (STA) and dynamic (DTA) traffic assignment models are insufficient: strategic applications on large-scale congested networks. The paper demonstrates how the model overcomes shortcomings in STA and DTA modelling approaches in the strategic context by describing its concept, methodology and solution algorithm as well as by presenting model applications on (small) theoretical and (large) real-life networks. The STAQ model captures flow metering and spillback effects of bottlenecks like in DTA models, while its input and computational requirements are only slightly higher than those of STA models. It does so in a very tractable fashion, and acquires high-precision user equilibria (relative gap < 1E-04) on large-scale networks. In light of its accuracy, robustness and accountability, the STAQ model is discussed as a viable alternative to STA and DTA modelling approaches.

Keywords: STAQ, Traffic assignment, strategic planning, large scale, congested networks, model, static

This chapter is a version of the following publication:
Brederode, L., Pel, A., Wismans, L., de Romph, E., Hoogendoorn, S., 2019. Static Traffic Assignment with Queuing: model properties and applications. Transportmetrica A: Transport Science 15, 179–214. https://doi.org/10.1080/23249935.2018.1453561

CRediT author statement:
Luuk Brederode: Conceptualization, Methodology, Software, Investigation, Writing – original Draft, Writing – Review & editing, Visualization, **Adam Pel**: Conceptualization, Writing – Review & editing, **Luc Wismans**: Conceptualization, Writing – Review & editing, **Erik de Romph**: Writing – Review & editing, **Serge Hoogendoorn**: Writing – Review & editing

3.1 Introduction

Since the late 1950's strategic transport models are used to assess the long-term impact of transport policies and the design and management of transport systems. Since then, road traffic congestion has become a common and structural part of many transport systems around the world. However, strategic transport models differ strongly with respect to how such structural congestion and the effects thereof are accounted for within the model used in the traffic assignment (TA) step. The TA model uses the travel demand and network supply as input and usually solves a user equilibrium (UE) problem determining the routes that travelers choose as well as the resulting traffic state on the network (i.e. traffic conditions, including congestion). The TA model is often the most computational expensive component of the model system because an iterative approach is required to solve the UE problem. Given the above, we argue that there is a need for computationally efficient TA models in strategic transport models for large-scale transport systems with structural congestion.

From the combined perspectives of policy makers and TA model users, the authors argue that apart from computational efficiency and the ability to accurately capture the effects of structural congestion, TA models should also be based on input data that can be forecasted with sufficient certainty for (distant) future years, and should produce accurate, robust and accountable model results for all vehicle classes and for both urban roads and motorways upon assessing policy, design and management measures for transport systems. These desired properties for TA models are in line with (Bliemer et al., 2013; Flötteröd and Flügel, 2015; Flügel et al., 2014) and are defined below.

The ability to capture congestion effects pertains to how bottlenecks lead to flow metering and spillback as well as how it affects route choice. Robustness and accountability are desired properties, because when comparing model outputs of different scenarios (e.g. sets of policy measures), we aim to single-out differences only caused by or related to the different scenario inputs. Hence, differences caused by random variables (e.g. due to stochastic processes in the model) or because the model output does not (sufficiently) represent a stable system state should be negligible or non-existent. Accountability also means that it should be possible to pinpoint and size the contribution of each of the different model components in terms of scenario outputs. This requires model components that can be isolated and that are mathematically tractable (i.e.: all calculations can be verified given the theory behind it). Finally, computational efficiency, low input requirements and applicability allow for fast calibration and application of the model on any network. These desired properties for TA models within large-scale strategic transport models are summarized in Table 3.1 for later reference.

A quick-scan of strategic transport model systems of large urban areas in Western Europe shows that in general two types of traffic assignment models are being used. Most strategic transport model systems use traditional static traffic assignment (STA) models (e.g. Paris, Berlin, Amsterdam, Lisbon, Vienna, Copenhagen, Rotterdam, The Hague). These models are computationally efficient, have low input requirements and are robust, tractable and accountable. However, they are not sufficiently accurate (and thus applicable) in congested conditions, because they do not capture flow metering and spillback effects due to congestion (Flötteröd and Flügel, 2015). To the best of the authors' knowledge, there are three (quasi dynamic) traffic assignment models in use in strategic transport model systems that, to some extent, capture flow metering and spillback effects: QBLOK (Bakker et al., 1994) used solely in the Dutch national models system, Saturn (Hall et al., 1980) used in e.g. London Highway Assignment Models, and the blocking back assignment in PTV VISUM (Bundschuh et al., 2006) used in e.g. the UK west midlands PRISM model and Flemish strategic traffic models.

Although they provide more accuracy than traditional STA models, all three models suffer from a solid theoretical basis, as they are merely presented as algorithms, while the underlying mathematical problem formulation and assumptions are not specified. This leads to poor accountability and makes calibration of parameters using observed data cumbersome and model-specific. Furthermore, queues and delays predicted by these models are not consistent with (simplified) kinematic wave theory (Lighthill and Whitham, 1955; Richards, 1956; or Newell, 1993), causing poor mathematical tractability.

Property	**Definition**
Tractability	The extent to which calculations in each model component can be verified using the theory behind the component or submodel
Accuracy under congested conditions	The extent to which flow metering, spillback and route choice effects caused by congestion are included in the model
Accountability	The extent to which different model components can be isolated and verified
Robustness (1)	The extent to which the model is free from random variables that affect its outcomes
Robustness (2)	The extent to which the model converges to a defined and meaningful stable state
Computational efficiency	The extent to which run times and memory requirements are acceptable for calibration and application of large scale models
Input requirements	The extent to which input requirements are available with acceptable uncertainty for distant future scenarios
Applicability	The extent to which the model is applicable for all vehicle classes and for both urban roads and motorways

Table 3.1: desired properties and criteria for traffic assignment models within large-scale strategic transport models

Over the last decades, there has been much emphasis on development of dynamic traffic assignment (DTA) models and their application in the operational (and sometimes tactical) context. However, as suggested by (Peeta and Ziliaskopoulos, 2001; Szeto and Lo, 2006; Transportation Research Board, 2014) DTA models lack the convergence properties that are needed for applications within the strategic context. This means that the robustness and accountability of these models is insufficient to be used in strategic transport model systems. Indeed, researchers and practitioners state that a duality gap value (DG, the metric most used to measure the level of disequilibrium) of 1E-04 or lower is needed in strategic context (Boyce et al., 2004; Brederode et al., 2016a; Caliper, 2010; Han et al., 2015), whereas, to the best of the authors' knowledge, no DTA algorithms exist that can converge to this level on realistic congested networks of reasonable size. Furthermore, the existence of a time dimension within DTA models is a major contributor to their high computational cost and memory usage and therefore limited scalability. The time dimension also causes DTA models to require much more input data in comparison with STA models, because demand-matrices (or demand models that can deliver these), traffic counts and route choice parameters (may) become time dependent. This input data is often not available, especially for longer-term scenarios (i.e. 5-20 years into the future). A quick-scan of DTA models in the strategic context (especially in the US) confirms that these model applications are all relative small-scale (<1300 centroids) and most do not converge well.

Based on these considerations, we first of all argue that traditional STA models are insufficiently accurate to be applied on strategic transport model systems with structural congestion, whereas the accountability and robustness of existing quasi-dynamic assignment models is questionable, and their calibration cumbersome due to the lack of a solid theoretical basis. Second of all, we argue that DTA models are sufficiently accurate to describe congestion effects, but their low computational efficiency, high input requirements and poor robustness

and accountability prohibit application in large-scale strategic models. To overcome these shortcomings, STAQ (Static Traffic Assignment with Queuing): an assignment model for road traffic within strategic transport models was developed as an alternative to the traditional STA model, providing more accuracy on congested networks without reducing robustness, applicability and accountability and without increasing input requirements, whilst keeping computational requirements to acceptable levels. This makes the model suitable for applications where both static and dynamic assignment models may fail, i.e. strategic applications on large-scale congested networks.
STAQ consists of two submodels, both consisting of several components. For each component variations are possible which, combined, result in a large set of possible implementations of STAQ. We shall first describe the concept, methodology and implementation using STAQ in its most accurate (or 'reference') form. Thereafter, the role and performance of variations applied in this paper will be described (from subsection 3.2.5 onwards). All model variations represent simplifications, thus leading to lower accuracy, but at the same time benefitting from equal or higher tractability, accountability, robustness or efficiency, or equal or lower input requirements compared to the reference form.
The mathematical problem formulation of STAQ, its theoretical advantages over STA and DTA models as well as earlier versions of its solution algorithm have been described before by (Bliemer et al., 2013, 2012; Brederode et al., 2010). Since the most recent publication, there have been a few minor methodological improvements, and various STAQ variants have successfully been tested and put to practice on several large-scale real-life strategic models. Now that mathematical development and conceptual testing of the model is completed, this paper focuses on the key aspects of practical model applications. The main contributions of the paper are 1) to provide a complete and up-to-date description of the model concept, methodology and implementation, 2) to (explicitly) show how, and to what extent the model addresses the shortcomings of STA and DTA models in practice in the strategic context, and 3) to demonstrate the model performance in terms of the desired properties listed in Table 3.1, both for the reference model form and several model variations, thereby helping model users to choose the variation best suited for their application. The latter is done using (small) theoretical and (large) real-life model applications.
The remainder of this paper is organized as follows. Section 3.2 describes the models concept and methodology, and section 3.3 describes the algorithmic implementation of its reference form. Throughout both sections, where appropriate, we discuss how STAQ qualitatively overcomes the shortcomings of STA and DTA models and adheres to the desired properties. Then section 3.4 demonstrates the model performance also quantitatively based on recent real-life model applications conducted in the past five years, and presents how different model variations affect the desired model properties. We end with discussion and conclusions in section 3.5.

3.2 Concept and methodology of STAQ

This section describes the concept underlying STAQ as well as its methodology and variations. STAQ is implemented in C++ and available and applied for policy makers as a part of OmniTRANS transport planning software since early 2015. Subsections 3.2.1 till 3.2.4 provide insight into how STAQ combines assumptions from static and dynamic assignment models to satisfy the desired properties for strategic transport model systems (Table 3.1) and form a prerequisite for the sections afterwards. Section 3.2.5 describes the STAQ variations used in this paper.

3.2.1 General concept and properties

STAQ achieves the desired properties in Table 3.1 by combining some (implicit) assumptions from STA models with some assumptions from DTA models. In order to include flow metering and spillback effects of congestion, its network loading submodel (subsection 3.2.3) respects strict capacity (maximum flow) and storage (maximum density) constraints respectively. Note that strict capacity constraints can be added straightforwardly as mathematical constraints to the STA model formulation (e.g. Larsson and Patriksson 1999), but when added for all links, this yields unrealistic (equilibrium) solutions without congestion, because all links are forced into free flow regime. Furthermore, solving the model becomes much more tedious (Nie et al., 2004). DTA models on the contrary simulate the full on-set and off-set of congestion due to the flow and density constraints, but calculate much more (dynamic information) than required in the strategic context at the cost of computational efficiency, convergence and scalability properties. STAQ resolves this trade-off by including strict capacity and storage constraints (as in DTA models), but excluding the time dimension by assuming stationary travel demand throughout the study period (as in STA models) and instantaneous propagation of unconstrained flow (as STA models assume for all flow). It does this in a way maintaining most of the robustness, accountability and low level of computational and input requirements from STA models. It uses a concave two-regime fundamental diagram for the relation between speed, flow and density on link level (subsection 3.2.3.1), and uses an explicit node model to describe merging, diverging and crossing flow interactions on node level (subsection 3.2.3.2). Additionally to the node model, to allow for application in the urban context, STAQ has a junction modeling component, taking into account capacity and delay effects on the level of turning movement caused by e.g. traffic rules, geometry and/or signal schemes on junctions (subsection 3.2.3.3) allowing for model application on both urban roads and motorways. Furthermore, STAQ allows for multi-user-class assignment, where each vehicle class has its own route choice parameters, free flow speed and set of network restrictions making the model applicable for all vehicle classes.

The specific assumptions in STAQ are beneficial for its purpose to overcome the shortcomings of STA and DTA models in the strategic context, but also have consequences for its usage and interpretation of its outcomes. First (contrary to STA and similar to DTA), its strict capacity constraints and explicit node model can lead to residual traffic: traffic that cannot reach its destination within the studied period. Second (similar to STA and contrary to DTA), its omission of a time dimension means that all model results (e.g. flows, travel times, densities) are averages over all travelers departing in the study period. Third (similar to STA and contrary to DTA), it forces the modeler to make an assumption on the network state before and after the study period, as there are no warm up or cool down periods to take care of this. On the one hand, just like static models implicitly assume, STAQ assumes an empty network before and zero demand after the study period. On the other hand, all travel time (and contributions to density and flow) of traffic that departed within the study period is accounted for in the average outputs, also when part of a trip takes place after the end of the study period (the latter cannot occur in STA models).

3.2.2 Modelling framework

The assignment model is split into two submodels: network loading, and routing. The network loading submodel uses route-specific travel demand to compute the resulting (route) travel times, whereas the routing submodel uses route travel times to compute the resulting travel demand per route. As shown in (Bliemer et al., 2012; Brederode et al., 2010), the network loading submodel of STAQ can be seen as a static version of the generalized link transmission model of (Gentile, 2010).

Therefore, STAQ is categorized using the framework for macroscopic DTA models displayed in Figure 3.1, adapted from (Cascetta, 2009). Note that STAQ uses a route submodel that is common to macroscopic DTA models, but has a very different network loading submodel. Note that the unit of demand is the number of car equivalents (in case of single-user-class assignment) or the number of vehicles per user-class (in case of multi-user-class assignment).
The remainder of this section describes the model components within both the network loading submodel (further elaborated in section 2.3) and the route submodel (as used in the case studies in section 2.4). Note that mathematical definitions of the different model components are omitted in this paper, as these have been described before in other publications (Bliemer et al., 2014; Raadsen et al., 2016). Instead, we provide references to those publications, and here conceptually elaborate how the various components are combined within the model and its variations. The model variations that are used and/or tested in section 3.4 are described in subsection 3.2.5.

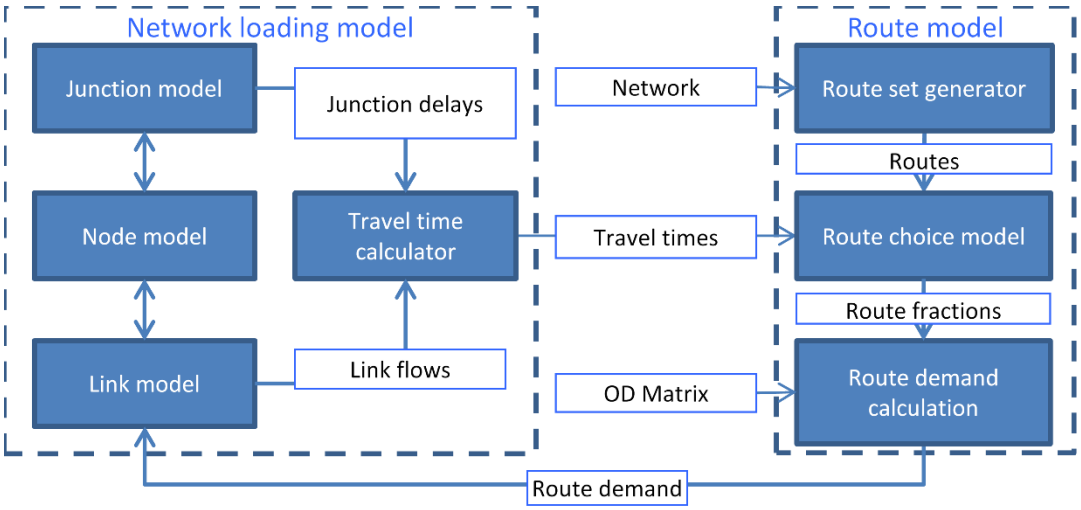


Figure 3.1: STAQ modeling framework (adapted from Cascetta, 2009)

3.2.3 Network loading submodel

The network loading submodel of STAQ consists of two phases that both use the same node and junction model components, but use a different link model component as to the adopted fundamental diagram.
First, *the squeezing phase* models the effect of the flow metering of bottlenecks using the path-based network loading model with strict capacity constraints as described in (Bliemer et al., 2014). This model assumes a fundamental diagram with in the density-flow plane a concave free-flow branch and a linear horizontal congested branch in the link model (Figure 3.2, middle) implying vertical queues on nodes for which the node model calculates active capacity constraints. Note that the squeezing phase implicitly assumes instantaneous flow propagation for all flow that is not held up in queues just like STA models assume for all flow (in free flow and congested state).
Second, *the queuing phase* models the effect of the spillback and secondary effects of bottlenecks using the event-based generalized dynamic link transmission model described in (Raadsen et al., 2016), assuming stationary demand and initial in- and outflow rates and fixed turn-fractions derived from the turn flows calculated in the squeezing phase. This model assumes a concave free-flow branch and a linear downward-sloping congested branch in the link model (Figure 3.2, right) implying storage constraints, while the node model calculates the effects of changes of in- and outflow rates on adjacent links. Note that although no ‘normal’ time dimension exists, the queuing phase uses a time dimension internally (referred to as

'queuing time') to capture the amount of spatial interaction between all the different spillback and flow metering effects. A specific queuing time however, cannot be related to, or interpreted as, a specific moment in time because the queuing phase starts with the instantaneously propagated flow rates from the squeezing phase, and demand is assumed to be stationary. Only the (demand averaged) flow rates and travel times are consistent with the assumptions in STAQ and as such form the primary output.
The most important reason for splitting the algorithm into two phases is to maintain scalability when calculating spillback and secondary effects of bottlenecks. Additional reasons are that the squeezing phase compensates for the lack of a pre-study-period warm-up and that flow metering and spillback effects can be analyzed separately.

3.2.3.1 Link model

Figure 3.2 illustrates the density-flow relation of the fundamental diagrams of STAQ (middle and right) with the BPR-type travel-time functions that are typically used in STA models (left). In the figure, the free-flow branch of each diagram is blue and the congested branch is red. Note that the travel time functions in STA models have no capacity constraint. Hence their fundamental diagram does not contain a congested branch. Considering the fundamental diagram of STAQ– squeezing (middle): it has a free-flow branch very similar to that of the STA model, but it has a congested branch that satisfies the capacity constraint on maximum flow and as such accounts for flow metering. However, because there is no constraint on maximum density, vertical queues are implied (i.e. point queues with infinite density). The fundamental diagram of STAQ – queuing (right) has the same concave free-flow branch as STAQ – squeezing, and a congested branch that complies with both the capacity constraints on maximum flow (accounting for flow metering) and maximum density (accounting for spillback). The mathematical formulation of this Quadratic-Linear (QL) fundamental diagram can be found in (Bliemer et al., 2014).

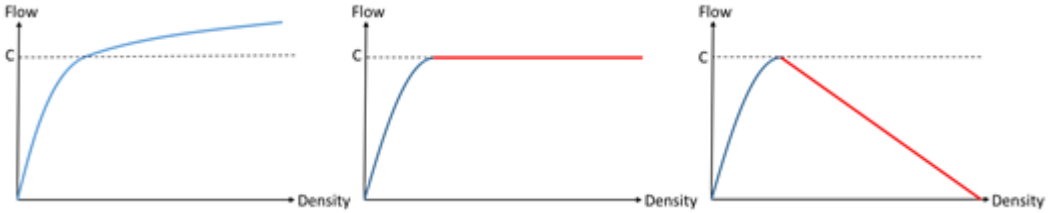


Figure 3.2: Fundamental diagrams: used in static (left), used in STAQ – squeezing (middle), used in STAQ – queuing (right)

3.2.3.2 Node model

The node model seeks for a consistent solution in terms of flows transferred over the intersection, assuming individual flow maximization and accounting for all demand and supply constraints of the adjacent links. This means that the node model can transfer the effect of capacity restrictions on downstream links to upstream links and can transfer the effect of changes in demand on upstream links to downstream links. STAQ uses the node model proposed in (Tampère et al., 2011) and (Flötteröd and Rohde, 2011) which complies to a set of generic requirements for first order macroscopic node models described in the first paper. Later, (Smits et al., 2015) generalize all feasible supply distribution schemes complying to the requirements of (Tampère et al., 2011) into a family of macroscopic node models, of which the model used in STAQ is a member.

3.2.3.3 Junction model

The junction model is an extension of the node model. It has two purposes in STAQ. Firstly, it accounts for the effect of limited supply due to conflict points on the junction itself (i.e. crossing

flows), since the node model itself only accounts for flow restrictions due to merge and diverge interactions between flows leaving in-links and entering out-links to the node. The junction model thus imposes further constraints onto the node model. Secondly, the junction model calculates travel-time delays due to passing the junction, caused by conflicts on turning-movement level depending on junction type (e.g. roundabout, prioritized or signalized). In the current implementation, the junction model uses the method described in (Bovy, 1991) for roundabouts and the Highway Capacity Manual (Awan and Solomon, 2000) for other junction types[5]. The junction model first calculates effective turn capacities given the local demand, and then derives turn delays using these capacities. These turn delays consist of deceleration and acceleration delays when approaching and leaving the node, and delays due to direct interference of other traffic or signaling on the node itself. Note that delays as a result of queuing are excluded from the junction model because its turn capacities are used in the node model that potentially triggers the link model to account for queuing.

3.2.3.4 Travel-time calculator

The travel-time calculator is used to derive travel times from the output as calculated by the link, node and junction models. The travel-time calculator has two functions. Firstly, it uses cumulative inflow and cumulative outflow curves created by the link model of each link to derive the link travel time (e.g. (Long et al., 2011)). Note that in this way, the effects of queues and spillback as a result of demand and (internal) supply constraints imposed by the node and junction models are automatically accounted for. Secondly, it translates these link-based travel times into route-based travel times, and includes delays from the junction model. It calculates the travel time of a route from an origin to a destination; flow averaged over all car equivalents *departing* within the study period. It includes the travel time experienced after the study period by car (equivalents) that did not reach their destination within the study period. This is achieved by setting outflow from all centroids to zero after all demand is put on the network, and letting the queuing phase continue until all traffic has reached its destination.

3.2.4 Route submodel

The advantages of STAQ are derived more from its unique network loading submodel than its route submodel, and hence the latter is interchangeable. Nevertheless, for sake of completeness and clarity we describe the route submodel here briefly.

3.2.4.1 Route set generator

The route set generator creates routes based upon a digitized transport network. It uses the Dijkstra algorithm to find the shortest path between each origin-destination (OD) pair. By use of a repeated random sampling process on free flow link travel times using a gamma distribution known as the accelerated Monte Carlo method (Fiorenzo-Catalano, 2007), alternative routes are generated. Route filters are applied after the repeated random sampling process to reduce route overlap, remove irrelevant routes and restrict the size of the set of potential routes.

3.2.4.2 Route choice model and convergence criterion

The route choice model uses the generalized route costs (based on the network loading submodel) to compute route fractions for all route alternatives between an OD pair. Here we

[5] The junction modelling component is currently being updated to also include the US HCM2010, the German HBS2015 and other state of the art junction models as part of the research described in (Bezembinder et al., 2015).

assume random utility maximization with perception errors, and hence use the multinomial logit (MNL) model to calculate route choice probabilities, such that route demand f_p is defined by:

$$f_p = \exp(-\mu_{od} c_p) / \sum_{p' \in P_{od}} \exp(-\mu_{od} c_{p'}) D_{od}\,, \tag{3.1}$$

where c_p is the route cost on route p, D_{od} is the travel demand for OD pair od and μ_{od} is the scale parameter describing the degree of travelers' perception errors on route travel times (where perfect knowledge is assumed when μ_{od} approaches infinity). Here (and in most real world applications) μ_{od} is determined using a global scale parameter μ normalized over ODpairs by $\mu_{od} = \mu / \min_{p \in P_{od}} c_p^0$, where c_p^0 is the free flow cost on route p. This normalization ensures that the relative effect of perception errors is the same on all OD pairs (regardless of their average route travel time).

Together with the feedback loop in Figure 3.1 and an averaging scheme, this leads to flow assignment complying to the stochastic user equilibrium (SUE). To check for convergence, we use the adapted relative duality gap as derived in (Bliemer et al., 2013) that accounts for perception errors and thus reaches zero upon convergence when using the MNL route choice model:

$$G = \frac{\sum_{(o,d)} \sum_{p \in P_{od}} f_p (c_p + \mu_{od}^{-1} \ln f_p - \psi_{od})}{\sum_{(o,d)} D_{od} \psi_{od}}\,, \tag{3.2}$$

where $\psi_{od} = \min_{p \in P_{od}} [c_p + \mu_{od}^{-1} \ln f_p]$ represents the minimum stochastic path cost.

3.2.4.3 Route demand calculation and averaging scheme

The route demand calculation component has two functions. Firstly, it computes the travel demand at route level, based on the OD-demand and route fractions. Secondly, it enforces and speeds up convergence by averaging route demands over iterations. STAQ uses the method of self-regulating averages (SRA) to average route demands over iterations. SRA complies to the convergence conditions derived by (Blum, 1954) stating that the influence of priori iterations must decrease in every subsequent iteration. SRA is described in detail in (Liu et al., 2009) and tends to provide fast convergence with high precision. The concept of SRA is to let the influence of prior iterations decrease with either a larger or smaller step size depending on the difference in levels of disequilibrium (in terms of 'excess' vehicle hours) between the last and second-to-last iteration.

3.2.5 STAQ variations

As mentioned in section 3.1 model components can relatively easily be exchanged or adapted thereby creating STAQ variations. A variation is a STAQ model application in which one or more of the components described in subsections 3.2.3 and 3.2.4 are replaced or altered. Variations are applied to change the balance between accuracy and applicability on the one hand and input requirements, tractability, accountability, computational efficiency and convergence properties on the other hand. Below, the five variations that will be used in section 3.4 are described. Note that each variation can be applied in combination with other variations; e.g. in subsection 3.4.3, three variations are used to construct the twelve different combinations listed in Table 3.3. Further note that more variations are feasible (and have been implemented), but are omitted here for reasons of relevance and brevity.

The first variation mainly influences the balance between accuracy and convergence properties by omitting the queuing phase until equilibrium has been reached, and then apply it only in the

last iteration to translate the equilibrated vertical (point) queues into horizontal (spatial) queues. When applying this variation, route choice is based on vertical queues, and effects of horizontal queues are only included in the final network traffic states (i.e. link flows, speeds and densities). This variation is tested in subsections 3.4.3 and 3.4.4. It is expected to improve convergence and thereby computational efficiency at the expense of accuracy and applicability, especially around heavy bottlenecks where in reality spillback would influence route choice.
The next two variations also mainly influence the balance between accuracy and convergence properties and are related to the junction model. Firstly, flow restrictions due to junction modeling can be omitted, in which case only the turn delays are taken into account in the travel time calculator. Secondly, junction modeling can be omitted entirely, in which case no additional flow constraints are imposed on the node model nor are turn delays considered in the travel time calculator. Both variations are tested in subsections 3.4.3 and 3.4.4.
The fourth variation is to increase model tractability at the expense of convergence properties by applying the MSA, instead of SRA, averaging scheme. Because MSA uses predefined fixed step sizes that are independent of results of previous iterations it is much easier to verify its outcomes. The effect on convergence properties (and thereby computational efficiency) is discussed in subsections 3.4.3 and 3.4.4.
The fifth variation is also to increase model tractability and relates to the form of the fundamental diagram. Instead of the QL diagram, the triangular fundamental diagram proposed by (Newell, 1993) can be used. This diagram implies no delays in the free flow branch, which means that it is less accurate in these circumstances. The diagram is especially useful to demonstrate tractability, because a flow/density tuple can easily be calculated using simple geometric algebra as will be shown in subsection 3.4.1.

3.3 Model implementation

This section describes the implementation of STAQ in terms of input, algorithm and output. Recall from section 3.1 that all variations are simplifications of the reference form. This section thus describes the normative input requirements, most advanced algorithm and most accurate output of the model. In line with section 3.2, mathematical or pseudo-code representation of the model is omitted here, as these have been provided before in publications to which we shall refer.

3.3.1 Model input

STAQ needs less input than DTA models and only slightly more input than STA models. Therefore, we first describe model input required for STA models, and then describe the additional input required for STAQ.
In STA models, the infrastructure (supply) is described by a (graph) network of the study area consisting of centroids, directed links and nodes. Centroids represent aggregated trip origins and destinations. Links represent road segments and have attributes pertaining to the free flow speed and the theoretical link capacity. Nodes represent merges, diverges and intersections. Only those nodes where junction modeling is applied have attributes, which pertain to the junction type, approach and exit lane configuration and dimensions and optionally the traffic light schema. Travel demand is assumed stationary during the study period and described for each origin-destination pair in a single OD matrix.
Most STA models have (implicit or explicit) link-flow propagation functions that only describe a free-flow branch of the fundamental diagram. To construct the fundamental diagrams for each link (Figure 3.2), STAQ uses the free-flow speed and capacity like in a STA model to determine the slope and height of the free-flow branch. Additionally, STAQ requires the jam density per

lane to determine the point of intersection of the congestion branch with the density axis, and requires the critical speed to determine the slope of the free-flow branch at capacity. Note that the critical speed can be derived from free-flow speeds from an existing STA network and jam density can be derived or assumed based on the average car length. Further note that STAQ does not need a link typing (as most STA models employ), since all link characteristics are derived from the (link specific) fundamental diagram. STAQ does not need any additional input on the travel demand.

Although STAQ needs little extra input compared to STA models, its strict capacity constraints put emphasis on the required level of precision and accuracy of the input data. Firstly, the strict capacity constraints make it necessary to define the stationary demand matrix more explicitly: it should contain all the traffic that chooses to *depart* in the study period, no matter if it reaches its destination within that study period. This means that when using traffic counts to calibrate the OD matrix, flow metering and spillback effects of congestion should be somehow taken into account (something that is usually *not* accounted for in matrix estimation procedures for static traffic assignment models). Another consequence of this more explicit definition of travel demand is that the modeler will have to think about the translation from the 'real' time-varying travel demand to an 'averaged' or 'peak' travel demand for the study time period, depending on the desired outcomes ('average' or 'peak' flows and travel times). This means the study period length and the static travel demand level should be defined consistently. Note that this is also the case with STA models, but the lack of strict capacity constraints prevents manifestation of erroneous choices[6].

Secondly, (similar to any macroscopic DTA model) when using a variation with the queuing phase and junction modeling, the strict capacity constraints require junctions to be modelled integrally using a single node, and not as an 'expanded node' (i.e. a constellation of short links and nodes that jointly represent a junction). In STA models, the latter is sometimes done to maintain (digital) network consistency with environmental models. Although not correct, the error introduced in the STA context is relatively small, because only the (additive) turn delays from junction modeling are used to influence route choice within the model. Therefore, the induced error could be traded-off for network consistency. However, because STAQ also uses the turn capacities from junction modeling as strict capacity constraints in the network loading submodel, this trade-off can no longer be made[7].

The effects of the capacity constraints described above can be considered a pain, but they do increase the accuracy of the model substantially by adding flow metering and spillback effects. Furthermore, they force the definitions of travel demand, study period length and junctions to be defined explicitly and more precise, thereby increasing the model accountability.

From the above, we conclude that with respect to the desired property of low input requirements, STAQ requires more input with a higher accuracy than STA models. However

[6] Because of the lack of strict capacity constraints, no queuing occurs in STA models, which means that the relation between demand and (modelled) delay due to congestion is much less sensitive compared to models with strict capacity constraints, preventing manifestation of erroneous choices.

[7] Because capacity is not additive, each path using the junction will only be affected by the first turn on the path that forms an active constraint. If this is a turn on a node originating from a 'junction-link' a queue will form on the junction-link, whereas in reality this would be prohibited (on signalized junctions), impossible (on junctions without mid verges) and/or would only occur when a queue formed downstream of the junction spills-back onto the junction. In the first two situations, a queue that in reality would form on the upstream links of the junction is modelled on the junction itself, potentially blocking other turns on the junction. Because junction-links are relatively short, spillback on these links occurs rapidly causing almost instant gridlock on the junction, whereas this would not happen in reality.

these requirements are very modest compared to those of DTA models, and most of the additional input can be derived and refined from STA model input. Hence, STAQ requires much less (precise) input compared to DTA models.

3.3.2 Algorithm description

Below the algorithms underlying the STAQ network loading submodel (left part of Figure 3.1) are described using flow charts. A full mathematical description of the squeezing and queuing algorithms can be found in (Bliemer et al., 2014; Raadsen et al., 2016) respectively.

The squeezing phase (Figure 3.3) primarily detects the locations and severity of active bottlenecks in the network, given the demand for all routes from the route submodel. It calculates a consistent set of reduction factors on turning movement ('turn') level that express the fraction of flow that can traverse the turn, given the capacities of the turn itself (as defined by the junction model), the capacity of its downstream link (as defined in the link attributes of the network) and all the reduction factors upstream from the turn (on routes that use the considered turn). The algorithm initializes reduction factors at a value of 1 (so no reduction) on all turns, and continues iterating[8] until on all turns the difference between the flow of the previous and current iteration is small enough. At this stage the final link (in)flows and turn flows are known, and (not shown in flowchart) vertical queues (on turn and node level) and link and route travel times can be derived using the final reduction factors and the route demand. Note that (Bliemer et al., 2014) have proven that the squeezing phase converges to a unique fixed point under very mild assumptions.

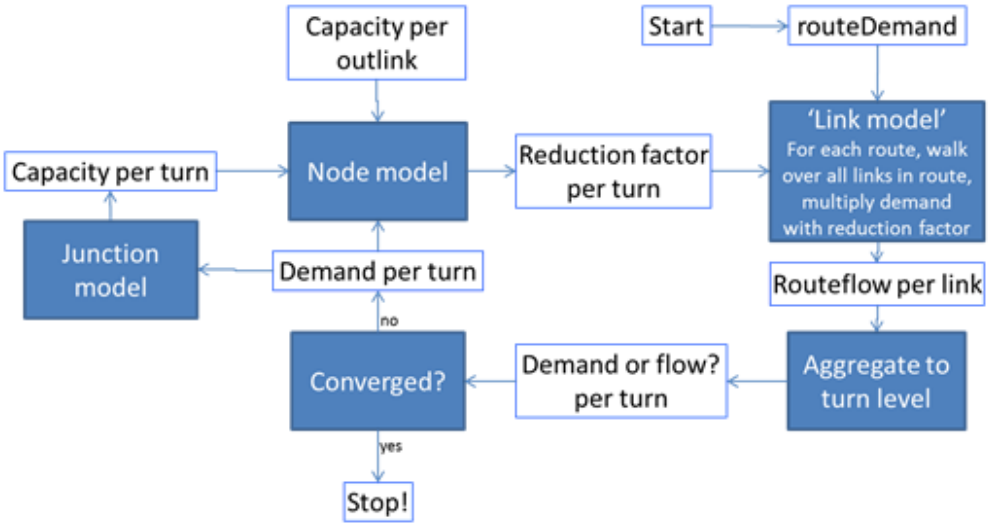


Figure 3.3: flowchart of squeezing phase

The queuing phase (Figure 3.4) adds spillback and secondary interaction effects between queues on the network. It tracks shockwaves through space using link discretization as in the link transmission model (LTM, Yperman, 2007), but does so in continuous time (using events) starting at the beginning of the study period. The queuing phase initializes by storing splitting rates derived from the turn flows from the squeezing phase and by translating the reduction factors from the squeezing phase into trigger events containing the flow rate upstream and downstream from the shockwave it represents. Then, the algorithm loop starts by running the link model for each trigger event. The link model updates the cumulative in- or outflow curve of the considered link and uses these to apply simplified kinematic wave theory (Newell, 1993)

[8] Note that these are iterations within the network loading submodel (inner loop), not to be confused with iterations between the network loading and route submodel (outer loop).

to calculate the release event time: the expected arrival time of the resulting shockwave at the other end of the link. After all trigger events are handled, the release events are placed on the event list that is then sorted ascending by time. Then, the first event is selected and its event time is validated. Validation is needed, because whilst the selected event was on the event list, other events on the same link may have updated its cumulative in- and/or outflow curve. If it is valid, time is set to the event time and the node model of the corresponding link end is run, given the updated in- or outflow rate from the event and the splitting rates stored during initialization. This generates new trigger events at links adjacent to the node, which closes the loop. If it is invalid, the event time is either updated (when other events have sped up or slowed down the shockwave) or the event is deleted (when other events have reversed the direction of the shockwave).

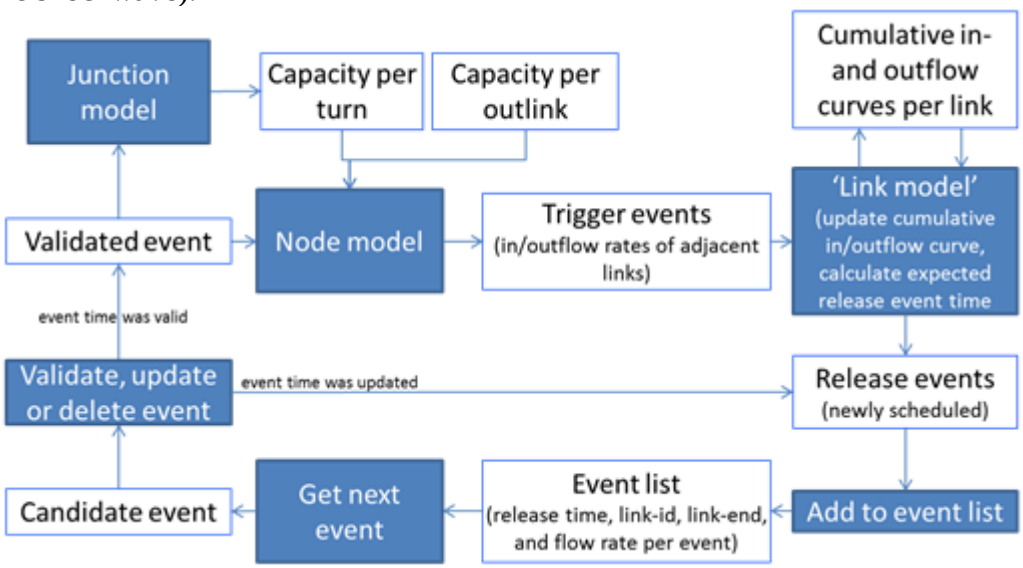


Figure 3.4: flowchart of queuing phase

The assumption of zero demand after the study period (section 3.2.1) is implemented by artificial trigger events at time T carrying zero flow on all upstream ends of links connecting origins to the network (not shown in flowchart). The algorithm stops when there are no more scheduled events on the event list, which means that the network is empty.

Note that the number of events in the queuing phase can become quite large in large networks, mainly due to forward moving shockwaves that spread out according to the turn fractions causing the change of flow rate between upstream and downstream end of the shockwave to approach zero quite quickly. To reduce the computational burden at the cost of model precision, the queuing phase can be configured to skip processing trigger events for which the difference between the updated flow rates from the node model are smaller than some threshold value epsilon. (Raadsen et al., 2016) discuss the effect of different epsilon values and conclude that a value of 5.0 veh/h provides a good trade-off between computation speed and precision. Throughout this paper we use a far more conservative value of 1.0 veh/h, for which negligible effects are reported in the same paper. Note that because the queuing phase is an event based algorithm, it only does calculations when and where needed. This makes the algorithm much faster compared to regular LTM implementations that evaluate all links in the network for each time step.

3.3.3 Model output

The primary output of STAQ consists of average flows, speeds and densities on link- and turn-level. All primary output is derived from the cumulative in- and outflow curves that are created in the queuing phase in a way that is consistent with simplified kinematic wave theory (Newell,

1993) and the assumptions of STAQ as described in subsection 3.2.1. This is illustrated using the example of cumulative flow curves for a link displayed in Figure 3.5 in which the dents in the cumulative flow curves correspond to the events in the queuing phase leading to an increase or decrease in the in- or outflow rate. This figure exhibits four phenomena directly related to the assumptions from subsection 3.2.1.

Firstly, the assumption of instantaneous propagation of unconstrained flow means that the initial inflow rate (the angle of the cumulative inflow curve at queuing time 0) is equal to the route flow per link from the squeezing phase (as defined in Figure 3.3). It also means that at queuing time 0, this flow rate applies to the entire link, from start to end. This means that the cumulative inflow curve does not start at zero, but at a value equal to the inflow rate times the free flow travel time on the link, to reflect that traffic has reached the link end before the queuing phase starts.

Secondly, due to the strict capacity constraints, the initial outflow rate (the angle of the initial cumulative outflow curve) may be lower than the initial inflow rate due to a vertical queue at the downstream side of the link, which means that it is equal to the route flow per link from the squeezing phase multiplied by the reduction factor of this link.

Thirdly, due to the strict storage constraints, density (the difference between cumulative in- and outflow at any point in queuing time) can never be larger than jam density, and the actual densities and changes in flow rates through queuing time are consistent with simplified kinematic wave theory (Newell, 1993).

Fourthly, the assumption of stationary travel demand during a single time period implies that the assignment is finished when on all links, the cumulative outflow curve has reached the unconstrained travel demand for the respective link (i.e. the total demand using this link according to the estimated demand matrix during the study period duration and route choice model). Considering Figure 3.5, the cumulative inflow curve shows that only after t_1 all travel demand has entered the link. Because $t_1 > T$, demand for this link is being held up by active bottlenecks upstream or due to spillback of the link itself (indeed the cumulative inflow curve shows periods where the inflow rate is decreased). Similarly, the last vehicle leaves the link at t_2, which includes the delay of all active bottlenecks upstream, delay due to spillback caused by the considered link but also any congestion on the link itself that does not lead to spillback.

Note that the squeezing and queuing phases both yield flows and speeds, where the output of the squeezing phase is predominantly used internally in STAQ, while the output of the queuing phase forms the primary model output. Further note that output of both phases is consistent with the route choice model, and that the squeezing phase does not yield densities because there exists no (internal) time dimension in this phase.

Other STAQ output consists firstly of vertical queues on turn-level and node-level, as calculated by the squeezing phase. These queues are defined as the number of car-equivalents that depart within the study period and have not yet exited the queue at the end of the study period. Secondly, the junction model yields effective turn capacities and turn delays on turn-level. And thirdly, the route choice model yields all common output on the route-level consisting of route fractions and costs.

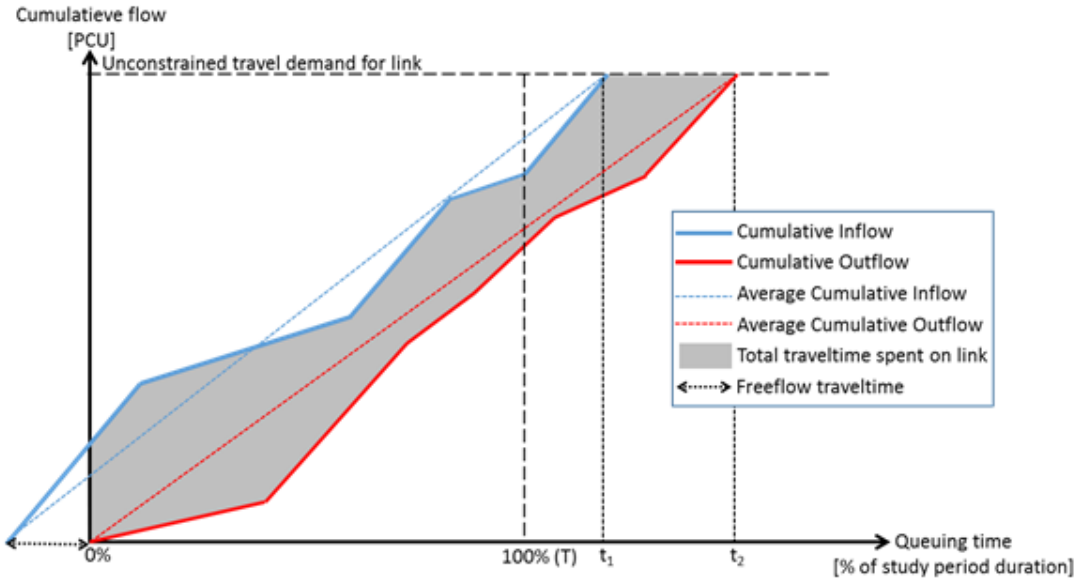


Figure 3.5: example of cumulative flow curves of a link as calculated in the queuing phase

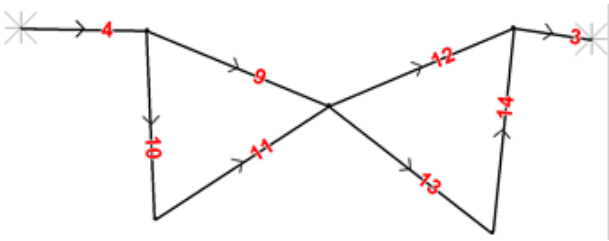
3.4 Demonstration of model properties using case study examples

In this section the properties of STAQ are demonstrated using several model applications, and discussed with respect to the desired properties from Table 1. In sections 4.1 till 4.4 we subsequently discuss: tractability, accuracy in congested conditions and accountability, robustness, and computational efficiency. The sixth desired property regarding input requirements is already discussed in subsection 3.3.1. The seventh desired property regarding applicability is already briefly mentioned in subsection 3.2.1, but also plays a role in subsections 3.4.2.2 and 3.4.3.

3.4.1 Tractability

Recall from section 3.1 that we have defined tractability as the extent to which the calculations in each of the components can be verified using the methodology underlying the component or submodel. In this subsection, we demonstrate the tractability of STAQ using the illustrative network displayed in Figure 3.6, by showing that all calculations can be done and understood using only the law of flow conservation and the shape of the fundamental diagram as underlying methods. For the reader to more easily verify the calculations, in this section the triangular fundamental diagram of Newell is used as a variation on the quadratic-linear (QL) fundamental diagram used by STAQ. Because only the shape of the fundamental diagram (one of the two inputs for demonstrating tractability) of the model variant is different to the reference form, conclusions drawn in this section will also hold for the reference form itself and thus for all variations (since these are simplifications of the reference form).

In the illustrative network, all links are unidirectional and have a length of 2 kilometers and a free flow speed of 100 km/h. Capacities per link are displayed in the middle part of the figure, jam density is set to 180 veh/lane. There is only one OD-pair that has its origin top left and destination top right carrying a stationary travel demand of 8000 veh/h. Four routes exist in this network, shown in the right part of the figure.



linknr	#lanes	road capacity
3	2	2000
4	4	8000
9	2	3000
10	2	5000
11	2	4000
12	1	2500
13	2	3000
14	1	2000

route	Links	cost (freeflow)
1	4,9,12,3	0.08
2	4,9,13,14,3	0.1
3	4,10,11,12,3	0.1
4	4,10,11,13,14,3	0.12

Figure 3.6: network with link numbers (left), link capacities (middle) and free flow travel times per route (right) of toy network

First we show the mathematical tractability of the multinomial route choice model. Assuming $\mu=1/0.14$ and given the free-flow travel-times derived from link lengths, $\mu_{od}\approx89.28$. Then applying equation (3.1) yields most vehicles (5867) choosing the shortest route 1, fewer vehicles choose routes 2 and 3 (984 vehicles each) and the longest route 3 is used the least (165 vehicles).

Given these route demands, the squeezing phase (Figure 3.3) detects that there are potential bottlenecks at the turning movements towards link 9 (demand: 6851 (5867+984), capacity: 3000), link 12 (demand: 6851, capacity: 2500) and link 3 (demand: 8000, capacity: 2000). For the sake of briefness, we only consider the first potential bottleneck here: the diverge upstream from link 9. Without going into details of the node model, one can apply the law of conservation of vehicles here to see that 3851 vehicles will be left in the vertical queue not able to enter link 9 yielding a reduction factor of 0.44 for all vehicles leaving link 4. Because of this queue at link 4 another 646 vehicles on route 3 and 4 towards link 10 are also caught in the same vertical queue, due to the conservation of turning fractions (one of the properties of the node model described in 3.2.3.2). Further iterations of the squeezing phase yield flows and vertical queues displayed in the left part of Figure 3.7, where one can verify that for each node, the summation of flow on its incoming links is equal to the summation of flow on its outgoing links plus the vertical queue on the node, proving tractability of the squeezing phase.

Given the flows and vertical queues, the queuing phase (Figure 3.4) starts out with three initial backward shockwaves. Shock 1 starts from the downstream end of link 12, shock 2 starts from the downstream end of link 9 and shock 3 starts from the downstream end of link 4. The conservation law implies that that shockwave speed is equal to the difference in flows divided by the difference in density in front and behind the shockwave. Using this and the link lengths, one can verify that shock 1 is the first to arrive to its upstream link end (after 446 seconds), whereas shockwave 3 arrives at its upstream link end after 576 seconds, and shockwave 2 arrives at the upstream end of link 9 after 792 seconds. From this moment onwards, links 4, 9 and 12 are spilling back, whereas the other links are in free flow state and derivative shocks are cycling through the two loops in the network. Shock 2 cycles through links 10 (forward), 11 (forward) and 9 (backward), whereas shock 1 cycles through links 12 (backward), 13 (forward) and 14 (forward). After one hour, inflow on all routes is set to 0, triggering a forward shockwave in link 4 that empties the network. Due to the heavy congestion (more than half of the demand is already being held up at the first bottleneck), it takes another 3 hours before the last vehicle has left link 3.

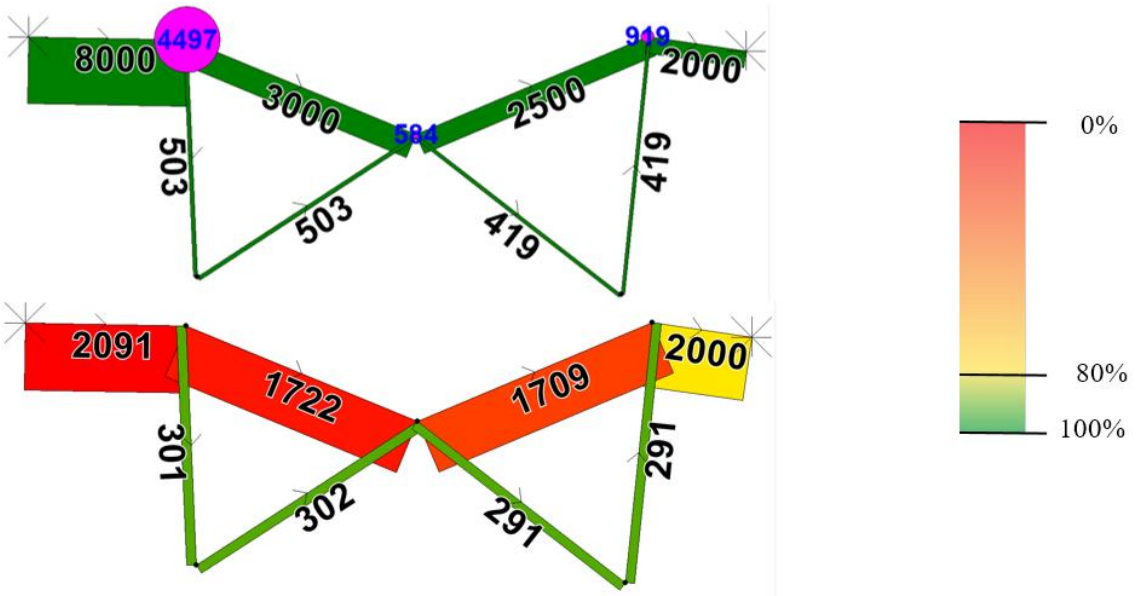


Figure 3.7: results of iteration 1; inflows (bandwidths / black font) and vertical queues (pie charts / blue font) from squeezing phase (left); outflows (bandwidths) and relative speeds (colours, see legend) from queuing phase (right)

To demonstrate how the node and link models work together we analyze shock 1 through time by looking at the cumulative in- and outflow curves of link 12 (Figure 3.8).

(1) At time 0 the shockwave starts at the downstream end (the slope of the cumulative outflow curve is lower than the slope of the cumulative inflow curve at this time).
(2) 446 seconds later (which is exactly the link length divided by the backward wave speed) the shockwave arrives at the upstream end (the slope of the cumulative inflow curve *decreases*), triggering an update of the node model at the upstream end. Because link 12 is now in spillback state, it can process *less* flow and thus has a *lower* effective capacity.
(3) Because link 12 is the normative link, this means that the reduction factor on link 9 is *decreased*, which also causes *less* flow towards link 13 (due to the conservation of turning fractions) and *less* inflow into link 14 at time 518.
(4) This leads to *less* demand from link 14 to link 3 at time 590, which causes the node model between these links to assign *more* flow from link 12 to link 3 and thus *increasing* outflow (the slope of the cumulative outflow curve slightly *increases*).
(5) The *increased* outflow triggers a backward shockwave, and the events described in step 2 till 5 are repeated, but now starting with the opposite effect causing all words in italics to be replaced by their respective opposites.

Note that each cycle of shockwave 1 corresponds to a downstream event followed by an upstream event on link 12. These events always occur 446 seconds apart (the time that a backward wave traverses the link), as can be derived from Figure 3.8. Due to the linear free flow branch of the fundamental diagram, the travel time for shockwaves to move forward through links 13 and 15 is also fixed at 144 seconds (as can be derived from Figure 3.8 by comparing durations between subsequent event times on the up- and downstream end of link 12). When using the QL fundamental diagram, or when other routes would influence this cycle, these time intervals would vary. Note that from t = 2218 onwards, no more events occur on link 12. This means that the differences between updated flow rates from the node models due to the shockwave that is cycling through links 12, 13 and 14 have become smaller than the epsilon value of 1.0 veh/h. Indeed, the differences in flow rate (the slope of the cumulative in- and

outflow curves) in Figure 3.8 before and after the last events where the epsilon is still greater than 1.0 (upstream at t = 2218 and downstream at t = 1771) is already very small. The cumulative curves in Figure 3.8 also show that the last vehicle enters link 12 at t=14125 and leaves the link at t = 14435.

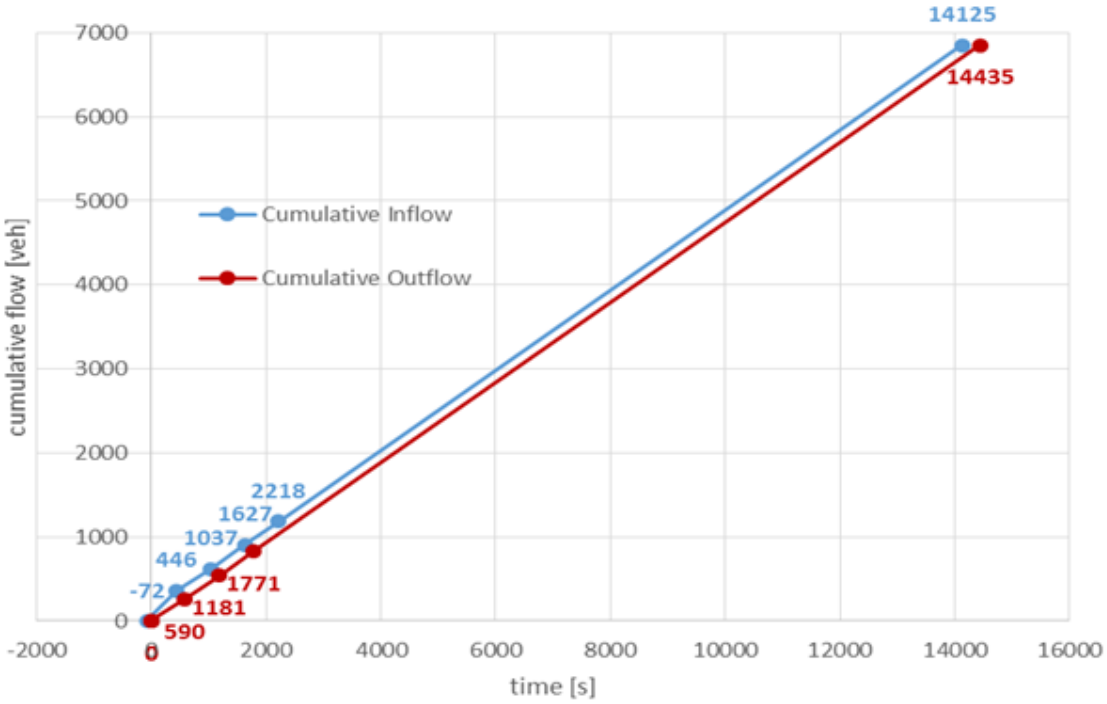


Figure 3.8: cumulative in- and outflow curves for link 12. Dots represent events in the queuing phase

To demonstrate how the average cumulative outflow curve (the red dashed line in the example of Figure 3.5) is used to calculate the link outflows as displayed in the right part of Figure 3.7, we acknowledge that only routes 1 and 3 make use of link 12, yielding an unconstrained demand of 6851 vehicles for link 12. From the cumulative outflow curve, we can see that the 6851th vehicle leaves the link at time 14435, which means that the average outflow per hour is equal to 6851/14435 * 3600 = 1709 veh/h, which corresponds to the outflow rate displayed on link 12 in the lower part of Figure 3.7.

In this section we have demonstrated that given a network, all calculations within the route submodel and the network loading submodel and the interaction between these components can be verified using only the specification of the route choice model, the law of flow conservation and the shape of the fundamental diagram. Such a level of tractability is matched by STA models (using a shortest path algorithm, some link delay function and an averaging scheme), and in theory also by non-heuristic DTA models (e.g.: CTM, LTM). However, in practice, DTA models cannot easily be traced in this way, mainly because they use time discretization requiring all time steps to be traced individually and sequentially requiring very large amount of calculations, even on small networks. Furthermore, time discretization implies discretization errors that make outcomes of these models dependent on the level of precision of their implementation. From this we conclude that STAQ satisfies the desired property of tractability both in theory and practice (whereas only some DTA models do in theory).

3.4.2 Accuracy in congested conditions and accountability

In section 3.1, we defined model accuracy in congested conditions as the accuracy of flow metering and spillback effects as well as route choice effects due to congested conditions. In the same section, we defined accountability as the extent to which different submodels can be

isolated. To assess both properties, we first isolate the flow metering and spillback effects by comparing congestion patterns (location and severity of queues) on a corridor network without route choice with observed congestion patterns and patterns from STA and DTA models (subsection 3.4.2.1). Thereafter, we add route choice effects by looking at congestion patterns and route choice effects in a case study on an urban network with route choice (subsection 3.4.2.2). This way, we isolate how the different model components capture the different mechanisms that occur in the transportation network, thus demonstrating the accountability of STAQ. Finally, in subsection 3.4.2.3 we show the impact of the model accuracy on the societal value of the measures taken in the same urban network as used in subsection 3.4.2.2.

3.4.2.1 Accuracy of network loading submodel on A12 Gouda – Den Haag

In the following analysis, loop detector data of the A12 morning peak on a representative workday in 2006 are used to qualitatively compare observed congestion patterns with model outputs from STAQ. For reference we also compare these with model outputs from an STA model and a second-order macroscopic DTA model (MaDAM, (Raadsen et al., 2010)). We stress here that the DTA model is of second order, meaning that anticipation (deceleration) and relaxation (acceleration) effects are accounted for in this model.

Figure 3.9 shows the A12 corridor network in which there are no route choice alternatives. Also, given that all network nodes are simple on-ramps and off-ramps, no junctions exist in this network. Hence, application on this network focuses on the link and node model within the network loading submodel. The OD-matrix has been calibrated on the observed demand just downstream from knooppunt Gouwe (on the motorway) and on all on-ramps indicated in Figure 3.9.

The congestion patterns are displayed in Figure 3.9, showing three active bottlenecks: 1) spillback from the traffic lights around Centrum Zuid, 2) the weaving section between Prins ClausPlein and off-ramp Voorburg and 3) the merge from on-ramp Zevenhuizen. Furthermore, the entire stretch of road between Zevenhuizen and Prins Clausplein is congested due to spillback from bottleneck 2, meaning that any potential bottlenecks along this stretch of road cannot clearly be identified from the data.

The first bottleneck (centrum Zuid) is not reproduced by any of the assignment models because spillback from outside the network is not modeled.

The second bottleneck (Voorburg) is identified by both STAQ and the DTA model. However, both models identify the merge from Prins ClausPlein as the only problem, whereas in reality the weaving section between Prins ClausPlein and Voorburg also causes problems that are not being picked up by STAQ nor the DTA model. The STA model wrongly identifies multiple links downstream from the true bottleneck as a bottleneck, because there is no flow metering in this model.

The third bottleneck (Zevenhuizen) is identified by the DTA model, causing a flow metering effect that results in a free-flow section between Zoetermeer and Zoetermeer Centrum that is not present in the observed data. STAQ does not detect the bottleneck at Zevenhuizen, although the capacity between Zevenhuizen and Bleiswijk and the demand from Knooppunt Gouwe and on-ramp Zevenhuizen is exactly the same. This must mean that the second order effects due to traffic merging from on-ramp Zevenhuizen lowers the effective capacity causing this bottleneck in reality and the DTA model. The omission of this bottleneck by STAQ causes activation of a downstream bottleneck at Zoetermeer Centrum. The STA model gives some delay at the link downstream from the bottleneck, although capacity has not been reached; meaning that the definition of the BPR function causes this link to be identified as a bottleneck.

Based on this comparison, we conclude that STAQ, contrary to the STA model, successfully detects and models primary bottlenecks, but may overlook bottlenecks that are activated due to

second-order and lane-distribution effects. These conclusions hold on any network, since they are a direct result of properties of the network loading submodel.
Although second-order and lane-distribution effects cannot be directly modelled using a first order network loading submodel such as STAQ[9], they could be added to the assignment model by decreasing the link capacities on weaving sections and merges following guides like the US Highway capacity manual (Awan and Solomon, 2000) or the Dutch CIA (Rijkswaterstaat, 2015). This could be done before the assignment, using merging and weaving proportions from the OD matrix assuming free-flow route choice, or incorporated within the assignment model using the actual proportions from the previous iteration. Note that this problem will mainly occur on motorways, because bottlenecks on urban roads typically occur at intersections.

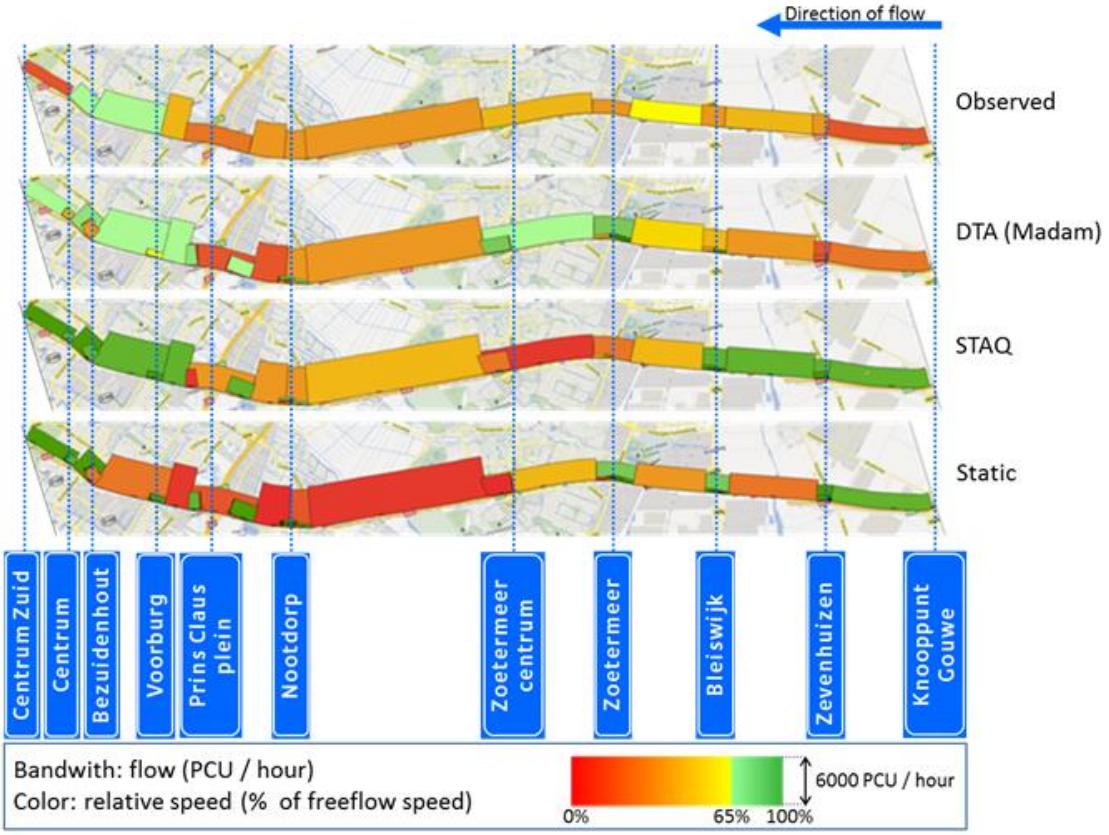


Figure 3.9: comparison of observed and modelled congestion patterns on the A12 motorway between Gouda and Den Haag

3.4.2.2 Accuracy and accountability of assignment model on case Den Bosch

In this section, the accuracy of STAQ compared to STA models is further analyzed using a bottleneck location close to the city of Den Bosch in the Netherlands. During the AM peak period the bottleneck manifests itself on the A59 motorway from Den Bosch towards Oss around the off-ramp Rosmalen (indicated by the black circle in the left part of Figure 3.10). In the reference situation, the STAQ results (right side of Figure 3.10) show a vertical queue between the off and on-ramp and a second, much smaller, vertical queue at the end of the off-ramp, together causing a queue spilling back all the way onto motorway intersection Empel (the upper left of the network cut out area displayed in the figures), whereas the static results only exhibit minor speed drops directly on the bottleneck links.

[9] Note that some second order DTA models (e.g. METANET) contain a correction term for merging sections

Figure 3.10: assignment results for reference scenario; static (left, black circle indicates bottleneck location) vs STAQ (right). Bandwidth colours: modelled speed as ratio of free flow speed; Bandwidth widths: modelled flow; Blue circles in STAQ results (right): vertical queues (radius indicates queue size)

For sake of analyses, we consider a network variant in which the capacity of the intersection at the end of the southern off-ramp is increased and an extra lane between the southern off- and on-ramp is added, leading to the assignment results displayed in Figure 3.11.

Figure 3.11: assignment results of the network variant; static (left) vs STAQ (right)

These assignment results lead to the following findings (demonstrating accuracy) and mechanisms (demonstrating accountability) for which the STA and STAQ model results are similar:

(1) The two bottlenecks around the off-ramp are effectively removed as a result of the capacity increase. In STAQ this finding is a result of the removal of an active supply constraint in the node model of the node connecting the motorway and the southern off-ramp and the removal of supply constraints of the junction model of the node at the end of the southern off-ramp.
(2) As a result of 1, the on-ramp itself and all arterial roads towards it are used more (i.e. higher flows). In STAQ this finding is a result of decreased travel times on turning movements over, and links around, the nodes mentioned in bullet 1, which cause the route submodel to increase route-fractions of routes using the on-ramp and adjacent arterial roads.
(3) The southbound traffic crossing the A59 returns from alternative routes to the arterial that uses the intersection with the considered off-ramp (indicated by the increased southbound flow on the arterial from the original bottleneck location). The mechanism causing this is thus the same as in finding 2.

Findings that the STA model results omit, but the STAQ model results do correctly show, thereby demonstrating its better accuracy under congested conditions, are:

(4) On the A59, the queue spilling back from the considered bottleneck towards the northwest is much shorter because the squeezing phase predicts the bottleneck to be much smaller and further downstream, which causes the shockwaves calculated in the

queuing phase to travel at a lower speed and over a longer distance towards the northwest. Furthermore, due to increased flow from this direction (calculated by the route choice model), a new bottleneck is activated at the merge of the motorway intersection Hintham (just west of the original bottleneck location).

(5) On the A59, downstream from the removed bottleneck, the existing bottlenecks intensify, and a new small bottleneck activates at the next off-ramp. This is caused by the increase of the reduction factor at the original bottleneck location as calculated by the squeezing phase in combination with the increase of flow due to the route choice model reacting to lower travel times for eastbound traffic on the motorway.

Comparing the STA and STAQ results we conclude that only effects on the links and nodes where measures were taken and some of the route choice effects of the network variant are captured by the STA model, whereas STAQ also captures the effects up- and downstream from the removed bottleneck. This leads to the conclusion that the addition of flow metering and spillback effects strongly improves the accuracy and realism under congested conditions. This conclusion holds on any network, because it is a direct result of properties of the network loading submodel. Furthermore, we have shown that the STAQ results can be related to (combinations) of model components, demonstrating its accountability. With respect to accountability we conclude that STAQ includes effects of route choice, flow metering and spillback; whereas STA models only include route choice effects. And furthermore, accountability of STAQ is still on a level that makes the results explainable on a level comparable to that of STA models. Also, this section has shown that STAQ is applicable on networks containing both urban roads and motorways.

3.4.2.3 Accuracy and its impact on the predicted societal value of the measures of case Den Bosch

To demonstrate that the differences between the assignment methods may also (substantially) change the outcomes of a (social) cost benefit analysis, we compare the effect of the network variant in terms of vehicle loss hours per road type for both STA and STAQ assignment results (Table 3.2). Note that these results are only for illustrative purposes, since no calibration has been performed on either model.

	Usage [veh*km]	**Experienced delay [vehicle loss hours]**					
	Both cases	Reference		Network variant		Difference	
Roadtype	Both assignments	Static	STAQ	Static	STAQ	Static	STAQ
Motorways	~57%	334	1130	314	1052	-20	-79
Non-motorways	~43%	1761	339	1717	283	-44	-55
Total	100%	2095	1469	2031	1335	-64	-134

Table 3.2: vehicle loss hours for reference and network variant for both static and STAQ assignment

Analysis of this table leads to the following findings:
Although route choice does vary among the two networks and assignment methods (see analysis above), the usage per road type in veh*km is (approximately) the same.

(6) In the STA model most delay occurs on the non-motorways, whereas in STAQ most delay occurs on the motorways. Given the usage and location of bottlenecks (both are concentrated on the motorways in this network) STAQ results are more consistent with the model input, than results from the STA model are.

(7) Both assignment models yield a reduction in vehicle loss hours as a result of the measures taken in the network variant. However, when using STAQ, the reduction is more than twice as large compared to the STA model output (a reduction of 134 vehicle

loss hours in the STAQ assignment versus a reduction of 64 vehicle loss hours in the STA model).

For illustrative purposes, the annual societal value of the travel time savings during the morning peak hour induced by the network variant is calculated. Following (Kouwenhoven et al., 2014) we assume an average value of time of €9,- per hour and an average reliability ratio of 0.6. Furthermore, we assume that per year 260 of these average morning peak hours occur. This means that the societal value of the network variant would approximately be €240.000,- according to the STA model output and €500.000,- according to the STAQ output, an increase of 108%. These findings show that choosing an assignment method that accounts for flow metering and spillback effects has substantial effects on the outcomes of a cost benefit analysis for study areas with structural congestion.

3.4.3 Robustness

As defined in section 3.1, we consider a model to be robust when there are no random variables in the model and when it converges to a defined and meaningful stable state. From section 3.2 we know that the model does not contain random variables or stochastic processes. Therefore, in this section we will only look at the convergence of STAQ towards Wardrops' conditions of user equilibrium (Wardrop, 1952) (which we consider a meaningful stable state indeed[10]) using the adapted relative duality gap as described in subsection 3.2.4.2.

Key components within STAQ are chosen or defined to maximize convergence properties. In the route submodel, the stochastic user equilibrium is chosen as the route choice paradigm which means that in each iteration traffic is distributed over all routes (instead of choosing one route in the deterministic user equilibrium), leading to better convergence properties on the route level (Bliemer et al., 2013). In the network loading submodel, the node model complies with the two invariance principles described in (Lebacque and Khoshyaran, 2013) ensuring that its outcomes are stable under constant link boundary conditions (a numerical example of how this ensures stability is given in (Tampère et al., 2011)). Furthermore, the link model contains no discretization over space or time. This means that its solutions are exact, avoiding any discretization errors as shown in a numerical example in (Raadsen et al., 2016).

In the remainder of this section, the convergence of STAQ is assessed using several congested networks taken from strategic transport model systems that normally use an STA model. Largely neglecting the required level of precision and accuracy of the input data for STAQ (described in subsection 3.3.1), the travel demand matrices used where taken directly from the original transport models systems, whereas the networks where only refined slightly on locations where effective capacities where incorrect (these errors did never manifest itself in the STA model due to the lack of strict capacity constraints). For each model, a hundred iterations where run for all twelve combinations of the STAQ variations that are known to have substantial influence on the convergence (Table 3.3). These twelve combinations are built up from two variations regarding the averaging scheme (MSA or SRA), three variations regarding junction modelling (no junction modelling ('NoJM'); take only calculated turn delays into account ('Delays'); take both calculated turn delays and turn flow restrictions into account ('JM')), and two variations regarding spillback effects (see subsection 3.2.5 for variation definitions).

[10] Note that uniqueness of the solution is only guaranteed when the TA model uses an (implicit) cost function that is strictly increasing (theorem 1.8 in (Nagurney, 1993)). Just like DTA models, the strict capacity constraints within STAQ cause a violation of this requirement. However, empirical tests show that STAQ approximates the same equilibria in terms of link flows, no matter the start solution.

1 MSA-NoJM-NoSpillb	4 MSA-NoJM-Spillb	7 SRA-NoJM-NoSpillb	10 SRA-NoJM-Spillb
2 MSA-Delays-NoSpillb	5 MSA-Delays-Spillb	8 SRA-Delays-NoSpillb	11 SRA-Delays-Spillb
3 MSA-JM-NoSpillb	6 MSA-JM-Spillb	9 SRA-JM-NoSpillb	12 SRA-JM-Spillb

Table 3.3: numbering of the twelve combinations of model variations tested

An overview of the strategic transport model systems tested is given in Table 3.4. The models are all strategic, but range from relatively coarse motorway oriented models (Leuven, NRM-West and NVM), to more fine-grained regional models (BBMB, Vlaams Brabant) and urban models (Breda, Haaglanden). Besides Leuven, all models classify as large-scale by the definition from section 3.1. Note that the digitized networks of Vlaams Brabant, NVM and NRM-West do not contain modelled junctions (no junction definitions set), and therefore, only combinations 1, 4, 7 and 10 where run for these models.

Model	Major cities in study area	model type	Links	Nodes	Junctions	Centroids
Leuven	Leuven (BE)	motorway	2698	1833	587	430
NRM-West	Amsterdam, Rotterdam, The Hague, Utrecht (NL)	motorway	86783	56739	0	3392
BBMB	Eindhoven, Tilburg, Breda, Den Bosch, Helmond (NL)	regional	142336	106780	15979	3321
Breda	Breda (NL)	urban	147253	107984	16241	6043
Haaglanden	The Hague (NL)	urban	140277	94159	3539	5845
Vlaams Brabant	Brussels, Leuven (BE)	regional	34239	23241	0	2999
NVM	all of the Netherlands (NL)	motorway	159920	65272	0	6102

Table 3.4: properties of models tested, models sorted by size measured in number of routes

For each model, the appendix contains a graph that shows the relation between the calculation time[11] and the adapted relative duality gap for each of the STAQ variations tested. Besides showing the trade-off between computational time and convergence, the total computational time needed to do 100 iterations can also be derived from the graphs in the appendix by looking on the vertical axis at the point where the curve stops.
Recall from section 3.1 that the adaptive relative duality gap should be lower than 1E-04 for the assignment mode to produce outcomes that are suitable to be used in the strategic context. From the graphs in the appendix we conclude that that almost all runs without spillback converge sufficiently within 100 iterations when using the SRA averaging scheme. Models Vlaams Brabant and NVM are the only exceptions, however their duality gap curves do suggest that they would reach 1E-04 when some more iterations would have been conducted. Both models show a lot of bottlenecks and a high percentage of routes affected by them (77% and 91% of all routes respectively; see also Table 3.5). Further investigation shows that the networks of model Vlaams Brabant and NVM are relatively coarse in relation to its density in urban areas, which can be seen when looking at the number of centroids and especially the number of links in relation to the number of inhabitants in the study area. This causes (artificial) problems on locations where centroids representing large and densely populated areas are connected to the network with only a limited number of connectors. This happens especially in the city of Brussels in model Vlaams Brabant, and in the larger cities in NVM. This causes the high number of blocking nodes and large proportions of routes being affected. In turn, this causes high sensitivity of route cost to changes in route demand and thus poor convergence

[11] All runs conducted on a Core I7-950 3.07 Ghz machine with 24 GBytes of memory running Win7

properties. Refining the network around these areas would very likely lead to much better convergence properties.
When using MSA, only the BBMB model converges sufficiently within the first 100 iterations (but only just), all models consistently show a well-known property of MSA: its convergence slows down considerably with higher iteration numbers, which happens long before convergence has been reached. Note that, although far from sufficiently converged, in the initial 10 to 15 iterations, MSA generally outperforms SRA. However, after these initial iterations, broadly when MSA approaches duality gap values between 1E-03 and 1E-02, the convergence properties of runs using SRA are clearly much better; leading to better convergence using far less calculation time.
We now consider the effect of junction modelling on models that have junctions defined in the network and for model variations that have proven to converge without junction modelling (i.e.: variations without spillback and using SRA). The graphs in the appendix show that enabling junction modelling, but neglecting its flow restrictions (thus only adding delays from junction modelling to the route cost) deteriorates high precision convergence properties, but does not prevent any model for reaching the required convergence rate, nor does it increase required calculation times significantly. Applying full junction modelling however does break convergence for Leuven, Breda and Haaglanden. Although the duality gap curve of Leuven suggests that it would reach 1E-04 when some more iterations would have been conducted. Further investigation showed that on the Breda network, a single junction that flip flops from under- to oversaturation causes the oscillations in duality gap values around 1E-04 that can be seen in its graph. Similar observations were made on the Haaglanden network, although in this model, not a single, but several (clustered) junctions showed oscillating under- and oversaturation. These observations suggests that methods similar to diagonalization (Dafermos, 1980) might resolve this problem; e.g. smoothing or less frequent updating of the flow restrictions from junction modelling.
The appendix also shows that on all models except Leuven, model variations with spillback do not converge sufficiently within 100 iterations: the duality gap keeps oscillating and never drops below 1E-04. Spillback effects are thus the most important cause for non-convergence, which makes sense when realizing that spillback is likely to cause the cost of routes that use link(s) affected by this spillback to become diagonally non-dominant (i.e.: the demand for such a route itself is no longer the main contributor to its cost; instead demand on other routes is), whereas the route choice model and averaging scheme do not anticipate for this. Note that the one run with spillback that does converge to below 1E-04 is a variation with SRA and without junction modelling on model Leuven. Further investigation shows that the Leuven model has relatively low demand (thus violating the requirement of an accurate definition of stationary demand as stated in subsection 3.3.1) due to demand matrix calibration conducted in a static context using observations in congested conditions thus causing spillback effects to only limitedly occur.
The findings described above suggest that the model variation #8 (SRA-Delays-NoSpillb) in Table 3.3 has the best accuracy whilst still converging sufficiently on all tested models. In some cases/models, full junction modelling (variation #9) can be used without losing sufficient convergence. Also, this section has shown that STAQ is applicable on networks ranging from fine grained urban to coarse motorway networks.

3.4.4 Computational efficiency

In section 3.1 we defined computational efficiency as the extent to which run times and memory requirements are acceptable for calibration and application of large scale models. Although no formal criteria exist, a general guideline is that it should be possible to run an assignment for all modes and for all modelled periods in a strategic transport model overnight. Assuming that a single car assignment takes up around 25% of the total computational effort, this means that

any assignment should not take longer than three to four hours. With respect to memory consumption we assume that it should be possible to run the assignment on a regular high-end desktop computer with 16 Gigabytes of RAM. In the remainder of this section we look at calculation times and memory usage for the STAQ model variation #8 (SRA-Delays-NoSpillb) on the models in Table 3.4 as it was selected as the most balanced model variation combination in section 3.4.3.

Since in the considered model variation combination, the queuing phase is only performed in the last iteration, calculation time per iteration is roughly equal to the calculation time for the squeezing phase. Given the mathematical problem solved by the squeezing phase (Bliemer et al., 2014), calculation time to run the network loading submodel is mainly proportional to the following variables (column names of variables included in Table 3.5 in parenthesis): the number of routes *(#routes*), the number of active bottleneck locations *(#blocking nodes*) and their usage *(% of routes blocked*), the severity of active bottleneck locations (e.g.: local demand to capacity ratio per active bottleneck location) and the strength of the relationships between those active bottleneck locations (e.g.: the number of shared routes per active bottleneck location). Note that the severity and strength of relationships per bottleneck location are omitted from Table 3.5 since they are hard to capture in a single indicator.

	Run properties					Calculation time			Mem usage	
Model	Peak period	#routes	#Itera-tions	#blocking nodes	% of routes blocked	total [hh:mm:ss]	per iter [mm:ss]	per route per iter [ms]	Total [Mb]	per route [Kb]
Leuven	PM	74697	49	74	21%	0:01:50	0:02	0.03	404	5.41
NRM-West	AM	1241762	31	863	56%	0:37:19	1:12	0.06	2935	2.36
BBMB	AM	1272227	14	470	27%	0:22:53	1:38	0.08	2245	1.76
Breda	PM	2069672	46	940	53%	3:02:58	3:59	0.12	6470	3.13
Haaglanden	PM	2854246	18	255	32%	1:36:56	5:23	0.11	7631	2.67
Vlaams-Brabant	PM	3109173	>100	1354	77%	7:35:37	4:33	0.09	7181	2.31
NVM	AM	4057235	>100	8390	91%	13:43:16	8:14	0.12	9418	2.32

Table 3.5: calculation times and peak memory usage of model variation combination #8 for all tested models

Looking at the calculation time per iteration, we see indeed that it is roughly proportional to the variables mentioned above yielding calculation times varying from 0.03 ms to 0.12 ms per route per iteration for the models tested, which translates to about 30 seconds to 2 minutes per iteration for every million routes. From Table 3.5 however, no relationship between the number of iterations required and other run properties can be identified, whereas total calculation time is roughly[12] proportional to the number of iterations between route and network loading submodel required for convergence (#Iterations) since the route submodel forms a loop around the network loading submodel (Figure 3.1).

To explain why no relationship is found between the required iterations and other run properties in Table 3.5, we look again at the adapted duality gap graphs in the appendix. In these graphs, some models and model variation combinations show strongly oscillating curves (e.g. combinations #8 and #9 of BBMB (after 30 minutes of calculation time) and combination #9 on both the models of Breda (after 2 hours of calculation time) and Haaglanden (after 1 hour of calculation time), which slows down and/or prevents further convergence. Analysis of the

[12] This holds only roughly, since later iterations contain fewer active bottlenecks, yielding less calculation time required for the network loading submodel.

adapted duality gap values per OD for these models (leaving out the summation over OD pairs in equation (3.2)) confirms that the least converging OD pair contributes the most to poor gap values. Using this knowledge, the cause of the oscillations could be traced to a limited set of OD pairs and even to a limited set of bottleneck locations. These bottleneck locations proved to be switching between an active and inactive state over (sets of) iterations. Often, by removing only one of such bottlenecks in the network, the duality gap graph could drop substantially (factors of 10 to 1000's at equal calculation times). This extends the finding in subsection 3.4.3 that not only single (clusters) of flip flopping junctions can cause oscillating duality gap values, but that it can also occur on bottleneck nodes not being modelled as a junction. Although identified, this phenomenon may substantially delay or even prevent reaching the required level of convergence and it also prevents formulation of a relationship between the run properties and expected total calculation time in Table 3.5.

To analyse the computational efficiency of the different model components, the share of calculation time per model component for model variation #8 for six of the tested models is displayed in Figure 3.12. This figure shows that the network loading submodel (link, node, junction modals and travel time calculator) take up most (54%-64%) of the calculation time. This share is much lower than the share of the network loading submodels within DTA models, demonstrating the high computational efficiency of the network loading submodel of STAQ. This also indicates, that efforts to further improve computational efficiency might need to be put into the route choice model. This component now claims a relative large proportion of calculation time (between 32% and 41%), which will only increase when using more advance route choice models than the relatively simple MNL route choice model used here.

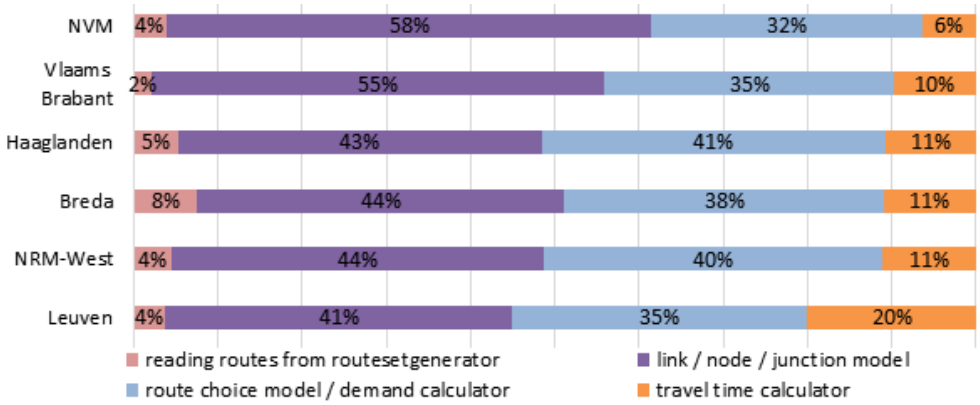


Figure 3.12: calculation time shares per model component

Comparing the total calculation times of the different models with the upper bound of three to four hours, we see that all models except for Vlaams Brabant and NVM exhibit acceptable calculation times. Although not further investigated, probably, the coarseness of these networks in relation to their density described in section 3.4.3 is likely to be the cause for its poor convergence.

With respect to memory usage, Table 3.5 indicates that that it is also proportional to the number of routes. On average, the peak memory usage per route is around 3 Kilobytes, which roughly translates to around 3 Gigabytes needed for every million routes, which means that the largest model tested here (NVM with more than 4 million routes) requires 9.4 Gigabytes of RAM, thereby easily meeting the requirement of maximum 16 Gigabytes of RAM.

3.5 Conclusions and discussion

In this paper, we have provided a complete description of the concept and implementation of the assignment model STAQ and several variations, along with insight into how the model addresses the shortcomings of STA and DTA models in the strategic context for large congested networks. In line with literature we have defined seven desired properties for strategic transport models for large congested networks, and have shown the performance of STAQ and its variants for each of these seven properties in comparison with STA and DTA models.

3.5.1 Main conclusions

The different mechanisms that occur in a transportation network when applying STAQ can all be isolated and verified using only the law of flow conservation and the shape of the fundamental diagram as underlying methods, proving that tractability and accountability of STAQ is comparable to that of STA models and amply exceed that of DTA models.

With respect to the accuracy under congested conditions, we conclude that, contrary to STA models, STAQ successfully detects and models flow metering and spillback effects of primary bottlenecks, with the limitation that STAQ may overlook bottlenecks that are activated due to second-order and lane-distribution effects. STAQ allows for assignment of different vehicle classes and the junction modelling component allows application on both urban roads as well as motorways.

Furthermore, we conclude that when evaluating network scenarios, STA models only capture effects on links and nodes where network changes occur and include some of the route choice effects, whereas STAQ also captures the effects up- and downstream from network changes. It was shown that the addition of these effects causes large differences in terms of vehicle loss hours and thus societal benefits of policy measures influencing travel times of travellers. This clearly demonstrates that the addition of flow metering and spillback effects strongly improves the accuracy and realism under congested conditions and that choosing an assignment method that accounts for these effects will have substantial effects on the outcomes of a cost benefit analysis for study areas with structural congestion.

Based on analysis of twelve different model variations on seven large scale strategic transport models of largely congested regions we conclude that STAQ with spillback in the last iteration, full junction modelling and the self-regulating averaging scheme proved to be the optimal variation, providing sufficient realism and convergence (duality gap values below 1E-04) within well acceptable calculation times for five of the seven models tested (ranging from 23 minutes up to 3 hours to achieve equilibrium on a regular desktop pc). A limitation of this model variant is that spillback effects are not included in the route choice behavior. Adding these effects is possible, but at the expense of convergence. The network of the models Vlaams Brabant and NVM prove to be too coarse in relation to its density, creating artificial congestion locations causing high sensitivity of route cost to changes in route demand and thus poor convergence properties. Refining the network in densely populated areas would very likely lead to better convergence properties for both models.

Input requirements of STAQ are much lower than those of DTA and only slightly higher than those of STA models. Although STAQ needs little extra input compared to STA models, its strict capacity constraints put emphasis on the required level of precision and accuracy of the input data. Most importantly, the definition of the study period and the level of stationary demand in the matrices should be consistent, flow metering and spillback effects in observed data should be taken into account while calibrating the OD matrices, and the hard capacity constraints in STAQ require more accurate capacity values on links and junctions to be coded as a single node. Based on the above, we conclude that STAQ is a viable alternative to the traditional STA model, providing more accuracy on congested networks without reducing

robustness and accountability and without increasing input requirements, whilst keeping computational requirements to acceptable levels (as opposed to DTA models). This makes the model suitable for applications where both STA and DTA models may fail: strategic applications on large-scale congested networks.

3.5.2 Recommendations and further research

Based on this research, several improvements in the way STAQ and its variations are being applied are proposed. Most importantly, the development of a STAQ based matrix estimation method that takes flow metering and spillback effects on observed data into account. A first attempt for such a method is described and applied in (Brederode et al., 2017, 2014) respectively. When in place, model systems can properly be calibrated using STAQ which enables more thorough validation of the assignment model comparing its outcomes with observed flows, congestion patterns and travel times for a large urban region. Furthermore, when thoroughly validated, the societal value of the model should be determined by comparing a full cost benefit analysis of one or more existing projects using an STA model and STAQ.

As described in subsection 3.4.3, there is still room for improvement on the speed and level of convergence of the model, especially for model variations with full spillback enabled. Several research directions are worth mentioning here. Firstly, the parameters that control the step sizes used within the self-regulating averaging scheme (subsection 3.4.3) should be calibrated (now the default values from (Liu et al., 2009) are used). Secondly, in subsection 3.4.4 we have already briefly mentioned that the causes for poor convergence can be traced down towards (sets of) bottleneck locations which is in line with findings in (Levin et al., 2015) for DTA models. This provides a starting point for various possible algorithmic enhancements that try to decrease the changes in demand per iteration for these locations by e.g. constraining changes in demand on OD pairs using sensitive bottlenecks through the route choice model and/or averaging scheme (note that some of these enhancements where already tested as described in (Brederode et al., 2016b)). From this same starting point, it might be possible to develop a method to calculate a rough estimate of the expected convergence properties of a model given its network and level of OD demand.

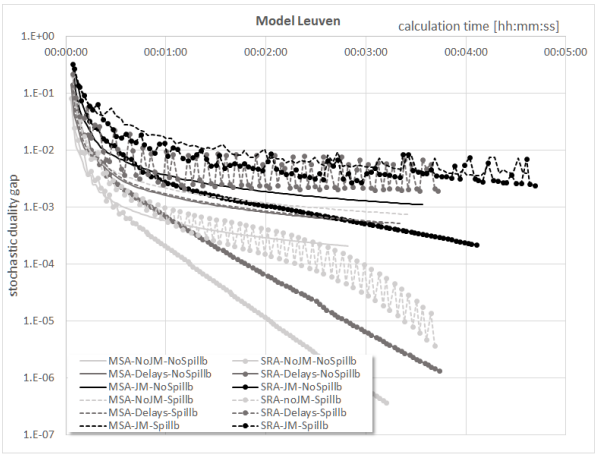
As pointed out in subsection 3.4.4, the calculation time per model component indicate that the network loading submodel of STAQ is relatively fast, such that efforts to further improve computational efficiency of STAQ are better put into other model components, primarily the route choice model.

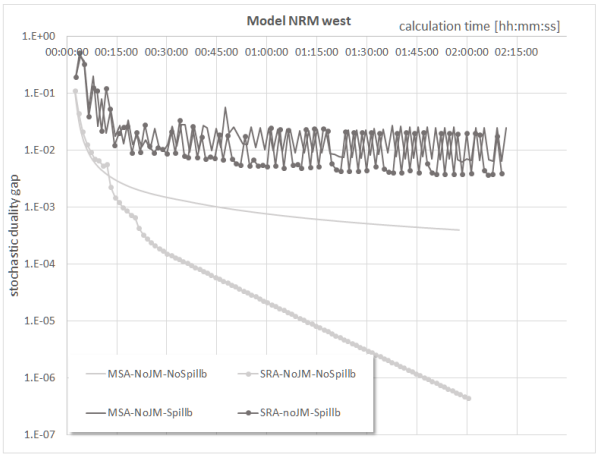
With respect to the route choice model, the paired combinatorial logit model (PCL, Pravinvongvuth and Chen 2005) is implemented as a STAQ variation. PCL adds support for route overlap and therefore allows inclusion of more relevant routes and thus is expected to improve convergence. To be able to test this hypothesis an adaptation of the duality gap for PCL (as has been done for MNL in equation (*3.2*)) needs to be derived.

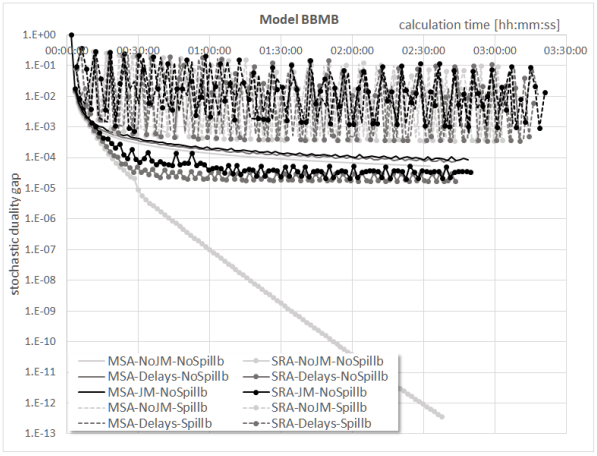
Finally, a recommendation with respect to the concept of STAQ. In its current form, STAQ effectively adds strict capacity constraints to STA models. However it still assumes stationary demand during a single time period. This means that the ‘true’ demand should always be averaged or aggregated in some way over the time period. To reduce averaging errors, an extension to STAQ that allows for multiple time periods would be needed. This would close the gap with DTA models further, however at the same time most likely will introduce new problems, such as more input requirements, poor convergence properties and longer calculation times. If these can be accepted or overcome, it would require for residual traffic to be transferred from one period to the next period. Such a mechanism would also solve another problem: residual traffic due to trip durations longer than the duration of the single time period, which can occur when dealing with large networks and/or short time periods.

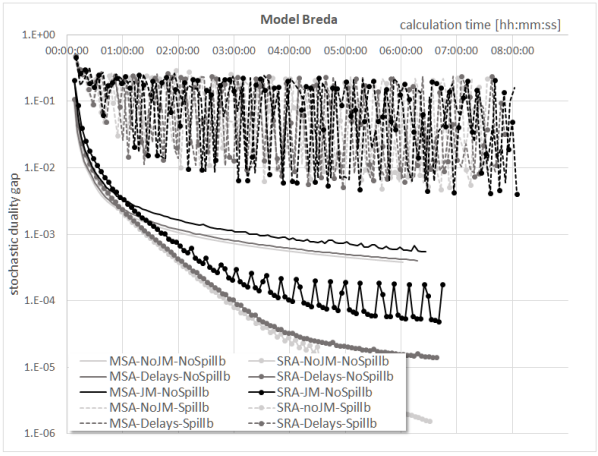
A. APPENDIX. DUALITY GAP VS. CALCULATION TIME FOR ALL TESTED MODELS AND RUNS

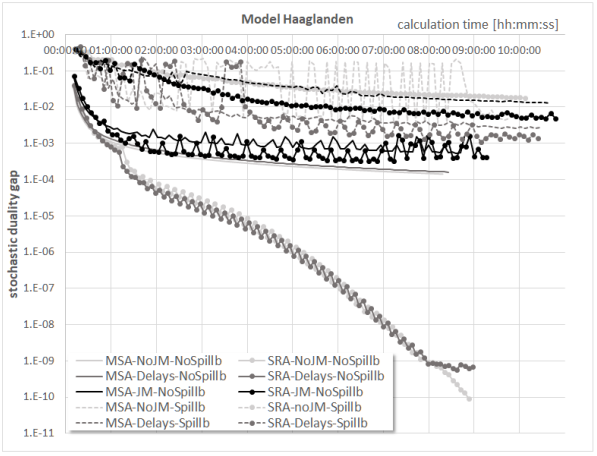
On the next page, the empirical relation between calculation time (on a Core I7-950 3.07 Ghz machine with 24 GBytes of memory) and convergence is displayed for all 7 models (Table 3.4) and all 12 model variations per model (Table 3.3). Each graph shows the runs on one model, and each curve in a graph represents a specific model run, its colour and shape indicate the combination of model components tested as displayed in the legend. The reds represent runs using the MSA averaging scheme, the greens represent runs using the SRA averaging scheme. Dashed curves represent runs where spillback is enabled, and continuous curves represent runs without spillback. The three different shades of both reds and greens represent the three different options for junction modelling.

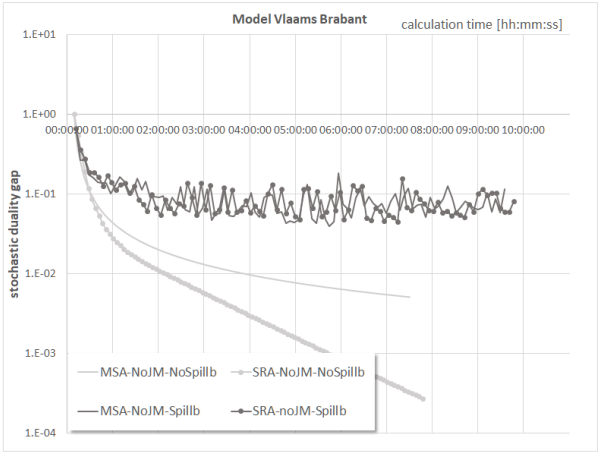

Model Leuven
calculation time [hh:mm:ss]
stochastic duality gap
MSA-NoJM-NoSpillb
SRA-NoJM-NoSpillb
MSA-Delays-NoSpillb
SRA-Delays-NoSpillb
MSA-JM-NoSpillb
SRA-JM-NoSpillb
MSA-NoJM-Spillb
SRA-noJM-Spillb
MSA-Delays-Spillb
SRA-Delays-Spillb
MSA-JM-Spillb
SRA-JM-Spillb

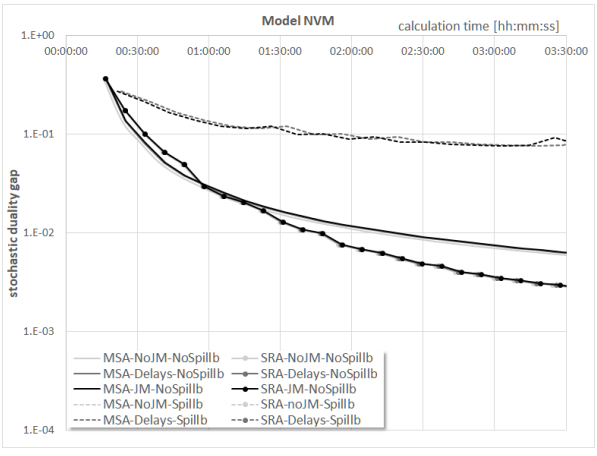

Model NRM west
calculation time [hh:mm:ss]
stochastic duality gap
MSA-NoJM-NoSpillb
SRA-NoJM-NoSpillb
MSA-NoJM-Spillb
SRA-noJM-Spillb


Model BBMB
calculation time [hh:mm:ss]
stochastic duality gap
MSA-NoJM-NoSpillb
SRA-NoJM-NoSpillb
MSA-Delays-NoSpillb
SRA-Delays-NoSpillb
MSA-JM-NoSpillb
SRA-JM-NoSpillb
MSA-NoJM-Spillb
SRA-noJM-Spillb
MSA-Delays-Spillb
SRA-Delays-Spillb


Model Breda
calculation time [hh:mm:ss]
stochastic duality gap
MSA-NoJM-NoSpillb
SRA-NoJM-NoSpillb
MSA-Delays-NoSpillb
SRA-Delays-NoSpillb
MSA-JM-NoSpillb
SRA-JM-NoSpillb
MSA-NoJM-Spillb
SRA-noJM-Spillb
MSA-Delays-Spillb
SRA-Delays-Spillb


Model Haaglanden
calculation time [hh:mm:ss]
stochastic duality gap
MSA-NoJM-NoSpillb
SRA-NoJM-NoSpillb
MSA-Delays-NoSpillb
SRA-Delays-NoSpillb
MSA-JM-NoSpillb
SRA-JM-NoSpillb
MSA-NoJM-Spillb
SRA-noJM-Spillb
MSA-Delays-Spillb
SRA-Delays-Spillb


Model Vlaams Brabant
calculation time [hh:mm:ss]
stochastic duality gap
MSA-NoJM-NoSpillb
SRA-NoJM-NoSpillb
MSA-NoJM-Spillb
SRA-noJM-Spillb


Model NVM
calculation time [hh:mm:ss]
stochastic duality gap
MSA-NoJM-NoSpillb
SRA-NoJM-NoSpillb
MSA-Delays-NoSpillb
SRA-Delays-NoSpillb
MSA-JM-NoSpillb
SRA-JM-NoSpillb
MSA-NoJM-Spillb
SRA-noJM-Spillb
MSA-Delays-Spillb
SRA-Delays-Spillb

Chapter 4

Extension of a static into a semi-dynamic traffic assignment model with strict capacity constraints

Abstract
This paper presents a straightforward extension of a static capacity constrained traffic assignment model into a semi-dynamic version. The semi-dynamic model is more accurate than its static counterpart as it relaxes the empty network assumption, but, unlike its dynamic counterpart, maintains the stability and scalability properties required for application on large scale strategic transport model systems. Applications show that semi-dynamic queue sizes and delays are very similar to dynamic outcomes, only the congestion patterns differ due to omission of spillback. Static model outcomes do not resemble the semi-dynamic nor dynamic model on size, temporal nor spatial distribution of queues and delays. The static and semi dynamic models can reach user equilibrium conditions, whereas the dynamic model cannot. On a real-world transport model, the static model omits up to 76% of collective losses. It is therefore very likely that the empty network assumption influences (policy) decisions based on static model outcomes.

Keywords: STAQ, semi-dynamic, traffic assignment model; user equilibrium; large-scale

This chapter is a version of the following publication:
Brederode, L., Gerards, L., Wismans, L., Pel, A., Hoogendoorn, S., 2023. Extension of a static into a semi-dynamic traffic assignment model with strict capacity constraints. Transportmetrica A: Transport Science 0(0), pp.1–34. https://doi.org/10.1080/23249935.2023.2249118

CRediT author statement:
Luuk Brederode: Conceptualization, Methodology, Software, Investigation, Writing – original Draft, Writing – Review & editing, Visualization. **Lotte Gerards**: Investigation, Formal analysis, Writing – Review & editing, **Luc Wismans**: Writing – Review & editing, **Adam Pel**: Writing – Review & editing, **Serge Hoogendoorn**: Writing – Review & editing

4.1 Introduction

Strategic traffic assignment (TA) models are used to assess the long-term impact on route choices of transport policies and the design and management of transport systems. As road congestion has become a structural problem in ever more regions around the world, TA model accuracy in congested conditions has become more important.

Because strategic TA models are used for long term forecasting, their outcomes should represent stable conditions in which travelers have adapted their route choice behavior to the forecasted scenario. Stability conditions in TA models are mostly operationalized by imposing user equilibrium conditions, where research suggests that a duality gap value (DG, the metric most used to measure the level of disequilibrium) of 1E-04 or lower is needed in strategic context (Boyce et al., 2004; Brederode et al., 2019, 2016a; Caliper, 2010; Han et al., 2015; Patil et al., 2021). Imposing equilibrium conditions on large scale TA models involves iterative solution algorithms that are computationally expensive.

For strategic TA models, there is a clear trade-off between stability and computational requirements on the one hand and accuracy on the other hand (Bliemer et al., 2013; Brederode et al., 2019; Flötteröd and Flügel, 2015). For each type of TA model the trade-off is made differently. In this paper, the framework described in (Bliemer et al., 2017) is used to define and classify the level of accuracy for different types of TA models. By only considering equilibrium models, the three dimensional framework from (Bliemer et al., 2017) simplifies into the two dimensional framework depicted in Figure 4.1. In this framework, the accuracy of TA models is classified by their spatial and temporal assumptions, where static unrestrained TA models are the least accurate, while dynamic capacity and storage constrained TA models are the most accurate. Below the effects of the different spatial and temporal assumptions on the accuracy of TA models are summarized, for a thorough description of the assumptions themselves the reader is referred to (Bliemer et al., 2017).

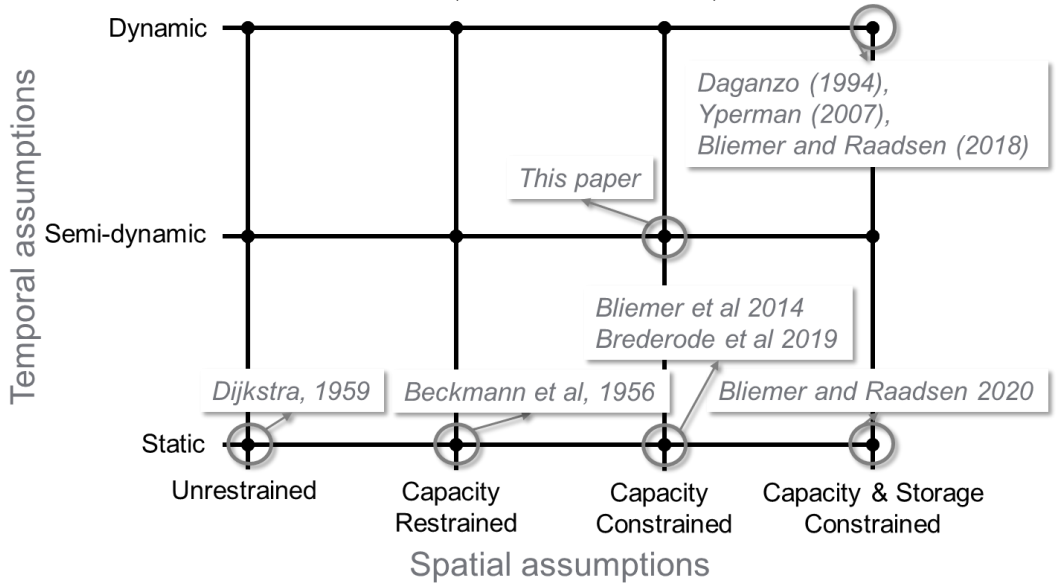


Figure 4.1: simplified framework for classification TA models

The spatial assumptions consider the effect of limited supply (capacity) on network usage and conditions. In unrestrained models (e.g.: All-Or-Nothing assignment), limited supply has no effect on the model outcome, whereas in capacity restrained models (e.g.: traditional static assignment models using BPR functions) route choice changes may occur due to limited supply although demand is still allowed to exceed supply. In capacity constrained models, route choice changes and vertical queues (and hence reduced flow downstream) may occur, whereas in

capacity and storage constrained models, route choice changes and horizontal queues (and hence spillback upstream and reduced flow downstream) may occur.
The temporal assumptions consider the effect that traffic that has departed but not arrived in previous time periods (residual traffic) has on network conditions (and hence usage) in the current time period. In static models, residual traffic has no influence on network conditions (i.e.: the model assumes an empty network at the start of the considered time period), whereas in semi-dynamic models residual traffic is transferred to the next time period (i.e.: the model considers the residual traffic that is on the network at the start of each time period). In dynamic models, period durations are very small (causing almost all traffic to be residual traffic), and network conditions are updated on link (or even cell-) level, thereby implicitly 'transferring' traffic and network start conditions to the next time period.

4.1.1 Research contribution, motivation and paper outline

This paper extends the static capacity constrained TA model described in (Bliemer et al., 2014; Brederode et al., 2019) to a semi-dynamic capacity constrained TA model and compares it to its static and dynamic counterparts on theoretical networks as well as a large scale realistic transport network. The motivation for the choice of a semi-dynamic capacity constrained TA model is described below.
For the last decades, emphasis has been mainly on the transition from *static* capacity restrained TA models to *dynamic* capacity and storage constrained TA models, where the most notable incarnations used in practice are the cell- (Daganzo, 1994) and link- (Yperman, 2007) transmission models. Although substantial progress for this model class has been made, both on computational efficiency (Bliemer and Raadsen, 2019; Canudas-de-Wit and Ferrara, 2018; Himpe et al., 2019, 2016; Petprakob et al., 2018; Simoni and Claudel, 2020) as well as stability (Ge et al., 2020), as far as the authors are aware, there are still no examples where the stability requirement for strategic applications (duality gap values below 1E-04) is met.
More recently, research into *static* capacity and storage constrained TA models shows that also these models fail to meet the stability requirement (Bliemer and Raadsen, 2020; Brederode et al., 2019; Smith, 2013) whilst their computational requirements are not (yet) on acceptable levels for practitioners (Raadsen and Bliemer, 2019a).
Hence, adding storage constraints to any model (either static or (semi-)dynamic) breaks convergence, which is in-line with theoretical findings in e.g. (Han et al., 2015; Szeto and Lo, 2006) that show that spillback can cause a non-continuous route cost function leading to non-convergence (Friesz and Han, 2019). These findings are confirmed on multiple networks in (Brederode et al., 2019) where on several large scale networks a dynamic capacity and storage constrained TA model assuming stationary demand did not converge below duality gap values below 1E-02 whereas a static capacity constrained TA model on the same networks converged to duality gap values well below 1E-04.
As shown in (Bliemer et al., 2014; Brederode et al., 2019), static capacity constraint TA models already greatly improve accuracy in congested conditions compared to capacity restrained TA models while maintaining scalability and stability properties required for strategic applications. To further increase accuracy without losing stability and/or low computational requirements, a shift from static capacity constrained towards semi-dynamic capacity constrained TA model seems to have the most potential.
Additionally, there are two application types for TA models that would greatly benefit from a shift from a static to a semi-dynamic TA model. Firstly, because semi-dynamic TA models do not assume an empty network at the start of the assignment, they provide a much better application context to add departure time choice models to the transport model system than their static counterparts do. Secondly, in the context of demand matrix estimation, a semi-

dynamic TA model allows to account for observed link flows that are partially composed of traffic that departed prior to the considered time period.
In conclusion, the semi-dynamic capacity constrained TA model is selected because, contrary to capacity and storage constrained TA models, it is expected to meet stability and computational requirements whilst it is expected to substantially improve accuracy compared to its static capacity constrained counterpart, especially in the application contexts of departure time choice modelling and demand matrix estimation.
The remainder of this paper is organized as follows. Section 4.2 positions the proposed semi-dynamic capacity constrained TA model in the field, whereas section 4.3 describes the algorithms used to solve it and to derive collective losses and average delays from its outcomes. In section 4.4, the accuracy, stability and scalability properties of the model are evaluated using applications on two theoretical and one real scale transport model instance. In section 0 properties and limitations specific to the semi-dynamic TA model are discussed. Finally, conclusions are drawn in section 4.6 along with recommendations for further research.

4.2 From static to semi-dynamic: relaxing the empty network assumption

To further position the semi-dynamic TA model subject of this paper, this section describes how it is derived from its static capacity constrained counterpart described in (Bliemer et al., 2014) and compares it to semi-dynamic TA models in literature.

4.2.1 From a single to multiple time periods with stationary travel demand

To extend the static capacity constrained TA model described in (Bliemer et al., 2014) to the semi-dynamic capacity constrained TA model used in this paper, the model assumption of a single time period with stationary travel demand is relaxed into the assumption that there are multiple time periods with stationary travel demand. Other model properties and assumptions are maintained, most importantly: the node model from (Tampère et al., 2011) and a fundamental diagram with horizontal hypercritical branch are used. With respect to flow propagation, instantaneous forward propagation of vehicles is assumed on uncongested links, whereas the horizontal hypercritical branch of the fundamental diagram imply backward wave speeds of zero and hence vertical queues.
The instantaneous forward propagation assumption means that all traffic that is not held up in queues by definition arrives at its destination within the duration of the considered time period. Strictly adhering to this assumption, the semi-dynamic TA model in this paper only transfers traffic held up in queues to the subsequent time period where it may re-evaluate its route choice. By doing so, the favorable scalability and stability properties of the static capacity constrained TA model from (Bliemer et al., 2014) are maintained.
To conclude: the relaxation from a single into multiple time periods in combination with the other (unchanged) model assumptions effectively means that it is no longer assumed that the network is empty at the start of each time period, but already contains traffic that was held up in queues in the previous time period.

4.2.2 Semi-dynamic traffic assignment in literature

Table 4.1 provides an overview of literature on semi-dynamic[13] traffic assignment models.

[13] By the definition from (Bliemer et al., 2017); some papers use the term quasi-dynamic instead

Publication	TA model type	Queues	Location and amount of residual traffic	Removal of flow downstream from bottleneck	user equilibrium type
Nakayama et al., 2012	Capacity restrained	none	based on total travel time from speed flow function	no	Deterministic
Bui et al., 2019; Chan et al., 2021; Fusco et al., 2013; Koike et al., 2022; Nakayama and Connors, 2014				yes, as a post processing step to the TA model	Deterministic
Bell et al., 1996	Capacity constrained	Vertical at downstream end of bottleneck link	based on location and size of vertical queues	no	Stochastic
Lam et al., 1996				no	Deterministic
Lam and Zhang, 2000				yes, part of assignment model	Deterministic
This paper		Vertical at node affected by capacity constraint(s)			Stochastic

Table 4.1: semi-dynamic traffic assignment models in literature

Some papers have deliberately been left out of this overview because they are either in Japanese (Akamatsu et al., 1998; Fujita et al., 1989, 1988; Kikuchi and Akamatsu, 2007; Miyagi and Makimura, 1991; Nakayama, 2009) or they describe models merely as algorithms without a mathematical problem formulation (Davidson et al., 2011; Taylor, 2003; Van Vliet, 1982), which makes it hard to compare them to the model used in this paper.

Based on the type of TA model used, two types of approaches can be distinguished within Table 4.1. Approaches using a capacity restrained TA model (the first two rows in Table 4.1) omit modelling of queues, but derive the location and amount of residual traffic by comparing the total travel time to the duration of the considered time period, assuming a uniform distribution of departure times within the considered period. This type of approach requires a post processing step to remove flow from links downstream from the location where residual flow was transferred to the next time period. This post processing step reduces flows and hence increases speeds on downstream links, thereby introducing a feedback loop around the user equilibrium within each period itself. Early approaches (Nakayama et al., 2012 and some of the papers written in Japanese) omit removal of flow from downstream links, thereby removing the need for the feedback loop. More recent approaches (Bui et al., 2019; Chan et al., 2021; Fusco et al., 2013; Koike et al., 2022; Nakayama and Connors, 2014) do remove downstream flow and focus on solving the optimization problem that arises from it.

Approaches using a capacity constrained TA model (the last four rows in Table 4.1) do model queues, and only transfers traffic that is held up in them to the next time period. To enforce capacity constraints, the TA models in (Bell et al., 1996; Lam et al., 1996; Lam and Zhang, 2000) employ exit link capacities, thereby assuming vertical queues on the downstream end of bottleneck links, whereas the TA model used in this paper (last row in) more accurately puts the vertical queue on the upstream end of the bottleneck link. Just like (Nakayama et al., 2012), the approaches in (Bell et al., 1996; Lam et al., 1996) omit removal of queued flow from downstream links, whereas in the approach in this paper as well as (Lam and Zhang, 2000) queued flow is removed from downstream links as part of the TA model itself.

4.3 Solution algorithm

To extend the static capacity constrained TA model described in (Bliemer et al., 2014) into a semi-dynamic version, our previous static capacity constrained TA model implementation (Brederode et al., 2019) was not altered. Instead, as shown in Figure 4.2, it is set in a loop with

a residual traffic transfer module and post processing modules that update cumulative in- and outflow curves and derive travel times from those curves are added.
The static capacity constrained TA model is described in subsection 4.3.1. It is applied for each time period $k \in K$ with start time t_k, yielding a set of inflows for all links $\mathbf{u}_k$, route choice probabilities ($\boldsymbol{\pi}_k$) for each route in the route set and a set of flow acceptance factors $\boldsymbol{\alpha}_k$ (representing the proportion of flow that is not held up in a queue) for all links that have a vertical queue on its downstream node.
The residual traffic transfer module (subsection 4.3.2) uses the route choice probabilities and link flow acceptance factors to transfer flow that is held up in queues ($\mathbf{Q}_k$) to the travel demand matrix of the next time period and to update the route set to include routes from the location of vertical queues to the original destinations of routes traversing these queues.
Finally, modules that update cumulative in- and outflow curves (subsection 4.3.3) and conduct delay calculations on link-, route- and network level (subsection 4.3.4) are added to derive collective losses and average delays per (departure) time period from link inflows and flow acceptance factors. Note that the loop consisting of the static capacity constrained TA model and residual traffic transfer together form the semi-dynamic TA model. Further note that the cumulative flow updating and delay calculation modules are optional post process modules and that (repetitive) delay calculation on route level also takes place as part of the static capacity constrained TA model using the delay formulation from (Bliemer et al., 2014). The difference is that the delay calculation within the TA model (subsection 4.3.1) assumes an empty network after the current time period (hence, it uses 'static delays'), whereas the semi-dynamic delay calculation (subsection 4.3.4) takes traffic on the network in the next time period into account (hence it derives 'semi-dynamic delays').

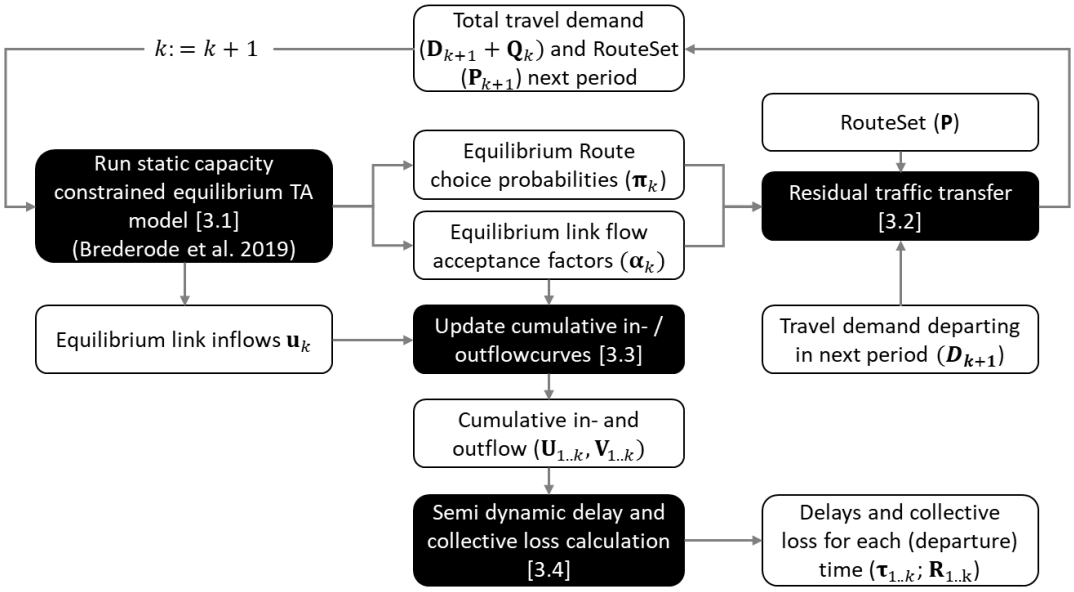


Figure 4.2: overview of the proposed solution algorithm (subsection numbers between squared brackets)

4.3.1 Static capacity constrained TA model

The TA model STAQ described in (Brederode et al., 2019) is the central module in the solution algorithm proposed in this paper. Several variations of the propagation model within STAQ are described in (Brederode et al., 2019), varying with respect to the nature of queues (horizontal or vertical), the fundamental diagram (triangular or Quadratic-Linear) and the inclusion of junction modelling (disabled, only turn delays or turn delays and turn flow restrictions).

The STAQ variation with vertical queues is used to converge towards the stochastic user equilibrium (Fisk, 1980), after which a single iteration with horizontal queuing is conducted to translate equilibrium queues horizontally. The quadratic-linear fundamental diagram (Figure 4.3, right) is selected, which -while equilibrating based on vertical queues- simplifies to quadratic-horizontal fundamental diagram (Figure 4.3, left). With respect to junction modelling, both turn delays and turn flow restrictions are included.

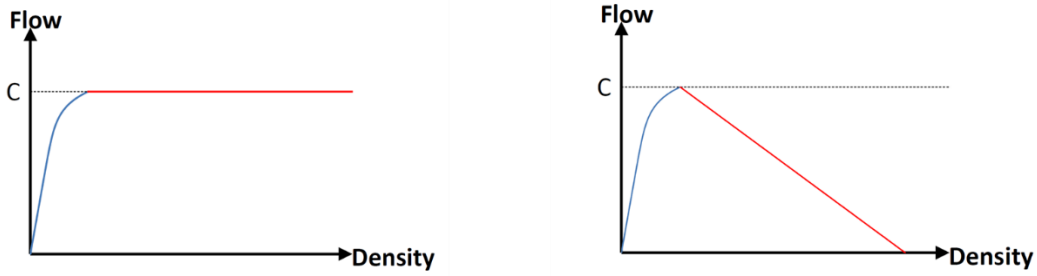


Figure 4.3: left: quadratic-horizontal fundamental diagram (used while equilibrating route demands); right: quadratic-linear fundamental diagram (used while translating vertical equilibrium queues into horizontal queues)

A pre-generated route set **P** is used, derived from the digitized transport network combining the Dijkstra algorithm to find shortest routes with the repeated random sampling process on free flow link travel times from (Fiorenzo-Catalano, 2007) to generate alternative routes. To reduce route overlap, remove irrelevant routes and restrict the size of the route set, route set filters are applied.

Route choices are modelled through the multinomial logit (MNL) model such that route choice probability $\pi_{p,rs}$ for route p on OD-pair rs is defined by

$$\pi_{p,rs} = \exp(-\mu_{rs} c_p) / \sum_{p' \in P_{rs}} exp(-\mu_{rs} c_{p'}) \tag{4.1}$$

where c_p represents the route cost on route p, μ_{rs} is the scale parameter describing the degree of travelers' perception errors on route travel times and P_{rs} is the set of routes between r and s. Note that for brevity in this subsection the argument (k) is omitted from all variables, because within the static capacity constrained TA model, there are no relationships with other time periods. Note that, without loss of generality, but at the cost of computational efficiency, the MNL route choice model could be replaced by more advanced route choice models (from e.g.: Prato, 2009; Smits et al., 2018).

To enforce and speed up equilibration of route demands, for non-theoretical test applications, the self-regulating average (Liu et al., 2009) is used to average route choice probabilities over iterations. To check for convergence to conditions of the stochastic user equilibrium conditions, the adapted relative duality gap as derived in (Bliemer et al., 2013) is used, which accounts for perception errors and thus reaches zero upon convergence when using the MNL route choice model:

$$DG = \frac{\sum_{rs \in RS} \sum_{p \in P_{rs}} \pi_p^{rs} D_{rs} (c_p + -\mu_{rs}^{-1} \ln (\pi_p^{rs} D_{rs} - \zeta_{rs})}{\sum_{rs \in RS} D_{rs} \zeta_{rs}} \tag{4.2}$$

where RS is the set of OD-pairs, D_{rs} the demand on OD-pair rs and $\zeta_{rs} = \min_{p \in P_{rs}} [c_p + \mu_{rs}^{-1} \ln \pi_p^{rs} D_{rs}]$ represents the minimum stochastic route cost on OD pair rs. In line with (Boyce et al., 2004; Brederode et al., 2019; Han et al., 2015; Patil et al., 2021), for non-theoretical test applications in this paper, a threshold value of 1E-04 is used as the stop criterion for the traffic assignment model.

Next to link inflows **u** and route choice probabilities **π,** the TA model provides flow acceptance factors α_{ij} for each turning movement (turn) from inlink i to outlink j. Flow acceptance factors represents the proportion of flow that passes that turn, the remainder of flow using the turn (i.e. proportion $1-\alpha_{ij}$) is left on the turn as a vertical queue. To ensure that all traffic reaches its destination according to the route definitions in **P** and route probabilities **π**, the node model within STAQ assumes the first-in-first-out (FiFo) principle (Daganzo, 1995), which in the context of STAQ means that flow acceptance factors for all turns sharing an inlink are equal, i.e.: $\alpha_i = \alpha_{ij}$ for all outlinks j connected to the node considered.

4.3.2 Residual traffic transfer

The goal of the residual traffic transfer module is to transfer travel demand that was held up in queues in the TA model for period k to the travel demand matrix for period $k+1$ such that it resumes its route from the location of the queue to its original destination.

The set of acceptance factors on turns with residual queues for period k is denoted by $\boldsymbol{\alpha}_k$. The acceptance factors are used together with the OD-demands $\mathbf{D}_k = \{D_{rs}(k)\ \forall rs \in RS\}$ and the route choice probabilities $\boldsymbol{\pi}_k = \{\pi_{p,rs}(k) \forall p \in P\}$ to calculate the amount of residual traffic $Q_{ij,s}(t_k)$ between turn ij and destination s using:

$$Q_{ij,s}(t_k) = \big(1-\alpha_i(k)\big) \sum_{p \in P_{ij}} \pi_{p,rs}(k) D_{rs}(k) \prod_{i'j' \in IJ_{p,i}} \alpha_{i'}\,(k) \tag{4.3}$$

where P_{ij} represents the set of routes traversing considered turn ij and $IJ_{p,i}$ represents the set of turns on route p up to but excluding the turn from link i to link j. The first term in equation (4.3), represents the proportion of demand that is held up in the queue on turn ij, whereas the second term represents the amount of demand that arrives at link i, taking reductions due to upstream queues into account.

The amount $Q_{ij,s}(t_k)$ of traffic is then transferred to the travel demand matrix of the next time period using a centroid connected to the upstream node of the link with the vertical queue as origin and the original destinations of the paths in P_{ij}. Furthermore, partial routes between this new centroid and the destinations of routes passing it are added to the route set.

Note that the centroid could also be directly connected to the node from which the vertical queue was transferred, allowing to steer the priority of the transferred demand over demand departed in the current period, by altering the capacity of the connector link. This idea will be further discussed in subsection 4.6.2, but for the sake of scalability and simplicity of the algorithm (it would require bookkeeping of transferred demand), its implementation and analysis is left for further research.

4.3.3 Updating cumulative in- / outflow curves

After each time period, for each link i on which a vertical queue remains or has remained in previous time periods, static cumulative in- and outflow curves are constructed using the duration of the time period and the link inflows and acceptance factors from the assignment (Figure 4.4, left). From these curves, the semi-dynamic piece-wise linear cumulative in- and outflow curves (Figure 4.4, right) are updated using:

$$\begin{aligned} U_i(t_k) &= U_i(t_{k-1}) + u_i(k)(t_k\text{-}t_{k-1}) - Q_i(t_{k-1}) \\ V_i(t_k) &= V_i(t_{k-1}) + \alpha_i(k) u_i(k)(t_k\text{-}t_{k-1}) \end{aligned} \tag{4.4}$$

where $U_i(t_k)$ and $V_i(t_k)$ represent the cumulative inflow and outflow respectively for link i at the end of period k, and $u_i(k)$ represents the inflow rate of link i during period k. Note that to

avoid double-counting, as illustrated by the dashed cumulative inflow curve in right part of Figure 4.4, the residual traffic transferred on link i from the previous period, $Q_i(t_{k-1})$, is subtracted from the static cumulative inflow curve since $u_i(k)$ includes residual traffic transferred from period k-1. For the interested reader, two expressions to derive $Q_i(\cdot)$ are shown in appendix A.

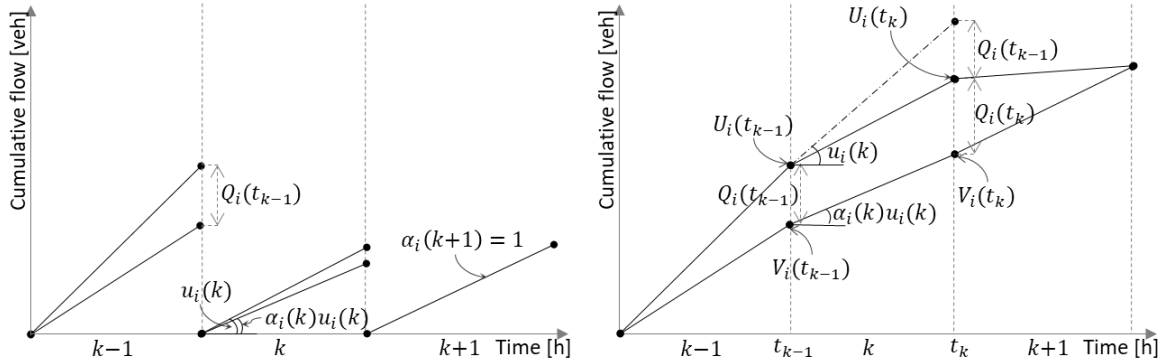


Figure 4.4: example of consecutive static cumulative flow curves (left) and corresponding semi-dynamic cumulative flow curves (right)

As first described in (Lo and Szeto, 2002), horizontal distances between cumulative inflow and outflow curves represent the link travel time at a given point in time. Because STAQ assumes instantaneous forward flow propagation (subsection 4.2.1), here the distance between the cumulative inflow curve and the cumulative outflow curve represents the link delay instead of the travel time. Additionally, the stationary travel demand assumption in STAQ causes the cumulative flow curves to be piece wise linear with tipping points only occurring at the start time of the time periods defined for the semi-dynamic TA model, allowing for easy calculation of time averaged delays by dividing the surface between the curves (representing collective losses) by the number of vehicles experiencing this collective loss.

4.3.4 Calculating collective loss and average delays

Collective travel time losses can be derived either from the perspective of the network operator or the traveler. From the network operator's perspective, $\widetilde{R}_i(k)$ represents the collective time loss of vehicles *using* link i during period k (Figure 4.5, left), whereas from the traveler's perspective $R_i(k)$ represents the collective time loss of vehicles using link i that have *departed* (or resumed as residual traffic from a queue) during period k (Figure 4.5, right).

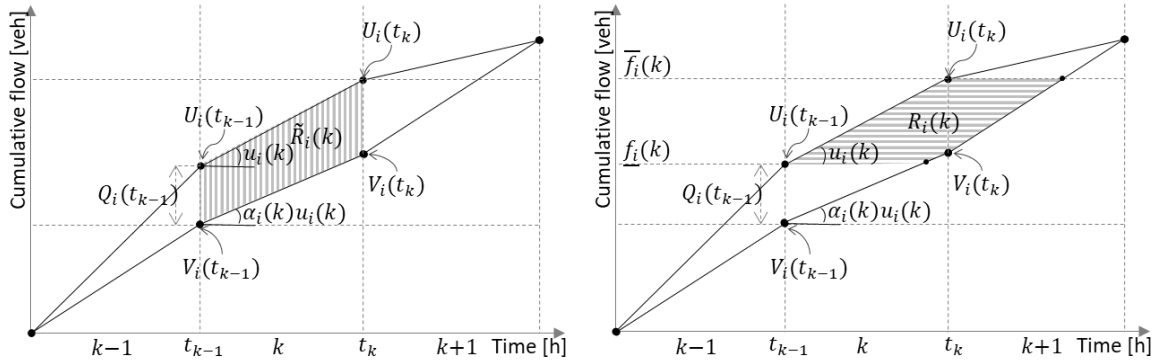


Figure 4.5: example of cumulative link flow curves for three time periods. Left: shaded area represents the collective loss from the network operator's perspective. Right: shaded area represents the collective loss from the traveler's perspective (for a link that is not used by routes affected by upstream bottlenecks in time periods *k*-1 and k)

Which of the perspectives is most appropriate depends on the application type. Table 4.2 summarizes typical application types for collective time loss and average delay calculation for both perspectives and for different levels of aggregation (link-, route- and network level).

	Collective loss / average delay experienced by vehicles…	
	… on the network within a given time period **Network operator's perspective**	…departed within a given time period **Traveler's perspective**
Link-level	• Application type: To determine where and when collective loss occurs • Collective loss on link i during period k: $\tilde{R}_i(k)$ - equation (4.5) • Average delay per vehicle on link i during period k: $\tilde{\tau}_i(k)$ - equation (4.6)	• Application type: To derive the od-link incidence indicator within semi-dynamic matrix estimation • Collective loss of vehicles using link i departed during k: $R_i(k)$ - equation (4.14) • Average delay per vehicle departed during period k on link i: $\tau_i(k)$ - equation (4.17)
Route-level	• Not relevant from network operator's perspective, as for travelers, only the non-instantaneous travel time to complete a route is of importance.	• Application type 1: route cost calculation within the semi-dynamic user equilibrium; • Application type 2: to use in a departure time choice model; • Average delay per vehicle departed during period k using route p upto link i: o for application type 1: $\dot{\tau}_{p,i}(k)$ - equation (4.19) o for application type 2: $\tau_{p,i}(k)$ - equation (4.20)
Network-lvl	• Application type: To quantify network performance over time • Network wide collective loss during period k: $\tilde{R}(k)$ - equation (4.7)	• Application type: To quantify network performance per departure time • Network wide collective loss for vehicles departed during period k: $R(k)$ - equation (4.18)

Table 4.2: Application types for travel time calculation per aggregation level for network operator's and traveler's perspectives

4.3.4.1 Network operator's perspective

On the link level, a typical application type for the network operator's perspective is to determine on which links and during which time periods collective losses occur. Illustrated in the right part of Figure 4.5, the surface representing the collective loss in network operator's perspective may be calculated using simple geometry:

$$\tilde{R}_i(k) = \frac{(t_k\text{-}t_{k-1})(U_i(t_{k-1})\text{-}V_i(t_{k-1}) + U_i(t_k)\text{-}V_i(t_k))}{2} \quad (4.5)$$

where $\tilde{R}_i(k)$ represents the collective loss for traffic on link i during period k. Average delay is derived by dividing collective loss by the number of vehicles using the link during period k:

$$\tilde{\tau}_i(k) = \frac{\tilde{R}_i(k)}{U_i(t_k)\text{-}V_i(t_{k-1})} \quad (4.6)$$

where $\tilde{\tau}_i(k)$ represents the average delay per vehicle on link i during period k.

On the network level, a typical application type for the network operator's perspective is to quantify network performance over time. This is done by simply summing all link level collective losses:

$$\tilde{R}(k) = \sum_i \tilde{R}_i(k), \quad (4.7)$$

where $\tilde{R}(k)$ is the network wide collective loss during period k.

Note that equations (4.5) and (4.6) are the equivalent to the average delay function from the 'longitudinal semi-dynamic perspective' in (Raadsen and Bliemer, 2019b), albeit that their formulation represents the average delay experienced by traffic *flowing out of* link i during

time period k (it divides collective loss by the amount of traffic that has flown out), whereas our formulation for the network operator's perspective represents the average delay experienced by traffic *on link* i during time period k (equation (4.6) divides collective loss by the amount of traffic that was on the link). Further note that on the route level, no application type for the network operator's perspective could be identified, so expressions for this aggregation level are omitted.

4.3.4.2 Traveler's perspective – link and network level

On the link level, a typical application type for the traveler's perspective is to derive the od-link incidence indicator within (semi-dynamic) matrix estimation on observed link travel times. To determine collective loss and average delays from the traveler's perspective, three steps are conducted. First, the range of cumulative flow that departed in the considered departure time period is derived. Then the points in time on which the first and last vehicles entered and exited the considered link are calculated and added to the set of relevant points in time. Finally, the collective loss and link delays for traffic that departed in the considered time period are derived.

4.3.4.2.1 Calculation of cumulative flow levels of first and last vehicle departed in considered time period

For reasons of clarity, two types of links are distinguished when determining the cumulative flow range. The first type are links that in the previous time period were only used by traffic on routes unaffected by active bottleneck(s) upstream (i.e.: links that were only used by paths for which $\prod_{i'j' \in IJ_{p,i}} \alpha_{i'}(k) = 1$). Illustrated in the right part of Figure 4.5, for these links the cumulative flow that departed during period k is identified as

$$\left[\underline{f_i}(k)..\overline{f_i}(k)\right] = [U_i(t_{k-1})..U_i(t_k)], \tag{4.8}$$

where $\underline{f_i}(k)$ and $\overline{f_i}(k)$ represent the cumulative flow levels corresponding to the first and last vehicle that departed during period k respectively. Note that this range excludes traffic $Q_i(t_{k-1})$ as it departed during the previous period but was held up in a vertical queue on the considered link i itself.

The second type are links that are used by at least one route that was affected by active bottleneck(s) upstream (i.e.: links that were (also) used by paths for which $\prod_{i'j' \in IJ_{p,i}} \alpha_{i'}(k) < 1$). Illustrated in Figure 4.6, for these links, demand that was held up during the previous time period in bottlenecks upstream ($\bar{Q}_i(t_{k-1})$) is deducted, while the demand that was held up during the current time period in bottlenecks upstream ($\bar{Q}_i(t_k)$) is added to the range of cumulative flow that has departed in the considered time period, thus defining the relevant cumulative flow levels as:

$$\left[\underline{f_i}(k)..\overline{f_i}(k)\right] = [U_i(t_{k-1})+\bar{Q}_i(t_{k-1})..U_i(t_k)+\bar{Q}_i(t_k)]. \tag{4.9}$$

Note that $\underline{f_i}(k+1) = \overline{f_i}(k)$, hence in each time period, only $\overline{f_i}(k)$ needs to be computed.

The amount of demand held up in upstream bottleneck(s) is derived by summing residual traffic from queues on upstream turns $i'j'$ on routes that use link i using:

$$\bar{Q}_i(t_k) = \sum_{p \in P_i} \sum_{i'j' \in IJ_{p,i}} Q_{i'j',s_p}(t_k), \tag{4.10}$$

where P_i is the set of routes using the considered link i and s_p the destination of route p. Alternatively, $\bar{Q}_i(t_k)$ can also be derived on link level by deducting the link inflow from the link demand:

$$\bar{Q}_i(t_k) = \sum_{p \in P_i} \pi_{p,rs}(k) D_{rs}(k) - u_i(k)(t_k - t_{k-1}). \tag{4.11}$$

One might argue that, when $\bar{Q}_i(\cdot)$ partly consists of residual traffic that was held up in an upstream bottleneck in the current time period after it was transferred from a bottleneck even further upstream from a previous time period, range (4.9) would not be continuous. However, it is assumed that it is, in order to avoid bookkeeping of transferred demand (subsection 4.3.2), thereby maintaining computational efficiency and simplicity.

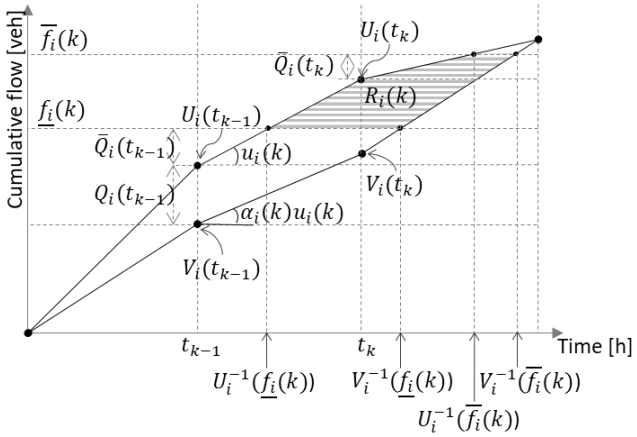


Figure 4.6: example of collective loss calculation from a traveler's perspective on a link where at least one route was affected by upstream bottleneck(s) during periods *k*-1 and *k*

4.3.4.2.2 Determination of the set of relevant points in time

As illustrated in Figure 4.6, the inverse functions of equation (4.4) are used to determine the entry and exit times of the first vehicle as $U_i^{-1}(\underline{f_i}(k))$ and $V_i^{-1}(\underline{f_i}(k))$ respectively, and the entry and exit times of the last vehicle as $U_i^{-1}(\overline{f_i}(k))$ and $V_i^{-1}(\overline{f_i}(k))$ respectively, together forming the set of entry/exit times:

$$\mathcal{F} = \left\{ U_i^{-1}\left(\underline{f_i}(k)\right), V_i^{-1}\left(\underline{f_i}(k)\right), U_i^{-1}\left(\overline{f_i}(k)\right), V_i^{-1}\left(\overline{f_i}(k)\right) \right\}. \tag{4.12}$$

From the set of all start times $\mathcal{T} = \{t_k \forall k \in K\}$, the start times that lie between the first entry and last exit time are added such that the set of relevant points in time is defined as:

$$\mathcal{G} = \mathcal{F} + \left\{ t \in \mathcal{T} : U_i^{-1}\left(\underline{f_i}(k)\right) < t < V_i^{-1}\left(\overline{f_i}(k)\right) \right\}. \tag{4.13}$$

After ordering from low to high, elements in $\mathcal{G}$ are referred to as t_n with $n = 1..|\mathcal{G}|$.

4.3.4.2.3 Calculation of collective loss and average delay per departure time period

The points in time in $\mathcal{G}$ together with the associated cumulative flow levels describe the vertices of the triangles and quadrilaterals that, when summed, form the total surface representing the collective loss for traffic departed in a considered time period k:

$$R_i(k) = \sum_{n \in 1..|\mathcal{G}|} r_i(k, t_n), \tag{4.14}$$

with

$$r_i(k, t_n) = \frac{(t_n - t_{n-1})\left(\overline{U_i}(k, t_{n-1}) - \underline{V_i}(k, t_{n-1}) + \overline{U_i}(k, t_n) - \underline{V_i}(k, t_n)\right)}{2} \tag{4.15}$$

and

$$\overline{U_i}(k, t_n) = \min\left[U(t_n), \overline{f_i}(k)\right] \tag{4.16}$$

$$\underline{V_i}(k, t_n) = \max\left[V(t_n), \underline{f_i}(k)\right]$$

Note that equation (4.15) is a special form of equation (4.5) that restricts the surface to within cumulative flow levels of the first and last vehicle departed in the considered time period.
Finally, average delay is derived by dividing collective loss by the number of vehicles using the link during period k:

$$\tau_i(k) = \frac{R_i(k)}{\overline{f_i}(k) - \underline{f_i}(k)} \tag{4.17}$$

where $\tau_i(k)$ represents the average delay per vehicle on link i that departed during period k.
The network wide collective loss from the traveler's perspective for traffic departed in period k ($R(k)$) is derived by taking the sum of collective losses per link per departure time from equation (4.14):

$$R(k) = \sum_i R_i(k) \tag{4.18}$$

4.3.4.3 Traveler's perspective – route level (for use in user equilibrium algorithms)

On the route level, the most important application type for the traveler's perspective is the evaluation of route delays within iterative solution algorithms imposing equilibrium conditions. This application type requires that the route delays are calculated using only information from the previous and current time period, as information for the next time period is dependent on the solution for the current time period and thus not available.
This means that, in contrast to the calculation of delay on link level from the traveler's perspective (subsection 4.3.4.2), the conditions of the next time period are not taken into account, as they are unknown. This is consistent with the definition of our semi-dynamic TA model in which only residual traffic is transferred to the next time period (subsection 4.2.1) and the definition of semi-dynamic TA models in general in (Bliemer et al., 2017). Also, as most travelers use information from route-planners that provide current instead of expected future route-delays, it probably leads to more realistic choice behavior.
The travel time on a route experienced by the traveler is equal to the summation of travel times on all links in the route, where for each link, the travel time at the time of entering the link is used. Hence, a trajectory through space (the links in the route) and time must be used to determine the travel time, given a specific departure time.
The instantaneous forward flow propagation assumption (subsection 4.2.1) in the static capacity constrained TA model causes demand on a route without upstream bottlenecks to arrive earlier at a specific link (i.e.: instantaneously) than demand on a route that experiences delay from some upstream bottleneck, whereas, as illustrated in the left part of Figure 4.4, the stationary travel demand assumption causes queues to be non-stationary, but growing (Bliemer et al., 2014; Gentile et al., 2015).
The combination of these two assumptions cause demand on a route without upstream bottlenecks to experience less delay on a specific link than demand on a route that experiences delay on some upstream bottleneck, which means that the delay on some bottleneck link i (i.e. $\alpha_i(k) < 1$) varies for different routes using the link, depending on the delay that was experienced in upstream bottleneck links. In other words, route queuing delays are non-separable over links whenever there's more than one bottleneck on the route. This was recognized in (Bliemer et al., 2014), who derived the following expression for the average route queuing delays on route p using only information from period k:

$$\dot{\tau}_{p,i}(k) = \frac{t_k - t_{k-1}}{2}\left(\frac{1}{\prod_{ij \in IJ_{p,i}} \alpha_i\,(k)} - 1\right). \tag{4.19}$$

4.3.4.4 Traveler's perspective – route level (for use in e.g. departure time choice models)

For application types in which information from time periods after the current time period can be used, more accurate (semi-dynamic) route delays can be derived. Typical application types would be the application of a departure time choice model based on travel times fed back from the semi-dynamic TA model and travel demand matrix estimation procedures that simultaneously handle multiple time periods.

For such application types, queuing delays on route p up to link i may be calculated by creating trajectories through (discretized) space and time, i.e.:

$$\tau_{p,i}(k) = \sum_{i' j' \in IJ_{p,i}} \tau_{i'}(k) \tag{4.20}$$

where $\tau_i(k)$ the delay on link i in period k calculated using equation (4.17). Note that although it allows to include information from time periods after the current time period, this approach ignores the non-separability of route queuing delays (discussed in subsection 4.3.4.3), just as (van der Gun et al., 2020) does. The implications that this inconsistency has on the application range are further discussed in 4.5.2.

Further note that, conceptually, equation (4.20) is in line with the approach to determine the amount of residual traffic adopted by the models in the first two rows of Table 4.1: comparing the total travel time to the duration of relevant time periods, assuming stationary travel demand (i.e.: a uniform distribution of departure times within each time period). However, equation (4.20) accounts for queues from the presented semi-dynamic capacity constrained TA model, whereas the other approach does not account for queues.

4.4 Applications

To give insights in the accuracy and stability of the proposed semi-dynamic TA model compared to its static and dynamic counterparts, this section compares its outcomes on two theoretical model instances (subsections 0 and 4.4.2) with those from the static capacity constrained TA model from subsection 4.3.1 (Bliemer et al., 2014) and the dynamic TA model described in (Bliemer and Raadsen, 2019). Furthermore, to give insights in the scalability and the effect of the empty network assumption, the comparison with the static TA model is repeated for a third, real scale model instance in subsection 4.4.3. Because the dynamic TA model is not able to converge to SUE conditions a comparison with it is omitted for this model instance.

Because the focus of the comparisons in this section is on differences in the TA models and not their inputs, the travel demand defined for the semi-dynamic model instances is also used as input to the static and dynamic TA model instances. For the static TA model instances this means that a sequence of static assignments (one for each period defined in the semi-dynamic travel demand) is run, whereas for the dynamic TA model instances this means that demand is kept stationary and during (relatively large) semi-dynamic time periods. The effect of this choice is that differences between the three types of models in this section are smaller than they typically will be in real world transport model applications, as these use a single (and larger) time period for static and more (and smaller) time periods for dynamic TA models.

The specific static and dynamic TA models are selected for comparison, as these are the closest related TA models that run on real scale transport model systems. This is illustrated in Table 4.3, which lists only the properties (using the definitions from (Bliemer et al., 2017)) for which the models differ. The differences with respect to the travel demand and traffic transfer relate to the temporal interaction assumptions whereas the other differences relate to the spatial interaction assumptions from Figure 4.1. Note that the differences with respect to the fundamental diagram and the constraints (last two columns) are considered a result of the

assumption with respect to the hypercritical wave speed. Because the static and semi-dynamic TA models assume a zero hypercritical wave speed, vertical queues are assumed, which means that: 1) the hypercritical branch of the FD becomes flat (as displayed in the left part of Figure 4.3); and 2) there cannot be storage constraints. This means that only three assumptions in Table 4.3 (travel demand, traffic transfer and hypercritical wave speeds) are the true drivers of all differences.

	Temporal interaction assumptions		Spatial interaction assumptions			
			Wave speeds			
Publication	**Travel demand**	**Traffic transfer**	**Hypo-Critical**	**Hyper-Critical**	**Fundamental diagram**	**Constraints**
Bliemer and Raadsen 2019	Dynamic	All residual traffic	Kinematic	Kinematic	Concave-linear	Capacity+storage
This paper	Semi-Dynamic	Queues only	Infinite	Zero	Concave-flat	Capacity only
Brederode et al 2019, Bliemer et al 2014	Static	None	Infinite	Zero	Concave-flat	Capacity only

Table 4.3: differences between the dynamic, semi-dynamic and static TA models compared in this section

4.4.1 Evaluating accuracy: corridor network with two bottlenecks

The corridor network with two bottlenecks was constructed to demonstrate differences in accuracy (of traffic conditions and collective losses) from network operator's perspective in a situation where both flow reduction downstream and spillback upstream occurs. Figure 4.7 displays the link capacities in pcu/h, whereas the top left graph of Figure 4.9 displays the travel demands for each of the seven defined one-hour time periods. Link lengths are 1 km for all three links. For the semi-dynamic and dynamic TA models, free flow speeds are 120km/h and (to simplify the examples) a triangular (instead of QL) fundamental diagram is used. For the dynamic TA model, jam density is set to 180pcu/km for each lane of 2000 pcu/h. Note that these inputs imply initial queues of 1000 pcu/h in front of the second and third links.

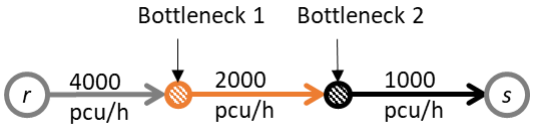


Figure 4.7: link capacities of the corridor network with two bottlenecks

Figure 4.8 summarizes link flows, conditions and vertical queues per time period as calculated by each of the three different TA models. Italics indicate the size of vertical queues at the end of the time period, dashed lines indicate congested links. In the dynamic TA model results, time averaged flows are reported, values starting with a tilde are rounded to the nearest 50 pcu/h. The upper right and bottom graphs in Figure 4.9 display the corresponding cumulative collective losses from the network operator's perspective. The cumulative collective loss (on link or network level) is the summation of collective losses from start of the simulation up to and including the considered time period.

These figures indicate that in the static TA model, the total collective loss amounts 2000 pcu-hours occurring only in $\{t_1, t_2\}$, and only due to queues that had built up in t_1 and shrink in t_2, whereas in the semi-dynamic and dynamic TA models, total collective loss amounts to some 7000 pcu-hours occurring in $\{t_1..t_5\}$ for both models. Furthermore, in the static and semi-dynamic TA models, there is no spillback, whereas in the dynamic TA model, during $\{t_1..t_5\}$ the queue from the second bottleneck spills back onto the first bottleneck location whilst during $\{t_1..t_3\}$ the queue spills back further onto the origin, where it is modelled as a vertical queue.

This comparison shows that the size and temporal distribution of queues and collective losses from the semi-dynamic and dynamic TA models are very similar, but the spatial distribution (i.e.: the congestion pattern) is different as the semi-dynamic TA model ignores spillback.

Furthermore it shows that the static TA model does not resemble the other two TA models on size, temporal nor spatial distribution of queues and collective losses.

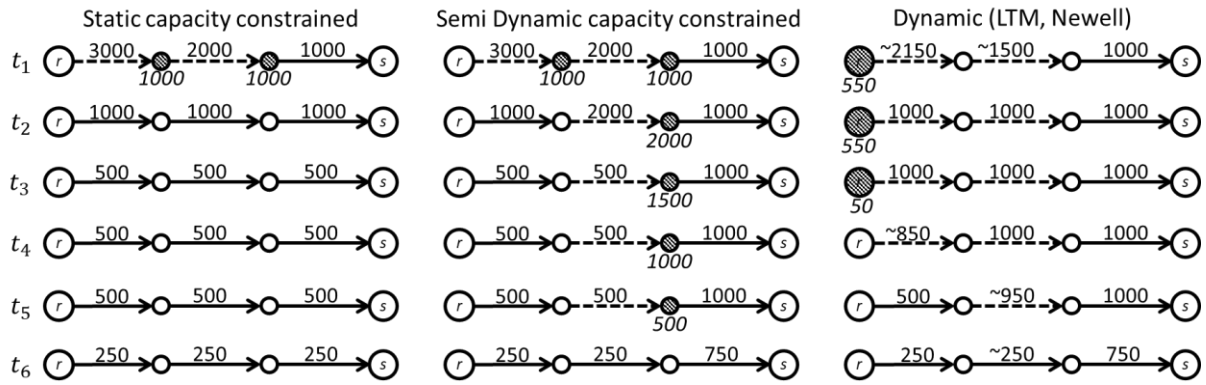


Figure 4.8: link flows [pcu/h] (regular font) and vertical queues [pcu*h] (italics below shaded nodes) per time period from the three TA models. Links conditions are indicated by continuous (uncongested) and dashed (congested) arrows.

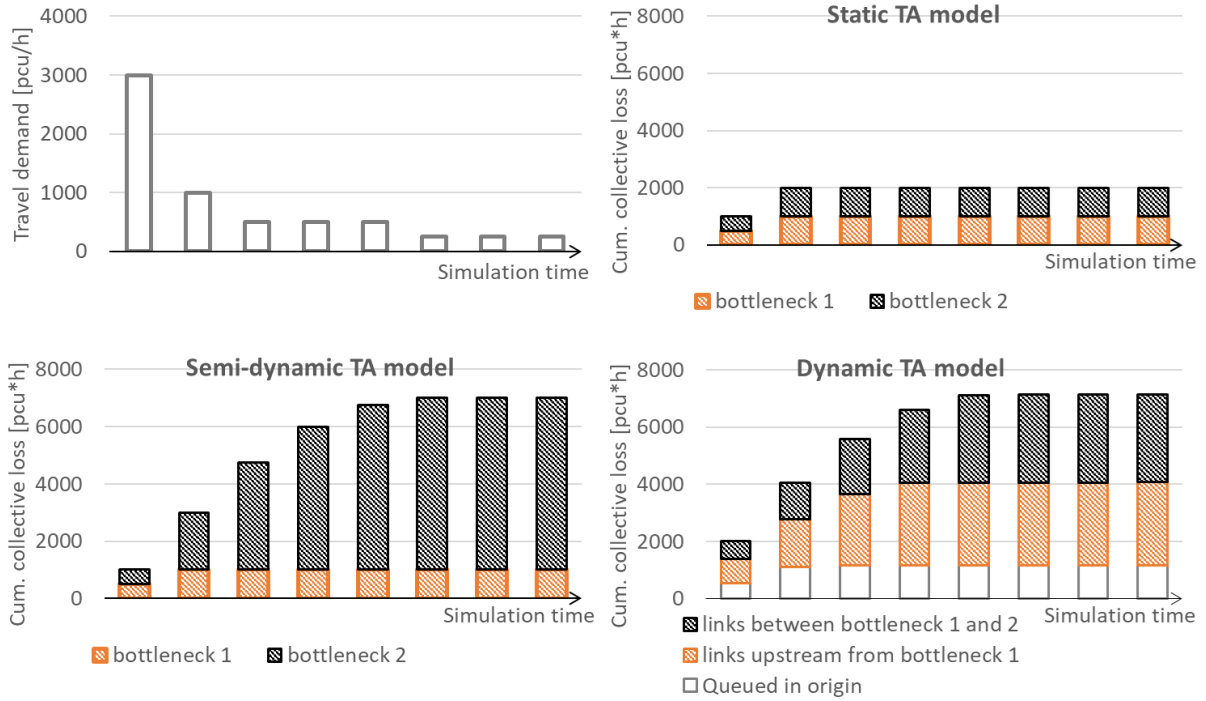


Figure 4.9: travel demand (top left) and corresponding cumulative collective vehicle losses from network operator's perspective per time period for the static (top right), semi-dynamic (bottom left) and dynamic (bottom right) TA models.

4.4.1.1 Detailed comparison between static and semi-dynamic TA model outcomes

Considering the static and semi-dynamic TA model outcomes in t_1, the collective losses are 1000 pcu-hours for both models, comprised of 500 pcu-hours (the average of a new queue growing to 1000 pcu in one hour) on both bottlenecks. The indifference between static and semi-dynamic outcomes is due to the empty network assumption (subsection 4.2.1) which in t_1 automatically holds for both models.

Considering the static and semi-dynamic TA model outcomes in t_2, the static TA model 'forgets' about the residual traffic on the network, causing equation (4.5) to only account for the collective loss due to the dissolving queues (500 pcu-hours for both bottlenecks), whereas in the semi-dynamic TA model, the residual traffic from t_1 (1000 pcu on both bottlenecks) and

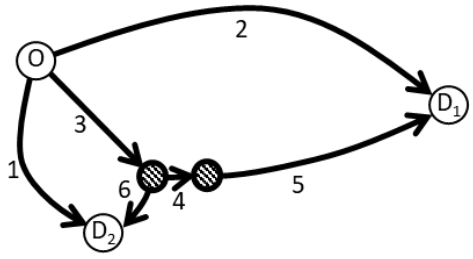
demand departed during t_2 causes a dissolved queue at the first bottleneck and a queue of 2000 pcu at the second bottleneck at the end of t_2. This effectively means that demand departed in t_2 equals the network outflow, i.e. the semi-dynamic outcomes include an additional stationary queue of 1000 pcu. In $\{t_3..t_6\}$, the static TA assumes no queues, whereas the semi-dynamic TA, the remaining queue shrinks until it has dissolved at the end of t_6, at which time the difference between static and semi-dynamic cumulative collective losses has risen to 5000 pcu-hours.

4.4.1.2 Detailed comparison between semi-dynamic and dynamic TA model outcomes

Considering the semi-dynamic and dynamic TA model outcomes, given the equal model inputs, all differences are related to spillback. Once the queue from the second bottleneck spills back, the effective capacity of the upstream links are reduced to 1000 pcu/h, whereas in the semi-dynamic model, they remain to their original values throughout the simulation. As inputs of the dynamic TA model imply a backward wave speed of ~12 km/h, spillback onto the first bottleneck occurs in t_1 after ~5 minutes, whereas spillback onto the origin occurs in t_1 after 15 minutes. Because traffic is confronted with a reduced capacity further upstream, losses during $t_1$are twice as high in the dynamic TA model. During t_2 however, the collective losses are practically equal, as for this time period, the residual demands from the previous time period in the semi-dynamic TA model are equal to the queues stored on the upstream links in the dynamic TA model, while in both models only bottleneck 2 is active. Although spatially distributed differently, total collective losses during $\{t_3..t_6\}$ are practically equal.

4.4.2 Evaluating stability: two congested OD pairs sharing a bottleneck

To demonstrate the stability (i.e.: the extent to which stochastic user equilibrium conditions are attained), the theoretical network displayed in Figure 4.10 is selected from (Brederode et al., 2016b). The static demand for this network was extended to the semi-dynamic demand for three time periods as displayed in Table 4.4.



Link #	Length [km]	Free speed [km/h]	Capacity [veh/h]
1	5	100	2000
2	10	100	2000
3	4	120	4000
4	0.5	120	2000
5	5.5	120	1000
6	1	120	1500

Figure 4.10: theoretical network (links 4, 5 and 6 are active bottlenecks in SUE conditions)

OD-pair	t_1	t_2	t_3
$O-D_1$	1750	2000	1750
$O-D_2$	2750	3000	2750

Table 4.4: Semi-dynamic Od demands for the network with two congested OD-pairs sharing a bottleneck

This network is hard to equilibrate for two reasons. Firstly, route costs on this network are strongly inseparable due to a shared active bottleneck at the downstream node of link 3 and due to spillback from the second bottleneck (at the downstream node of link 4) towards the origin. Secondly, the network is very sensitive, because travel demand, link capacities and free flow speeds are chosen in a way that in SUE conditions, the (vertical) queues on the network are very small (Table 4.5). This means that bottlenecks may (de-)activate from only small

changes in route choice probabilities that occur during equilibration causing sudden changes in route costs.

	Queue on link 3			Queue on link4		
	t_1	t_2	t_3	t_1	t_2	t_3
Static	10.39	15.61	10.39	0.00	0.14	0.00
Semi-dynamic	10.39	15.74	10.60	0.00	0.18	0.00
Difference	0.00	0.13	0.21	0.00	0.04	0.00

Table 4.5: sizes of vertical queues [pcu] in static and semi-dynamic TA models in SUE conditions

All three TA models were run for 100 iterations using the settings described in subsection 0 and the method of successive averages to equally average route fractions over iterations. Furthermore, in the dynamic TA model, route choice moments at the start of each time period are applied, to align it with the route choice moments in the static and semi-dynamic TA models. Figure 4.11 displays the adapted relative duality gap (equation (4.2)) per iteration for each of the three tested TA models. This figure indicates that stability conditions are maintained from the static to the semi-dynamic TA model (reaching the required 1E-04 threshold after about 45 to 50 iterations), whereas they are not met by the dynamic TA model (as it does not reach values below 1E-02).
The convergence graphs of the static and semi-dynamic TA models are very similar. This is due to the relatively small vertical queues in SUE conditions (Table 4.5), which means that the amount of residual traffic and hence the difference between the static TA model (which ignores it) and the semi-dynamic TA model (which transfers it) is also very small.

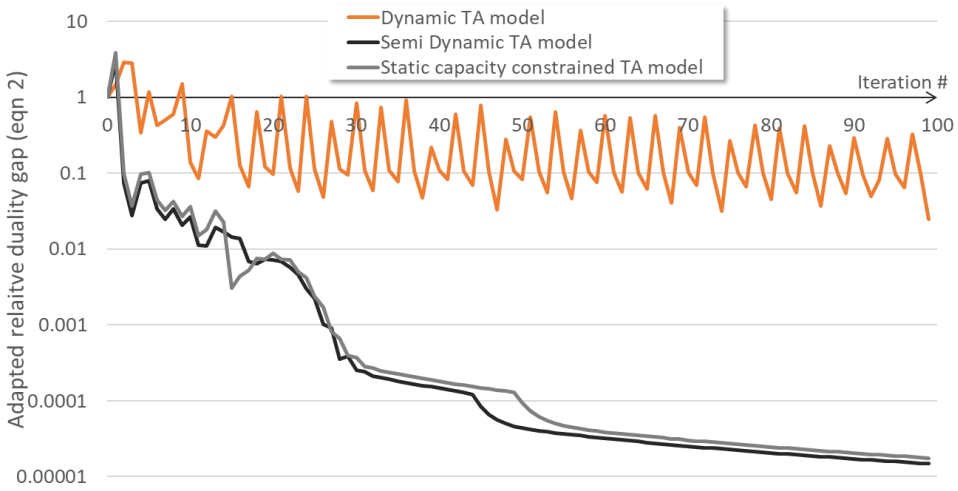


Figure 4.11: adapted relative duality gap per iteration for the three TA models

The oscillations visible in the first 20 to 30 iterations in the static and semi-dynamic TA model are caused by the averaging scheme overshooting and thereby temporary de-activating bottleneck(s) and thus become inconsistent with the network state under SUE conditions. This mechanism was recognized in (Brederode et al., 2016b), which referred to it as the 'unstable phase'. After this phase, the correct network state is maintained smoothening convergence.
Closely related to this, note that although the semi-dynamic TA model assigns slightly more demand, its convergence is slightly better than its static counterpart. This is because the additional demand increases the size of the vertical queues (last row in Table 4.5) and therefore causes the network state in SUE conditions ('the stable phase') to be attained in less iterations. This specific network shows that a more congested network does not always correspond to a less stable network, although in most cases (and hence in in real scale transport networks) this will be the case, as illustrated by the applications in (Brederode et al., 2019). In general, it is

expected that on real scale transport networks, convergence of the semi-dynamic TA model is expected to be slightly worse compared to the static TA model.

4.4.3 Evaluating scalability and the effect of the empty network assumption on a real scale network

To give insights into the scalability of the semi-dynamic TA model, but also in the order of magnitude of the effect of relaxing the empty network assumption (subsection 4.2.1) on outcomes in a realistic transport model context, the static and semi-dynamic TA models were applied on the large-scale strategic transport model of the province of Noord-Brabant (abbreviated in Dutch to 'BBMB'). The network and prior OD demand for road traffic of the base year (2015, version S107) of the BBMB was used. This network contains 1425 centroids, 145.269 links and 103.045 nodes.

Trip-purpose specific OD matrices (including freight) from the BBMB describing demand on an average workday were split up into 24 time periods of 1 hour using purpose specific split factors and then summed up over trip purposes. Within these matrices, 1.590.247 OD pairs with nonzero demand in one or more time periods existed. During assignment 4.019.425 unique routes were generated and used, yielding 2.52 routes per OD pair on average.

4.4.3.1 Scalability

The static and semi-dynamic TA models were both run for each of the 24 time periods until equilibrium using a threshold value of 1E-04 on the duality gap (equation (4.2)). As there are no queues in the semi-dynamic TA model on the BBMB network before 6:00 and after 20:00; the empty network assumption holds for both TA models during these periods. This means that results and calculation times for these periods are exactly the same. Therefore, these time periods are left from the analysis below.

Figure 4.12 displays calculation times per time period for both TA models. The numbers above the bars for assignment indicate the number of iterations required to reach user equilibrium conditions in the considered time period. Both TA models were run on a machine with AMD Ryzen 9 3900X CPU (12 cores) @3.79 Ghz and 128GB of RAM. Calculation times per time period in the static TA model vary between 1:08h and 1:30h requiring 8 up to 11 iterations, whereas the semi-dynamic TA model requires up to 1:44h (+16%) and 13 iterations (+18%) for the assignments themselves. On top of that, the transfer of residual traffic within the semi-dynamic TA model requires an additional 5 minutes up to 1:30h (so up to +50% compared to the static TA model).

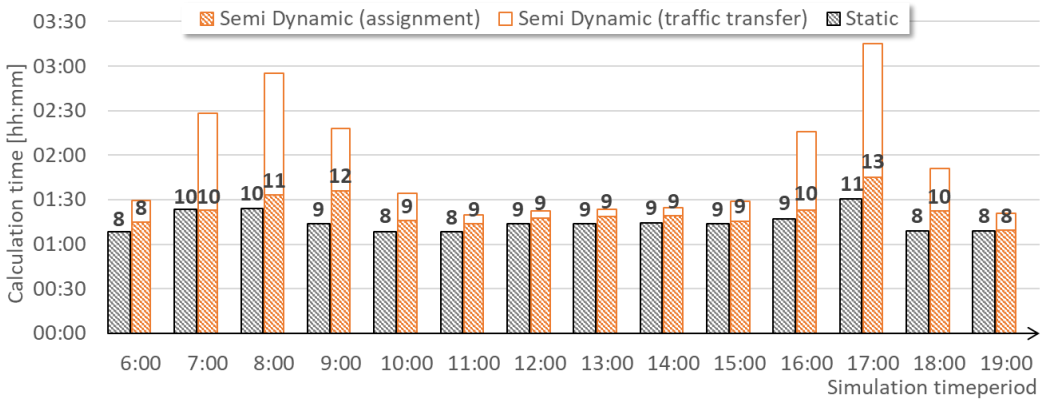


Figure 4.12: calculation times (bars) and #iterations (numbers) per time period for static and semi-dynamic TA models on BBMB

Table 4.7 compares (time) averaged and total calculation times. This shows that in time periods with queues, the semi-dynamic TA model requires on average 51% more calculation time,

predominantly due to calculation time spent by the traffic transfer module. It should be noted however that the assignment model implementation is optimized C++ code, whereas the residual traffic transfer module is a prototypical implementation in Ruby using file-based data exchange with the assignment model. Given its low computational complexity, it is expected that the additional calculation time for the residual traffic transfer module could easily be reduced to less than 10% when its implementation would be merged with the assignment model code into a single code base.

	Static	Semi-dynamic			Difference
	Assignment	Assignment	Traffic transfer	Total	
Average 6h-20h	01:14	01:21	00:31	01:53	51%
Total 6h-20h	17:27	19:06	07:19	26:26	51%
Total 24h	28:33	30:13	07:29	37:43	32%

Table 4.6: comparison of minimum, maximum, average and summed calculation times in [hh:mm]

4.4.3.2 Effect of the empty network assumption on collective loss and link flows

The black bars in Figure 4.13 and Figure 4.13 compare the network wide collective losses per hour from the network operator's perspective between the static and semi-dynamic TA models. This comparison shows that the relaxation of the empty network assumption by the semi-dynamic TA model yields more demand and hence more collective loss in time periods starting with residual traffic from a previous time period. On the BBMB, this yields up to 76% more collective losses during the peak periods and also extension of especially the AM peak period. Considering the entire 24h period, the semi-dynamic TA model yields 54% more collective loss. These substantial differences indicate that using a static TA model (i.e.: assuming an empty network at the start of each assignment) severely under-estimates delays on congested networks. It is therefore very likely that the empty network assumption in static TA models influences (policy) decisions based upon queue size and delay related model outcomes on congested networks.

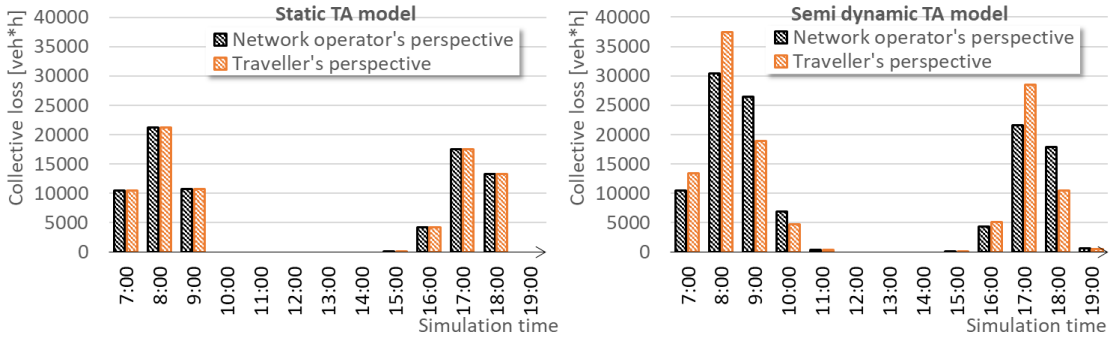


Figure 4.13: network wide collective loss per hour from network operator's and traveler's perspective from static (left) and semi-dynamic (right) TA model application on Noord-Brabant model.

The orange bars in Figure 4.13 display the network wide collective losses from traveler's perspective. Because the static TA model (left graph) omits residual traffic transfer, the cumulative flow corresponding to the first and last vehicles departed in a time period are equal to the cumulative in- and outflow levels of that period (e.g.: in Figure 4.6, $\overline{f_i}(k) = U_i(t_k)$ and $\underline{f_i}(k) = V_i(t_{k-1})$), which causes the collective loss from the traveler's perspective (equations (4.14)-(4.16)) to be equal to that from the network operator's perspective (equation (4.5)).
Comparing collective losses from traveler's perspective with those from network operator's perspective for the semi-dynamic TA model (right graph) demonstrates the difference between looking at when collective losses occur on the network versus looking at when travelers

experiencing losses have departed. The surpluses of orange bars represent losses occurring later on the network, whereas surpluses of black bars represent losses occurring due to demand departed in earlier time periods.

	Static TA model		Semi-dynamic TA model		Difference	
	Period	collective loss	Period	collective loss	absolute	relative
AM peak	07:00 - 10:00	42554	07:00 - 11:00	74774	32220	76%
PM peak	16:00 - 19:00	35156	16:00 - 19:00	44560	9404	27%
24h period	00:00 - 24:00	77710	00:00 - 24:00	119334	41624	54%

Table 4.7: collective losses from network operator's perspective and peak period durations from static and semi-dynamic TA model applications.

4.5 Discussion

In this section, properties and limitations specific to the semi-dynamic TA model are discussed. These in particular pertain to (potential) accuracy improvements with respect to handling and prioritization of residual traffic (subsections 4.5.1 and 4.5.3), the definition of time period durations (subsection 4.5.4) and the route delay calculation method (subsection 4.5.2).

4.5.1 Omitting transfer of traffic based on total travel time

The accuracy of the approach in this paper could be further improved by not only transferring traffic held up in queues, but to also transfer traffic for which the calculated travel time is longer than duration of the considered time period (following the approach from the models in the first two rows in Table 4.1). The authors deliberately choose to leave this topic for future research because 1) we see no means to integrate it into the TA model and hence it would require a post processing step; and 2) it would introduce an optimization problem around the equilibrium from the capacity constrained TA model. Both consequences would severely reduce tractability and stability whilst increasing computational requirements. Furthermore, in congested networks, including the effects of queuing is considered more important than including the effects of traffic not reaching its destination within a certain time period because a) route choices are not affected, as these are still based on complete travel time, even if it exceeds the period duration; and b) the proportion of trips with a longer duration than a typical time period duration in semi-dynamic TA models is relatively small.

4.5.2 Separable vs inseparable route delays

As mentioned in subsection 4.3.4.3, the flow propagation assumptions within the static capacity constrained TA model yields route queuing delays that are non-separable over links (Bliemer et al., 2014). To remove non-separability, during route delay calculation (van der Gun et al., 2020) (implicitly) assume instantaneous flow propagation for both congested and uncongested links. By doing so, the realism of individual route delays may be increased, but this does introduce an inconsistency between the route delay calculation (assuming instantaneous propagation, even when in a queue) and the TA model (assuming zero propagation when in a queue).

Because such an inconsistency reduces stability and convergence properties of the model, we recommend to use the route delay calculation from subsection 4.3.4.3 (yielding inseparable route delays) for any model application in which user equilibrium conditions are important. For model applications in which user equilibrium conditions are not important, the approach by (van der Gun et al., 2020) may be used to increase realism when comparing delays on individual OD pairs.

4.5.3 Influencing priority of transferred traffic

In this paper, for the sake of scalability and simplicity, residual traffic is transferred to a centroid connected to the upstream node of the link with the vertical queue as origin (subsection 4.3.2), and, during the updating of cumulative flow curves, this is compensated for by subtraction of $Q_i(t_{k-1})$ in equation (4.4). This means that transferred traffic from a previous time period has equal priority to demand from the current time period, and only higher priority to current demand that has encountered an upstream queue.

This subsection discusses an alternative method, in which the model user can steer the priority of transferred to current demand in situations where a queue remains in the current time period. In this method, residual traffic from a vertical queue in period k on node n_{ij} between links i and j is transferred into the OD matrix for period $k+1$ using an origin centroid that is directly connected to n_{ij}. By doing so, it is implicitly assumed that the queue discharge rate of the transferred demand instantly changes from $\alpha_{ij}(k)u_i(k)(t_{k+1}\text{-}t_k)$ into:

$$\frac{\alpha_{i'j}(k+1)\sum_{s\in S_{ij}} Q_{ij,s}(t_k)}{(t_{k+1}\text{-}t_k)}, \tag{4.21}$$

where i' represents the connector from centroid storing the residual traffic to node n_{ij} and S_{ij} is the set of destinations of the routes using turn ij in period k.

Now, the modeler can use the capacity of connector i' to influence to what extent the transferred traffic experiences delay traversing node n_{ij} in the next time period. When the capacity of the connector is set to a high value, $\alpha_{i'j}(k+1)$ approaches 1, hence the transferred traffic does not experience a queue while traversing node n_{ij}, whereas when it is set to a very low value, the same mechanism causes the transferred traffic to experience a very large queue while traversing node n_{ij}.

4.5.4 Transitions from queue discharge to departure rates

The method from subsection 4.5.3 allows to influence the priority of residual traffic only when the queue on n_{ij} persists in the next time period. But when the amount of (departed plus residual) traffic arriving at n_{ij} in the next time period is smaller than the normative capacity constraint on that node, $\alpha_{i'j}(k+1)$ equals 1, hence there is no queue and the transferred traffic does not experience a delay no matter what capacity is set for connector i'. Instead, in this case, the queue discharge rate instantly changes to a stationary 'departure' rate of

$$\frac{Q_{ij,s}(t_k)}{t_{k+1} - t_k}. \tag{4.22}$$

This demonstrates that the assumption of stationary travel demand within each time period in semi-dynamic TA model can result in residual traffic that 'departs' with a stationary departure rate depending on the duration of the next time period from the queue instead of with the queue discharge rate from the previous time period. This effectively means that in the semi-dynamic TA model, the FiFo condition still holds within each time period, but may be violated across time periods.

This inconsistency is part of the very definition of the semi-dynamic TA model, and therefore cannot be removed. Instead, effects can only be reduced by shortening time periods. The optimal duration of the next time period should be a function of the discharge time of each queue according to the flow rates of the current time period. As the optimal duration is time period specific, such approach will result in varying period durations within the semi-dynamic

TA model. Note that only the discharge rate is considered relevant, because for a shrinking queue, a too long duration causes it to prematurely dissolve, whereas for a growing queue it only causes an instantaneous change in queue density.

4.6 Conclusions and recommendations

4.6.1 Findings and conclusions

This paper presents a straightforward extension from the static capacity constrained TA model from (Bliemer et al., 2014; Brederode et al., 2019) into a semi-dynamic version. This effectively removes the empty network assumption, yielding a TA model that is more accurate than its static counterpart whilst, unlike its dynamic counterpart, it maintains the favorable stability and scalability properties required for application on large scale strategic transport model systems. To the best of the authors knowledge, the presented approach is the only semi-dynamic TA model that places vertical queues at the correct location (on the upstream node of the link affected by capacity constraint(s)) and also removes flow downstream from bottlenecks as part of the assignment model.

The solution algorithm consists of the static TA model from (Bliemer et al., 2014; Brederode et al., 2019) set in a loop with a residual traffic transfer module. Collective losses and average delays on network, route and link level from the network operator's perspective (quantifying delay within a time period) and the traveler's perspective (quantifying delay within a *departure* time period) are determined from cumulative in- and outflow curves as a post processing module.

Outcomes from the semi-dynamic TA model were compared to its most closely related static (Bliemer et al., 2014; Brederode et al., 2019) and dynamic (Bliemer and Raadsen, 2019) counterparts.

With respect to model accuracy, the comparison showed that the size and temporal distribution of queues and collective losses from the semi-dynamic and dynamic TA models are very similar, but that the spatial distribution is different as the former model ignores spillback. Furthermore, it shows that the static TA model does not resemble the other two models on size, temporal nor spatial distribution of queues and collective losses.

Application of the static and semi-dynamic TA models on the large-scale strategic transport model of Noord-Brabant showed that the empty network assumption in the static TA model causes omission of 27% (PM peak) up to 76% (AM peak) of collective losses in busy periods and 54% when considering the 24h period. It is therefore very likely that the empty network assumption in static TA models influences (policy) decisions based upon queue size and delay related model outcomes on congested networks.

With respect to model stability, the comparison showed that stability is maintained from the static to the semi-dynamic TA model (reaching the required 1E-04 duality gap threshold), whereas it is broken for the dynamic TA model (it does not reach the required duality gap threshold). This means that only the static and semi-dynamic TA models are suitable for strategic applications.

With respect to model scalability, the semi-dynamic TA model in its current (prototypical) form requires on average 51% more calculation time in time periods with queues, predominantly due to calculation time spent by the traffic transfer module. However, it is expected that the additional calculation time for the residual traffic transfer module could easily be reduced to less than 10% when its implementation would be merged with the assignment model code into a single code base. Authors argue that the additional calculation time is a worthwhile inconvenience to bear, given the substantial amount of collective loss being omitted by the static TA model due to its empty network assumption.

4.6.2 Recommendations

In this subsection, recommendations for further research are described, in order of priority from the authors' point of view.

With the shift from a static to semi-dynamic TA model, the question arises how the study period time dimensions (the list of start- and end time of each time period) should be defined. This could be done using an analytical approach that minimizes e.g. the number of transitions from queue discharge to departure rate due to time period boundaries (subsection 4.5.4), or using a more empirical approach in which sensitivity analysis is conducted on the effect of different time dimensions on the model outcomes compared to observed data and/or outcomes of a dynamic TA model.

To determine the representativeness and robustness of the findings with respect to the effect of the empty network assumption (subsection 4.4.3.2), sensitivity analysis on the time dimensions (see previous paragraph) should be combined with a sensitivity analysis on the level of travel demand (what would happen if the temporal distribution of demand of the BBMB model would widen or narrow?) and ideally such analysis should be repeated on other realistic strategic transport models (especially with different levels of urbanization and geographical density distributions).

As a follow up to the suggestion in subsections 4.3.2 and 4.5.3 to directly connect the residual traffic centroids to the queue location to be able to steer the priority of residual traffic, authors recommend to develop a method to automatically determine the values of the capacity of the connector links, such that the violation of FiFo across time periods is minimized.

Because the effects are expected to be negligible, but more importantly, to maintain good tractability, stability and computational properties, the proposed semi-dynamic TA model omits transfer of traffic for which the calculated travel time is longer than duration of the considered time period (subsection 4.5.1). To approximate the effect of this omission on model accuracy, the primary effect per time period could be determined by comparing the amount of traffic that would have been transferred based on total travel time (by post processing the outputs of the model application from 4.4.3) to the amount of traffic transferred by the proposed semi-dynamic TA model. Because the effects of the omission are larger for smaller time periods, this analysis could also lead to recommendations with respect to a lower bound value for time period durations in the semi-dynamic TA model.

One of the motivations to shift to a semi-dynamic TA model, is that it is expected that especially departure time choice modelling and travel demand matrix estimation would greatly benefit from it (subsection 4.1.1). With respect to the prior, authors recommend to do research in how to embed it in a travel demand model containing a departure time choice model. With respect to the latter, the most straightforward way to continue would be to extend the matrix estimation method described in (Brederode et al., 2023) to an estimation period covering the whole day, allowing to include observed flows, delays and congestion patterns on any temporal aggregation level. In line with the discussion from subsection 4.5.2, for both application types, a point of attention would be the effects of employing inseparable route delays in the TA model (equation (4.19)) and separable route delays (equation (4.20)) in the departure time choice model or matrix estimation method.

The application in subsection 4.4.3 suggests that the empty network assumption in static TA models most likely has an effect on (policy) decisions based upon queue size and delay related model outcomes on congested networks. Authors recommend to more broadly investigate such effects by looking at the societal value of the shift from the static to the semi-dynamic TA model. This could be done by redoing a cost-benefit analysis for some (set of) policy measure(s) such that differences in model outputs are translated into differences in benefits and hence difference in policy decisions. Another case with respect to the societal value of the proposed method would be to re-evaluate a set of measures with respect to environmental and

noise standards. Again, differences could also be translated into their (expected) effects on policy decisions.

Finally, authors recommend to further formalize the boundary between semi-dynamic and dynamic TA models. With respect to travel demand the boundary is somewhat defined in (Bliemer et al., 2017) who state "In a semi-dynamic model, multiple time periods are considered [..] e.g. 1 hour time slices" and "[dynamic TA models] usually consider many smaller time periods (e.g. time slices of 15 minutes)". With respect to the definition of residual traffic (Bliemer et al., 2017) loosely define it as "[a semi-dynamic TA model] takes the result from a previous period (such as vehicles in a queue) into account, for example by passing on residual traffic to the next period', whereas the literature review from subsection 4.2.2 suggests that not only vehicles in queue, but also travelling vehicles can be considered residual traffic. The authors current point of view on this is that whenever a TA model transfers not only demand and/or link flows but also link states (i.e. speeds and densities) it becomes dynamic, whereas a clear boundary with respect to travel demand has yet to be defined.

A. APPENDIX: TWO DERIVATIONS FOR DETERMINATION OF QUEUE SIZE AT LINK LEVEL

The amount of traffic held up on the queue of a link i can be directly derived from the cumulative flow curves of the link by:

$$Q_i(t_k) = \big(1 - \alpha_i(k)\big)u_i(k)(t_k - t_{k-1}) \tag{4.23}$$

or equivalently by summing residual traffic transferred from this link from equation (4.3):

$$Q_i(t_k) = \sum_{j \in J_i} \sum_{s \in S_{ij}} Q_{ij,s}(k) \tag{4.24}$$

where J_i is the set of outlinks from link i and S_{ij} is the set of destinations for which in time period k residual traffic was transferred from a vertical queue on turn ij.

Chapter 5

Travel demand matrix estimation methods integrating the full richness of observed traffic flow data from congested networks

Abstract
Road travel demand matrix estimation fuses prior or synthetic travel demand matrices with observed flow data. Due to technological advances, ever more observed link flows, speeds and densities are available, whereas rising congestion levels trigger the urgency to use robust and sound estimation procedures on them.
This paper addresses difficulties when estimating travel demand using link flows observed on congested networks. Active bottlenecks on these networks influence flow values both upstream (queues) and downstream (flow reduction). This implies that, depends on the specific traffic conditions in the network, observed link flow values may represent either 1) the unconstrained travel demand for that link, 2) a proportion of the capacity of a set of upstream links, 3) the capacity of the normative downstream link; or 4) a combination of these quantities.
If the used traffic assignment model does not strictly adhere to link capacity constraints, flow reduction (2) is not accounted for and all traffic is considered unaffected (1), thereby forcing incorrect assumptions upon the estimation. Current practice is to derive unconstrained link demand values from flows affected by congestion (2, 3 or 4) and then, instead of the actual observed flows, use these link demand values during matrix estimation. As such, these methods exhibit poor tractability and robustness and do not integrate any information from the assignment model about the composition of routes on the observed links.
This paper describes and compares three novel demand matrix estimation methods for large scale strategic congested transport models that use assignment models that strictly adhere to link capacity constraints and explicitly consider the conditions under which link flows are observed. It compares these methods to the current practice and gives practical insights from applications, demonstrating that these methods are more tractable and robust and allow for usage of observed congestion patterns and travel times from (big) data sources. Furthermore, these methods reveal data inconsistencies, allowing the modeler to correct the model network and other matrix estimation input.

Keywords: demand matrix estimation, congested networks, strict capacity constraints, big data, floating car data, ANPR data, Bluetooth data

This chapter is a version of the following publication:
Brederode, L., Verlinden, K., 2019. Travel demand matrix estimation methods integrating the full richness of observed traffic flow data from congested networks. Transportation Research Procedia, Modeling and Assessing Future Mobility Scenarios. Selected Proceedings of the 46th European Transport Conference 2018, ETC 2018 42, 19–31.
https://doi.org/10.1016/j.trpro.2019.12.003

CRediT author statement:
Luuk Brederode: Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Kurt Verlinden**: Writing – review & editing.

5.1 Introduction

In strategic transport models, road travel demand matrices are usually estimated using estimation methods that fuse prior or synthetic travel demand matrices with flow data observed on individual roads ('links') in the network. On the one hand, ever more data on flows, speeds and/or densities on link level is available, driven by technological advances (e.g. PnD's, smartphones, IoT), trends in transport policy towards smarter usage instead of expansion of the network and the smart mobility concepts arising from them. On the other hand, the urgency of robust and sound estimation procedures is triggered by rising congestion levels on these networks that are at an all-time high. In this paper we address the known difficulties when estimating travel demand using link flows observed on a network with high levels of congestion.

5.1.1 Interpretation of observed flows under different network conditions

Congested networks incorporate at least several active bottlenecks, which influence flow values both upstream (queues will form) and downstream (flow is metered). This implies that, on such a network, observed link flow values may represent either 1) the unaffected travel demand for that link, 2) a proportion of the capacity of (a set of) upstream link(s), 3) the capacity of the normative (in terms of capacity deficit) downstream link or 4) a combination of these quantities.

These four conditions are illustrated in a unidirectional corridor network with two active bottlenecks in Figure 5.1, where:

- The **unaffected links** (continuous black arrows) are unconstrained by active bottlenecks. This means that on links 1 and 2, link outflow equals the demand from centroid 1, whereas link outflow form link 3, equals the demand from centroids 1 and 2.
- The outflow on **flow metered links** (continuous grey arrows) is determined by active bottlenecks upstream. This means that the outflow on bottleneck link b1 equals the capacity of this link, whereas the outflow on bottleneck link b2 and link 11 equals the capacity of link b2. The outflow on flow metered links 6 and 7 equals the capacity of bottleneck b1 multiplied by the turn proportion from link b1 to link 6, whereas the outflow on link 6a equals the capacity of b1 multiplied by the turn proportion towards link 6a.
- The outflow on **links in queue** (short dashed black arrows) is determined by the normative downstream link. This means that on links 4 and 5, outflow equals the capacity of b1, whereas outflow of link 10 equals the capacity of b2. Note that for illustrative purposes, in this example bottleneck b1 affects two upstream links, whereas bottleneck b1 only affects one upstream link. In reality, the influence of downstream bottlenecks depends on the severity of the bottleneck in relation to the flow towards it and the buffer capacity on the links upstream from the bottleneck.
- The outflow on **partially metered links** 8 and 9 (long dashed black arrows) is a combination of unconstrained link demand from centroid 3 and the capacity of active bottleneck (metered link) b1 reduced by the turn proportion towards link 6. Note that combinations of unaffected links (1) and links in queue (3) and combinations of flow metered links (2) and links in queue (3) are not included in this example. In practice, these situations can occur, but only when the considered link has at least one outlink that is not affected by the bottleneck causing the queuing to occur. This requires that exit lanes allowing traffic towards the unconstrained outlinks to freely traverse the queue must exist on the link. These conditions are outside of scope of this paper, as in macroscopic traffic assignment models these conditions cannot occur due to the first-in-first-out (FIFO)

assumption in these models, which is required to maintain the route choice of travelers on the network.

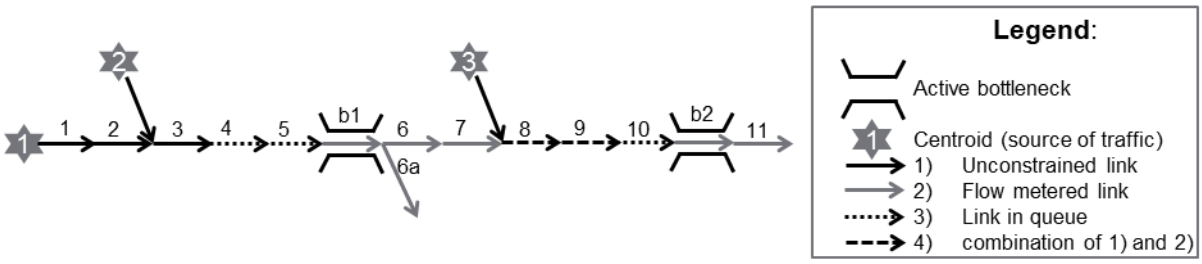


Figure 5.1: example network illustrating different quantities that link (out)flow may represent

As illustrated in the example from Figure 5.1, the specific traffic conditions in the network define which quantity each observed link flow represents. Note that only flows measured under conditions (1) or (4) contain information about the absolute level of traffic demand, whereas flows measured under conditions (2) or (3) only contain information about network capacities and bottleneck locations (hence a lower bound on the level of traffic demand).

5.1.2 Problem formulation and current practice

Demand matrix estimation methods use a traffic assignment model to assess the relationship between travel demand and link flow in intercept information. In current (strategic transport modelling) practice, intercept information is provided by traffic assignment models that cannot distinguish between the different conditions, because they do not strictly adhere to link capacity constraints. Therefore, flow metering (2, 4) nor queuing effects (3) of bottlenecks are taken into account and all traffic is implicitly considered to be unaffected (1), thereby forcing incorrect assumptions upon the estimation. Therefore, matrix estimation methods using these models should only be applied on observed flows values that are unaffected (1), rendering them mostly useless on networks with high congestion levels. Note that by nature these assignment models should not be applied on study areas with congestion altogether.

5.1.3 Contributions

This paper describes existing demand matrix estimation methods for large scale strategic congested transport models that use assignment models that strictly adhere to link capacity constraints, allowing them to explicitly consider the conditions under which link flows are observed. It compares these methods to the current practice and gives practical insights from applications of methods that are already implemented and applied, thereby demonstrating that these methods allow for usage of (big) data sources such as floating car data and congestion patterns (used in methods 2 and 3) and (route) travel time observations from e.g. Bluetooth or ANPR data (intended to be used in method 3).

5.2 Methodologies

All methodologies are described in the bi-level optimization framework summarized in equation (5.1), where in the upper level, the origin-destination (OD) demand matrix is altered to minimize differences between observed and modelled link flows and between the prior and modelled OD demand matrix, while in the lower level a traffic assignment model is used solving a user equilibrium problem translating the new OD demand into modelled link flows.

$$\begin{aligned} \mathbf{D}^* = \underset{\mathbf{D}}{\text{Argmin}}(F) = \underset{\mathbf{D}}{\text{argmin}}[(\mathbf{D} - \mathbf{D_0})^2 + (\mathbf{y}(\mathbf{D}), \tilde{\mathbf{y}})^2] \\ \text{s.t.: } \mathbf{y}(\mathbf{D}) = assign(\mathbf{D}), \\ \mathbf{D} > 0, \end{aligned} \tag{5.1}$$

where F denotes the upper level objective function to be minimized, $\mathbf{D}^*$, $\mathbf{D}$ and $\mathbf{D}^0$ denote vectors containing posterior, current and prior (or observed) OD demand respectively for all OD pairs, $\mathbf{y}(\mathbf{D})$ and $\tilde{\mathbf{y}}$ denote vectors of estimated and observed link flows. Furthermore, we define $L = \{L_1, L_2, L_3, L_4\}$ as the set of observed links split up into the four different traffic condition types, to be used in the remainder of this paper.

5.2.1 Assignment model classes

In the lower level, the function $assign$ represents the assignment model used. The method from current practice (subsection 5.2.2) requires a static capacity restrained traffic assignment (SCRTA) model, whereas the other methods (subsections 5.2.3 through 5.2.5) require a static capacity constrained assignment (SCCTA) model. The essential difference between these model classes is that the SCCTA strictly respects link capacity constraints (link flow can never be larger than link capacity), whereas in the SCRTA model, only the route choice is influenced by capacity constraints (and link flow can be larger than link capacity). We refer to (Bliemer et al., 2017) for concise definitions of these assignment model classes.

5.2.2 Solution method used in current practice

Current practice to use observed flows affected by congestion (conditions 2, 3 or 4) is to estimate unconstrained link demand values from the observed flow values, for example using the 'Tonenmethodiek' (Transpute, 2003) used in the Dutch LMS/NRM models, or similar techniques that shift observed flows to upstream unconstrained links. Then, instead of the actual observed flows, the post-processed link demand values are used during OD demand matrix estimation. These models do not make use of any information from the assignment model about the network conditions on the observed links. Therefore, even flow metered observations (which by definition only contain information on network capacity) are erroneously used in the demand estimation instead of network supply calibration. For these reasons, these methods exhibit poor tractability and robustness.
The objective function of this method is defined as

$$F = w_1 \sum_{od \in OD} (D_{od}^0 - D_{od})^2 + w_2 \sum_{l \in L} \left(y_l^{SCRTA}(\mathbf{D}) - f(\tilde{y}_l) \right)^2, \tag{5.2}$$

where f denotes a function (like the 'Tonenmethodiek') that estimates corresponding unconstrained link demand values from observed link flows, $y_l^{SCRTA}(\mathbf{D})$ represents the link flows as calculated by a SCRTA assignment model and w_1 and w_2 represent parameters that express the relative importance of the prior demand component in relation to the link flow component in the objective function.
Note that although an SCRTA model must be used to provide the link flows in the upper level, the final assignment of the estimated OD demand matrix can be done using a SCCTA model to increase accuracy. This is effectively being done by the assignment model QBLOK in the LMS/NRM model system, which uses capacity constraints model for route choice and the final assignment results but omits capacity constraints to determine the link flows used in the upper level.

5.2.3 Solution method 1: Using SCCTA instead of SCRTA model

This method uses the SCCTA model to isolate unmetered from total demand and apply the upper level only on the unmetered demand. To this end, metered demand is subtracted from both the observed and modelled flows, yielding the following objective function:

$$F = w_1 \sum_{od \in OD} (D_{od}^0 - D_{od})^2 + w_2 \sum_{l \in \{L_1, L_4\}} (\alpha_l(\mathbf{D}) y_l^{SCCTA}(\mathbf{D}) - \alpha_l(\mathbf{D}) \tilde{y}_l)^2 \tag{5.3}$$

where α_l denotes the proportion of flow that arrives at link l unaffected by any upstream bottleneck(s). Proportion factors α_l are derived from od specific proportion factors α_l^{od} outputted by the SCCTA model using:

$$\alpha_l = \frac{\sum_{od \in OD} \delta_l^{od} \alpha_l^{od} D^{od}}{\sum_{od \in OD} \delta_l^{od} D^{od}} \tag{5.4}$$

where δ_l^{od} is the link-od incidence indicator which equals one if link l is used by od pair od, and zero otherwise.

Note that this method (correctly) only estimates demand using observed flow on links in $\{L_1, L_4\}$, but does not use the information on bottleneck locations that can be derived from observed flows on links in $\{L_2, L_3\}$. This method was initially applied in the 2018 version of the transport models of Noord-Brabant, but due to the omission of information on L_3 links, queue lengths where structurally underestimated and not all bottleneck locations where modelled. This led to adoption of method 2 (described in the next subsection) in these transport models.

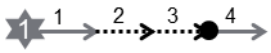
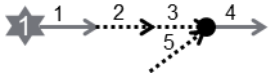
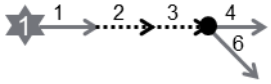
5.2.4 Solution method 2: Adding information on bottleneck locations

This method is an extension of the method described in 5.2.3 and adds usage of information from speeds observed on links in $\{L_2, L_3\}$ to determine bottleneck locations. In the applications presented, observed speeds from floating car data where used. To extract bottleneck locations from these speeds, first all links for which the observed speed v_{fcd} is lower than its critical speed v_{crit} are identified as being in queue (i.e.: part of set L_3). The required critical speeds can be derived from the speed limit that applies on the considered link. Once set L_3 has been defined, bottleneck locations can be identified as the node between the last (most downstream) link from a (spatial) sequence of L_3 links and the first (most upstream) link in a (spatial) sequence of other links.

The queue on the (spatial) sequence of L_3 links upstream from the bottleneck location can be translated into an excess demand (D_l^{exc}), which, added to the capacity (C_l) of the first link downstream from the bottleneck location, is treated as a (indirect) observation of demand just upstream from the bottleneck link. Note that, by definition, this first link must belong to L_2.

The method requires excess demand D_l^{exc} to be calculated for all L_3 links in the network. To do so, the observed speeds and the fundamental diagram of each link can be translated into the density that, according to the fundamental diagram, would apply on that link. The densities and lengths of al links in the considered sequence of L_3 links together with the capacity of the bottleneck link can then be translated into the excess demand D_l^{exc}. Alternatively, the set of links in queue (L_3) may be derived from annual summaries of daily traffic reports (e.g. the File Top 50 in the Netherlands (VID, 2017)). These annual summaries provide observed bottleneck locations along with observed queue lengths and durations. Using a bottleneck model these queue lengths can be translated into excess demand, either assuming some value for the density

in queue, or using the density values derived from the observed speeds and fundamental diagrams as described above.
Note that both sources for location and excess demand estimation may be combined, allowing the modeler to choose the most accurate source available to be used. For example, annual summaries of traffic reports may provide more accurate bottleneck locations, but they are typically not available for lower order roads, whereas accuracy of the densities derived from observed speeds may provide better estimates for excess demand than the bottleneck model would. This might lead the modeler to choose to use observed bottleneck locations on the higher order roads and derived bottleneck locations on the lower order roads.
Figure 5.2 illustrates the procedure for bottleneck location detection and excess demand estimation on a corridor, merge and diverge network. The formulae for the corridor and merge cases indicate that the observed link speeds provide enough information to estimate the (indirect) observed demand for the bottleneck link. However, the formula for the diverge network indicates that more information is required to determine the normative outgoing (bottleneck) link. This information cannot be derived from observations on link level and would require observations on node and turn level and modelling of traffic flow on lane and turning movement level. Such observations are not (widely) available yet, but more importantly, such a level of traffic flow modelling is beyond the scope of the macroscopic traffic assignment models used in strategic transport models.

Network		**(Indirect) observed demand**
Corridor		$\tilde{y}_3 = C_4 + D_2^{exc} + D_3^{exc}$
Merge		$\tilde{y}_3 + \tilde{y}_5 = C_4 + D_2^{exc} + D_3^{exc} + D_5^{exc}$
Diverge		$\tilde{y}_3 = C_4 + D_2^{exc} + D_3^{exc}$ or $\tilde{y}_3 = C_6 + D_2^{exc} + D_3^{exc}$

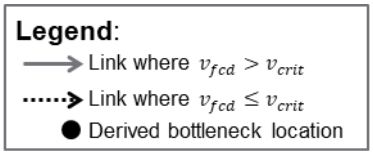


Figure 5.2: derivation of bottleneck locations and (indirect) observed demand from observed link speeds

Method 2 implies the following objective function:

$$F = w_1 \sum_{od \in OD} (D_{od}^0 - D_{od})^2 + w_2 \sum_{l \in \{L_1, L_4\}} (\alpha_l(\mathbf{D}) y_l^{SCCTA}(\mathbf{D}) - \alpha_l(\mathbf{D}) \tilde{y}_l)^2 + w_3 \sum_{b \in B} \left(C_{\underline{b}} + D_{\underline{b}}^{exc} - \sum_{\overline{b}} y_{\overline{b}}^{SCCTA} \right)^2 \quad (5.5)$$

where B denotes the set of bottleneck nodes, $C_{\underline{b}}$ and $D_{\underline{b}}^{exc}$ denotes the capacity and excess demand on the normative outlink (the link that has caused activation of the bottleneck location) of bottleneck node b respectively, $y_{\overline{b}}^{SCCTA}$ denotes the modelled flow on inlink $\overline{b}$ of bottleneck node b towards bottleneck link $\underline{b}$ and w_3 denotes the relative importance of the objective function component that covers the demand for bottleneck links.
This method was applied on the NRM-West: the Dutch regional strategic transport model of the Randstad Agglomeration (including the 4 largest cities in the Netherlands) as described in (Brederode et al., 2017) and is recently implemented in the 2018 version of the strategic transport models of the province of Noord Brabant.

5.2.5 Solution method 3: Adding sensitivities of proportion factors and travel times

Equations (5.3) and (5.5) indicate that the unmetered proportion factors α_l depend on the current OD demand matrix **D**. This means that any changes made to **D** in the upper level have an immediate effect on the value of the unmetered proportion factors, whereas these are considered constant in objective functions (5.3) and (5.5). For this reason[14], this method approximates the sensitivity of the proportion factors to changes in demand ($\partial\alpha_l/\partial\mathbf{D}$) using marginal simulation of the node model component within the assignment model and adds these sensitivities to the objective function assuming a first order Taylor approximation.

Equation (5.5) indicates that the bottleneck component in the objective function is competing with the prior demand and link flow components. This means that adding the bottleneck component reduces the chance that bottlenecks switch from active to inactive state during matrix estimation. Bottlenecks that switch from active to inactive or vice versa disturb the matrix estimation process is undesirable, because 1) it causes changes to the definition of sets L_1, L_2, L_3 and L_4, thereby non-convergence; and 2) because the (added) sensitivities of the proportion factors are point approximations which are only valid when the considered link remains in the state in which the sensitivity was estimated. For these reasons, this method removes the bottleneck component from the objective function and instead, adds it as a constraint to the optimization problem[15].

Lastly, this method adds, when available, travel times to the objective function, as these can also be expressed as a function of α_l. Observed travel times can be derived from e.g. floating car data on link level or from ANPR or Bluetooth measurements on route level. These changes and additions yield the following optimization problem:

$$\begin{aligned}\mathbf{D}^* = \underset{\mathbf{D}}{\operatorname{argmin}} \Bigg[& w_1 \sum_{od \in OD} (D_{od}^0 - D_{od})^2 \\ & + w_2 \sum_{l \in \{L_1, L_4\}} \left(\left[\alpha_l(\mathbf{D}) + \frac{\partial \alpha_l}{\partial \mathbf{D}} (\mathbf{D} - \mathbf{D^0}) \right] y_l^{SCCTA}(\mathbf{D}) - \alpha_l(\mathbf{D}) \tilde{y}_l \right)^2 \\ & + w_4 \sum_{p \in \tilde{P}} \left(\tau_p(\mathbf{D}) - \tilde{\tau}_p \right)^2 \Bigg] \end{aligned} \tag{5.6}$$

$$\text{s.t.:} \quad \mathbf{y}(\mathbf{D}) = assign(\mathbf{D}),$$

$$\mathbf{D} > 0,$$

$$\sum_{\underline{b}} y_{\underline{b}}^{SCCTA} = C_{\underline{b}} + D_{\underline{b}}^{exc} \quad \forall\, b \in B,$$

where $\tilde{P}$, $\tilde{\tau}_p$ and τ_p denote the set of paths with observed travel times, the observed travel time on path p and the modelled travel time on path p respectively. Weighing parameter w_4 expresses the relative importance of the travel time component of the objective function.

This method is a continuation of the method described in (Brederode et al., 2014) and is implemented in prototype form. The method has proven to outperform methods 1 and 2, both

[14] A more thorough argumentation for adding these sensitivities to the objective function derivative is given in subsection 6.2.1.2.

[15] By including the bottleneck components as constraints to the optimization problem, the solver can be used to nudge any initial $\mathbf{D}^0$ to a demand vector that satisfies all constraints before evaluating the Taylor approximation. This is further described in subsection 6.3.2.6.

in accuracy as well as speed of convergence on small test networks. The prototype is still under development as its runtimes make it currently not practically applicable on large networks.

5.3 Software

Section 5.2 describes four methodologies in terms of problem formulations and solution methods, which give insight in the theoretical added value of the different methods. However, the extent to which this theoretical value is translated into practical value is determined by its software implementation. Therefore, this chapter briefly describes the different software implementations used by the authors gaining insights in practical applications, before these insights are described in section 5.4. Table 5.1 summarizes the applications and software examined.

Method	Transport model used	Software lower level	Software upper level
Current practice	LMS/NRM models	QBLOK	AVVMAT
Method 1	models of Noord Brabant	STAQ	OtMatrixEstimation
Method 2	models of Noord Brabant	STAQ	OtMatrixEstimation
Method 2	NRM West (Randstad model)	STAQ	AVVMAT
Method 3	Various transport models	STAQ	MATLAB

Table 5.1: applications examined in this paper

Furthermore, in this chapter, requirements for alternative software implementations (not used by the authors) are given to allow readers to adopt methods from section 5.2 in their own preferred software.

5.3.1 Assignment model used in current practice

Authors have gained experience of current practice using the assignment model used in the LMS/NRM methodology of the dutch national and regional strategic transport models: QBLOK (Bakker et al., 1994). QBLOK is a deterministic equilibrium model that extends traditional SCRTA models on the following three points:

1) It not only models actual flow that uses the network within the study period, but also the flow that would have wanted to travel in the study period but did not reach its destination in time due to congestion.

2) It takes the network effects of congestion (flow metering and spillback) into account using a heuristic, but these effects are only included in link travel time calculation (and thus route choice), not in the modelled traffic flows.

3) It uses a fixed number of iterations and prefixed weights to approximate the user equilibrium, as convergence to equilibrium is infeasible within acceptable computation times.

5.3.2 Assignment model used in methods 1 through 3

Authors have gained experience of methods 1 through 3 using the SCCTA model STAQ described extensively in (Brederode et al., 2018). STAQ is implemented as a propagation model within the StreamLine framework in OmniTRANS transport planning software.

The model supports any concave, two regime fundamental diagram, but insights in this paper where gained using the quadratic linear diagram (QL) from (Bliemer et al., 2014). To describe interaction of flows on nodes STAQ uses the explicit node model from (Tampère et al., 2011), which allows to explicitly calculate and output OD specific proportion factors α_l^{od} used in methods 1 through 3. This node model is also used in method 3 for the marginal simulation that determines the sensitivity of the proportion factors to changes in demand ($\partial \alpha_l / \partial \mathbf{D}$). In

most studies, the node model was extended with the junction modelling component of OmniTRANS transport planning software to account for the effect of limited supply due to conflict points on the junction itself (i.e. crossing flows), and to calculate travel-time delays due to geometry of the node and conflicts on turning-movement level.
The assignment model can be used with different route choice models, but insights in this paper where gained using the multinomial logit (MNL) model with scale parameters set to one over 14% of the minimal route cost of the considered OD pair. This means that the route choice model is only sensitive to the ratio of different route costs, not their absolute values. In all three methods, the stochastic user equilibrium (SUE) is used as underlying route choice paradigm. The adapted relative duality gap derived in (Bliemer et al., 2013) that accounts for perception errors and thus reaches zero upon convergence when using the MNL route choice model is used as convergence criterion with a threshold value of 5E-04. The method of self-regulating averages described in (Liu et al., 2009) is used to average route demands over iterations providing fast convergence.
The assignment model makes use of pre-generated route sets. Insights in this paper where gained using the routeset generator from the StreamLine framework, which uses the Dijkstra algorithm to find the shortest path between each OD pair and then uses a repeated random sampling process on free flow link travel times using a gamma distribution known as the accelerated Monte Carlo method (Fiorenzo-Catalano, 2007) to generate additional alternative routes. Route filters are applied after the repeated random sampling process to reduce route overlap, remove irrelevant routes and restrict the size of the set of potential routes.
For methods 1 and 2, any SCCTA model that can output OD-specific proportion factors α_l^{od} can be used as an alternative for STAQ. For method 3, the assignment model must be suitable to be used to efficiently approximate the sensitivity of the link-based proportion factors to changes in demand ($\partial \alpha_l / \partial \mathbf{D}$)

5.3.3 Upper level solvers used in current practice and methods 1 and 2

In current practice and in the NRM West application of method 2, the AVVMAT software is used to solve the upper level. AVVMAT is based on the Combined Calibration matrix calibration program develop by Hague Consulting Group in the 1990's. AVVMAT assumes a multiplicative model in which each matrix cell is a function of its initial value and a set of parameters (count, trip ends, trip length class, etc.). Furthermore AVVMAT assumes that the parameters are statistical of nature and therefor have a level of reliability. AVVMAT assumes a Poison distribution and applies the BFGS algorithm. Derivation and implementation of the AVVMAT OD matrix estimator is described in (Lindveld, 2006) in more detail.
In the application of methods 1 and 2 on the models of Noord-Brabant OtMatrixEstimation is used to solve the upper level. OtMatrixEstimation is part of OmniTRANS transport planning software and uses a heuristic to iteratively scale relevant matrix cells in the prior matrix to better match with observed flows on link or screen line level. It threats trip ends and trip length distribution from the prior as constraints. A more extensive description of the heuristic can be found in (Smits, 2010).
For methods 1 and 2, any solver that can handle the convex quadratic objective function and the non-negativity constraint may be used. However, because OD demand matrix estimation problems in strategic transport models are usually very large and very sparse, the solver should be able to exploit the sparsity of the problem to be able to solve the problem within constraints of available computer memory and required computation time. For the same reason, the solver should not rely on finite differences to approximate the gradient. Instead, the solver should use an analytically calculated gradient.

5.3.4 Upper level solver used in method 3

The upper level and interaction with the lower level of method 3 is implemented in prototype form in Matlab and uses the fmincon interior point algorithm within the optimization toolbox of Matlab. The interior point algorithm uses a conjugate gradient descent method. To prevent memory issues, it is set to use the limited memory version of the BFGS algorithm (Nocedal, 1980) to approximate the Hessian. Furthermore, functions to calculate the gradient of the objective function and the gradient of the bottleneck constraints are included in the implementation and passed to fmincon, to prevent it from doing finite difference analysis on every OD pair (which would take too much time).

Like method 2, method 3 can be solved using any solver that is suitable for large sparse quadratic optimization problems. However, it must also be able to include the linear (bottleneck) constraints. Furthermore, analytical calculation of the gradient for both the objective function and constraints is possible (Rijksen, 2018), but due to the inclusion of the sensitivity of the bottleneck proportion factors the calculations are more complex than for methods 1 and 2.

5.4 Practical Insights from applications

This section discusses insights from the practical applications listed in Table 5.1.

5.4.1 Insights in conditions on observed links

Preliminary analysis on the input data of the Noord-Brabant and NRM West models reveal using floating car data to identify the links in queue (L_3) and STAQ to determine the distribution over unaffected (L_1), metered (L_2) and partially metered (L_4) links. Results of this analysis reveal that the majority of observed link flows are unaffected or partially metered (i.e. they belong to $\{L_1, L_4\}$) and could thus be used for demand estimation using method 1.

To illustrate this, we describe results from the preliminary analysis for the AM peak period of the base year of the NRM-West, which describes the most congested region of the Netherlands. For this model, there were no flow metered observed link flows ($L_2 = \emptyset$) and only 6% of the count locations where observed in queue. Covering the other 94% of the count locations, the black line in Figure 5.3 shows the portion of flow unaffected by upstream bottlenecks per count location according to the assignment results of the prior OD demand matrices. In the graph, count locations are ordered by their portion of unaffected flow increasingly. This reveals that about 66% of the count locations where partly metered (the percentile where the black line hits 100%) and about 34% was unaffected (the remainder of the locations). These findings suggest that although most observed link flows are influenced by congestion, there are only few observed links that are not suitable to be able to apply method 1. Since the NRM-West describes the most congested region in the Netherlands, other Dutch models are expected to exhibit even lower proportions of link flows unsuitable for use with method 1.

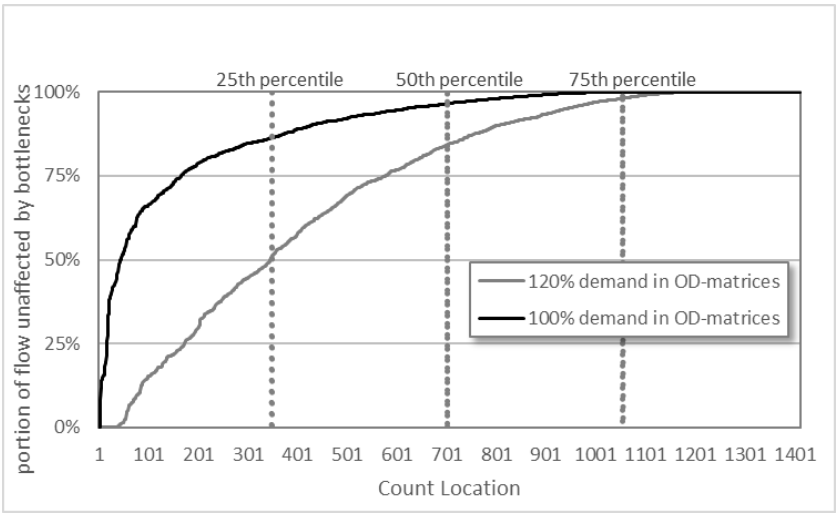


Figure 5.3: portion of flow unaffected by bottlenecks per count location in NRM-West model, AM peak

To determine robustness of these findings, sensitivity analysis was carried out in which the prior OD demand matrix was increased by 20%. The result is displayed as the gray line in Figure 5.3. In this case, around 3% of the count locations that where not in queue became flow metered, whereas the share of partially metered count locations increased to about 76%, leaving 21% unaffected. Although the 20% increase of demand yields slightly more links to become unsuitable for application of method 1, it is still only a small minority. Since methods 2 and 3 follow the same underlying principle but add (indirect) estimation using observations on links in queue (L_3) these insights about applicability holds to an even greater extent for those methods. For these methods no more than 3% of observed link flows is metered and could therefore not be used in demand matrix estimation methods 2 and 3.

5.4.2 Insights from method used in current practice

The method used in current practice (as described in subsection 5.2.2) circumvents computational issues that arise from inclusion of strict capacity constraints in OD demand matrix estimation methods for strategic transport models by projecting the (estimated) effect of capacity constraints on the input of the methodology (the observed flows), rather than adapting the methodology itself to include the constraints. This approach allows for the usage of proven technology: SCRTA models and (upper level) solution methods widely available since the 1990's; see e.g. (Abrahamsson, 1998) for an overview. Because of the (desirable) mathematical properties of the SCRTA model and its corresponding (upper level) problem, solutions are found relatively easily and fast.

However, using (estimated) link demands instead of observed link flows as primary input gives rise to the following myriad of problems all related to the fact that link demand is a quantity that cannot be measured. Firstly, this means that the accuracy of methods that estimate link demands (e.g. Tonenmethodiek) cannot be determined directly. Instead, only the accuracy of the solution method as a whole can be evaluated by comparing the result of a capacity constrained assignment of the estimated OD demand matrix with the observed flows. Differences between these modelled and observed link flows can be caused by either errors in the method used to estimate link demands, the assignment model or the solution method. Formulated differently: although the methodology minimizes differences between observed and modelled link demand, it does not necessarily minimize differences between observed and modelled link flows. This means that calibration of the parameters of the matrix estimation method and the assignment model, as well as finding and fixing input errors needs to take place in a single process. In practice, this leads to extensive estimation procedures that aim to provide

acceptable outcomes using (structured) trial and error. This causes high and uncertain lead times for projects including OD demand matrix estimation with only reasonable outcomes.
Secondly, the process described in the previous paragraph is highly sensitive to changes in input. This means that a process that has produced acceptable outcomes for a set of observed link flows representing a certain study area or base year might not give acceptable outcomes for set of observed link flows representing another study area or base year. This reasoning also holds for different sets of parameters for the assignment model and/or upper level solution method. In practice, this requires estimation procedures to be changed when the input data or parameter set of the considered project gives rise to it. This causes expensive matrix estimation projects with poor tractability and comparability of model outputs.

5.4.3 Insights from application using method 1

By replacing the SCRTA model with a SCCTA model and considering only the unmetered demand in the upper level, the problems related to the usage of link demands described in section 5.4.2 are effectively removed. Method 1 allows to directly compare modelled flows with observed link flows and to isolate effects of changes in parameters of the upper level solution method and SCCTA model. Furthermore, there is no need to change the estimation procedure when input or parameter sets change.
Although the share of observed link flows in queue is only small (section 5.4.1), these links are most important for a transport model to describe accurately. However, as mentioned in section 5.2.3, method 1 does not use information on links in queue, hence it neglects observed queues. Instead, the hypothesis behind method 1 is that demand estimation on the other (majority) of the count locations will cause the correct demand on the queued links as well. This hypothesis proved wrong, as it turns out that fitting flows on unconstrained or partially constrained links only does not (substantially) improve the fit of link demand for links in queue.
The reason for this is explained using the example in Figure 5.4. Assume that in this network observed flows are available for links 2 and 3. Method 1 would not use any information from link 3 (as this link is in queue) and thus would only try to minimize differences between modelled and assigned flow on link 2. Assume that modelled flow on link 2 is underestimated. Method 1 would then evenly increase demand on all OD pairs using link 2, neglecting the effect that demand on OD pairs towards link 6 would have on the queue on link 3, whereas a different (more uneven) distribution over OD pairs could effectively improve the fit on link 3 with the same improvement of fit on link 2.

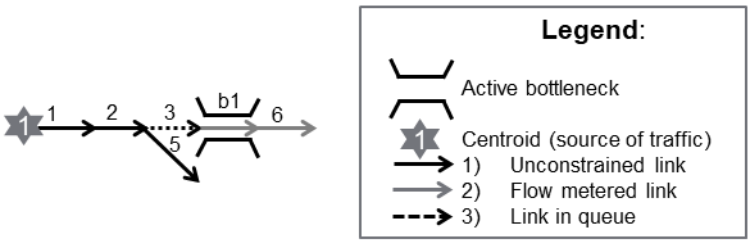


Figure 5.4: example where estimating demand matrices using method 1 and a single count on link 2 does not imply a positive effect on the fit on link 3

The example shows that the OD demand matrix estimation problem has too many degrees of freedom (too many OD pairs to choose from) to expect a method to improve the overall fit on links that the method does not explicitly consider. For method 1, this means that the level of congestion in the final assignment results will mainly be determined by the level of congestion

in the assignment result of the prior demand matrix[16], as was seen in its application in the transport models of Noord-Brabant.
Another potential issue of method 1 is that it contains no mechanism that prevents bottlenecks to switch from active to inactive or vice versa during the matrix estimation process. As described in subsection 5.2.4, this causes poor convergence. In the Noord-Brabant application, this issue did not clearly manifest itself, probably because the prior demand matrices generally underestimated the congestion levels causing a limited set of active bottlenecks and hence a small chance of state switches during estimation. However, the more theoretical tests described in (Brederode et al., 2014; Frederix, 2012) clearly demonstrate this issue.

5.4.4 Insights from applications using method 2

In addition to the findings described in (Brederode et al., 2017) the application of method 2 on the strategic transport model of the Randstad Agglomeration proved that the addition of (indirect) observation of demand just upstream from the bottleneck link allows for accurate representation of observed queues while maintaining the fit on unconstrained and (partly) flow metered links, thereby solving the problems described in subsection 5.4.3.
The application also demonstrated that by changing the ratio between weights w_2 and w_3, inconsistencies between model link capacities and observed congestion patterns and inconsistencies between count values can be isolated, allowing the modeler to correct the model network and other matrix estimation input. Often errors with respect to the exact bottleneck location, its normative outlink or the combination of observed flows and observed link speeds from different data sources proved to be the cause of these inconsistencies. The methodology proved an asset in removing these errors and inconsistencies.
However, the application also showed that weighing parameter w_3 needs to be set carefully. It should be high enough to ensure and maintain activation of the correct bottlenecks throughout the estimation procedure, but low enough to allow accurate representation of unaffected and partly metered link flows near bottleneck locations.

5.4.5 Insights from applications using method 3

As mentioned in subsection 5.2.5, method 3 outperforms methods 1 and 2, both in terms of accuracy as well as convergence properties. On top of that, removes the issue of choosing w_3 by replacing this component of the objective function with an equivalent constraint. Furthermore, it supports observed travel times as an additional input data type.
However, the prototype is still under development as its runtimes make it currently not applicable on large networks. This is mainly caused by large sparse matrix multiplications that are required to translate the sensitivities of the proportion factors from the marginal node model runs to their effect on the objective function. Until this implementation issue is fixed, the method is best applied excluding the sensitivities of the proportion factors but including the constraints that ensure and maintain correct bottleneck states.

5.5 Conclusion and recommendations

Active bottlenecks in congested networks influence observed link flow values both up- as well as downstream. In strategic transport models, this means that an observed link flow value is either unaffected, metered, in queue or partially metered due to active bottlenecks. Flow observed on unaffected or partially metered links contains information about travel demand

[16] Assuming that no wide-spread unidirectional changes to the demand matrix are being made by the estimation method

that can be directly used for OD demand matrix estimation, whereas observed flow on links in queue is only useful when supplemented by observed link speeds or queue lengths. Flows observed on metered links only contain information on network supply and can therefore not be used for travel demand estimation.

Data and sensitivity analysis on the transport model describing the most congested region of the Netherlands indicates that it is highly unlikely that more than 3% of observed link flows of any Dutch strategic transport model is flow metered, meaning that more than 97% contains information on OD demand. This data can be used, provided that observed link speeds (from e.g. floating car data) or observed queue locations (from e.g. daily traffic reports) are available and that a method that supports partial metered and links in queue is used.

The most common OD demand matrix estimation method used for strategic transport models can only handle observed flows that are unaffected by active bottlenecks (which is the case for 34% or less of the observations), and therefore needs to translate observed link flows into estimated link demands to account for bottleneck effects. This approach allows for the usage of conventional SCRTA models and relative quick solution of the matrix estimation problem. However, the use of input that is estimated rather than measured, makes the method non-transparent and input sensitive resulting in poor tractability, comparability and transferability of estimation processes. This has led to high and uncertain project lead times with outcomes of only reasonable accuracy.

Therefore, this paper assessed three methods that take bottleneck effects into account by replacing the SCRTA model with a SCCTA model. Method 1 can handle observed flows on partially metered links in addition to unaffected links and allows for direct use of and comparison with observed link flows. Methods 2 and 3 additionally provide support for observed queue lengths on links in queue, thereby integrating the full richness of traffic flow data on congested networks.

For network diverges, methods 2 and 3 require information on the normative outgoing link which demands for observations on node and turn level and modelling of traffic flow on lane and turning movement level. Such observations are not (widely) available yet, but more importantly, such a level of traffic flow modelling is beyond the scope of the macroscopic traffic assignment models used in strategic transport models as it would violate the first-in-first-out assumption. Although (Wright et al., 2017) describe a node model that would allow for such violations, development in this direction will (further) degrade on mathematical properties that are desirable in the strategic context: existence and uniqueness of the SUE solution of the assignment model.

Compared to method 2, method 3 provides greater accuracy and faster convergence, removes the need to set a sensitive w_3 parameter and it supports observed travel times as an additional input data type. However, its implementation is still in prototype form limiting its applicability on large networks.

The upper level problem of all three methods can be solved using widely available software packages for large sparse quadratic programming problems with linear constraints. Calculation of the gradient efficiently and correctly is a point of attention when implementing these methods.

Currently, the authors are working on extension of methods 2 and 3 to support estimation of OD demand matrices that cover multiple period(s), which should eventually lead to a method that supports 24 hour estimation. This requires the SCCTA model to be extended to be able to transfer residual traffic (traffic that has not reached its destination within a previous time period) to the next time period, and an upper level extension that can simultaneously estimate matrices for all considered time periods. Both extensions are viable from a methodological point of view, but especially implementation of the latter is expected to create new challenges.

Chapter 6

Travel demand matrix estimation for strategic road traffic assignment models with strict capacity constraints and residual queues

Abstract
This paper presents an efficient solution method for the matrix estimation problem using a static capacity constrained traffic assignment (SCCTA) model with residual queues. The solution method allows for inclusion of route queuing delays and congestion patterns besides the traditional link flows and prior demand matrix whilst the tractability of the SCCTA model avoids the need for tedious tuning of application specific algorithmic parameters.
The proposed solution method solves a series of simplified optimization problems, thereby avoiding costly additional assignment model runs. Link state constraints are used to prevent usage of approximations outside their valid range as well as to include observed congestion patterns. The proposed solution method is designed to be fast, scalable, robust, tractable and reliable because conditions under which a solution to the simplified optimization problem exist are known and because the problem is convex and has a smooth objective function.
Four test case applications on the small Sioux Falls model are presented, each consisting of 100 runs with varied input for robustness. The applications demonstrate the added value of inclusion of observed congestion patterns and route queuing delays within the solution method. In addition, application on the large scale BBMB model demonstrates that the proposed solution method is indeed scalable to large scale applications and clearly outperforms the method mostly used in current practice.

Keywords: demand matrix estimation, static traffic assignment model, capacity constrained, congestion patterns, route travel times, prior OD demand matrix, large scale, strategic, mathematical properties

This chapter is a version of the following publication:
Brederode, L., Pel, A.J., Wismans, L., Rijksen, B., Hoogendoorn, S.P., 2023. Travel demand matrix estimation for strategic road traffic assignment models with strict capacity constraints and residual queues. Transportation Research Part B: Methodological 167, 1–31. https://doi.org/10.1016/j.trb.2022.11.006

CRediT author statement:
Luuk Brederode: Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Adam Pel**: Conceptualization, Supervision, Writing – review & editing. **Luc Wismans**: Supervision, Writing – review & editing. **Bernike Rijksen**: Investigation. **Serge Hoogendoorn**: Supervision, Project administration.

6.1 Introduction

Traditionally, travel demand origin-destination (OD) matrix estimation for road traffic is a process in which a prior demand matrix specifying travel demand between origin and destination nodes in the road network is enriched using observed flows on link level. It is a bi-level optimization problem where in the upper level, the OD demand matrix is altered to minimize differences between observed and modelled link flows and between the prior and modelled OD demand matrix, while in the lower level a traffic assignment model is used solving a user equilibrium problem translating the new OD demand into modelled link flows.
Different traffic assignment models can be used in the lower level, varying by capability and complexity. In this paper, we shall use the classification of assignment models described in (Bliemer et al., 2017) when referring to specific assignment model types. The majority of strategic transport model systems used today use static capacity restrained traffic assignment (SCRTA) models. SCRTA models assume separable monotonously increasing link travel time functions, yielding computationally fast and scalable models with desirable convergence properties needed for strategic large-scale transport model systems. Matrix estimation methods using SCRTA models have been studied extensively and are readily available, see e.g. (Cascetta, 2009) and references herein.
However, link flows and speeds from SCRTA models do not correspond to empirically supported macroscopic traffic flow theory that describes the relation between flow, speed and density (the fundamental diagram). This is caused by the lack of a strict capacity constraint and omission of a congested branch and storage constraints in travel time functions used in SCRTA models.
The most common solution for this problem is to is to switch to a (macroscopic) dynamic capacity and storage constrained traffic assignment (DCSTA) model. This model class does incorporate capacity constraints and hence physical effects of congestion and is more realistic compared to the SCRTA model class. However, the dynamic nature of DCSTA class models causes the mapping from OD demand to link flows to be extended with a temporal dimension, causing temporal correlations between model variables which makes estimation of OD demand matrices much more tedious and are therefore limited to small sized networks (see e.g.: Toledo et al., 2015).

6.1.1 Contribution, positioning and outline

This paper presents an efficient solution method for the matrix estimation problem using a static capacity constrained traffic assignment (SCCTA) model with residual queues. This type of assignment model is described in e.g. (Bliemer et al., 2014; Brederode et al., 2019; Bundschuh et al., 2006; Lam and Zhang, 2000; Smith, 2013), implemented in OmniTRANS Transport planning Software, PTV VISUM and Aimsun Next and applied in various contexts (e.g., Brederode et al., 2016a; Huang et al., 2020; Tajtehranifard, 2017; Tsanakas et al., 2020).
The solution method is developed to be used in the context of strategic transport models where the user wants to refine prior OD demand from a demand model with information on link and route level (e.g.: loop detector and floating car data).
Note that this paper only considers SCCTA models that incorporate capacity constraints within the link cost functions, thereby allowing for explicit residual queues in model outcomes, whereas SCCTA models in which the capacity constraints are only added as upper bounds on link flows (often referred to as ‘extended Beckmann’ or ‘capacitated Beckmann’; e.g. (Correa et al., 2004; Larsson and Patriksson, 1999; Nie et al., 2004; Yang and Yagar, 1994)) do not allow for residual queues and are therefore outside of scope of this paper. For reasons of brevity, in the remainder, we shall simply refer to ‘the SCCTA model’ (i.e.: omitting ‘with residual queues’) when referring to the TA model type considered in this paper.

By using a SCCTA model, the solution method combines the favorable properties of SCRTA and DCSTA models in the context of matrix estimation. Similar to DCSTA models, the strict capacity constraints of SCCTA models account for flow metering effects of active bottlenecks, which allows direct comparison and usage of observed flows that are reduced by upstream bottlenecks. The strict capacity constraints also extend the supported set of datatypes for estimation with observed (link- or route-) travel times and observed congestion patterns because queues are explicitly modelled. Similar to SCRTA models, the static nature of SCCTA models removes the temporal dimension in the relation between link flows and OD-demands, which allows demand estimation at a time-aggregate level. This avoids temporal correlations between model variables which causes the solution method to be relatively fast and suitable for large scale networks. [17]

Note that there is a big difference in what is considered a large network in the DSCTA compared to the SCCTA context. In the DSCTA context, networks containing in the order of tens of thousands OD pairs are considered large scale (e.g: Castiglione et al., 2021; Osorio, 2019a), whereas SCCTA models are typically applied on networks containing in the order of millions OD pairs (Brederode et al., 2019). The proposed solution method presented In this paper targets networks that are considered large in the SCCTA context. To the best of the authors' knowledge, the sheer size of these networks prevent usage of any DSCTA based method. Therefore, this paper takes the SCCTA model as a starting point for the matrix estimation method and does not include a comparison between DSCTA and SCCTA models. Instead, the interested reader is referred to (Brederode et al., 2019).

Further note that, similarly to the methodology proposed in this paper, the quasi dynamic approaches by (Marzano et al., 2018; Van der Zijpp, 1996) also employ time-aggregation on observed variables, but they do so to reduce -not avoid- temporal correlations between model variables. Furthermore, the meta model approach in (Osorio, 2019b), the computational graph approach in (Ma et al., 2020; Wu et al., 2018) and various SPSA-based approaches (e.g: Qurashi et al., 2020) show promising results in handling and reducing temporal correlations that exist in DCSTA models.[18] The approach proposed in this paper is different because it is only applicable to SCCTA (and SCRTA) models, which means that it solves a less complex problem which should make it more efficient compared to the more generic approaches developed for DCSTA models.

The remainder of this paper is organized as follows. Section 6.2 defines the matrix estimation problem for SCCTA models, the SCCTA model itself, solution methods to similar problems currently used in practice and the proposed solution method. Section 6.3 elaborates on the proposed solution method that uses a combination of analytical and approximated relationships in the lower level as well as the mathematical properties of its solution(s). In section 6.4 the added value of the proposed solution method is demonstrated using several test case applications on the small Sioux Falls model, whereas section 6.5 presents application of the proposed solution method on a large scale strategic transport model, demonstrating its performance and scalability. We end with discussion and conclusions in section 6.6.

[17] Note that in reality (and therefore also in time-aggregated observed variables), temporal correlations do occur, and -just like with any other static traffic assignment model- this knowledge must be taken into account when assessing the SCCTA models outcomes.

[18] Note that (Osorio, 2019) reports that her meta model approach should be transferable to any traffic assignment model, thereby making it applicable to the strategic context, but to the best of the authors knowledge, this has not been tested yet (Wu et al., 2018).

6.2 The Matrix estimation problem for SCCTA models

In this section, the travel demand matrix estimation problem is defined for road traffic using input data consisting of a prior demand matrix, observed link flows and route queuing delays. Note that to the best of our knowledge, the inclusion of route queuing delays in the context of static traffic assignment models is novel, and it is only possible because strict capacity constraints are included. In subsection 6.2.3.3 we shall present another novelty in the context of static traffic assignment models which is to include observed congestion patterns in the optimization problem.

6.2.1 Problem formulation

Consider a general network $G = (N, L)$ where N denotes the set of nodes n and L denotes the set of directed links l. Let $R \subset N$ and $S \subset N$ be the set of origins r and destinations s respectively and $RS = R \times S$ the set of all OD-pairs rs. Furthermore, let $\tilde{L} \subset L$ be the set of links for which flow has been observed ('observed links'). Then, the bi-level matrix estimation problem using a prior OD matrix, observed link flows and observed route queuing delays is defined as:

$$\begin{aligned} &\mathbf{D}^* = \underset{\mathbf{D}}{\operatorname{argmin}}(F) = \underset{\mathbf{D}}{\operatorname{argmin}}[f_1(\mathbf{D}, \mathbf{D}_0) + f_2(\mathbf{y}(\mathbf{D}), \tilde{\mathbf{y}}) + f_3(\boldsymbol{\tau}(\mathbf{D}), \tilde{\boldsymbol{\tau}})] \\ &\text{s.t. } \mathbf{y}(\mathbf{D}), \boldsymbol{\tau}(\mathbf{D}) = assign(\mathbf{D}), \\ &\quad \mathbf{D} \geq \mathbf{0}, \end{aligned} \tag{6.1}$$

where F denotes the upper level objective function to be minimized, $\mathbf{D}^*$, $\mathbf{D}$ and $\mathbf{D}_0$ denote vectors containing posterior, current and prior (or observed) OD demand respectively for all OD pairs in RS, $\mathbf{y}(\mathbf{D})$ and $\tilde{\mathbf{y}}$ denote vectors of current and observed link flows in $\tilde{L}$, $\boldsymbol{\tau}(\mathbf{D})$ and $\tilde{\boldsymbol{\tau}}$ denote vectors of current and observed route queuing delays (for the set $\tilde{P}$ of routes for which travel time has been observed), while f_1, f_2 and f_3 denote distance functions measuring the differences between observed (or prior) and current OD demand, link flows and route queuing delays respectively. In the lower level, the function $assign$ represents the traffic assignment model used (i.e. here the SCCTA model described in subsection 6.2.2).

Note that $\tilde{\mathbf{y}}$, $\mathbf{D}_0$ and $\tilde{\boldsymbol{\tau}}$ contain aggregate variables observed over some period(s) of time. Therefore, the observed values in these vectors are in fact instances of some probability distribution. Although, when known, these distributions can be considered when solving the upper level, this is not subject of this paper. In the remainder we therefore choose the least squared error as distance function for all three components since it does not require any additional data on the distribution of the observed flow values, prior matrix or route queuing delays. Furthermore, we introduce parameters that allows for weighing and normalization of the three components in the objective function. Using least squared errors and weighting parameters w_1,w_2 and w_3, the objective function now reads:

$$F = w_1 \sum(\mathbf{D} - \mathbf{D_0})^2 + w_2 \sum(\mathbf{y}(\mathbf{D}) - \tilde{\mathbf{y}})^2 + w_3 \sum(\boldsymbol{\tau}(\mathbf{D}) - \tilde{\boldsymbol{\tau}})^2 \tag{6.2}$$

6.2.1.1 Decomposition of observed link flows

As described in Chapter 5, active bottlenecks in a network influence flow values both upstream (queues will form) and downstream (flow is metered). This means that an observed link flow value represents either 1) the unaffected travel demand for that link, 2) a proportion of the capacity of (a set of) upstream link(s), 3) the capacity of the normative (in terms of capacity deficit) downstream link or 4) a combination of these quantities.

Only flows measured under conditions (1) or (4) contain information about the absolute level of traffic demand, whereas flows measured under conditions (2) or (3) only contain information about network capacities and bottleneck locations (hence a lower bound on the level of traffic demand).
Because strict capacity constraints are lacking in SCRTA models, matrix estimation methods using these models do not take flow metering (2, 4) nor queuing effects (3) of bottlenecks into account. Instead, all traffic is (implicitly) considered to be unaffected (1), thereby forcing incorrect assumptions upon the estimation (Brederode and Verlinden, 2019). Formulated differently: SCRTA models and their matrix estimation methods assume that travel demand on route level equals route flow by definition, whereas in SCCTA models, route flow is lower than route demand on links downstream from the first bottleneck on the route. This advocates for the use of an SCCTA model and a different matrix estimation method, such that the conditions under which link flows are observed are considered during the estimation.

6.2.1.2 Solution methods: current practice

In SCRTA model context, bi-level problem (1) is typically solved by iteratively assigning the OD matrix from the upper level into the lower level to determine the relationships between link flows and OD-demands (assignment matrix $\mathbf{A}(\mathbf{D})$ of size $\left|\tilde{L}\right|$ x $|RS|$) and the relationships between route queuing delays and OD-demands and then use these relationships to solve the upper level. In the SCRTA model context the relationships are considered constant while solving the upper level yielding the following response function for link flows:

$$\mathbf{y}(\mathbf{D}) = \mathbf{A}(\mathbf{D}_{\mathrm{k}-1})\mathbf{D} \tag{6.3}$$

where $\mathbf{D}_{\mathrm{k}-1}$ represents the OD demand from the previous upper level solution.
For SCRTA models, numerous solution algorithms using constant response functions have been proposed and successfully applied as summarized by e.g. (Abrahamsson, 1998). For SCCTA models, to the best of our knowledge, only the following three solution approaches described in Chapter 5 have been proposed and/or applied. Below, these approaches are summarized, an extensive description of the practical implications for all three methods can be found in (Brederode et al., 2017; Brederode and Verlinden, 2019).
The longest and most widely used approach to estimate demand for SCCTA models is to estimate unconstrained link demand values from observed link flow values and apply a traditional SCRTA-based solution algorithm assuming equation (3) in the upper level. By using (estimated) link demands instead of directly observed link flows as input, this approach constructs a synthetic matrix estimation problem in which all observations adhere to condition 1 from subsection 6.2.1.1, allowing usage of an SCRTA model. However, this approach does not use any information about local network conditions on observed links as the SCRTA model cannot provide it. Instead, this approach relies on the unconstrained link demand values that are derived using heuristics based on generic model-wide temporal demand distributions. This means that using this approach will yield an OD matrix that fits to the estimated link demand values, but it does not guarantee that the final assignment of this OD matrix using the SCCTA model will yield link flows that fit to the observed link flows. Because of this, these methods exhibit poor tractability and robustness.
The second approach is to use the SCCTA model to determine the assignment matrix and then apply matrix estimation only on unmetered demand, while assuming equation (6.3) in the upper level. This is the first solution approach that does not require usage of estimated link demands. This approach was tested on the Dutch regional strategic transport model of the Randstad Agglomeration (Brederode et al., 2017) and is implemented in the 2018 version of the strategic transport models of the Dutch province of Noord Brabant.

The last solution approach described in Chapter 5 refers to an early version of the solution method that is described in this paper. To the best of our knowledge this is the only estimation method for SCCTA models that can include observed queuing delays and/or congestion patterns in the estimation. Note that all but the proposed solution approaches for SCCTA models assume equation (6.3) in the upper level. By doing so, these methods all treat the matrix estimation problem as if it were a Cournot-Nash game by omitting any sensitivities in the response function, whereas (Frederix et al., 2013; Maher et al., 2001; Yang, 1995) point out that it intrinsically is a Stackelberg game. Although incorrect in theory, given the widespread usage with SCRTA models, this appears to not be a problem in practice in this context.
However, strict capacity constraints cause the true response function to be more sensitive and less separable, which means that to solve it, $\mathbf{A}(\mathbf{D})$ should no longer be considered constant while solving the upper level. Instead, the sensitivity of link flows (and the sensitivity of route queuing delays) for changes in OD-demand need to be included in the response functions.
This was recognized in the DCSTA model context for which most approaches in literature use either direct finite differences (e.g: Djukic et al., 2017; Frederix et al., 2013; Shafiei et al., 2017; Toledo and Kolechkina, 2013) or some form of the Simultaneous Perturbation Stochastic Approximation (SPSA) method (e.g: Antoniou et al., 2015; Cantelmo et al., 2017; Cipriani et al., 2013; Tympakianaki et al., 2015) to approximate the sensitivity of link flows to changes in OD-demand $\boldsymbol{\partial \mathbf{y}(\mathbf{D})/\partial \mathbf{D}}$ and use it in a first order Taylor expansion around the current solution yielding the following response function for link flows:

$$\mathbf{y}(\mathbf{D}) = \mathbf{y}(\mathbf{D}_{k-1}) + \left.\frac{\partial \mathbf{y}(\mathbf{D})}{\partial \mathbf{D}}\right|_{\mathbf{D}_{k-1}} (\mathrm{D} - \mathbf{D}_{k-1}). \tag{6.4}$$

In these studies, minimization of the upper level objective function in DCSTA context is done using a solver that can handle the quadratic objective function specified by (6.2) and (6.4) in combination with the linear constraints on link flows (6.4) and the bound constraints enforcing non-negativity in (6.2). Both the direct finite difference and SPSA approaches require additional assignment model runs in the lower level to determine the sensitivity of the link flows and therefore exhibit large calculation times and thus poor scalability. Also, SPSA-based approaches entail tedious tuning of application specific algorithmic parameters (e.g.: Cipriani et al., 2011).

6.2.2 SCCTA model formulation

This section describes the mathematical relationships within an SCCTA model, as these will be used by the solution method that will be proposed in subsection 6.2.3. An SCCTA model consists of two submodels: a network loading submodel and a route choice submodel (Figure 6.1). The network loading submodel uses route demand $\mathbf{Q}$ from the route choice submodel to calculate route travel times which are used by the route choice submodel to calculate route choice probabilities $\boldsymbol{\psi}$ to distribute OD demand over routes.

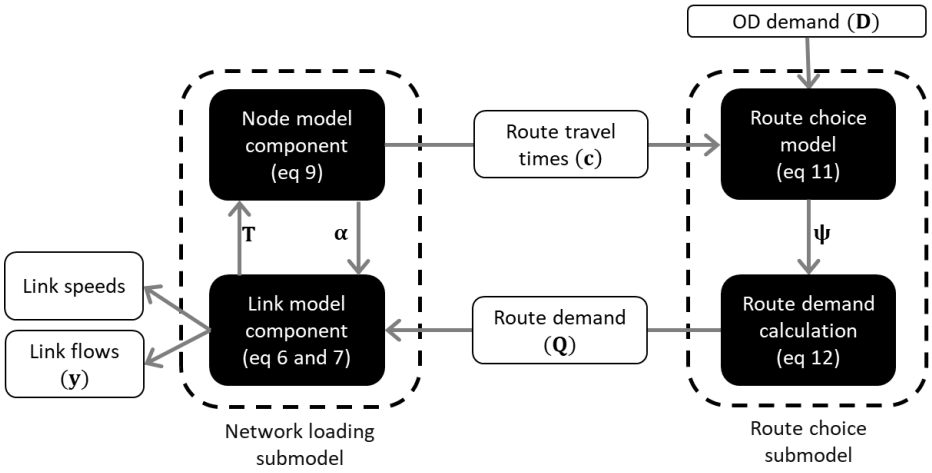


Figure 6.1: framework for SCCTA models

The network loading submodel uses this route demand to compute the resulting link flows and speeds and thereby (route) travel times. The most important components within the network loading submodel are the node model component that calculates flow acceptance factors **α** on links entering nodes with active capacity constraints, and the link model component that applies these factors to the route demands yielding turn demands **T,** which by aggregation yield link flows **y**. Note that, as mentioned in section 1, some SCCTA models do not have a node model, but use link exit capacities instead. Further note that besides link flows (subsection 6.2.3.1) the acceptance factors **α** also define the queuing delays (subsection 6.2.3.2) and the congestion patterns (subsection 6.2.3.3) used in the demand estimation.
The mathematical definition of the route choice submodel depends on the chosen traffic assignment problem formulation. In this paper the stochastic user equilibrium (SUE, Fisk, 1980) is chosen, which leads to the route choice submodel described in subsection 6.2.2.4. Other (non-equilibrium and/or deterministic) assignment problem formulations may also be used with the SCCTA network loading submodel but are not described here, because fixed route choice probabilities are assumed in the upper level (i.e.: route fractions are assumed to be locally constant), and, similar to approaches used for SCRTA models, it is assumed that iterations between lower and upper level will solve the consistency problem between route choice probabilities and OD demands. The remainder of this subsection describes each of the components from the SCCTA model framework in more detail.

6.2.2.1 Link model component

The link model component determines link flows taking into account reductions due to active bottlenecks in the form of flow acceptance factors per link $\alpha_l \; \forall \; l \in L$, calculated by the node model component (subsection 6.2.2.2). These flow acceptance factors are aggregated to the route-link level using:

$$\hat{\alpha}_{lp} = \prod_{ij' \in IJ_{p,il}} \alpha_{ij'} \, . \tag{6.5}$$

where $\hat{\alpha}_{lp}$ denotes the acceptance factor due to upstream bottlenecks at link l on route p and $IJ_{p,il}$ represents the set of turns on route p up to (and including) the turn from link i to link l. Note that the matrix $\hat{\boldsymbol{\alpha}}$ containing acceptance factors for all route-link combinations is actually the route-level equivalent of assignment matrix **A** introduced in subsection 6.2.1.2. Further note that we define $\hat{\alpha}_{lp} = 0$ for all l not used by p, such that it also doubles as a route-link incidence indicator.

Given route demand Q_p from the route choice submodel (6.2.2.4), route specific link inflows are calculated using:

$$y_{lp} = Q_p \hat{\alpha}_{lp} \quad \forall l \in L, \forall p \in P_{rs}, \forall rs \in RS. \tag{6.6}$$

The route specific link inflows are used to determine turn demands T_{ij} from inlink i to outlink j used as input for the node model component by:

$$T_{ij} = \sum_{\boldsymbol{RS}} \sum_{\boldsymbol{p \in P_{rs}}} \sigma_{jp} y_{ip} \quad \forall i \in I_n, \forall j \in J_n, \forall n \in N, \tag{6.7}$$

where $\sigma_{jp} \in \{0,1\}$ indicates if route p uses link j.

6.2.2.2 Node model component

The node model component in SCCTA models determines which nodes in the network form an active bottleneck. Bottlenecks are activated on nodes where the demand for one or more of the 'outlinks' or turning movements ('turns') exceeds the capacity of the respective outlink(s) or turn(s) (Figure 6.2). On nodes that represent an active bottleneck, the node model component also determines how the available supply on its outgoing links and turns is distributed over the competing ingoing links ('inlinks').

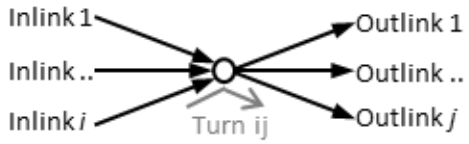


Figure 6.2: inlinks, outlinks and turns associated with a node

Any first order node model can be used, as long as it complies to a set of seven requirements[19] for first order macroscopic node models described in (Tampère et al., 2011). One of these requirements is that the node model should comply with local supply constraints, which is the very reason that SCCTA models obey strict link capacity constraints. Below, a coarse outline of the workings of such node models is sketched; we refer to (Bliemer et al., 2014; Flötteröd and Rohde, 2011; Smits et al., 2015; Tampère et al., 2011) for a more thorough description and solution algorithms for specific node models.

Consider a node n connected to a set of inlinks I_n and a set of outlinks J_n forming the set of turn movements using the node $IJ_n = I_n \mathrm{x} J_n$. Furthermore, define the set of outlinks directly related to inlink i as $J_i = \{j | T_{ij} > 0\}$ and the set of inlinks directly related to outlink j as $I_j = \{i | T_{ij} > 0\}$. For all $n \in N$ a node model $\Gamma_n(\cdot)$ is defined that calculates the vector of turn acceptance factors $\boldsymbol{\alpha}_\mathrm{n}$ reducing turn flows traversing n as a function of the vector of travel demand for each turning movement on the node ($\mathbf{T}_n$), the vector of link capacities of inlinks ($\mathbf{C}_n$) and the vector of supply constraints on the outlinks of the node ($\mathbf{R}_n$) defined by either the capacity, or (in case of spillback) the outflow of the outlink.

[19] These are: 1) general applicability (not just merges and diverges), 2) no holding back of flows, 3) non-negativity, 4) conservation of vehicles, 5) satisfying demand and supply constraints, 6) obeying the conservation of turning fractions, 7) satisfaction of the invariance principle)

This yields:

$$\begin{aligned} \boldsymbol{\alpha}_n &= \Gamma_n(\mathbf{T}_n, \mathbf{C}_n, \mathbf{R}_n) \\ \text{where: } \boldsymbol{\alpha}_n &= \{\alpha_{ij} \; \forall \; ij \in IJ_n\}, \\ \mathbf{T}_n &= \{T_{ij} \; \forall \; ij \in IJ_n\}, \\ \mathbf{C}_n &= \{C_i \; \forall \; i \in I_n\} \text{ and} \\ \mathbf{R}_n &= \{R_j \; \forall j \in J_n\}. \end{aligned} \tag{6.8}$$

Note that one of the requirements from (Tampère et al., 2011) is the first-in-first-out (FiFo) assumption. It means that traffic flows out of an inlink and into different outlinks in the same order they reached the end of the inlink. In the context of an static traffic assignment model without time-varying traffic flows, this assumption causes the flow acceptance factors for all turns on an inlink of a node to be equal by definition, thereby also defining the relation between turn based and link based flow acceptance factors as $\alpha_l = \alpha_{ij} \Leftrightarrow i = l$. Further note that since we are using an SCCTA model (hence: without storage constraints), spillback cannot occur, and $R_j = C_j$, whereas in models with storage constraints, due to spillback, R_j can also be equal to the outflow of link j.

6.2.2.3 Fixed point problem and travel time calculation

As Figure 6.1 and subsections 6.2.2.1 and 6.2.2.2 suggest, turn demands and flow acceptance factors are mutually dependent, and iterations between the node (equation (6.8)) and link model (equations (6.6) and (6.7)) are required to reach a fixed point. This fixed point problem was identified by (Bliemer et al., 2014), who have proven that its solution is unique under very mild conditions, whereas (Raadsen and Bliemer, 2019a) provide a more general and capable solution scheme for the problem.

Once the fixed point is reached, route travel times are calculated using:

$$c_p = \sum_{l \in L_p} \frac{L_l}{\dot{v}_l} + \tau_p, \tag{6.9}$$

where L_l and $\dot{v}_l$ are the length and maximum speed on link l respectively and τ_p represents the route queuing delay. The route queueing delays are a function of all turn based flow acceptance factors on the route as derived in (Bliemer et al., 2014):

$$\tau_p = \frac{T}{2}\left(\frac{1}{\hat{\alpha}_p} - 1\right), \tag{6.10}$$

where T represents the study period duration considered by the assignment model. Note that $\hat{\alpha}_p$ represents the same variable as in equation (6.5), but subscript l was removed because in this context it is the last link of the route p by definition. Further note that, without loss of generality, delay occurring on links in free flow conditions could be added to the first term by using the speed specified by the fundamental diagram in the free flow branch instead of the maximum speed on the link.

6.2.2.4 Route choice submodel for SUE

Within the route choice submodel, the route choice model uses the route travel times from the network loading submodel to compute route fractions for all route alternatives between an OD pair. The SUE assignment model assumes random utility maximization with perception errors, hence a multinomial logit (MNL) model to calculate route choice probabilities:

$$\psi_{rs,p} = \exp(-\mu_{rs} c_p) / \sum_{p' \in P_{rs}} \exp(-\mu_{rs} c_{p'}) \tag{6.11}$$

where $\psi_{rs,p}$ denotes the probability of choosing route p for demand on OD pair rs and μ_{rs} is a scale parameter describing the degree of travelers' perception errors on route travel times (where perfect knowledge is assumed when μ_{rs} approaches infinity). Note that μ_{rs} is determined using a global scale parameter μ (which can be estimated using the variance in observed data on route choices), normalized over ODpairs by $\mu_{rs} = \mu / \min_{p \in P_{rs}} \sum_{l \in L_p} \frac{L_l}{\dot{v}_l}$. This normalization ensures that the relative effect of perception errors is the same on all OD pairs (regardless of their average route travel time). Furthermore, the SUE is approximated using route choice iterations between the network loading and route choice submodels. In each route choice iteration, new route demands are calculated using:

$$Q_p = \psi_{rs,p} D_{rs} \tag{6.12}$$

where Q_p denotes the demand on route p. Note that in practical applications, convergence to SUE conditions is enforced and sped up by averaging the route choice probabilities between the route choice iterations using a smart averaging scheme (in this case the self regulating average (Liu et al., 2009) is used). The way this is done in the test case applications will be described in section 6.4. Further note that, without loss of generality, other discrete route choice models may be used (e.g. path size logit (Ben-Akiva and Bierlaire, 1999), C-logit (Cascetta et al., 1996) or paired combinatorial logit (Chu, 1989)), but this is outside the scope of this paper.

6.2.3 Proposed solution method

The proposed solution method solves bi-level problem (6.1) using first order Taylor approximated response functions to replace the SCCTA model to solve a series of simplified optimization problems. The simplified optimization problem (subsection 6.2.3.5) includes the sensitivity of link flows (subsection 6.2.3.1) and route queuing delays (subsection 6.2.3.2) for changes in OD-demand, but, contrary to the methods from current practice, avoids performing costly additional assignment model runs in the lower level to determine these sensitivities. Because the sensitivities used are point approximations, link state constraints are added to prevent their use outside their valid range. These constraints are also used to include observed congestion patterns in the matrix estimation problem (subsection 6.2.3.3).

6.2.3.1 Response function for observed link flows

To determine the response function for link flows we express link flow as a function of OD demand by substitution of (6.12) into (6.6) and summing over OD pairs:

$$y_l(\mathbf{D}) = \sum_{\boldsymbol{rs} \in \boldsymbol{RS}} \sum_{p \in P^{rs}} \hat{\alpha}_{lp}(\mathbf{D}) \psi_{rs,p} D_{rs} \tag{6.13}$$

Following (Frederix et al., 2013), we use the first order Taylor approximation around the current solution $\mathbf{D}_{k-1}$ as the response function for link flows yielding:

$$
\begin{aligned}
y_l(\mathbf{D}) &= \sum_{\boldsymbol{rs}\in\boldsymbol{RS}} \sum_{p\in P_{rs}} \hat{\alpha}_{lp}(\mathbf{D}_{k-1})\psi_{rs,p}D_{rs} + \sum_{\boldsymbol{rs}\in\boldsymbol{RS}} \frac{\partial y_l(\mathbf{D}_{k-1})}{\partial D_{rs}}\left(D_{rs} - D_{k-1,rs}\right) \\
&= \sum_{\boldsymbol{rs}\in\boldsymbol{RS}} \sum_{p\in P_{rs}} \hat{\alpha}_{lp}(\mathbf{D}_{k-1})\psi_{rs,p}D_{rs} \\
&\quad + \sum_{\boldsymbol{rs}\in\boldsymbol{RS}} \frac{\partial \sum_{\boldsymbol{rs}\in\boldsymbol{RS}} \sum_{p\in P^{rs}} \hat{\alpha}_{lp}(\mathbf{D})\psi_{rs,p}D_{rs}}{\partial D_{rs}}\left(D_{rs} - D_{k-1,rs}\right) \\
&= \sum_{\boldsymbol{rs}\in\boldsymbol{RS}} \sum_{p\in P_{rs}} \hat{\alpha}_{lp}(\mathbf{D}_{k-1})\psi_{rs,p}D_{rs} \\
&\quad + \sum_{\boldsymbol{rs}\in\boldsymbol{RS}} \left(D_{rs} - D_{k-1,rs}\right)\left[\sum_{\boldsymbol{rs'}\in\boldsymbol{RS}} \sum_{p'\in P_{rs}} \psi_{rs',p'} \frac{\partial \hat{\alpha}_{lp'}(\mathbf{D})}{\partial D_{rs}}\bigg|_{\mathbf{D}_{k-1}} D_{k-1,rs'}\right]
\end{aligned}
\tag{6.14}
$$

Or, in vector-matrix form:

$$
\mathbf{y}(\mathbf{D}) = \hat{\boldsymbol{\alpha}}\boldsymbol{\psi}\mathbf{D}_{k-1} + \frac{\partial \hat{\boldsymbol{\alpha}}}{\partial \mathbf{D}}\boldsymbol{\psi}\mathbf{D}_{k-1}(\mathbf{D} - \mathbf{D}_{k-1}) \tag{6.15}
$$

where $\mathbf{y}(\mathbf{D})$ is the vector of link flows of size $\tilde{L}$ x 1, $\hat{\boldsymbol{\alpha}}$ is the assignment matrix on route level (size $\tilde{L}$ x P) determined by assigning $\mathbf{D}^{k-1}$, $\boldsymbol{\psi}$ is a matrix of route choice probabilities of size P x RS and $\partial\hat{\boldsymbol{\alpha}}/\partial\mathbf{D}$ is the sensitivity of the assignment matrix on route level (size $\tilde{L}$ x RS x P).

6.2.3.2 Response function for observed queuing delays.

We propose to add observed travel times on observed routes $\tilde{p} \in \tilde{P}$ to the optimization problem. As reflected in equation (9), the average travel time on a route consists of the free-flow travel time and the queueing delay. Being a constant, the free-flow component per route, optionally including any delay occurring on links in free flow conditions (subsection 6.2.2.3), can be deducted from the observed route travel times to derive an approximated observed route queuing delay $\tau_{\tilde{p}}$.

Through equation (6.8) $\tau_{\tilde{p}}$ is a function of $\mathbf{T}_n$, which means that, through equations (6.6) and (6.7), is it also a function of $\mathbf{D}$. This means that a response function for observed queuing delays can be included into the optimization problem. Analogue to the approach taken for link flows, the first order Taylor approximation is derived for route queuing delays as:

$$
\begin{aligned}
\tau_{\tilde{p}}(\mathbf{D}) &= \frac{T}{2}\left(\frac{1}{\hat{\alpha}_{\tilde{p}}(\mathbf{D}_{k-1})} - 1\right) + \sum_{rs\in RS} \frac{\partial \tau_{\tilde{p}}(\mathbf{D})}{\partial D_{rs}}\left(D_{rs} - D_{k-1,rs}\right) \\
&= \frac{T}{2}\left(\frac{1}{\hat{\alpha}_{\tilde{p}}(\mathbf{D}_{k-1})} - 1\right) - \frac{T}{2}\sum_{rs\in RS} \frac{\partial \hat{\alpha}_{\tilde{p}}(\mathbf{D})/\partial D_{rs}\big|_{\mathbf{D}_{k-1}}}{\hat{\alpha}_{\tilde{p}}(\mathbf{D}_{k-1})^2}\left(D_{rs} - D_{k-1,rs}\right),
\end{aligned}
\tag{6.16}
$$

or, in vector-matrix form:

$$
\boldsymbol{\tau}(\mathbf{D}) = \frac{T}{2}\left(\frac{1}{\tilde{\boldsymbol{\alpha}}(\mathbf{D}_{k-1})} - 1\right) - \frac{T}{2}(\mathbf{D} - \mathbf{D}_{k-1})^T\left(\frac{\partial \tilde{\boldsymbol{\alpha}}(\mathbf{D})}{\partial \mathbf{D}} \cdot \frac{1}{\tilde{\boldsymbol{\alpha}}^{\mathbf{2}}(\mathbf{D}_{k-1})}\right) \tag{6.17}
$$

where $\boldsymbol{\tau}(\mathbf{D})$ and $\tilde{\boldsymbol{\alpha}}$ are vectors of size (1 x $\tilde{P}$) containing route queuing delays and flow acceptance factors on route level respectively, and $\partial\tilde{\boldsymbol{\alpha}}(\mathbf{D})/\partial\mathbf{D}$ is a matrix of size (RS x $\tilde{P}$) containing the sensitivity of the acceptance factors on route level. Note that $\tilde{p}$ may be any non-

cyclical combination of adjacent directed links in the network. Further note that because only the travel time of the whole route $\tilde{p}$ is relevant, $\tilde{\boldsymbol{\alpha}}$ only contains flow acceptance factors for the last link for each $p \in \tilde{\mathbf{P}}$ (removing the l subscript on the $\hat{\alpha}_{\tilde{p}}$ variable), whereas $\hat{\boldsymbol{\alpha}}$ contains flow acceptance factors for each link l within each route $p \in \mathbf{P}$. Therefore, $\tilde{\boldsymbol{\alpha}} \subseteq \hat{\boldsymbol{\alpha}}$ and $\partial\tilde{\boldsymbol{\alpha}}(\mathbf{D})/\partial\mathbf{D} \subseteq \partial\hat{\boldsymbol{\alpha}}(\mathbf{D})/\partial\mathbf{D}$, which means that calculation of the response function for queueing delay does not require any derivation of additional acceptance factors or sensitivities.

6.2.3.3 Link state constraints for observed congestion patterns

To include observed congestion patterns, link state constraints are used. Link state constraints enforce and preserve all links in a state known to coincide with an observed congestion patterns and are defined as:

$$\chi_j \left(\sum_{i \in I_j} T_{ij}(\mathbf{D}) - \delta_j R_j \right) \leq 0 \;\forall\, j \in L, \tag{6.18}$$

where χ_j indicates the state of link j, which is either constraining (χ_j =-1) or not constraining (χ_j =1) and δ_j represents the minimum size of the deficit (when χ_j =-1) or surplus (when χ_j =1) of supply at link j expressed as the ratio between demand for link j and its supply R_j. The response function for turn demands is derived by including (out)link-route incidence indicator σ_{jp} in both terms of (6.14) yielding:

$$\begin{aligned} T_{ij}(\mathbf{D}) = & \sum_{\boldsymbol{RS}} \sum_{\boldsymbol{p} \in \boldsymbol{P^{rs}}} \sigma_{jp}\, \hat{\alpha}_{ip}(\mathbf{D}_{k-1}) \psi_{rs,p} D_{k-1,rs} + \\ & \sum_{\boldsymbol{rs} \in \boldsymbol{RS}} \left(D_{rs} - D_{k-1,rs}\right) \left[\sum_{\boldsymbol{rs'} \in \boldsymbol{RS}} \sum_{p' \in P^{rs}} \psi_{rs',p'} \sigma_{jp} \left. \frac{\partial \hat{\alpha}_{ip'}(\mathbf{D})}{\partial D_{rs}} \right|_{\mathbf{D}^{k-1}} D_{k-1,rs'} \right] \\ & \forall i \in I_n, \forall j \in J_n, \forall n \in N, \end{aligned} \tag{6.19}$$

Contrary to observed link flows and route queuing delays, congestion patterns are not included as an objective function component but as linear constraints to the optimization problem. The reason for this is that the strict capacity constraints in the node model cause discontinuities in $\boldsymbol{\alpha}_n(\mathbf{T}_n)$ whenever a change in $\mathbf{T}_n$ causes an outlink from node n to switch from unconstrained to supply constrained or vice versa. This is illustrated in Appendix B using a numerical example. When such a link state switch would occur during matrix estimation, an update of $\hat{\alpha}_{lp}$ for all routes using this bottleneck would be necessary, as all downstream count locations change from sensitive to insensitive or vice versa. Furthermore, all (gradient approximation) calculations done so far with respect to these routes would become useless, since they are no longer valid after the state-change of the (potential) bottleneck. Simply updating $\hat{\alpha}_{lp}$ for all p using l after a link state switch would practically mean starting over the matrix estimation process with an altered prior matrix, causing unnecessary bias from the original prior matrix, wasted calculation time and probably non-convergence of the bi-level optimization problem.

The issue described above is present in all matrix estimation methods using an assignment model with strict capacity constraints. It has been described before in the context of matrix estimation using DTA models by (Frederix, 2012) who referred to it as "*Non-convexity* [of the upper level objective function] *due to congestion dynamics"*. Frederix suggests that any transitions between traffic regimes during matrix estimation should be avoided, meaning that the link states for all potential bottleneck links should be consistent with the start solution (i.e.: the link states from the assigned prior demand matrix) and that this state should be maintained

during matrix estimation. These suggestions are operationalized in the proposed solution method by addition of link state constraints (6.18) to the simplified optimization problem.
To specify $\boldsymbol{\chi} = \{\chi_j \; \forall \; j \in L\}$ the link states from assignment of the prior OD matrix could be used when this congestion pattern sufficiently corresponds to the observed congestion pattern or when no observed congestion patterns are available. Alternatively, $\boldsymbol{\chi}$ could be derived by determining the regime of all $j \in L$ by comparing observed link speeds (from e.g. floating car data or loop detectors) with critical link speeds from the fundamental diagram. When the observed speed is lower than the critical speed, the link is in congested state, otherwise the link is in free flow state. Then, set $\chi_j = -1$ on links that are in free flow state and have one or more inlinks that is in congested state and $\chi_j = 1$ for all other links. In case of diverges with one or more congested inlinks and more than one uncongested outlinks, additional data or knowledge is needed to determine which of the outlink(s) is actively constraining the inlink(s). Note that link state information from floating car data (on observed links) and from prior demand assignment results (on unobserved links) may be combined, hence the proposed solution method does not require observed link state information for all links in the network.
The minimum capacity surpluses (on non-constraining outlinks $\delta_j < 1 \; \forall \; \{j \in L \mid \chi_j = 1\}$) and deficits (on constraining outlinks $\delta_j > 1 \; \forall \; \{j \in L \mid \chi_j = -1\}$), act as a buffer around the discontinuity in $\boldsymbol{\alpha}_n(\mathbf{T}_n)$ and should be set to a value as close to one as possible, but sufficiently far away from one to prevent unintentional regime switches when running the lower level.
Note that there are three reasons to include link state information as constraints instead of objective function components. Firstly, constraints guarantee that transitions between traffic regimes during matrix estimation can indeed not occur. Secondly, since a link state is a binary variable, it is more natural to include it as a constraint. Thirdly, under the hood, any analytical gradient based solver will use some sort of barrier function to penalize constraint violations, which is effectively the same as including it in the objective function, but only now the solver (instead of the user) determines sufficiently large weight values.
Further note that the capacity deficits $\delta_j > 1$ may also be used to include observed capacity deficits (but then as a lower bound) derived from the prior assignment, by setting:

$$\delta_j = \frac{\sum_{i \in I_j} T_{ij}(\mathbf{D^0})}{R_j} \; \forall \; \{j \in L \mid \chi_j = -1\}, \tag{6.20}$$

or alternatively by setting to values derived from observed queue lengths in front of the link using a simple point queue model (Brederode et al., 2017). Note that the observed queuing delays from routes in $\tilde{P}$ represent the observed size of capacity deficits on the links it traverses. Therefore, to prevent overspecification (and the risk of infeasibility) of the optimization problem this may only be done for constraining outlinks that are not traversed by routes in $\tilde{P}$.

6.2.3.4 Normalization of weights

Introduced in subsection 6.2.1, weighing parameters w_1, w_2 and w_3 are used to define the relative importance of the three objective function components f_1 (prior OD demand), f_2 (link flows) and f_3 (route queuing delays). Typically, these weights are set proportional to the relative level of confidence associated with the three types of observed data. However, since these types of data have a different scale (a summation of OD demand differences versus a summation of link flow differences versus a summation of route queuing delay differences) they must be normalized to allow the weighting parameter to be given a meaningful interpretation expressing the relative importance on a scale of zero to one. Note that we choose

to separate normalization and weighting for sake of tractability; alternatively, one could combine the normalization and weighting into a single weight value for each component.
To normalize different objective function components a method described in e.g. (Alpcan, 2013) is used to determine the range between the optimal (so called *Utopia*) and pseudo-worst (so called *Nadir*) points in objective space for each component of the objective function. Using these points, the scale of each component relative to the other can be calculated and used for normalization within the objective function.
In our case, the value of the Utopia points $f_{1,U}$, $f_{2,U}$ and $f_{3,U}$ are all zero, occurring when $\mathbf{D} = \mathbf{D}_0$, $\mathbf{y}(\mathbf{D}) = \tilde{\mathbf{y}}$ and $\boldsymbol{\tau}(\mathbf{D}) = \tilde{\boldsymbol{\tau}}$ respectively. Nadir point $f_{1,N}$ is defined as the summation of quadratic differences between either the prior demand and zero or the prior demand and its upper bound. Nadir point $f_{2,N}$ is defined as the summation of quadratic differences between either the observed link flow and zero or the observed link flow and the links capacity. Nadir point $f_{3,N}$ is approximated by the summation of quadratic differences between either the observed route queuing delays and zero or the observed route queuing delays and route queuing delays from an assignment of the upper bound on OD demand. For all three Nadir points, for each element, the largest quadratic difference is chosen. Equations (6.21) summarize the Nadir point definitions described above.

$$\begin{aligned} f_1^N &= \sum_{rs \in RS} \max\left[D_{rs,0}^2, \left(\overline{D}_{rs} - D_{rs,0}\right)^2\right] \\ f_2^N &= \sum_{l \in \tilde{L}} \max[\tilde{y}_l^2, (C_l - \tilde{y}_l)^2] \\ f_3^N &= \sum_{p \in \tilde{P}} \max\left[\tilde{\tau}_p^2, \left(\tilde{\tau}_p - \tau_p(\overline{\mathbf{D}})\right)^2\right] \end{aligned} \tag{6.21}$$

After (arbitrary) normalization of f_2 and f_3 to f_1 the objective function in (6.1) reads:

$$\mathbf{D}^* = \underset{\mathbf{D}}{\operatorname{argmin}} \left(w_1 \sum(\mathbf{D} - \mathbf{D_0})^2 + \frac{w_2 f_{2,N}}{f_{1,N}} \sum(\mathbf{y}(\mathbf{D}) - \tilde{\mathbf{y}})^2 + w_3 f_{3,N}/f_{1,N} \sum(\boldsymbol{\tau}(\mathbf{D}) - \tilde{\boldsymbol{\tau}})^2 \right) \tag{6.22}$$

6.2.3.5 Simplified optimization problem

We aim to solve the bi-level problem defined in equation (6.1) by use of objective function (6.22) and the response functions for observed link flows and observed route queuing delays defined in equations (6.15) and (6.17) respectively. To avoid transitions between traffic regimes during estimation link state constraints are added in the form of inequality (6.18) on top of the non-negativity constraints. Furthermore, upper bounds on the OD demands $\overline{\mathbf{D}}$ are added which can be used to reduce the number of potential constraining links from L to $J_{\overline{\mathbf{D}}}$, thereby decreasing the calculation time of the solution method. Upper bounds $\overline{\mathbf{D}}$ are typically related to $\mathbf{D_0}$ reflecting a maximum (absolute or relative) allowed increase per cell. This yields the simplified optimization problem (6.23) displayed below.

Optimization problem (6.23) is a simplified version of the true optimization problem (6.1), because the link flows and link demands are approximated instead of determined by the assignment model, because the link state constraints restrict the solution space in order to be able to safely use the approximated variables, and because the vector of route fractions $\boldsymbol{\psi}$ is assumed constant. The simplified optimization problem has a quadratic objective function, linear inequality constraints and is typically very large (given the number of elements in $\mathbf{D}$ in real world transport models).

$$\begin{aligned}
\mathbf{D}^* = \underset{\mathbf{D}}{\operatorname{argmin}}\big(& w_1 \sum(\mathbf{D}-\mathbf{D_0})^2 + w_2 f_{2,N}/f_{1,N} \sum(\mathbf{y}(\mathbf{D})-\tilde{\mathbf{y}})^2 \\
& + w_3 f_{3,N}/f_{1,N} \sum(\boldsymbol{\tau}(\mathbf{D})-\tilde{\boldsymbol{\tau}})^2\big) \\
\text{Subject to: } & \mathbf{y}(\mathbf{D}) = \hat{\boldsymbol{\alpha}}\boldsymbol{\psi}\mathbf{D}_{k-1} + \frac{\partial\hat{\boldsymbol{\alpha}}}{\partial\mathbf{D}}\boldsymbol{\psi}\mathbf{D}_{k-1}(\mathbf{D}-\mathbf{D}_{k-1}) \\
& \boldsymbol{\tau}(\mathbf{D}) = \frac{T}{2}\left(\frac{1}{\tilde{\boldsymbol{\alpha}}(\mathbf{D}_{k-1})}-1\right) - \frac{T}{2}(\mathbf{D}-\mathbf{D}_{k-1})^T\left(\frac{\partial\tilde{\boldsymbol{\alpha}}(\mathbf{D})}{\partial\mathbf{D}}\cdot\frac{1}{\tilde{\boldsymbol{\alpha}}^2(\mathbf{D}_{k-1})}\right) \\
& \mathbf{0} \le \mathbf{D} \le \bar{\mathbf{D}} \\
& \chi_j\left(\sum_{i\in I_j} T_{ij}(\mathbf{D}) - \delta_j R_j\right) \le 0 \;\forall\, j \in J_{\bar{\mathbf{D}}}
\end{aligned} \tag{6.23}$$

6.3 Solution algorithm

This section describes the proposed solution algorithm. Subsection 6.3.1 provides an overview, whereas subsections 6.3.2 and 6.3.3 describe the algorithm details and some of its mathematical properties respectively.

6.3.1 Overview

The proposed solution algorithm is summarized in Figure 6.3: Overview of proposed solution approach. Each iteration consists of six steps to solve (an updated version of) the simplified optimization problem. Within an iteration, only a single SCCTA model assignment is run to determine the assignment matrix (subsection 6.3.2.1). Then, only for turns traversing an active bottleneck node, the local sensitivity of its bottleneck flow acceptance factor to local turn demand is approximated using finite differences, requiring one additional run of only the node model component (subsection 6.3.2.2). The resulting local sensitivities are used to construct the approximated sensitivity of the assignment matrix, as described in subsection 6.3.2.3. Furthermore, the approximated sensitivity of the assignment matrix is used to approximate gradients of the link flow and route section delay components within the objective function (subsection 6.3.2.4) and its linear constraints (6.3.2.5), which are used to efficiently solve the simplified optimization problem (subsection 6.3.2.6). Using locally approximated sensitivities, the computational cost of the proposed method is negligible compared to methods using full assignment runs to determine sensitivities.

The proposed solution method is fast, scalable, robust, tractable and reliable because conditions under which a solution to the simplified optimization problem exists are known and because the problem is convex and has a smooth objective function. These favorable mathematical properties are discussed in subsections 6.3.3.1, 6.3.3.2 and 6.3.3.3 respectively as they played an important role during the development of the solution method and the implementation of the solution scheme.

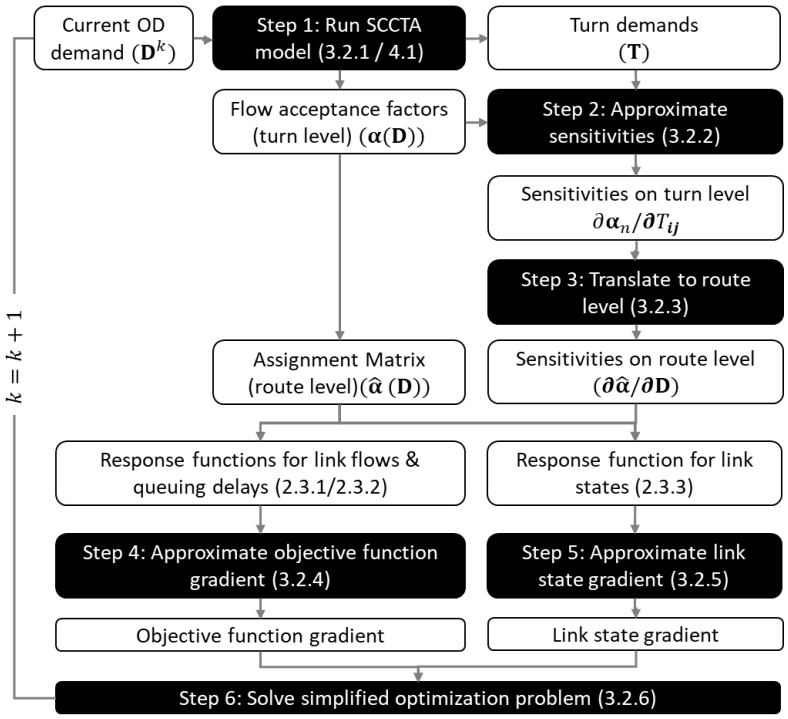


Figure 6.3: Overview of proposed solution approach

6.3.2 Solution scheme

Using knowledge about the SCCTA model (subsection 6.2.2) and the simplified optimization problem (subsection 6.2.3.5), in this section the solution scheme is presented. It consists of six steps, each of which is executed in each iteration. The sixth step includes a stop criterion to determine whether the solution to the true optimization problem has been found or if an additional iteration is required.

6.3.2.1 Step 1: Run SCCTA model and derivation of the assignment matrix

As a reference, we first describe how the assignment matrix relates to assignment model results for SCRTA class models before we describe this relationship is the case of SCCTA models. Because SCRTA class models lack strict capacity constraints, all traffic arrives at its destination within the study period. Therefore, an element in the assignment matrix from SCRTA class models merely describes the proportion of demand on an OD pair that has chosen a route using the considered observed link and can be derived from the route choice probabilities calculated by the route choice submodel only using:

$$A_{rs,l} = \sum_{p \in P_l} \psi_{rs,p}, \tag{6.24}$$

where $A_{rs,l}$ denotes the entry in the assignment matrix for observed link l and OD pair rs and P_l denotes the set of routes using link l.

When using a SCCTA class model, entries in the assignment matrix are reduced by the proportion of OD flow being held up by capacity constraints on links upstream from the considered link as calculated by the network loading submodel, yielding:

$$A_{rs,l} = \sum_{p \in P_l} \psi_{rs,p} \hat{\alpha}_{lp} \, , \tag{6.25}$$

where the route-level assignment matrix $\hat{\alpha}_{lp}$ is calculated using (6.5). Note that, because within each iteration constant route probabilities are assumed (subsection 6.2.2), elements in the route based assignment matrix ($\hat{\alpha}_{lp}$) are the driving variables in the lower level. They are used to approximate link flows (6.14), queuing delays (6.16) and turn demands (6.19) and to approximate the gradients of the objective function (subsection 6.3.2.4) and the link state constraints (subsection 6.3.2.5). Therefore, in the remainder of this paper we do show how $A_{rs,l}$ and its sensitivities are derived, but in the solution algorithm, its two components $\psi_{rs,p}$ and $\hat{\alpha}_{lp}$ are used separately.

6.3.2.2 Step 2: Approximate sensitivities on turn level

Analogue to the assignment matrix itself (subsection 6.3.2.1), the sensitivity of the assignment matrix (to be captured in $\partial\hat{\boldsymbol{\alpha}}/\partial\mathbf{D}$) is constructed from the sensitivities of the acceptance factors on turn level.

Using the node model, a point derivative of $\alpha_{ij}$to any T_{ij} can be approximated using finite differences. Only for turns that are actively constrained by an outlink or turn (i.e. $\alpha_{ij} < 1$), the local sensitivity of its flow acceptance factor to local turn demand needs to be approximated. This is being done using the (one sided) finite difference method around the solution obtained by the (single) full assignment run in the lower level:

$$\frac{\partial\boldsymbol{\alpha}_n}{\partial T_{ij}} = \frac{\boldsymbol{\alpha}_n^* - \Gamma_n(\mathbf{T}_n^-, \mathbf{C}_n, \mathbf{R}_n)}{\epsilon} \quad \forall \{ij \in IJ | \alpha_{ij}^* < 1\}, \tag{6.26}$$

where $\boldsymbol{\alpha}_n^*$ and $\mathbf{T}_n^*$ are the vectors of turn flow acceptance factors and turn demands from the solution calculated by the full assignment, ϵ is the step size used for the finite difference calculation, $\mathbf{T}_n^- = (\mathbf{T}_n^* \backslash \{T_{ij}^*\}) \cup \{T_{ij}^* - \epsilon\}$, the set of turn demands where the turn demand for the considered turn ij is lowered by ϵ for finite differences. This requires only one additional application of a node model for each actively constrained turn. These point derivatives are then used as an approximate of $\partial\alpha_{ij}/\partial T_{ij'}$ in the upper level. Note that these point derivatives approximate the function well if no discontinuity occurs. This is illustrated in the example from appendix B.

Note that by approximating derivatives we determine the partial derivatives for all turns on the network, but we choose to omit approximating secondary interaction effects. This means that we omit the fact that when simultaneously changing multiple elements in **D**, the effect on the set of route based flow acceptance factors $\hat{\alpha}_{lp}$ and therefore the effect on the assignment matrix **A** might not be simply the sum of the effects of changing D_{rs} sequentially per OD pair. Furthermore, note that although the proposed method would also work for nodes on which constraints imposed by geometry of the node itself exist, for the sake of simplicity, in this paper we assume these so called internal node constraints to be non-existent.

6.3.2.3 Step 3: Translate sensitivities to route level

The sensitivities on turn level (subsection 6.3.2.2) and the link states constraints (subsection 6.2.3.3) allow for calculation of the sensitivity of the assignment matrix on route level ($\partial\hat{\boldsymbol{\alpha}}/\partial\mathbf{D}$ in (6.23)). This is a three-dimensional matrix containing elements $\partial\hat{\alpha}_{lp}/\partial D_{rs'}$ that describe the sensitivity of link l used by route p for changes in demand on OD-pair rs'. We calculate these elements as follows.

First, we determine $\partial\hat{\alpha}_{lp}/\partial Q_{p'}$, the sensitivity of link l on route p for changes in demand on route p', by taking the derivative of equation (6.5) to **Q** (using the product rule), yielding:

$$\frac{\partial \hat{\alpha}_{lp}(\mathbf{Q})}{\partial Q_{p'}} = \left(\prod_{ij' \in IJ_{p,il}} \alpha_{ij'}(\mathbf{Q}) \right) \left(\sum_{ij' \in IJ_{p,il}} \left(\left. \frac{\partial \alpha_{ij'}}{\partial Q_{p'}} \right|_{\mathbf{Q}} / \alpha_{ij'}(\mathbf{Q}) \right) \right) \tag{6.27}$$

To simplify equation (6.27), we first define the following variables. Let ij_p^* be the first blocked turning movement on route p (i.e.: $\alpha_{ij_p^*} < 1$). Let $\overline{IJ_p^*}$ be the set of turns on route p located upstream from ij_p^* and $\underline{IJ_p^*}$ be the set of turns on route p downstream from turn ij_p^* until and including turn il. Furthermore let $T_{ij_{p'}''}$ be demand on turn $ij_{p'}''$ on route p' that influences $\alpha_{ij_p^*}$ (and hence must located on the same node as ij_p^*). Note that $\alpha_{ij_p^*}$ may be influenced by routes using the turn itself (i.e.: $ij_{p'}'' = ij_p^*$), but it can also be influenced by routes on other turns sharing their outlink with one of the turns that share their inlink with turn ij_p^* (in which case $ij_{p'}'' \neq ij_p^*$). Then, given that link state constraints will maintain to be satisfied, three properties of equation (6.27), all related to the strict capacity constraints in the assignment model, are considered:

1. On turns ij' located upstream from the first blocking turn on route p, by definition, all demand passes, hence $\alpha_{ij'} = 1 \; \forall \; ij' \in \overline{IJ_p^*}$ and $\partial \alpha_{ij'} / \partial Q_{p'} = 0 \; \forall \; ij' \in \overline{IJ_p^*}$;
2. On turns ij' located downstream from the first blocking turn on route p, due to the strict capacity constraints, the acceptance factor on the first blocking turn ij_p^* will neutralize any changes in demand on p, such that its downstream turns become insensitive: $\partial \alpha_{ij'} / \partial Q_{p'} = 0 \; \forall \; ij' \in \underline{IJ_p^*}$.

Incorporating these two properties into equation (27) yields:

$$\begin{aligned} \frac{\partial \hat{\alpha}_{lp}(\mathbf{Q})}{\partial Q_{p'}} &= \left(\prod_{ij' \in \left\{ ij_p^* \,;\, \underline{IJ_p^*} \right\}} \alpha_{ij'}(\mathbf{Q}) \right) \left. \frac{\partial \alpha_{ij_p^*}}{\partial Q_{p'}} \right|_{\mathbf{Q}} / \alpha_{ij_p^*}(\mathbf{Q}) \\ &= \left(\prod_{ij' \in \underline{IJ_p^*}} \alpha_{ij'}(\mathbf{Q}) \right) \left. \frac{\partial \alpha_{ij_p^*}}{\partial Q_{p'}} \right|_{\mathbf{Q}} \end{aligned} \tag{6.28}$$

To use the approximated point derivatives from 6.3.2.2, the third property is considered:

3. Analogue to the second property, when turn $ij_{p'}''$ on route p' (the turn for which its demand is influencing $\alpha_{ij_p^*}$) is located downstream from the first blocking turn on route p', acceptance factor $\alpha_{ij_p^*}$ will neutralize any changes in demand on p', such that $\frac{\partial \alpha_{ij'}}{\partial Q_{p'}} = 0 \; \forall \; \{p' : ij_{p'}'' \in \underline{IJ_{p'}^*}\}$. In other cases, $Q_{p'} = T_{ij_{p'}''}$ and thus $\partial \alpha_{ij'} / \partial Q_{p'} = \partial \alpha_{ij'} / \partial T_{ij_{p'}''} \; \forall \; \{p' : ij_{p'}'' \in \left\{ ij_{p'}^* \,;\, \overline{IJ_{p'}^*} \right\}\}$.

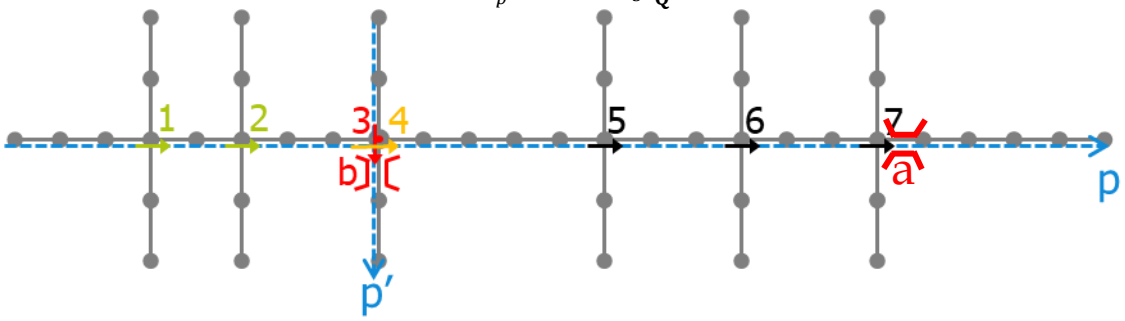
Incorporating the third property into equation (6.28) yields:

$$\frac{\partial \hat{\alpha}_{lp}(\mathbf{Q})}{\partial Q_{p'}} = \begin{cases} \left(\prod_{ij' \in \underline{IJ_p^*}} \alpha_{ij'}(\mathbf{Q}) \right) \left. \frac{\partial \alpha_{ij'}}{\partial T_{ij''_{p'}}} \right|_{\mathbf{Q}} & \forall \{p' : ij''_{p'} \in \left\{ ij^*_{p'} \, ; \, \overline{IJ^*_{p'}} \right\}\} \\ 0 & \forall \, \{p' : ij''_{p'} \in \underline{IJ^*_{p'}}\} \end{cases} \tag{6.29}$$

Example network

To further clarify these variables and properties, the figure below shows an example network (nodes and links in grey) with routes p and p' indicated by the blue dashed lines. Small arrows indicate turning movements used by the routes, each turning movement is labeled by a colored number. There are bottlenecks a and b causing turn 7 on route p and turn 3 on route p' to be blocking. Due to FiFo on the upstream node of the bottleneck link, this also causes turn 4 on route p to be blocking, thereby defining $ij_p^* = 4$, $\overline{IJ_p^*} = \{1,2\}$, $\underline{IJ_p^*} = \{5,6,7\}$ *and* $ij'' = 3$. *Considering route p, the first two properties indicate that only the sensitivity of the alpha on turn* $ij_p^* = 4$ *is relevant, whereas the third property indicates that demand on route p' is represented by turn demand on turn ij'' = 3. This means that:*

- *for links upstream from turn 3:* $\frac{\partial \hat{\alpha}_{lp}(\mathbf{Q})}{\partial Q_{p'}} = 0$;
- *for links between turns 4 and 7:* $\frac{\partial \hat{\alpha}_{lp}(\mathbf{Q})}{\partial Q_{p'}} = \left. \frac{\partial \alpha_4}{\partial T_3} \right|_{\mathbf{Q}}$, *and;*
- *for links downstream from turn 7* $\frac{\partial \hat{\alpha}_{lp}(\mathbf{Q})}{\partial Q_{p'}} = \alpha_7 \left. \frac{\partial \alpha_4}{\partial T_3} \right|_{\mathbf{Q}}$.



Equation (6.29) expresses equation (6.27) in terms of turn based acceptance factors that are output from the SCCTA model and a single partial derivative that can be derived using finite differences of its node model component. Interpretation of equation (6.29) shows that its second term represents the maximum sensitivity of route flows on p for demand on route p' whereas the first term propagates (and dampens) this sensitivity downstream from turn ij_p^* to turn ij on route p. The derivatives with respect to OD- (instead of route-) demand are defined as:

$$\frac{\partial \hat{\alpha}_{lp}}{\partial D_{rs}} = \sum_{p' \in P_{rs,l}} \psi_{rs,p'} \frac{\partial \hat{\alpha}_{lp}}{\partial Q_{p'}}. \tag{6.30}$$

and since route choice probabilities are fixed within a single upper level evaluation ($\partial \psi_{rs,p} / \partial D_{rs} = 0 \; \forall \, rs \in RS$), the sensitivity of an element in the assignment matrix can now be expressed as:

$$\frac{\partial A_{rs,l}}{\partial D_{rs'}} = \sum_{p \in P_{rs,l}} \psi_{rs,p} \frac{\partial \hat{\alpha}_{lp}}{\partial D_{rs'}}. \tag{6.31}$$

The sensitivities in (6.30) are used to approximate link flows (6.14), queuing delays (6.16) and turn demands (6.19) in the upper level using while the solver is evaluating a candidate vector of OD demands **D** and to approximate the gradients of the objective function (subsection 6.3.2.4) and the link state constraints (subsection 6.3.2.5).

6.3.2.4 Step 4: Approximate objective function gradient

Gradients can be derived for all three components (f_1, f_2 and f_3) of the objective function. Note that by doing so, the gradient of the total objective function (6.2) is also determined. The partial derivatives of the first part of objective function to OD-demands are given by:

$$\frac{\partial f_1}{\partial D_{rs}} = \frac{\partial}{\partial D_{rs}} \left(\sum_{rs \in RS} (D_{rs} - D'_{rs})^2 \right) = 2(D_{rs} - D'_{rs}), \quad \forall\, rs \in RS. \tag{6.32}$$

The partial derivatives of the second part of the objective function f_2 to OD-demands **D** are derived using the approximated sensitivity of the assignment matrix from subsection 6.3.2.3. First the gradient of f_2 is translated from OD to route level:

$$\frac{\partial f_2}{\partial D_{rs}} = \sum_{p \in P_{rs}} \psi_{rs,p} \frac{\partial f_2}{\partial Q_p} = \sum_{p \in P_{rs}} \psi_{rs,p} \sum_{l \in \tilde{L}} 2(y_l - \tilde{y}_l) \frac{\partial y_l}{\partial Q_p}. \tag{6.33}$$

To derive $\partial y_l / \partial Q_p$, first equation (6.13) for link flows is expressed on route level:

$$y_l(\mathbf{Q}) = \sum_{p \in P_l} \hat{\alpha}_{lp}(\mathbf{Q}_{k-1}) Q_p + \sum_{p' \in P} \left(Q_{p'} - Q_{k-1,p'}\right) \left[\sum_{p \in P_l} \left. \frac{\partial \hat{\alpha}_{lp}}{\partial Q_{p'}} \right|_{\mathbf{Q}_{k-1}} Q_{k-1,p} \right] \tag{6.34}$$

Taking the derivative to Q_p yields:

$$\begin{aligned} \frac{\partial y_l}{\partial Q_p} &= \frac{\partial}{\partial Q_p} \left(\sum_{p \in P_l} \hat{\alpha}_{lp} Q_p + \sum_{p' \in P} \left(Q_{p'} - Q_{k-1,p'}\right) \left[\sum_{p \in P_l} \left. \frac{\partial \hat{\alpha}_{lp}}{\partial Q_{p'}} \right|_{\mathbf{Q}^{k-1}} Q_{k-1,p'} \right] \right) \\ &= \hat{\alpha}_{lp} + \sum_{p' \in P} \left. \frac{\partial \hat{\alpha}_{lp}}{\partial Q_{p'}} \right|_{\mathbf{Q}_{k-1}} Q_{k-1,p'}, \quad \forall l \in L, \forall p \in P, \end{aligned} \tag{6.35}$$

where elements in the assignment matrix $\hat{\alpha}_p^l$ and their sensitivities are calculated using equations (6.5) and (6.29) respectively.

Finally, the partial derivatives of the third part of the objective function f_3 to OD-demands **D** are given by:

$$\begin{aligned}\frac{\partial f_3}{\partial D_{rs}} &= \frac{\partial}{\partial D_{rs}}\left(\sum_{\tilde{p}\in\tilde{P}}\left(\tau_{\tilde{p}}(\mathbf{D})-\tilde{\tau}_{\tilde{p}}\right)^2\right) = \sum_{\tilde{p}\in\tilde{P}} 2\left(\tau_{\tilde{p}}(\mathbf{D})-\tilde{\tau}_{\tilde{p}}\frac{\partial\tau_{\tilde{p}}(\mathbf{D})}{\partial D_{rs}}\right)\\ &= \sum_{\tilde{p}\in\tilde{P}} 2\left(\tau_{\tilde{p}}(\mathbf{D})-\tilde{\tau}_{\tilde{p}}\right)\left(-\frac{T}{2}\frac{\partial\hat{\alpha}_{\tilde{p}}(\mathbf{D})/\partial D_{rs}}{\hat{\alpha}_{\tilde{p}}(\mathbf{D})^2}\right)\\ &= \sum_{\tilde{p}\in\tilde{P}}\left(\tilde{\tau}_{\tilde{p}}-\tau_{\tilde{p}}(\mathbf{D})\right)\left(\frac{T\,\partial\hat{\alpha}_{\tilde{p}}(\mathbf{D})/\partial D_{rs}}{\hat{\alpha}_{\tilde{p}}(\mathbf{D})^2}\right)\end{aligned} \tag{6.36}$$

This section has shown that the gradient of the objective function can be approximated using information from a single assignment model evaluation (subsection 6.3.2.1) and a single additional node model evaluation for each turn 6.3.2.2 only.

6.3.2.5 Step 5: Approximate link state gradient

The gradient of the link state constraints is defined as the derivatives of equation (6.18) to OD demand:

$$\frac{\partial}{\partial\mathbf{D}}\left[\chi_j\left(\sum_{i\in I_j}T_{ij}-\delta_j R_j\right)\right] = \chi_j\sum_{i\in I_j}\frac{\partial T_{ij}}{\partial D_{rs}}\ \forall j\in L,\forall\, rs\in RS \tag{6.37}$$

To calculate $\partial T_{ij}/\partial D_{rs}$, we can use the same approach used for calculation of the gradient of the link flow part of the objective function (equations (6.34)-(6.35)), once we have established the relationship between link flow and turn demands. To do so we point out that in equation (6.7), the turn demand is expressed in terms of all route flows on the turns **in**link (y_{ip}), directed towards the considered outlink j (σ_{jp}), thereby excluding the acceptance factor on turn ij itself. This is shown when equation (6.6) is substituted in equation (6.7), yielding:

$$\begin{aligned}T_{ij} &= \sum_{\boldsymbol{RS}}\sum_{\boldsymbol{p}\in\boldsymbol{P^{rs}}}\sigma_{jp}\sigma_{ip}Q_p\prod_{ij'\in IJ_p^{ij}\{\setminus ij\}}\alpha_{ij'}\\ &= \sum_{RS}\sum_{p\in P_{rs}}\sigma_{jp}\sigma_{ip}Q_p\prod_{ij'\in IJ_{p,ij}}\alpha_{ij'}/\alpha_{ij}\\ &= \sum_{RS}\sum_{p\in P_{rs}}\sigma_{jp}y_{lp}/\alpha_{ij}.\end{aligned} \tag{6.38}$$

Realizing that turn demands are related to link flows through equation (6.38), approximations for $\partial\hat{\alpha}_{ip}/\partial Q_{p'}\ \forall i\in I_n,\forall n\in N$ can be derived by replacing the superscript l with i in equation (6.29) and remove any routes p' for which $ij'' = ij_p^*$ yielding:

$$\frac{\partial\hat{\alpha}_{ip}(\boldsymbol{Q})}{\partial Q_{p'}} = \left(\prod_{ij'\in\underline{IJ_p^*}}\alpha_{ij'}(\boldsymbol{Q})\right)\frac{\partial\alpha_{ij_p^*}}{\partial T_{ij''}}\bigg|_{\mathbf{Q}_{k-1}}\quad\forall p'\exists\ ij''\in\overline{IJ_{p'}^*} \tag{6.39}$$

which is the turn-demand equivalent of equation (6.29). These derivatives can be used to calculate

$$\frac{\partial T_{ij}}{\partial Q_p} = \sigma_{jp}\hat{\alpha}_{ip} + \sum_{p' \in P} \left.\frac{\partial \hat{\alpha}_{ip}}{\partial Q_{p'}}\right|_{\mathbf{Q}_{k-1}} Q_{k-1,p}\,, \quad \forall j \in L, \forall p \in P, \tag{6.40}$$

which is the turn-demand equivalent of equation (6.35) to be translated to OD level using:

$$\frac{\partial T_{ij}}{\partial D_{rs}} = \sum_{p \in P_{rs}} \psi_{rs,p} \frac{\partial T_{ij}}{\partial Q_p} \tag{6.41}$$

which is multiplied by its corresponding χ_j to yield the gradient of the link state constraint. Analogue to the approximation of the gradient of the objective function (subsection 6.3.2.4), this section has shown that the gradient of the link state constraints can be approximated using information from a single assignment model evaluation (subsection 6.3.2.1) and a single additional node model evaluation for each turn traversing an active bottleneck only (subsection 6.3.2.2).

6.3.2.6 Step 6: Solve simplified optimization problem

The simplified optimization problem can be solved by applying any solver that can handle such a problem. In this paper, the interior point algorithm described in (Waltz et al., 2006) is used in combination with the approximated gradients from subsections 6.3.2.4 and 6.3.2.5. Once solved, the estimated OD matrix is assigned using the SCCTA model as a step 1 of the next iteration, and the objective function value is evaluated and compared to a user defined threshold value for convergence.

To have a meaningful and comparable convergence criterion, the convergence threshold is defined in terms of average link flow and route delay deviations (equation (6.42)).[20] A run is considered converged when the solver has found a feasible solution and both conditions are met.

$$\frac{\sum_{a \in \tilde{\mathbf{A}}} |y_a - \tilde{y}_a| / \tilde{y}_a}{|\tilde{\mathbf{A}}|} \leq \varepsilon_{\mathbf{A}} \wedge \frac{\sum_{p \in \tilde{\mathbf{P}}} |\tau_p - \tilde{\tau}_p| / \tilde{\tau}_p}{|\tilde{\mathbf{P}}|} \leq \varepsilon_{\mathbf{P}} \tag{6.42}$$

For non-converging runs, criteria on objective function stability (absolute difference between the true objective function value of latest and previous iteration) and the maximum number of iterations are added.

If either the convergence or stability criterion is met, or when the maximum number of iterations is reached, the solution is accepted, and the algorithm stops. In other cases, the algorithm starts a new iteration by using the assignment results from the new OD matrix that was already assigned for objective function evaluation. Note that the difference between the objective function value of the simplified problem (known after step 6) and the objective function value after assignment (known after the assignment in step 1) can be used as an indicator for the size of the errors due to the first order Taylor approximations and the neglection of secondary interaction effects (as defined in subsection 6.3.2.2).

[20] Note that the proposed solution method does not capture secondary interaction effects (subsection 6.3.2.2) which means that the optimization problem is not fed with information to actively steer it towards solutions where combined changes in the OD matrix yield lower OD matrix deviations whilst still fitting it to observed network data. This means that it can only fit an OD matrix to observed network data by increasing the deviation from the prior OD matrix which means that it makes no sense to include a stop criterion on the deviations to the prior OD demand.

As described in subsection 6.2.3.3, for the solution algorithm to converge, links states should be consistent with the start solution (the prior demand matrix) and these states should be maintained during matrix estimation. Link state constraints (6.18) take care of maintaining link states, but do not enforce the start solution to be consistent. To prevent the sensitivity information in the first iteration to be based on a (possibly) inconsistent OD matrix and let the algorithm to take off on a false start, an (optional) nudging iteration is prepended to the solution scheme. In this nudging iteration, the same interior point algorithm is used to solve only the feasibility problem from the link state constraints by setting the objective function to a value of zero.

6.3.3 Mathematical properties of the simplified optimization problem

6.3.3.1 Feasibility

When the conditions under which a solution to the problem exists are known, the input can be adapted to satisfy these conditions. By doing so, the solver is guaranteed to find a feasible solution, thereby contributing to the reliability of the solution method. For the simplified optimization problem (6.23), feasibility is guaranteed when the non-negativity and link state constraints are satisfied by (assignment of) the prior OD matrix. This means that the prior OD matrix may not contain negatives, and that the link state constraints must be satisfied in the assignment results of the prior OD matrix. In all subsequent iterations, feasibility will be automatically maintained through the constraints themselves.

This means that feasibility can be guaranteed by adding a check on negatives in de prior OD matrix to prevent violation of the non-negativity constraints and to set the values for χ_j according to the assignment results of the prior OD matrix itself to prevent violation of the link state constraints. Alternatively, when using χ_j values from an exogenous source (i.e. observed congestion patterns), the simplified optimization problem (6.23), but with removed objective function, can be solved to nudge the prior OD matrix into the feasible region. Solving this problem is computationally very cheap as it is an instance of the "first phase problem" in the two phase simplex method (Murty, 1991, pp 60).

6.3.3.2 Convexity

Problems that are convex are likely to be solved using polynomial time algorithms which are relatively fast and scalable. Furthermore, any solution to a convex problem is a global minimum and when the problem is strictly convex this global minimum is unique, contributing to robustness and tractability of the solution method.

In appendix A it is proven that the first part of the objective function of the simplified optimization problem (6.23) is strictly convex, whereas the second and third part of the objective function are convex. Furthermore, all considered constraints are linear inequalities, and as such form a closed convex set which means that (6.23) as a whole is classified as a convex optimization problem.

6.3.3.3 Smoothness

Problems having a smooth (i.e.: twice differentiable) objective function may be solved using algorithms that exploit information from its gradient and Hessian, which are relatively fast and can provide first order optimality measure values. Formulated differently: smoothness of the objective function avoids the need to resort to derivative free algorithms, thereby improving the tractability of the solution method.

In subsection 6.3.2.4, first order derivatives of the objective function where calculated whereas in appendix A, it is shown that the second order derivatives (the Hessian matrices) of all three

objective function components can also be calculated. This means that the objective function is indeed twice differentiable in **D**, which was implicitly already concluded in subsection 6.3.2.5 when the objective function of the simplified optimization problem was classified as quadratic.

6.4 Application on a small network

In this section, the added value of the proposed matrix estimation methodology is demonstrated using four test case applications on the well-known Sioux Falls test network that gradually build up from the traditional approach used for SCRTA models towards the proposed solution method from section 6.3. First, the specifics of the used SCCTA model implementation (subsection 6.4.1) and network (subsection 6.4.2) are described. Then, the evaluation framework (subsection 6.4.3) and test case applications (subsection 6.4.4), are defined and results are presented (subsection 6.4.5).

6.4.1 SCCTA model implementation

The mathematical relationships in SCCTA assignment models for the SUE have already been described in subsection 6.2.2. In this section, specifics of the used SCCTA model implementation are briefly considered.

With respect to the network loading submodel, the SCCTA model STAQ (Static Traffic Assignment with Queueing, Brederode et al., 2019) is used which possesses all the favorable properties described in section 0. Note that in (Brederode et al., 2019), the assignment model used in this paper is referred to as STAQ - variation without spillback, but for brevity, in this paper we shall abbreviate this to STAQ. Findings in this paper with respect to the matrix estimation method apply to any (future) SCCTA class network loading submodel using an explicit node model described in subsection 6.2.2.

The used route choice submodel relies on a route set that is pre-generated from the digitized transport network using a route set generator. The route set generator used combines the Dijkstra algorithm to find the shortest path between each OD pair and the repeated random sampling process on free flow link travel times from (Fiorenzo-Catalano, 2007) to generate alternative routes. Route filters calibrated on GPS data (Fafieanie, 2009) are applied after the repeated random sampling process to reduce route overlap, remove irrelevant routes and restrict the size of the set of potential routes.

To check for convergence to SUE conditions, the adapted relative duality gap as derived in (Bliemer et al., 2013) is used, which accounts for perception errors and thus reaches zero upon convergence when using the MNL route choice model:

$$DG = \frac{\sum_{rs \in RS} \sum_{p \in P_{rs}} Q_p (c_p + -\mu_{rs}^{-1} \ln (Q_p - \zeta_{rs})}{\sum_{rs \in RS} D_{rs} \zeta_{rs}} \tag{6.43}$$

where $\zeta_{rs} = \min_{p \in P_{rs}} [c_p + \mu_{rs}^{-1} \ln Q_p]$ represents the minimum stochastic path cost on OD pair rs.

In line with (Boyce et al., 2004; Brederode et al., 2019; Han et al., 2015; Patil et al., 2021), for all test applications in this paper, a threshold value of 5E-05 is used as the stop criterion for the traffic assignment model. Note that (Bliemer et al., 2014) provides proof for existence and uniqueness of the user equilibrium solution for exactly this SCCTA model implementation.

As mentioned in subsection 6.2.2.4 convergence to the SUE is enforced and sped up by smartly averaging route demands about iterations. To this end, the method of self-regulating averages (SRA, Liu et al., 2009) is used which tends to provide fast convergence with high precision.

Note that, apart from using this more efficient averaging scheme, this paper uses the exact same SCCTA model (i.e.: it uses the same network loading submodel and reaches the same SUE conditions) as described in (Bliemer et al., 2014), who use the method of successive averages instead.

6.4.2 Network and observed input data

To evaluate the quality and convergence properties of the methodology synthetic test cases on the well-known Sioux Falls network are used which contains 24 centroids (that also serve as nodes) and 35 links. The network and OD matrix are downloaded from (Transportation Networks for Research Core Team, 2019) but the OD matrix was adapted because the original OD matrix represents extremely high levels of congestion: all or nothing assignment yielded 64% of the links having a volume/capacity (V/C) ratio greater than one, whereas 26% of the links had a V/C ratio greater than two.

Such high demand is not desired for test case applications because its model variables are not within a range that is representative for real world situations, because the delays caused by such high demands would force travelers to change their mode, departure time and/or trip frequency, effectively lowering demand for the mode and time period considered by the assignment model.

Dividing the OD matrix by a factor of two is in line with findings in (Chakirov and Fourie, 2014) and yield link demands within a more realistic -but still very high- level of congestion: 22% links with a V/C ratio higher than one and 3% links with a V/C ratio higher than two. Interpreting this OD matrix as the 'true' OD demand it was assigned using STAQ to generate 'true' observed flows, congestion patterns (constrained (out)links) and travel times. This 'true' OD matrix contains 528 OD pairs with nonzero demand, and during assignment 1430 unique routes where generated and used, yielding 2.71 routes per OD pair on average.

Note that in real world applications, observed flows, speeds, travel times and congestion patterns are observed in a more fine-grained time interval than a typical study period of SCCTA models, hence time-aggregation of observed values is required. In line with assumptions of static traffic assignment models, time- averaged values of observed flows, speeds and travel times should be used, and time average values for speeds or densities should be used to derive congestion patterns. It is however not possible to correct real world observed data to another assumption made in static traffic assignment models: the network is assumed empty before and after the study period. The solution to this problem would be to extend the approach to use a semi-dynamic capacity constrained traffic assignment model (SDCCTA) such that residual traffic is accounted for in both the model and the observations, which is therefore recommended in subsection 6.6.2.

The (arbitrary selected) set of count locations and the set of constrained outlinks as well as the two selected routes with observed queuing delays (one traversing a single vertical queue, the other one traversing two vertical queues) are displayed on the left hand side of Figure 4 whereas the prior assignment results (link flows, (vertical) queue sizes and link speeds as a percentage of maximum speeds) are displayed on the right hand side of Figure 6.4.[21]

[21] Note that In Figure 4, to indicate the turning movement from which the queuing delay on a routes last node is included, the route is defined to end halfway onto the last nodes downstream link.

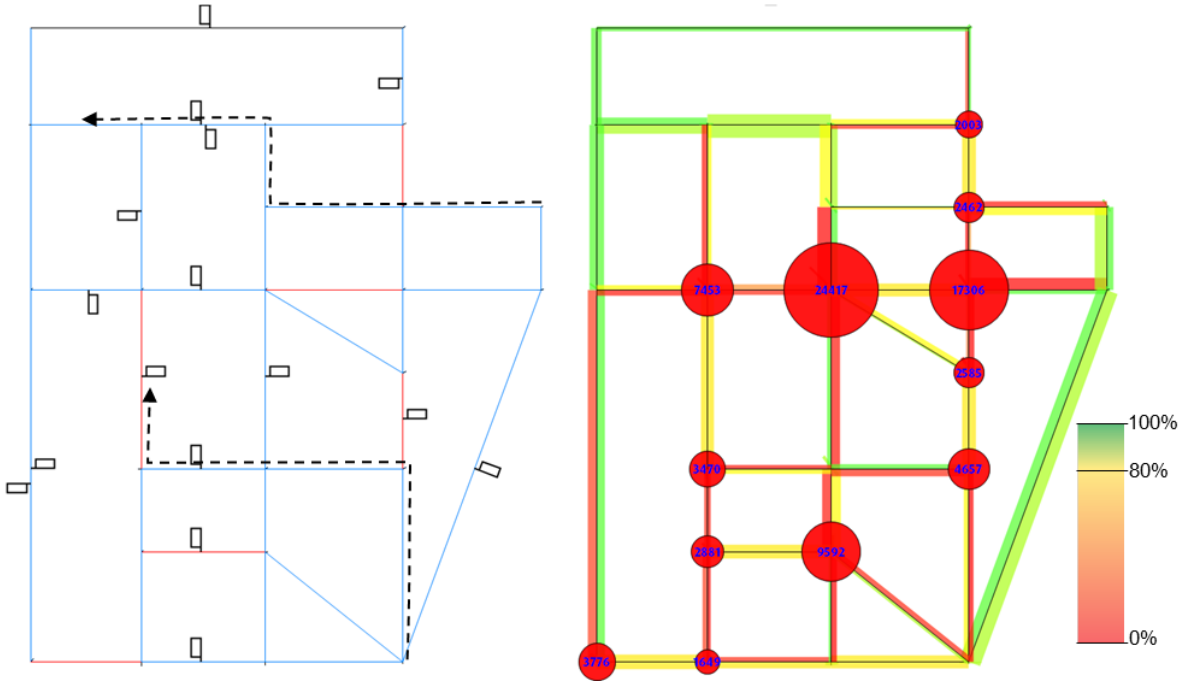


Figure 6.4 Sioux Falls network: left: count locations (black flags next to links on the side of direction of travel), observed route definitions (dashed arrows) and congestion patterns (blocking links in red); right: assignment results of 'true' OD matrix (width: flow; colour: speed as percentage of maximum speed; pie charts: number of vehicle hours lost in (vertical) queue))

6.4.3 Evaluation framework

The performance of a matrix estimation methodology relates to the difficulty of the problem it needs to solve, which is related to the amount and (in)consistency and sensitivity of its input data (observed flows, congestion patterns, prior OD matrix and travel times).

Inconsistencies in input data force the solution methodology to (implicitly) choose or average between inconsistent datapoints which deteriorates the quality of the output and the speed and end-level of convergence. Sensitivity of the model also has influence on the convergence of the lower level, as very sensitive models (i.e.: high levels of congestion) force the use of small step sizes and the use of smart step size calculation methods, which both increase calculation time. Furthermore, high sensitivities amplify the negative effects on convergence due to differences between the true (6.1) and simplified (6.23) optimization problem.

To evaluate the performance of the matrix estimation methodology, the evaluation framework from Figure 6.5 is used. For this synthetic application on the Sioux Falls network, a 'true' OD matrix ($\mathbf{D}_{true}$) is available, whereas in real model applications this is not the case. Therefore, $\mathbf{D}_{true}$ is perturbed and used as the prior OD matrix together with the 'true' observed link flows and travel times. Recognizing that inconsistencies are merely coincidences of inconsistent inputs each test case application is run repetitively 100 times with a differently perturbed prior OD matrix, thereby robustly evaluating the performance. Recognizing the effect of sensitivity of the model, the prior OD matrix (being perturbed around $\mathbf{D}_{true}$) as well as the other input represent a situation with much congestion (as shown in the right hand side from Figure 6.4).[22]

[22] Note that more inconsistencies could be introduced by also perturbing around the 'true' observed flow and travel time values. We leave this idea for further research.

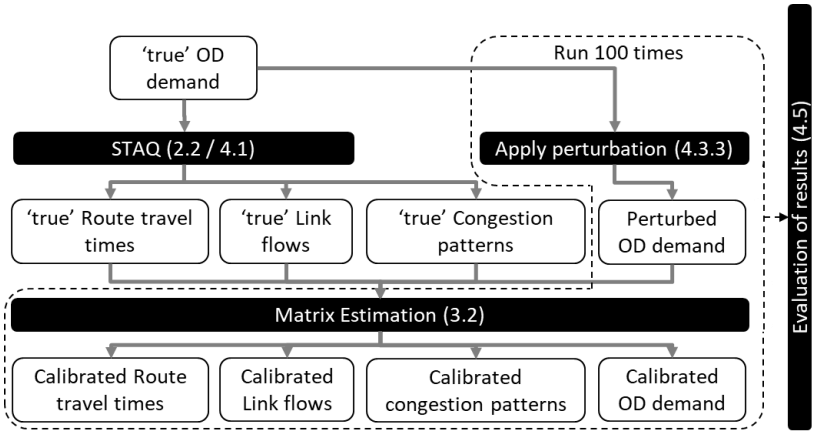


Figure 6.5 Evaluation framework used for test case applications 3 and 4

6.4.3.1 Stop criteria and performance indicators

For all applications in this paper, the thresholds for the convergence criterion (defined in subsection 6.3.2.6) are set to 1% for the average percentual deviation in link flow ($\varepsilon_{\mathbf{A}}$) and 5% for the average percentual route delay deviation ($\varepsilon_{\mathbf{P}}$). These values fall well below accuracy levels of observed link flows and route delay deviations in strategic transport models. The threshold of the stability criterion is set to zero, meaning that it is only met when the upper level yields the exact same OD matrix in two concurrent iterations. The maximum number of iterations is set to 10, because a solution method that would require that many iterations would be unsuitable for application on large scale networks due to computational requirements.

With respect to comparison of observed and calibrated link flows and route travel times, the convergence criterion effectively monitors both. Therefore, strictly spoken, there is no need to monitor these explicitly when evaluating application results. Instead, the number of iterations, the number of upper level function evaluations and the calculation time (on a machine with AMD Ryzen 9 3900X CPU (12 cores) @3.79 Ghz) required for convergence are monitored as the performance indicator for the match between link flows and route travel times. However, for clarity, average link flow and route delay deviations will also be included in the analysis of the results in subsection 6.4.5.

With respect to comparison of OD demands, the evaluation framework from Figure 6.5 allows to evaluate to what extent the matrix estimation method retrieves $\mathbf{D}_{true}$ when fed with congestion patterns, link flows and travel times that are consistent with it, but also how much deviation from the prior OD matrix it requires. The structural similarity index (SSIM, Djukic et al., 2013) is used to compare both the estimated and 'true' OD matrices as well as the estimated and prior OD matrices. More specifically, mean SSIM (mSSIM) values are used as performance indicators. Following suggestions by (Ros-Roca et al., 2018) mSSIM values are calculated by averaging SSIM values per row of the OD matrix considered. The mSSIM is an indicator for the similarity in matrix structure (by the definition from (Behara et al., 2020): the arrangement of the destinations from each origin). To also consider the actual differences in values per OD pair ((Behara et al., 2020) use the term 'mass'), for the comparison between estimated and prior OD matrices root mean squared error (RMSE) values are also presented.

Differences between observed and estimated congestion patterns are enforced to be nonexistent in the upper level by the link state constraints. But because the problem that is solved in the upper level (6.23) is a simplified version of the true problem (6.1), this does not guarantee that all link state constraints are satisfied after application of the SCCTA model in the last iteration. Therefore, the number of link state violations in the final assignment results of the lower level are explicitly monitored as a performance indicator of the match with observed congestion patterns.

6.4.3.2 Generation of perturbed OD matrices

Treating entries in an OD matrix as exponentially distributed random variables, differences between different OD matrices would be best governed by a Laplace distribution (Kotz et al., 2001). Therefore, a Laplace distribution is used to randomly draw the cell-by-cell perturbations applied to the true OD matrix to generate the 100 OD matrices used as priors.

To determine the relevant size range of the perturbations, peak-hour OD matrices derived from observed Dutch mobile phone data from Vodafone for all non-holiday workdays in March 2017 where used. First, for each OD pair having more than 10 observations in the considered peak hour during the considered days, differences between the cell's values over the different days and its average value where calculated and expressed as percentual differences to the average. This yielded a dataset of about 4.5 million relative differences per peak hour on which Laplace distributions where fitted yielding location parameters close to 0 and scale parameters around 0.135 for both peak periods. The estimated distributions are displayed in Figure 6.6 together with the distribution applied on the Sioux Falls test cases. To ensure that the test cases represent a worst-case scenario for the matrix estimation methods, the distribution applied on the Sioux Falls test cases uses a much higher and wider distribution by increasing the scale parameter to 0.3, which means that the structure of the OD matrix is severely changed by the perturbations, but the number of nonzero OD pairs remained 528 for all perturbed ODmatrices.[23] Furthermore, to make sure that the performance of the proposed solution method is tested in congested conditions, it was verified that the percentage of demand per observed link varied between 5% and 100% for all use cases.

Note that the fitted distributions imply that no structural bias is introduced during generation of perturbed OD matrices. Given the use case of the method (subsection 6.1.1: to refine prior OD demand from a demand model with data on link and route level), this a deliberate choice for the test case applications, because if in practice there would be a structural bias between prior assignment results and observed demand levels, authors suggest using a global scaling factor to remove it before applying the matrix estimation method, such that the prior structure is kept in-tact as much as possible.

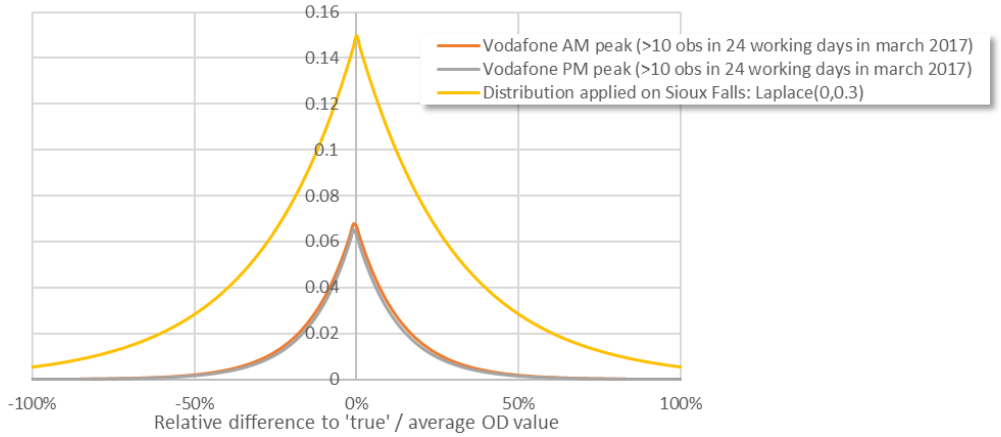


Figure 6.6: Probability mass functions of the Laplace distributions fitted to (relative) variations in Vodafone data and the Laplace distribution applied on the Sioux Falls test cases in this section

[23] Note that during application, the distribution was truncated such that OD pairs for which perturbations of less than -100% would be applied where assigned a value of 1 (to not introduce new OD pairs with value 0) while OD pairs with a true value of more than 500 and a perturbation of more than 100% where truncated to that value.

6.4.4 Test case applications

Four test case applications are defined that gradually add solution method features and support for additional types to the traditional approach used for SCRTA models until we arrive at the proposed solution method from section 6.3. Distinctive properties of the four test case applications are summarized in Table 6.1. In all test case applications, the same route set with 1430 routes was used yielding an average of 2.71 routes per OD pair (since all perturbed OD matrices have the same number of nonzero OD pairs). Furthermore, in all test case applications, the same pre-generated route set (subsection 6.4.1) is used.

Test case		**Response functions**		**Congestion patterns**				**Prior**		**Flows**		**Queuing delays**	
#	Referral	$\mathbf{y}(\mathbf{D})$	$\boldsymbol{\tau}(\mathbf{D})$	$\boldsymbol{\chi}$	δ_j (deficits)	δ_j (surpluses)	nudging	w_1	$\widetilde{\mathbf{D}}$	w_2	$\tilde{\mathbf{y}}$	w_3	$\tilde{\boldsymbol{\tau}}$
1	[REF]	eq (13)		∅	∅	∅		½	$\mathbf{D}_o$	½	from $\mathbf{D}_{true}$		∅
2	[+LS]	eq (13)		from $\mathbf{D}_{true}$	1.01	0.99		½	$\mathbf{D}_o$	½	from $\mathbf{D}_{true}$		∅
3	[+LS+S]	eq (15)		from $\mathbf{D}_{true}$	1.01	0.99		½	$\mathbf{D}_o$	½	from $\mathbf{D}_{true}$		∅
4	[+LS+S+QD]	eq (15)	eq (17)	from $\mathbf{D}_{true}$	1.01	0.99(94);0.9(6) (9)/100		⅓	$\mathbf{D}_o$	⅓	from $\mathbf{D}_{true}$	⅓	from $\mathbf{D}_{true}$

Table 6.1: Distinctive properties of the four test case applications

The first test case application (referred to as [REF]) employs the approach used in (Brederode et al., 2017),weighing (normalized) deviations from prior demand and (normalized) deviations from observed link flows equally ($w_1 = w_2 = ½$). [REF] acts as a reference in which sensitivities and link state constraints are omitted, in which case it is not possible to include observed queuing delays nor congestion patterns. Therefore, in [REF], problem (6.23) simplifies into:

$$\begin{aligned} &\mathbf{D}^* = \underset{\mathbf{D}}{\operatorname{argmin}}\left(w_1 \sum(\mathbf{D} - \mathbf{D_0})^2 + w_2 f_{2,N}/f_{1,N} \sum(\mathbf{y}(\mathbf{D}) - \tilde{\mathbf{y}})^2\right) \\ &\text{Subject to: } \mathbf{y}(\mathbf{D}) = \hat{\boldsymbol{\alpha}}\boldsymbol{\psi}\mathbf{D} \\ &\qquad\qquad \mathbf{0} \le \mathbf{D} \le \bar{\mathbf{D}} \end{aligned} \tag{6.44}$$

which resembles, apart from the traffic assignment used, any of the gradient based approaches described in (Abrahamsson, 1998).

The second test case application (referred to as [+LS]) adds link state constraints from observed congestion patterns derived from $\mathbf{D}_{true}$ by adding equation (6.18) to optimization problem (6.44). The minimum capacity surpluses δ_j (on non-constraining outlinks $\{j \in L \mid \chi_j = 1\}$) and deficits (on constraining outlinks $\{j \in L \mid \chi_j = -1\}$) that act as a buffer around discontinuities in $\boldsymbol{\alpha}_n(\mathbf{T}_n)$ are set to 0.99 and 1.01 respectively. By adding link state constraints, transitions between traffic regimes are avoided, which should improve convergence and reduce the link state violations. However, because exogenous congestion patterns (from $\mathbf{D}_{true}$) are used inconsistencies between the exogenous congestion patterns and the prior OD matrix may cause the objective function to become non-convex (as described in 6.2.3.3), hence reducing convergence in [+LS].

The third test case application (referred to as [+LS+S]) adds sensitivities (+S) to the response function for link flows to account for the sensitivity of the assignment matrix to changes in OD demand. To do so, the response function for link flows is restored to the first order Taylor approximation (equation (6.15)), effectively arriving at problem (6.23), but excluding observed queuing delays (i.e.: $w_3 = 0$). The inclusion of sensitivities should improve convergence as the upper level has more accurate information which should also lead to less link state violations and less unnecessary changes to the prior OD demand.

The fourth test case application (referred to as [+LS+S+QD]) adds observed queueing delays (+QD) derived from $\mathbf{D}_{true}$. This means that (normalized) route delay deviations are added to (6.23), weighted equally to both (normalized) OD demand deviations and (normalized) link flows ($w_1 = w_2 = w_3 = ⅓$). Because the queuing delays operate on the level of individual turning movements (instead of aggregations over inlinks (for link flow deviations) or outlinks (for link state constraints)), the solution candidates evaluated by the upper level contain relatively large changes to individual turn demands. This increases the likelihood that discontinuities in $\boldsymbol{\alpha}_n(\mathbf{T}_n)$ are crossed due to difference between the true function within the node model (6.8) and its linear approximate derived by finite differences (see subsection 6.3.2.2) used in the upper level. This mechanism is described in more detail in appendix C and it led to 9 runs that did not converge within the maximum number of (10) iterations. For these runs, a nudging iteration (section 6.3.2.6) was prepended to reduce the chance that the mechanism occurs in the first iteration, whereas for 6 of these 9 runs it was also necessary to increase the buffer around discontinuities in $\boldsymbol{\alpha}_n(\mathbf{T}_n)$ for one or two of the non-constraining outlinks by lowering the minimum capacity surpluses δ_j from 0.99 to 0.9 to prevent the mechanism to occur in later iterations.

6.4.5 Results

Results of all four test case applications are displayed in Figure 6.7. Because each test case is run a hundred times with a different prior demand matrix, all performance indicators are summarized as cumulative distributions of each indicator over the different runs. Recall from subsection 6.4.3.1 that stop criteria are defined for link flow and route delay deviations. Note however, that for [REF], [+LS] and [+LS+S], the stop criterion on route delay deviation is ignored, as in these applications, route delays are not included in the optimization. This means that for these applications convergence is reached when only link flow deviations meet the stop criterion, whereas for [+LS+S+QD] convergence is only reached when both criteria are met.

Considering the level of convergence of the bi-level problem (upper left graph), the number of converging runs is read by looking at the value at iteration ≤ 10. This shows that in [REF] 98/100 runs converge. Addition of link state constraints [+LS] causes a reduction to 96/100 converging runs, which shows that (at least on this network), the positive effect of added stability is outweighed by the negative effect of (potential) additional data inconsistencies. As expected in subsection 6.4.3.1, addition of sensitivities to the response function [+LS+S] increases the number of converging runs (to 99/100) as the upper level has more accurate information. Addition of queuing delays [+LS+S+QD] only slightly reduces the number of converging runs to 98/100. However, without the algorithmic enhancements (nudging and lowering the minimum capacity surpluses on specific link state constraints) the number of runs converged would have been 91/100, showing that addition of observed queuing delays without mitigating measures has the largest negative effect on the level of convergence. Note that the speed of convergence barely varies over the different test case applications; only the addition of queuing delays structurally lags about one iteration for runs requiring more than four iterations.

Note that for all four test case applications the cumulative distributions in the upper left graph indicate that no additional runs are converging beyond iteration 8 or 9. Additional test runs (not described in this paper) with the maximum number of iterations criterion increased beyond 10 confirm this observation. Analysis of individual non-converging data points within some of the non-converging runs point towards differences between the route choice probabilities from the ‘true’ and perturbed OD demand. These differences can cause certain combinations of observed link flows and/or route queuing delays to become inconsistent, causing the optimal, still feasible, solution to not satisfy the convergence criteria (subsection 6.4.3.1). That differences

due to route choice inconsistencies are indeed the cause is confirmed by the fact that additional test runs (not described in this paper) where route choice probabilities from $\mathbf{D}_{true}$ where kept fixed over iterations all converged within two to four iterations. Note that in practice, non-converging datapoints are easily detectable and may be resolved by increasing the difference tolerance on one or both datapoints or removing one of the datapoints.

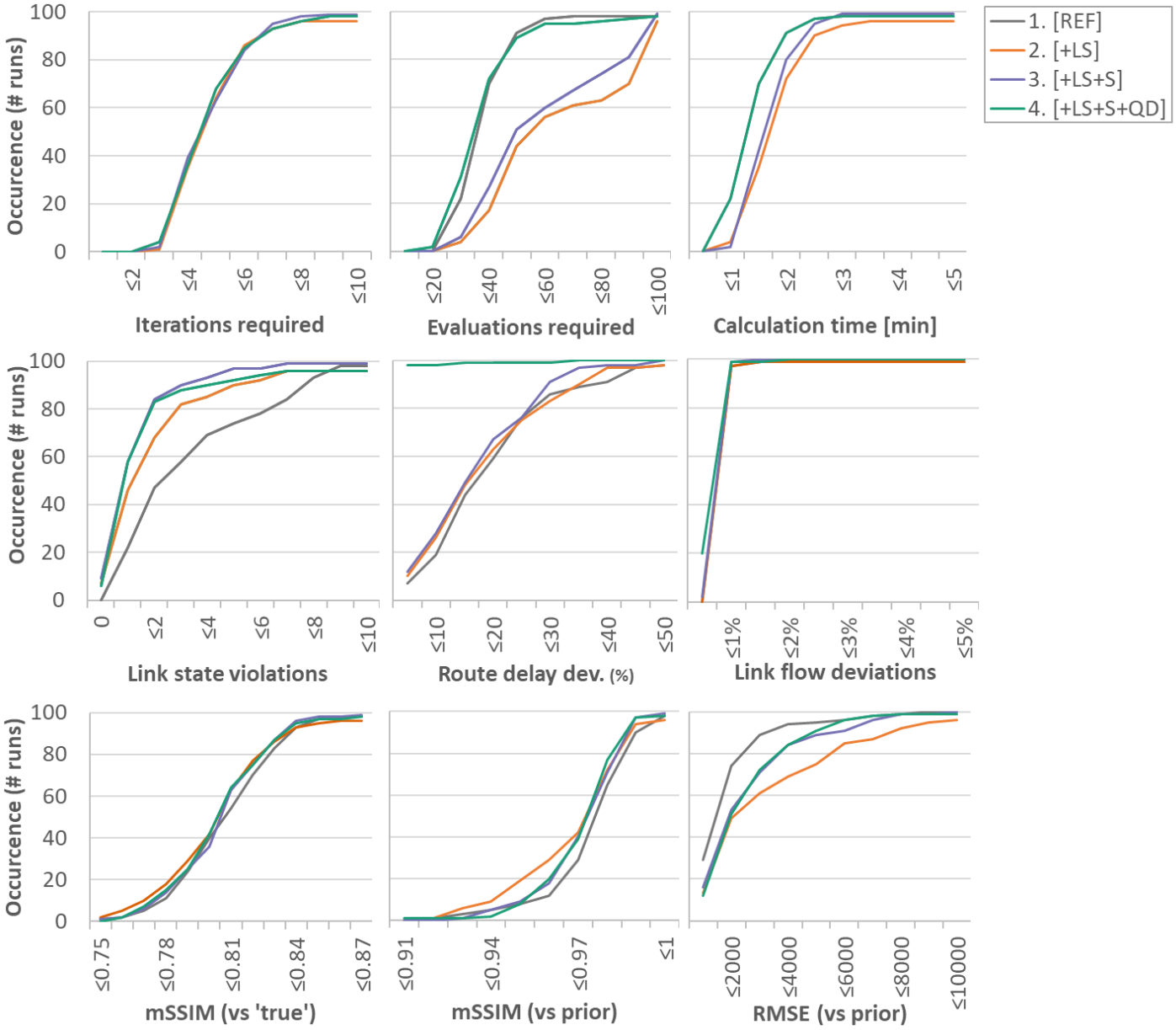


Figure 6.7: cumulative distributions of performance indicators for all four test case applications. Note that 40 runs fall outside the range of the vertical axis of the mid upper graph for test case application 4. Therefore it is noted here that the 95th percentile of the number of evaluations required for this test case application is 214.

From the number of upper level function evaluations required for convergence (upper mid graph) two mechanisms are derived. Firstly, adding data sources increases difficulty of the optimization problem, and thus requires more function evaluations, which is shown by comparison of [+LS] with [REF] for the effect of addition of congestion patterns and comparison of [+LS+S+QD] with [+LS+S] for the effect of addition of queuing delays. Secondly, enhancing the gradient information in the upper level by including sensitivities

increases effectiveness of the upper level solver and thus reduces the number of function evaluations required, which is shown by comparison of [+LS+S] with [+LS].
Considering the calculation time required for convergence (upper right graph) in relation to the previous two graphs shows that relatively small differences in the number of iterations required and the relatively large differences in the number of function evaluations required translate into relatively small differences in calculation time. This reveals that most time is still spent in the lower level (and within the lower level the SCCTA run takes up most of the time), whilst the upper level is relatively fast.
With respect to the number of link state violations (left graph on second row), comparison of [+LS] with [REF] shows the effectiveness of the link state constraints, whereas differences between these cumulative distributions and a (non-shown) vertical asymptote at 0 violations represent the number of violations caused by the difference between the simplified problem solved in the upper level and the true bi-level optimization problem. Results from test case application [+LS+S] compared to [+LS] show that addition of sensitivity information to the gradient decreases the number of link state violations, as expected in subsection 6.4.3.1, while adding queuing delay information (compare [+LS+S+QD] with [+LS+S]) does not have a clear effect. The latter observation makes sense because in these test case applications, congestion patterns and queuing delays are fully consistent, as these are both derived from $\mathbf{D}_{true}$.
With respect to the average route delay deviations (mid graph on second row), comparison of [+LS] and [+LS+S] with [REF] shows that inclusion of link states and sensitivity information only slightly improves the fit on observed route delays, whereas the proposed method [LS+S+QD] is required to include observed queuing delays. The limited effect of adding link state and sensitivity information on the fit on observed route delays shows that there is indeed relatively limited correlation between model variables, suspectedly because temporal correlations are avoided as the solution method employs a static (hence time aggregated) assignment model (recall from section 0).
For the sake of completeness, Figure 6.7 also includes a comparison of average link flow deviations (right graph on second row). This graph confirms that, except for the 7 non-converging runs already described above, the stop criterion of 1% average link flow deviations is met for all four test case applications.
Comparison of the estimated against the 'true' OD matrix (lower left graph) shows no notable differences between the different runs. Apparently, even though observed link flows, congestion patterns and queuing delays are all derived from $\mathbf{D}_{true}$, in all four hundred runs there is an abundant number of optimal solutions close to the prior. Additional test runs (not included in this paper) show that this remains the case, even when the search space is increased by excluding the prior demand component in the objective function (by setting w_1 to zero). This demonstrates that although the proposed solution method finds the global optimum to the simplified optimization problem for each iteration, this does not mean that it finds the global optimum (if it exists) to the true optimization problem.
Comparison of the estimated against the prior OD matrix (lower mid and right graphs) shows that adding congestion patterns leads to substantial larger deviations from the prior OD matrix compared to the results from test cases with added sensitivity information. This demonstrates the effect of the increased effectiveness of the upper level solver due to the sensitivity information on the quality of the estimated matrix. Compared to the mSSIM, the RMSE indicator shows larger differences, since the latter captures all differences, whereas the former only targets differences in matrix structure (subsection 6.4.3.1).

6.5 Application on a large network

In this section, results of a large scale application of the proposed solution algorithm on (data from) the strategic transport model of the province of Noord-Brabant, the Netherlands (Heynickx et al., 2016) are presented.

6.5.1 Transport model and observed input data

The network and prior OD demand for road traffic of the base year (2015, version S107) of the provincial model of Noord-Brabant (abbreviated in Dutch to 'BBMB') is used. This network contains 1425 centroids 145.269 links and 103.045 nodes. The prior OD matrix used describes the AM peak period (07:00-09:00) and contains 1.580.764 OD pairs with nonzero demand. During assignment 5.162.010 unique routes where generated and used, yielding 3.26 routes per OD pair on average.

With respect to observed input data, the full BBMB count-data set for the AM peak period is used, which contains observed link flows for 415 count locations, along with a set of observed travel times on 24 (highway) routes. Up until now, this set was only used for validation purposes, as the prevailing matrix estimation method of the BBMB-model is not capable of including observed queuing delays. Link state constraint values are derived from assignment results of the prior demand matrix. To reduce problem size, the upper bound on od demands (subsection 6.2.3.5) is set to $\bar{\mathbf{D}} = \mathbf{2D_0}$, yielding a set of relevant links ($J_{\bar{\mathbf{D}}}$) containing 21 constraining and 1583 non-constraining links (hence a reduction of 98.9% compared to the set L containing all links).

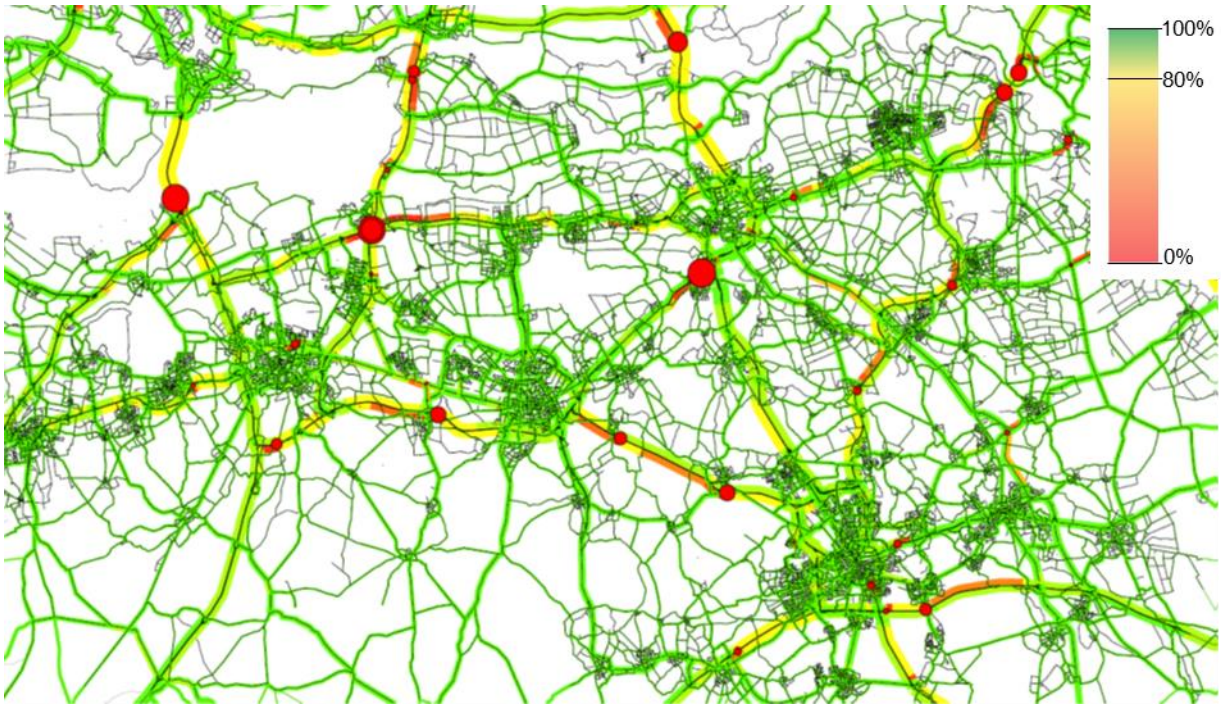


Figure 6.8: (study area of the) BBMB network: assignment results of prior OD matrix (width: flow; colour: speed as percentage of maximum speed; pie charts: number of vehicle hours spent in (vertical) queue))

Note that the BBMB model employs junction modelling, which means that its node models do not only account for constraints due to limited supply on outlinks (in the form of link capacities), but also for the effect of limited supply due to conflict points on the junction itself (i.e. crossing flows; in the form of turn capacities). To support this in the context of the proposed matrix estimation method, turn capacities are calculated using the junction modelling

component of OmniTRANS (Bezembinder and Brandt, 2016) and included as internal node constraints (Tampère et al., 2011) while running the SCCTA model (subsection 6.3.2.1) and approximating sensitivities (subsection 6.3.2.2).

6.5.2 Convergence and calculation time

Ten iterations of the proposed methodology where run on the BBMB model, after which all convergence indicators (solid lines in Figure 6.9) seem to have stabilized. The minimum capacity surpluses and deficits added to prevent unintentional regime switches when running the lower level (subsection 6.2.3.3) where both set to 1% (i.e.: $\delta_j = 0.99$ for non-constraining outlinks and $\delta_j = 1.01$ for constraining outlinks). The weighting parameters in the objective function where set to w_1=0.01 (prior), w_2=0.12 (link flows) and w_3=0.87 (queuing delays), and directly applied (i.e.: normalization as described in 6.2.3.4 was omitted). Also in Figure 6.9, dashed lines indicate minimum deviations yielded by the software currently used for matrix estimation in the BBMB, which employs the [REF] method described in 6.4.4.

Considering the average link flow deviation per count location, the upper left graph shows that these quickly reduce from 27% to around 5% and that it outperforms the reference methodology (which averages on 90 vehicles per count location) in iteration three. This graph also shows that for link flow deviations, to save calculation time, the algorithm could be stopped after iteration four, as results hardly improve afterwards.

The average route delay deviations (upper mid graph) show a reduction from around 39 to around 15 percent, which translates to a reduction from 112 seconds (in a range from 13 up to 241 seconds) to 43 seconds (in a range from 1 up to 114 seconds). Note that from iteration 7 onwards, the fit on link flows slightly deteriorates while the route delay deviations keep improving. During these iterations, the objective function keeps improving, which demonstrates the weighting of objective function components. Further note that the reference method does not consider route delay deviations, which causes an average deviation of 322 seconds per observed route, which is (much) larger than the average route delay deviation when assigning the prior demand.

Considering the congestion patterns (lower right graph) deviations reduce from a deficit of more than 1700 vehicles on four different locations in the first iteration to zero vehicles from the fifth iterations onwards. The reference method also converges to zero vehicles. Note that from the fifth iteration onwards, the graph still reports one location on which the congestion pattern is not matched. This is because the minimum capacity surpluses and deficits added to prevent unintentional regime switches when running the lower level (subsection 6.2.3.3) are included in this graph. Inspection of the assignment results showed that the concerning location is a constraining outlink for which the demand is indeed higher than its capacity, but lower than the capacity multiplied by δ_j.

Considering deviations to the prior OD demand, the lower right graph shows that, in correspondence to the link flow deviations, most changes to the OD matrix are done in the first four iterations. This is also confirmed by the objective function values (lower mid graph). The proposed method requires less than 0.1 trips per OD pair on average, which is less than the reference method, which required a change of 0.13 trips per OD pair on average.

Calculation times per iteration of the proposed solution method on an AMD Ryzen 9 3900X CPU (12 cores) @3.79 Ghz are displayed in Table 6.2, along with the total calculation time of the reference method.[24] The total calculation time of the proposed solution method (10 outer loop iterations) amounts 61 hours, of which 46% is spent in the lower and 54% is spent in the

[24] Note that the reference method does not use a solver, but a heuristic approach in the upper level, which explains why the number of solver iterations is left empty for this method.

upper level. Apart from the first iteration (in which the route set and mappings between OD-, route-, turn- and link level are generated), lower level calculation times show limited variation at around 2:45 hours per iteration. This is explained by realizing that most of it is spent during application of the SCCTA model (STAQ) whilst the number of STAQ iterations only varies between 12 and 14 iterations per (outer loop) iteration, translating to 12:19 up to 12:46 minutes per STAQ iteration. Upper level calculation times vary extensively between 45 minutes and 8 hours per (outer loop) iteration, translating to 1:18 up to 4:51 minutes per solver iteration. This means that not only the number of solver iterations, but also the calculation time per solver iteration varies extensively.

Both the upper level calculation times as well as the number of solver iterations indicate that most effort is put in the first three (outer loop) iterations which corresponds to the reductions of the objective function value per iteration and the amount of change in the OD matrix, which are largest in these first three iterations. As the default settings for the solver (Waltz et al., 2006) where used in this application, the number of solver iterations per (outer loop) iteration could probably be reduced and stabilized over iterations by tuning its parameters and stop criteria, but we leave this for future research.

Figure 6.9 shows that, compared to the minimum deviations from the reference method (dashed lines), the proposed method (solid lines) attains lower deviations on all objective function components from iteration three onwards. We therefore compare the calculation times of the reference method (19 hours) with the calculation time spent in the first three iterations of the proposed method (31 hours). This comparison shows that the proposed method spends 61% more time, but yields lower link flow and prior demand deviations and much lower route delay deviations, whilst performing equally on congestion pattern deviations.

Iteration#	# of STAQ iterations	calculation time lower level	# of solver iterations	calculation time upper level	total calculation time
1	12	03:15:12	100	08:04:42	11:19:54
2	14	02:52:26	100	07:43:52	10:36:18
3	13	02:43:45	100	06:30:09	09:13:54
4	13	02:42:54	35	01:44:18	04:27:12
5	13	02:42:54	71	04:01:08	06:44:02
6	13	02:45:07	46	02:11:37	04:56:44
7	13	02:46:04	35	01:02:40	03:48:44
8	13	02:44:35	35	00:45:30	03:30:05
9	13	02:43:48	35	00:45:36	03:29:24
10	13	02:54:04	0	00:00:00	02:54:04
Proposed method (10 iterations)	**130**	**28:10:49**	**557**	**32:49:32**	**61:00:21**
Proposed method (3 iterations)	**39**	**8:51:23**	**300**	**22:18:43**	**31:10:06**
Reference method (4 iterations)	**40**	**15:06:35**	**-**	**04:15:31**	**19:22:06**

Table 6.2 calculation times and related indicators from application of proposed and reference method on the BBMB model

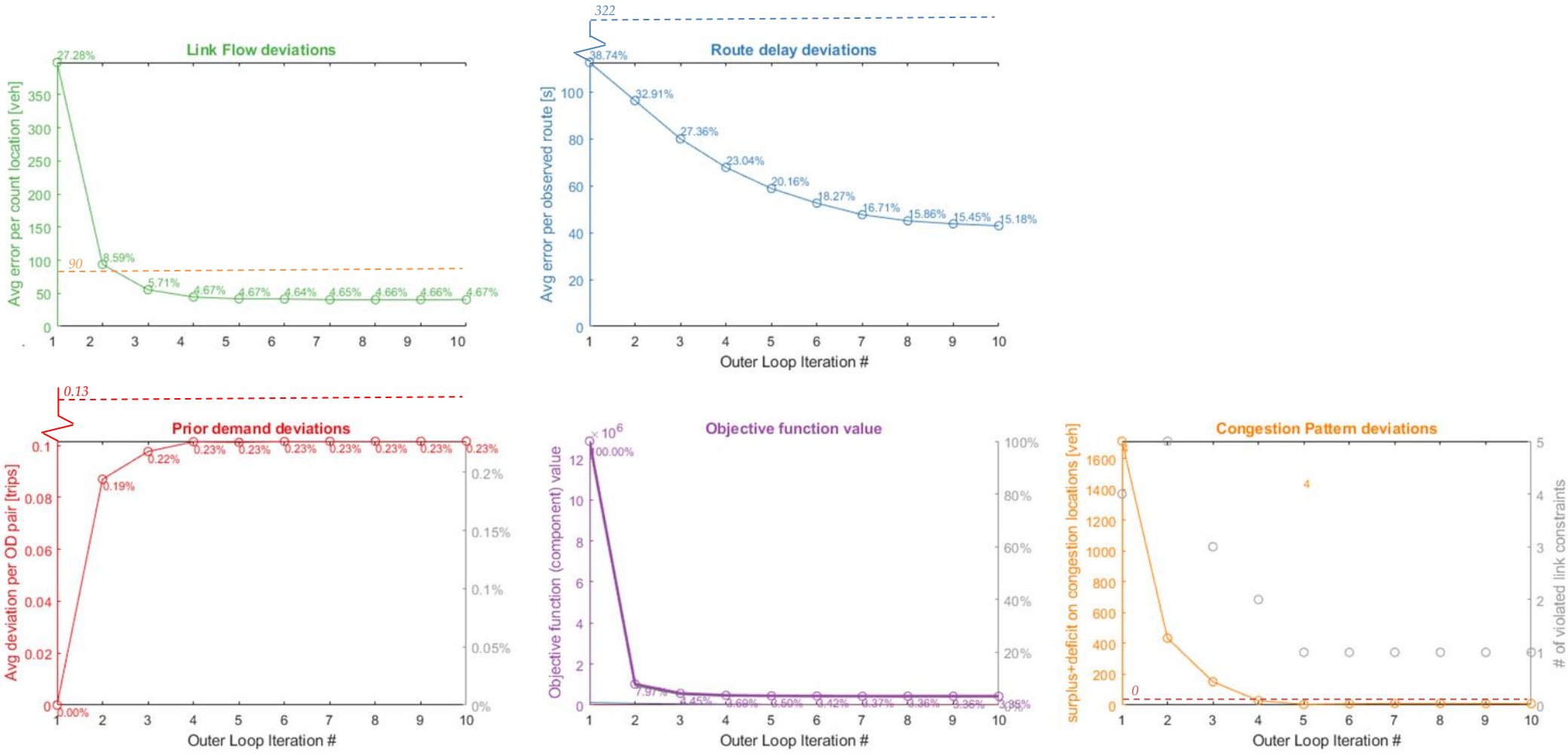


Figure 6.9: convergence of proposed methodology on the BBMB model in terms of average link flow deviations (upper left graph), route delay deviations (upper mid graph), prior demand deviations (lower left graph), congestion pattern deviations (lower right graph) and objective function value (lower mid graph). Dashed lines indicate minimum deviations yielded by the software currently used for matrix estimation in the BBMB, which employs the [REF] method described in 4.4

6.5.3 Findings on large network application

The most important finding from the application on the BBMB model is that the proposed solution method is indeed applicable to large scale transport models. The proposed method clearly outperforms the reference method and does so within feasible calculation times, but only because of the use of the following three problem size reducing features of the solution method.

Firstly, recall from subsection 6.4.5 that the added value of including sensitivities for observed link flows and link states proved very limited, whereas inclusion of sensitivities proved to be a requirement for observed route queuing delays. It is suspected that this is caused by the flow maximization property of the node model, which is one of the seven requirements for first order macroscopic node models (Tampère et al., 2011). This property causes that reductions in turn flow towards supply constrained outlinks (the source of all sensitivities) due to reduced demand are compensated for by increases of flow on other turns towards that outlink. This yields stable link flows on constrained outlinks, composed by unstable flows on turns towards the outlink. Therefore, being the only data source dependent of flows on turn instead of link level, observed queuing delays require inclusion of sensitivities, whereas other (link-level) data sources do not. This led to the insight that sensitivities for observed link flows and link states may be omitted altogether, which reduced the number of the required evaluations of equation (6.29) for application on the BBMB model by more than 95%.

Secondly, reducing the problem sizes for steps 2 through 6 in subsection 6.3.2, upper bounds on od demands are set to $\bar{\mathbf{D}} = \mathbf{2D_0}$ (subsection 6.5.1), reducing the number of links considered by 98.9%. Although this reduction proved sufficient for the application presented, the solution space may be widened, and/or the problem size number be further reduced by setting the upper bounds per OD pair allowing to combine absolute values and values relative to the prior demand.

Thirdly, not indicated earlier, the problem size in the upper level (steps 3 through 6 in subsection 6.3.2) is reduced by only including paths and OD pairs that use links with observed flow, state or queuing delay data. For the application on the BBMB model, this reduces the number of OD pairs considered from 1.58 million to 1.51 million (a reduction of 4%).

6.6 Conclusions, discussion and further research

In this paper, an efficient solution method for the matrix estimation problem is presented using a SCCTA model which combines the favorable properties of SCRTA and DCSTA models. The solution method allows for inclusion of route queuing delays and congestion patterns besides the traditional link flows and prior demand matrix, which is novel to the best of our knowledge. The proposed solution method uses response functions constructed from sensitivities on node level to solve a series of simplified optimization problems in the upper level, thereby avoiding costly additional assignment model runs of the lower level. Link state constraints are added to prevent usage of approximations outside their valid range as well as to include observed congestion patterns. The proposed solution method is robust, tractable and reliable because conditions under which a solution to the simplified optimization problem exist are known and because the problem is convex and has a smooth objective function.

Four test case applications on the small Sioux Falls model where conducted, each consisting of 100 runs with varied prior OD demands for robustness. These applications demonstrate the inclusion of observed congestion patterns and that adding sensitivities to the response function leads to more accurate results and slightly less computation time required.

Addition of queuing delays proved to be the biggest challenge, as these operate on the level of individual turning movements, and are therefore, contrary to link flows and link demands, not stabilized by the flow maximization property of the node model. This increases the likelihood

that discontinuities in $\boldsymbol{\alpha}_n(\mathbf{T}_n)$ are crossed during estimation. This caused 9/100 runs with queuing delays to not converge, but algorithmic enhancements (nudging and lowering the minimum capacity surpluses on specific link state constraints) resolved this problem, decreasing the number of non-converging runs to only 1.
The proposed solution method yields 99 converging runs whereas analysis shows that the single non-converging run is caused by a prior OD Demand matrix from which the SUE route choice probabilities cause incompatibility between two datapoints. In practice such inconsistencies can easily be detected and removed.
In addition to the Sioux Falls results, a large scale application on the BBMB model was conducted. Results show that when using its problem size reducing features, the solution method is indeed capable of solving large scale problems within feasible calculation time and while doing so, it attains lower deviations on all objective function components compared to the reference method.

6.6.1 Discussion

Although this paper shows the potential for SCCTA model in the context of travel demand matrix estimation, use of this type of assignment model is still very limited. Apart from STAQ (used in this paper), all SCCTA models of which the authors are aware of (Bakker et al., 1994; Bell, 1995; Bifulco and Chrisalli, 1998; Bundschuh et al., 2006; Köhler and Strehler, 2012; Lam and Zhang, 2000; Smith, 2012, 1987) are not directly suitable for two reasons. Firstly, they use link exit capacities that constrain flow through a link only at the downstream end of the link, thereby unrealistically modelling queues inside the bottleneck links contrary to upstream of the bottleneck. Secondly, all these models lack a node model satisfying the requirements posed by (Tampère et al., 2011). Fortunately, research on SCCTA models and its more advanced sibling SCSCTA models that adds storage constraints is still ongoing (Bliemer and Raadsen, 2019; Raadsen and Bliemer, 2019b).
The proposed solution approach is currently only applicable to SCCTA and SCRTA models as it solves a matrix estimation problem in which temporal correlations between model variables do not exist. Authors believe that the proposed solution approach would be extendible to the semi-dynamic capacity constrained case, where multiple time periods, each with its own stationary travel demand, are modelled and residual traffic is transferred in between (Bliemer et al., 2017) but do not believe that this approach is easily extendible to DTA context, as such models introduce temporal correlations.
Because the way in which the proposed solution derives and uses the assignment matrix ($\widehat{\boldsymbol{\alpha}}(\mathbf{D})$) and its sensitivity ($\partial\widehat{\boldsymbol{\alpha}}/\partial\mathbf{D}$) the approach is not dependent on conditions (if any) to which the assignment model results adhere. This was confirmed by test runs of the proposed method (not described in this paper for brevity) in which the SCCTA model was run for only one route choice iteration and test runs where the SCCTA model was run to equilibrium using the paired combinatorial instead of multinomial logit route choice model. This makes the approach suitable for application in a wide range of SCCTA model applications spanning from operational (disequilibrium) to strategic (equilibrium) and using any first order node model (see (Smits et al., 2015) for an overview of different node models proposed in literature).
The link state constraints in this paper effectively stabilize the solution method by maintaining the state of potential bottleneck links. However, constraints in the node model actually operate on the turn level, and any constraint state switch causes a discontinuity in $\boldsymbol{\alpha}_n(\mathbf{T}_n)$ (Brederode et al., 2014). This means that the link state constraints from subsection 6.2.3.3 are too simplistic as they do not specify the normative turning movement. However, refining link state constraints is not pursued any further for three reasons. Firstly, during development of the solution method, tests with constraints on turning movement have been conducted, showing that the turn

constraints in combination with fixed route choice probabilities constrain the simplified optimization problem too much, causing this version of the solution method to perform very poorly. Secondly, it has not proven to be a problem in any of the test case applications. Thirdly, deriving link state constraint values from observed data currently is already a challenge and deriving normative turning movements from some suitable data source would be even harder.
To increase robustness of the test results, an almost infinite number of additional test applications could be defined by introducing different inconsistencies and variations in the data by e.g. perturbing around the 'true' observed flow and travel time values, change the set of observed links, congestion patterns and routes, change the level of demand in $\mathbf{D}_{true}$, and vary the combinations of different observed types. This has not been described in this paper for two reasons. Firstly, authors feel that it is more useful to switch to true empirical data first (as done in section 6.5), to be sure that the correct range of input data is tested. Secondly, a multitude of test cases have been conducted in preparation of this paper, but due to limits to the length of a scientific paper only insights from those runs relevant to the four test case applications and the large network application have been used in subsections 6.4.5 and 6.5.3.
Subsection 6.2.1.2 stated that, although incorrect in theory, omitting sensitivities in response functions appears to not be a problem in practice, which seems to withhold practitioners from using methods that include sensitivities. Authors suspect that this is the case because the sensitivities are relatively very small compared to the direct effect of changing OD demand. This is especially the case when an SCRTA assignment model is used assuming SUE conditions, as in such models only route choice sensitivities exist (the lack of capacity constraints implies flow acceptance factors are non-existent) and traffic is spread out over routes the most, dampening any effects of changed route choice.
This same mechanism but then applied to flow acceptance factors is suspected to only partly explain why the added value of inclusion of sensitivities in the SCCTA context is very limited (compare results from testcases [LS+S] to [LS]). The other part of the explanation in SCCTA context was already discussed in subsection 6.5.3: the flow maximization property of the node model stabilizes link flows but does not stabilize turn flows. These hypotheses could be checked by comparing the effects of omitting sensitivities for the different types of observed data using an assignment model assuming stochastic user equilibrium conditions with an assignment model assuming all-or-nothing route choice behavior for both the SCRTA and SCCTA cases, but we leave this idea for further research.

6.6.2 Future research

In this section, recommendations for further research are described, in order of priority from the authors' point of view.
This paper shows that the proposed solution method converged in 99/100 test runs conducted on Sioux Falls, and that non-convergence occurs due to route choice probabilities being incompatible with some of the datapoints describing link flows and route queuing delays. This advocates for an extension to the solution method that accounts for sensitivities of the response function of route choice probabilities -just as the current method does for the response functions for link flows, queuing delays and congestion patterns- such that it can actively steer away from situations where route choice probabilities cause datapoints to become inconsistent.
Although not yet applicable to real sized transport model networks, authors believe that in time static models that take both capacity and storage constraints into account (Bliemer and Raadsen, 2020) will replace the role of SCCTA models in strategic transport model systems. Therefore, on the longer term, research into extension of the proposed solution algorithm to support storage constraints is desired.
The mSSIM performance indicator used for OD matrix comparison in this paper is known to be sensitive to the way it is averaged. Based on literature, in this paper, averaging per matrix

row was chosen. Other aggregates might reveal different insights. Given this sensitivity, use of another performance indicator for OD matrix comparison is recommended. The mean normalized Levenshtein distance as proposed by (Behara et al., 2020) seems a promising alternative.

A. APPENDIX: CONVEXITY

A.1 Convexity and uniqueness of first part of the objective function

To consider the first part of the objective function we look at problem (6.23) with $w_1 = 1, w_2 = 0, w_3 = 0$. To prove convexity of the first part of the objective function, we look at the Hessian (second partial derivatives to the OD demands), which is given by:

$$\frac{\partial}{\partial D_{rs'}}\frac{\partial f_1}{\partial D_{rs}} = \frac{\partial^2(\sum_{rs\in RS}(D_{rs} - D_{rs}^0)}{\partial D_{rs}\partial D_{rs'}} = \begin{cases} 0 & \forall rs' \neq rs \\ 2 & \forall rs' = rs \end{cases} \quad \forall rs', rs \in RS \tag{45}$$

Hence, in this case, the Hessian matrix $|RS|x\,|RS|$ has value two on all elements of its diagonal and zeroes in all other cells, which means it is positive definite. Therefore, the first part of the objective function is strictly convex and since all constraints are linear inequalities, and as such form a closed convex set, problem (6.23) always has a unique solution when $w_1 = 1$, $w_2 = 0$, $w_3 = 0$).

A.2 Convexity of second part of objective function

To consider the second part of the objective function we look at problem (6.23) with $w_1 = 0$, $w_2 = 1$, $w_3 = 0$. To prove convexity of the second part of the objective function, we look at the first order Taylor approximations of the link flows $y_l(\mathbf{D})$ which are linear functions of the form:

$$y_l(\mathbf{D}) = c_l + b_l^T D \text{ with } c_l \in \mathbb{R}, b_l \in \mathbb{R}^{|RS|}. \tag{46}$$

As such, $(y_{\boldsymbol{l}}(\mathbf{D}) - \tilde{y}_{\boldsymbol{l}})^2$ is a quadratic function:

$$(y_{\boldsymbol{l}}(\mathbf{D}) - \tilde{y}_{\boldsymbol{l}})^2 = (b_l^T D)^2 - 2b_l^T D(\tilde{y}_{\boldsymbol{l}} - c_l)^2 \tag{47}$$

With corresponding Hessian matrix $\nabla^2(b_l^T D)^2 = 2b_l b_l^T$ which is positive semidefinite. Indeed:

$$D^T b_l b_l^T D = (b_l^T D)^2 \geq 0 \; \forall D \in \mathbb{R}^{|RS|}, \tag{48}$$

which means that the second part of the objective function in (6.23) is convex. This means that there is no unique solution for problem (6.23) when $w_1 = 0$, $w_2 = 1$, $w_3 = 0$, but all local solutions are global minimizers.

A.3 Convexity of third part of objective function

To consider the third part of the objective function we look at problem (6.23) with $w_1 = 0$, $w_2 = 0$, $w_3 = 1$. To prove convexity of the third part of the objective function, we look at the first order Taylor approximations of the queuing delays on route level τ_p. Using the same reasoning as in A.2 it can be proven that its corresponding Hessian matrix is positive semidefinite which means that also the third part of the objective function in (6.23) is convex.

B. APPENDIX: DISCONTINUITIES IN FLOW ACCEPTANCE FACTOR FUNCTION

Because the node model adheres to strict link capacity constraints, a discontinuity in the relationship between OD demand and flow acceptance factors occurs whenever a change in demand causes a bottleneck to switch from an inactive to an active state or vice versa. To

illustrate this, consider the corridor network displayed in Figure 10 (top) where $C_1 = 3000$, $C_2 = 2000$ and $C_3 = 1000$ and the corresponding relations between demand on rs and flow acceptance factors (bottom left) and between demand on rs and link flows (bottom right). It is clearly visible that functions α_{23} and y_3 are discontinuous at $D_{rs} = 1000$ (link 3 switches state) whereas functions α_{12} and y_2 are discontinuous at $D_{rs} = 2000$ (link 2 switches state). The switch of state of link 2 at $D_{rs} = 2000$ causes a second discontinuity of $\alpha_{23}(D_{rs})$ due to the active bottleneck upstream, but the route based flow acceptance factor at link 3 ($\hat{\alpha}_p^3 = \alpha_{12}\alpha_{23}$) and thus y_3 do not have such a discontinuity.

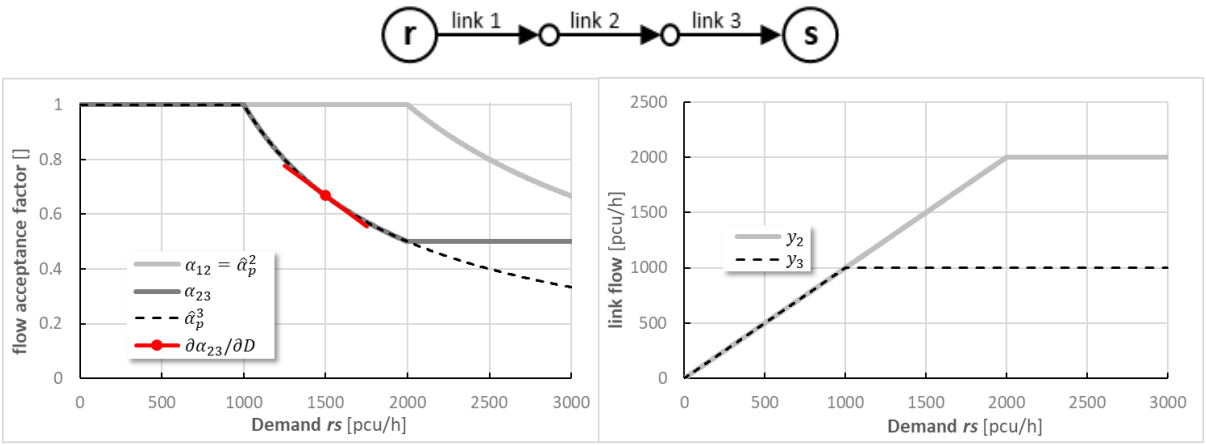


Figure 10: corridor network (top), flow acceptance factors (left) and link flows as function of demand (right)

From Figure 10 (right), it becomes apparent that elements in response function (6.4) become unresponsive to changes in demand whenever one or more upstream bottlenecks are active on the considered OD pair. This means that, in the upper level, these OD pairs cannot be used to directly influence flows downstream links, and as such become irrelevant. We therefore look at the sensitivity of link flows for different demand intervals in Table 6.3 and conclude that link 3 can only be influenced when $D_{rs} \leq 1000$, whereas link 2 can only be influenced when $D_{rs} \leq 2000$ and link 1 remains sensitive as long as $D_{rs} \leq 3000$.

Figure 10 (left) illustrates approximation for $\partial\alpha_{23}/\partial D_{rs}$ evaluated in point $D_{rs} = 1500$. Notice that because in this example and case $\alpha_{12} = 1$, the turn demand T_{23} is equal to OD demand D_{rs}, so translation from turn to route and OD level (subsection 6.3.2.3) is omitted in this example. Further note that the results in this example seem trivial, as they can be deduced by simply analysing the network, but this example is given as an introduction to the solution scheme described in subsection 6.3.2. A more complex case based on the numerical example described in (Tampère et al., 2011) is given in (Brederode et al., 2014).

Demand interval	**Flow acceptance factors**		**Sensitivity of link flows**		
D_{rs}	α_{23}	α_{12}	$\frac{\partial y_3(D_{rs})}{\partial D_{rs}}$	$\frac{\partial y_2(D_{rs})}{\partial D_{rs}}$	$\frac{\partial y_1(D_{rs})}{\partial D_{rs}}$
$D_{rs} \leq 1000$	1	1	>0	>0	>0
$1000 < D_{rs} \leq 2000$	($^1/_2$,1)	1	0	>0	>0
$2000 < D_{rs} \leq 3000$	½	($^2/_3$,1)	0	0	>0

Table 6.3: turn based flow acceptance factors and sensitivity of link flows on corridor network for different demand intervals

C. APPENDIX: APPROXIMATED SENSITIVITIES ON TURN LEVEL BREAKING CONVERGENCE

This appendix describes the mechanism that causes discontinuities in $\boldsymbol{\alpha}_n(\mathbf{T}_n)$ to be crossed during application of the proposed solution method. This may happen due to difference between the true function within the node model equation (6.8) and its linear approximate derived by

finite differences (see subsection 3.2.2) used in simplified optimization problem (6.23) applied in the upper level.
Continuing the numeric example from appendix B, but now the domain of interest is limited to $D_{rs} \leq C_2$. In this case $\alpha_{23}(\mathbf{T}_n) = \alpha_{23}(D_{rs})$ and the link state constraint (as defined in (6.18)) on link 3 may be written as:

$$\chi_3(D_{rs} - \delta_3 C_3) \leq 0. \tag{49}$$

The two graphs in Figure 10 display $\alpha_{23}(D_{rs})$ for the case where link 3 is not constraining (left; $D_{rs} = 900$) and constraining (right; $D_{rs} = 1100$), along with its linear sensitivity-approximations $\frac{\partial \alpha_{23}}{\partial D_{rs}}\Big|_{500}$ and $\frac{\partial \alpha_{23}}{\partial D_{rs}}\Big|_{1500}$ and link state constraints assuming $\delta_3 = 0.95$ (left) and $\delta_3 = 1.05$ (right).

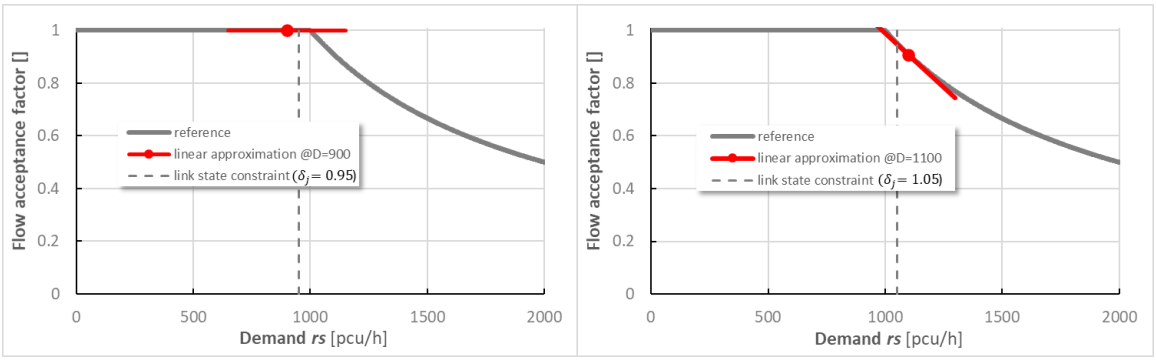


Figure 11: $\alpha_{23}(D_{rs})$, sensitivity approximations and link state constraints for not constraining (left) and constraining (right) cases in reference situation with effective turn capacity equals 1000 pcu/h

Consider a situation where this corridor network is just a part of a general network where demand on other OD pairs in the network has influence on the distribution of available supply of the outlinks of the node between links 2 and 3. Assume that the upper level solver changes the OD demand matrix such that the distribution of supply on the considered node in altered reducing the effective capacity of the turn from link 2 to link 3 to 750 pch/h. As displayed in Figure 12, this yields a shifted $\alpha_{23}(D_{rs})$ and different sensitivities, whereas the approximated sensitivities used in the simplified optimization problem are not updated. This means that in both cases the link state constraint is still satisfied, but

1. In the non-constraining case (Figure 12, left), the upper level uses an approximated sensitivity of zero, whereas it has become negative. This occurs in 6 runs of test case [+LS+S+QD] and breaks convergence. In the test case applications in subsection 6.4.4 this is prevented by lowering δ_j to 0.9 on the considered outlinks. In this theoretical example δ_j should be lowered to 0.75 or lower;
2. In the constraining case (Figure 12, right), the upper level assumes a slightly more negative sensitivity than its non-approximated counterpart. This probably occurs in some runs in some test cases in subsection 6.4.4, but does not break convergence, as the gradient information correctly assumes the sensitivity to remain smaller than zero (albeit that the approximate value is slightly off).

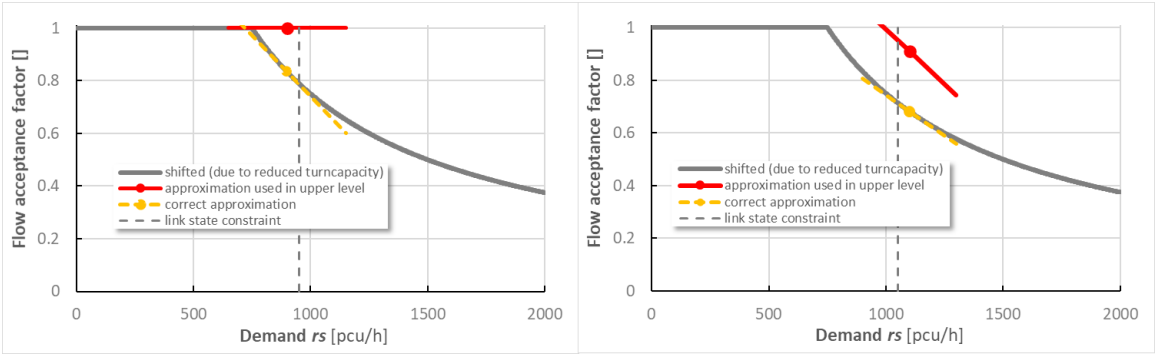


Figure 12: $\alpha_{23}(D_{rs})$, sensitivity approximations and link state constraints for not constraining (left) and constraining (right) cases when effective turn capacity is reduced to 750 pcu/h

Now assume that the upper level solver changes the OD demand matrix such that the distribution of supply on the considered node in altered increasing the effective capacity of the turn from link 2 to link 3 to 1250 pch/h[25]. As in cases 1 and 2, this yields a shifted $\alpha_{23}(D_{rs})$ whilst still satisfying the link state constraint, but

3. In the non-constraining case (Figure 13, left), the approximated sensitivity of zero used in the upper level remains correct. This probably occurs in some runs in some test cases in subsection 6.4.4, but it does not break convergence.
4. In the constraining case (Figure 13, right), the upper level assumes a negative sensitivity , whereas it has become zero. This apparently does not occur in the test cases in subsection 6.4.4, or at least not to the extent that it breaks convergence. However, this mechanism is likely to break convergence.

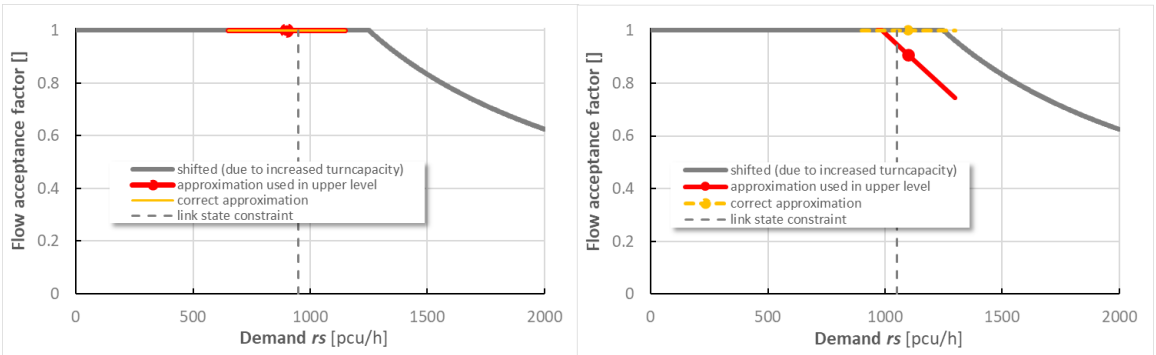


Figure 13: $\alpha_{23}(D_{rs})$, sensitivity approximations and link state constraints for not constraining (left) and constraining (right) cases when effective turn capacity is increased to 1250 pcu/h

Currently, it is unknown why the first case breaks convergence in the test case applications whereas the fourth test case does not (or does not occur). This is left for further research.

[25] This is not possible in the corridor network, because the turn capacity would be higher than the capacity of its outlink, but in a general network, the effective turn capacity can increase due to a change in the distribution of supply on a node, which is the mechanism that this example illustrates.

Chapter 7

Conclusions, implications and discussion

In this thesis, the development, implementation, testing and large scale applications of a new static and a new semi-dynamic capacity constrained TA model, as well as a travel demand estimation methodology that uses the SCCTA model are presented. Both TA models where developed to provide better accuracy in congested conditions compared to the SCRTA models that are the most widely used strategic TA models to date. The travel demand estimation methodology allows to include observed travel times, congestion patterns and observed flows affected by upstream bottlenecks (besides unaffected flows), which are all unique features in the large scale strategic context.
Besides these methodological contributions, this thesis also positions the two developed TA models and the travel demand estimation method in the field. For the TA models, a theoretical framework was developed that classifies all macroscopic first order TA models in terms of a genetic code with three genes and nine nucleotides consisting of four spatial, three temporal, and two behavioural assumptions. For demand estimation methods, a conceptual framework was developed that distinguishes four types of observed flow values based on how they are affected by active bottlenecks. Using this framework, three SCCTA model based solution strategies are identified and assessed by comparison to current SCRTA model based practice.
The remainder of this chapter draws conclusions on the SCCTA model (section 7.1), its semi-dynamic counterpart (section 7.5) and the travel demand estimation method (section 7.2) and concludes with implications for current and future strategic transport model systems (section 7.3) and a discussion (section 7.4).

7.1 Conclusions on the SCCTA model STAQ

Based on the comparison of the SCCTA model STAQ with STA and DTA models from Chapter 3 we conclude that STAQ possesses all of the desired properties for application on large scale strategic transport models with congested networks. Below, conclusions per model property are drawn, referring to the TA model related criteria from Table 1.2.
With respect to accountability (S7) and tractability (S8), it was shown that the different mechanisms that occur in a transportation network when applying STAQ can all be isolated and verified using only the law of flow conservation and the shape of the fundamental diagram, proving that tractability and accountability of STAQ are comparable to that of STA models and amply exceed that of DTA models.
With respect to model accuracy (R1), we conclude that, contrary to STA models, STAQ successfully detects and models flow metering and spillback effects of primary bottlenecks, but may overlook bottlenecks that are activated due to second-order and lane-distribution effects. STAQ allows for assignment of different vehicle classes and the junction modelling component allows application on both urban roads as well as motorways.
Based on analysis of twelve different model variations on seven large scale strategic transport models of largely congested regions (R2) we conclude that STAQ with spillback in the last iteration, full junction modelling and the self-regulating averaging scheme proved to be the optimal variation, providing sufficient realism while adhering to UE conditions (S1) within well acceptable calculation times (R3) ranging from 23 minutes up to 14 hours on a regular desktop pc (Core I7-950 3.07 Ghz). Being a macroscopic TA model (S3), STAQ does not contain

randomness (S2), thus satisfying all stability criteria. A limitation of the optimal model variant is that spillback effects are not included in the route choice behavior. It is possible to add these effects, but at the expense of convergence to UE conditions.
STAQ needs little extra input compared to STA models (S6), but its strict capacity constraints put emphasis on the required level of precision and accuracy of the input data. Most importantly, the definition of the study period and the level of stationary demand in the matrices should be consistent and the strict capacity constraints require more accurate capacity values on links and junctions to be coded as a single node.
Based on the above, we conclude that STAQ is a viable alternative to capacity restrained TA models, providing more accuracy whilst maintaining low complexity and sufficient stability. This makes the model suitable for applications where both static capacity restrained and dynamic TA models may fail: strategic applications on large-scale congested networks.

7.2 Conclusions on the travel demand estimation method

Chapter 6 presents an efficient solution method for the offline matrix estimation problem using STAQ (thus satisfying criterion S5 from Table 1.2). It allows for inclusion of observed route queuing delays and congestion patterns besides the traditional link flows and prior demand matrix, which is novel to the best of our knowledge. It thus satisfies criterion R4 from Table 1.2. It can also interpret and use observed flows affected by upstream bottlenecks, besides unaffected flows.
The proposed solution method uses response functions constructed from sensitivities on node level to solve a series of simplified optimization problems, thereby avoiding costly additional TA model runs. Link state constraints are added to prevent usage of approximations outside their valid range as well as to include observed congestion patterns. The proposed solution method is robust, tractable and reliable because conditions under which a solution to the simplified optimization problem exist are known and because the problem is convex and has a smooth objective function (S4). Although not described in Chapter 6 and not yet tested in practice, the gradient information that is explicitly calculated within the proposed travel demand estimator can be used to identify data inconsistencies (R5) by looking for OD pairs that have gradient values with a different sign for different datapoints.
Test case applications on the small Sioux Falls model demonstrated that the inclusion of observed congestion patterns and the addition of sensitivities to the response function leads to more accurate results and slightly less computation time. Addition of queuing delays proved to be the biggest challenge, as these operate on the level of individual turning movements, and are therefore, contrary to link flows and link demands, not stabilized by the flow maximization property of the node model. This increases the likelihood that discontinuities in the response functions are crossed during estimation. Algorithmic enhancements (nudging and lowering the minimum capacity surpluses on specific link state constraints) resolved this problem.
A large scale application on the large-scale strategic transport model of Noord-Brabant model demonstrated that using its problem size reducing features, the solution method is indeed capable of solving large scale problems within feasible calculation time and while doing so, it attains lower deviations on all objective function components compared to the reference method that used a (less advanced) SCCTA-based model based solution strategy in the conceptual framework from Chapter 5.

7.3 Implications

The combination of STAQ (Chapter 3) and the travel demand estimation method (Chapter 5) in their current form allows for deployment of a new generation of strategic transport model

sytems that, for the first time, are suited for application on large scale, structurally congested networks consisting of both highways and urban roads and for different road user classes, whilst maintaining the (low) complexity and (high) stability levels required for strategic applications. The improved accuracy of STAQ compared to SCRTA model outcomes causes large effects in terms of societal benefits of policy measures (subsection 3.4.2.3). This demonstrates that changing from capacity restrained TA models to capacity constrained TA models has substantial effects on the outcomes of a cost benefit analysis for networks with structural congestion.

Building upon this new generation of strategic transport model systems, an extension of the travel demand estimator to handle multiple time periods would be straightforward from a methodological point of view (Chapter 5). If such extension would be implemented and combined with the semi-dynamic version of STAQ (Chapter 4), it would yield a transport model system in which demand can be estimated using observed flows, delays and congestion patterns on any temporal aggregation level, ultimately leading to continuous (24 hour) estimation.

Besides providing the building blocks for new generations of strategic transport model systems, the software and methods developed in this thesis can also be used to incrementally improve current strategic transport model systems, as described below.

The presented SCCTA-based travel demand estimation method (Chapter 6) allows to directly use measured link flows, thereby removing the need to estimate link demands and use these as input (as required for SCRTA based estimation methods). This means that compared to SCRTA model based estimation methods, the presented method is much more transparent and less input sensitive, resulting in better tractability, comparability and transferability of the estimation process. This greatly reduces lead times for application of the estimation method, whilst delivering better accuracy (subsection 5.4.2 contains further elaboration on this topic).

Another potential incremental improvement to current strategic transport model systems is that the inclusion of observed queuing delays in the presented travel demand estimation method reduces the under-specification of the mathematical problem that it solves. This reduces differences between different application instances of the demand estimation method, thereby increasing consistency in estimation results for different instances. This ensures similar quality of results when e.g. updating the base year of a transport model system, or comparing base year outcomes of two different transport models systems.

Finally, the shift from the static capacity restrained to more advanced TA models has implications for policy makers as well. As shown in subsection 3.4.2.3, the addition of capacity constraints to the TA model (Chapter 3) has substantial effects on the outcomes of cost benefit analysis for study areas with structural congestion. On top of that, the relaxation of the empty network assumption that leads to the semi-dynamic version of STAQ (Chapter 4) causes large increases in collective losses compared to its static capacity constrained counterpart (up to 76% in peak periods; subsection 4.4.3.2). It is therefore very likely that the empty network assumption in static TA models (both capacity restrained and capacity constrained) influences (policy) decisions based upon queue size and delay related model outcomes on congested networks.

7.4 Discussion

This section discusses more subjective findings by the author on strategic TA models (subsection 7.4.1), strategic travel demand estimation (subsection 7.4.2) and the gap between scientific research on methodologies for strategic transport model systems (such as this thesis) and methods currently used by practitioners (subsection 7.4.3).

7.4.1 On traffic assignment models in strategic transport model systems

Based on the applications of the static and semi-dynamic TA models in subsections 3.4 and 4.4 the author concludes that semi-dynamic capacity constrained TA models are currently the most capable models (Figure 2.2) that still possess the stability and complexity properties required for the strategic application context (Table 1.2).

The author argues that the lack of stability in dynamic and/or storage constrained TA models is related to sensitivity of the cost function, summarized in its Jacobian, that contains sensitivities of route cost to changes in OD-demand for all route-OD-pair combinations that exist in the model system.

With respect to TA models with storage constraints, stability is lost due to diagonal indominance of the Jacobian, which means that cost of some routes are more sensitive to changes in demand on other OD-pairs than to changes in demand on the OD-pair that uses the considered route. Diagonal indominance causes the mathematical problem to be solved by storage constrained TA models to be non-convex (Dafermos, 1980), which means that it no longer has a unique solution. Because the averaging schemes employed here (MSA and SRA, subsection 3.2.4.3) were designed to solve the convex (deterministic and stochastic) UE problems described in (Beckmann et al., 1956) and (Fisk, 1980) they no longer work for non-convex storage constrained TA models. Alternative algorithms to enforce convergence for this type of problem do exist (Dafermos, 1980; Florian and Spiess, 1982; Lawphongpanich and Hearn, 1984), but unicity is lost and computational efficiency is relatively low. Note that diagonal indominance may also occur in TA models without storage constraints due to application of junction modelling (subsection 3.2.3.3) or even only the node model in capacity constrained TA models (subsection 3.2.3.2), but based upon the numerous applications conducted, this seems to be rarely the case in practice. This, in combination with the shift from focus from static to dynamic TA models might be the reason why research on these alternative algorithms seems to have stalled.

With respect to dynamic TA models, stability is lost due to their relatively sensitive (implicit) route cost function because these models employ short time periods from which all network conditions are transferred. As illustrated in the example in subsection 4.4.2, this causes much larger variablility in network conditions compared to static and semi-dynamic TA models that employ longer time periods and do not transfer traffic conditions or only transfer residual demand. It is expected that this problem can to some extent be solved by more enhanced algorithms (Brederode et al., 2016b describes a first attempt from the author), but at the cost of computational efficiency.

Note that almost all dynamic TA models used in practice also are storage constrained, hence suffering from both causes of instability. This yields insufficient stability and computational efficiency for application in the strategic context.

7.4.2 On matrix estimation methods for strategic transport model systems

When authorities decide to shift from a static capacity restrained to a static or semi-dynamic capacity constrained TA model, the assignment model no longer assumes that flow on all links is unaffected by active bottlenecks (subsection 5.1.2). The introduction of capacity constraints therefore requires a different solution method for travel demand estimation, as the relationship between OD demand and link flows is no longer strictly monotonic.

Current practice (subsection 5.2.2) is to maintain a matrix estimation method designed for capacity restrained TA models, but feed it with unconstrained link demand values estimated from the observed flows in a preprocessing step. The author strongly advices against this, as preprocessing methods assume general (exogenous) relationships between link demands and link flows that differ from the relationships calculated by the assignment model taking local (modelled) network conditions into account. This yields estimation results that fit well to the

preprocessed link demand estimates, but, after application of the capacity constrained TA model, fit poorly to the observed link flows.
To improve the fit, in practice, manual changes to the link demand estimates are conducted in an iterative fashion, causing high and uncertain lead times for projects including OD demand matrix estimation with only reasonable outcomes (subsection 5.4.2). Instead, the author advises to shift to a travel demand estimation method that makes use of the additional information which allows for direct estimation on observed link flows. Preferably, the estimation method from Chapter 6 (corresponding to method 3 from Chapter 5) is used, because compared to methods 1 and 2, it provides greater accuracy and faster convergence, removes the need to set a sensitive weight parameter on demands on bottleneck links and supports observed congestion patterns and travel times as additional input data types (section 5.5)[26].
Note that a shift to semi-dynamic capacity constrained TA models does not necessarily require a different travel demand estimation methodology, as long as travel demand is sequentially estimated for subsequent time periods. However, to include observed flows, delays and congestion patterns on any temporal aggregation level, (expected minor) extension of the method from Chapter 6 is required. This is left for further research.
As travel demand estimation using dynamic TA models involves spatio-temporal lag and assignment matrices with high dimensionality, authors argue that for such cases, research should focus on methods that do not rely on explicit assignment matrix and gradient calculation such as SPSA (e.g., Qurashi et al., 2020), the Kalman Filter (e.g., Castiglione et al., 2021) or the meta model approach from (Osorio, 2019b), hence steer away from the method presented in Chapter 6.

7.4.3 On bridging the gap between scientists and practitioners

All methodological advances presented in this thesis have been developed with practical applications in mind. This means that this thesis represents only part of the research output, the software implementations of STAQ, its semi-dynamic counterpart and the demand estimation method are an equally, if not more important result. At the time of writing, STAQ is included in OmniTRANS transport planning software and is used in eight different Dutch strategic transport model systems[27], while its semi-dynamic counterpart and the travel demand estimation method have already successfully been applied in pilots on full scale Dutch strategic transport model systems.
The author noted that during the last decades, besides steady advances on the route choice sub-model (Figure 2.1) for static capacity restrained TA models (Perederieieva et al., 2015 provides an overview), academic research has shown little development on the network loading submodel of strategic TA models. Instead, focus has been on dynamic storage and capacity constrained TA models and their application in the on-line / operational (and sometimes tactical) context. The author argues that this research may help to put more focus on development of the network loading submodels of strategic TA models for the following two reasons.

- Over the last decades, classification of macroscopic TA models has mainly been done using temporal interaction assumptions (static vs dynamic) and behavioral interaction assumptions (all or nothing vs equilibrium); see e.g. (Cantarella et al., 2019; Cascetta, 2009). In general, spatial interaction assumptions were only implicitly associated with temporal interaction assumptions: static implied a capacity restrained TA model whereas

[26] Note that subsection 5.4.5 mentions scalability issues which have already been resolved (as shown by the application in Chapter 6)
[27] The five strategic regional models in the province of Noord-Brabant, the strategic regional models of Overijssel and Arnhem/Nijmegen, and the tactical model of The Hague.

dynamic implied a capacity and storage constrained TA model. Because of this, practitioners and academics have been largely unaware of intermediate (conceptual) model classes, including the static and semi-dynamic capacity constrained TA models presented in Chapter 3 and Chapter 4.

- Given the stability required for strategic application and the lack of (known) alternatives described in previous bullet, accuracy of capacity restrained TA models might have been considered good enough. The author argues that this reasoning is no longer valid, as this thesis has shown that capacity constrained TA models improve accuracy on the effects of active bottlenecks and residual traffic (subsections 3.4.2 and 4.4.3 respectively), whilst maintaining stability. On top of that, Chapter 6 demonstrates that capacity constrained TA models allow to include data on observed queuing delays and congestion patterns into travel demand estimation.

The author argues that the frameworks presented in chapters 2 and 5 play an important role for adoption by practitioners of the methodological advances from chapters 3, 4 and 6. These frameworks can be used to aid practitioners to find a suitable modelling approach given their functional requirements, but also help to avoid confusion due to ambiguous (ab)use of terminology. Examples of such ambiguities are the use of the term quasi-dynamic (mostly used for static TA models with capacity and/or storage constraints, see subsection 2.2.2), the simplistic division of TA models mainly using only temporal interaction assumptions (first bullet in previous paragraph) and, less directly related to this research, the loose definitions of the different types of disaggregated and microscopic travel demand models (see e.g., Vovsha, 2019 and references therein).

Finally, the author acknowledges that the ongoing transition from the 'predict and provide' to the 'vision and validate' paradigm behind transport policy making reduces the importance of stability, as in the 'vision and validate' paradigm, uncertainty regarding the future of transport is seen as an opportunity for transport policymakers to play a part in shaping future society rather than a threat to the more reactive ('predict and provide') approach that responds to trends (Lyons and Davidson, 2016). This loosely corresponds to practitioners stating that the need for strategic TA models with sufficient accuracy and stability will remain, but focus of development should be on reduction of uncertainty in reference models (by use of more data driven modelling and more person- and household segments), whilst increasing computational efficiency to allow generation of large numbers of (uncertain) forecasts to sketch plausible policy paths within the vision and validate paradigm (Clerx, 2022; de Graaf, 2021; Hofman, 2018; van Vuren, n.d.). The author supports this transition and acknowledges the altered role of strategic transport model systems therein.

7.5 Conclusions on the semi-dynamic version of STAQ

The semi-dynamic version of STAQ presented in Chapter 4 is a straightforward extension of STAQ that effectively removes the empty network assumption, yielding a TA model that is more accurate than its static counterpart whilst still maintaining all properties discussed in section 7.1, except for the computational requirements (R3). Depending on the number and duration of the periods defined for the semi-dynamic TA model, calcultion times may exceed the criterion of 16 hours (R3).

To the best of the authors knowledge, Chapter 4 describes the only semi-dynamic TA model that places vertical queues at the correct location (on the upstream node of the link affected by capacity constraint(s)) and also removes flow downstream from bottlenecks as part of the assignment model. The solution algorithm consists of STAQ, set in a loop with a residual traffic transfer module. Collective losses and average delays on network, route and link level from the network operator's perspective (quantifying delay within a time period) and the traveler's

perspective (quantifying delay within a *departure* time period) are determined from cumulative in- and outflow curves as a post processing module.
Comparison of the model outcomes to STAQ and its closest dynamic counterpart (Bliemer and Raadsen, 2019) shows that the size and temporal distribution of queues and collective losses from the semi-dynamic and dynamic TA models are very similar, but that the spatial distribution is different as the former model ignores spillback. Furthermore, it shows that the static version of STAQ does not resemble the other two models on size, temporal nor spatial distribution of queues and collective losses.
With respect to model stability, the comparison showed that stability is maintained from STAQ to its semi-dynamic version (reaching the required 1E-04 duality gap threshold), whereas it is broken for the dynamic TA model (it does not reach the required duality gap threshold).
With respect to model scalability, the semi-dynamic TA model in its current (prototypical) form requires on average 51% more calculation time in time periods with queues, predominantly due to calculation time spent by the traffic transfer module. However, it is expected that the additional calculation time for the residual traffic transfer module could easily be reduced to less than 10% when its implementation would be merged with the assignment model code into a single code base. Authors argue that the additional calculation time is a worthwhile inconvenience to bear, given the substantial amount of collective loss being omitted by the static version of STAQ due to its empty network assumption.

Acknowledgements

This research has been largely funded by Goudappel transport and mobility consultants, The Netherlands. The research in Chapter 2 was supported by the Australian Research Council [grant number LP130101048] and the application of the presented travel demand estimation method on the large network in section 6.5 was partially funded by the Dutch province of Noord-Brabant. Furthermore, the strategic transport models used in Chapter 3 and Chapter 4 where kindly provided by the province of Noord-Brabant, the Dutch municipality of Den Haag and the Flemish traffic centre in cooperation with Goudappel.

References

Abrahamsson, T., 1998. *Estimation of origin-destination matrices using traffic counts - a literature survey* (IIASA Interim Report No. IR-98-021). International Institute for Applied Systems Analysis, Laxenburg, Austria. Available at: https://pure.iiasa.ac.at/5627 (accessed 7.27.23).

Akamatsu, T., Makino, Y., Takahashi, E., 1998. 'Semi-dynamic Traffic Assignment Models with Queue Evolution and Elastic OD Demand'. *Infrastructure Planning Review* 15, pp.535–545. https://doi.org/10.2208/journalip.15.535

Akcelik, R., 1991. 'Travel time functions for transport planning purposes: Davidson's function, its time dependent form and alternative travel time function'. *Australian Road Research* 21(3), pp.49–59.

Alpcan, T., 2013. 'A framework for optimization under limited information'. *J Glob Optim* 55(3), pp.681–706. https://doi.org/10.1007/s10898-012-9942-z

Antoniou, C., Azevedo, C.L., Lu, L., Pereira, F., Ben-Akiva, M., 2015. 'W–SPSA in Practice: Approximation of Weight Matrices and Calibration of Traffic Simulation Models'. *Transportation Research Procedia* 7, pp.233–253. https://doi.org/10.1016/j.trpro.2015.06.013

Awan, J., Solomon, N. (Eds.), 2000. *Highway capacity manual*. Transportation Research Board, National Research Council, Washington, D.C.

Bakker, D.M., Mijjer, P.H., Daly, A.J., Vrolijk, P.C., Hofman, F., 1994. 'Prediction and evaluation of the effects of traffic management measures on congestion and vehicle queues'*, in: Proceedings of PTRC Summer Annual Meeting*. Warwick, England.

Beckmann, M., McGuire, C.B., Winsten, C.B., 1956. *STUDIES IN THE ECONOMICS OF TRANSPORTATION*. Yale University Press, New Haven.

Behara, K.N.S., Bhaskar, A., Chung, E., 2020. 'A novel approach for the structural comparison of origin-destination matrices: Levenshtein distance'. *Transportation Research Part C: Emerging Technologies* 111, pp.513–530. https://doi.org/10.1016/j.trc.2020.01.005

Bell, M.G.H., 1995. 'Stochastic user equilibrium assignment in networks with queues'. *Transportation Research Part B: Methodological* 29(2), pp.125–137. https://doi.org/10.1016/0191-2615(94)00030-4

Bell, M.G.H., Lam, H.K., Lida, Y., 1996. 'A Time-Dependent Multi-Class Path Flow Estimator'*, in: Proceedings of the 13th International Symposium on Transportation and Traffic Theory*. Lyon, France.

Ben-Akiva, M., Bierlaire, M., 1999. 'Discrete Choice Methods and their Applications to Short Term Travel Decisions'*, in: Hall, R.W. (Ed.), Handbook of Transportation Science, International Series in Operations Research & Management Science*. Springer US, Boston, MA, pp. 5–33. https://doi.org/10.1007/978-1-4615-5203-1_2

Bezembinder, E., 2021. *Junction design rules* (PhD thesis). University of Twente, Enschede, The Netherlands. https://doi.org/10.3990/1.9789036552448

Bezembinder, E.M., Brandt, F., 2016. *Junction Modelling in OmniTRANS* (appendix to the OmniTRANS manual). DAT.Mobility, Deventer.

Bezembinder, E.M., Wismans, L.J.J., van Berkum, E.C., 2015. 'Using decision trees to determine junction design rules'*, in: Proceedings of the 94th Annual Meeting of the Transportation Board, 11-15 January 2015, Washington DC, USA. (on CD ROM)*. Transportation Research Board (TRB), pp. 1–16.

Bifulco, G., Chrisalli, U., 1998. 'Stochastic user equilibrium and link capacity constraints: Formulation and theoretical evidences. Paper presented at the proceedings of the European transport conference, Loughborough, UK, pp. 85–96.'*, in: Proceedings of the European Transport Conference*. Loughborough, UK, pp. 85–96.

Bliemer, M., Raadsen, M., 2020. 'Static traffic assignment with residual queues and spillback'. *Transportation Research Part B: Methodological, 23rd International Symposium on Transportation and Traffic Theory (ISTTT 23)* 132, pp.303–319. https://doi.org/10.1016/j.trb.2019.02.010

Bliemer, M., Raadsen, M., 2019. 'Continuous-time general link transmission model with simplified fanning, Part I: Theory and link model formulation'. *Transportation Research Part B: Methodological* 126, pp.442–470. https://doi.org/10.1016/j.trb.2018.01.001

Bliemer, M., Raadsen, M., Brederode, L., Bell, M., Wismans, L., Smith, M., 2017. 'Genetics of traffic assignment models for strategic transport planning'. *Transport Reviews* 37(1), pp.56–78. https://doi.org/10.1080/01441647.2016.1207211

Bliemer, M., Raadsen, M., De Romph, E., Smits, E.-S., 2013. 'Requirements for traffic assignment models for strategic transport planning: A critical assessment', *in: Proceedings of the 36th Australasian Transport Research Forum*. Presented at the 36th Australasian Transport Research Forum, Institute of Transport and Logistics Studies, ITLS, University of Sydney, Brisbane, Australia.

Bliemer, M., Raadsen, M., Smits, E.-S., Zhou, B., Bell, M., 2014. 'Quasi-dynamic traffic assignment with residual point queues incorporating a first order node model'. *Transportation Research Part B: Methodological* 68, pp.363–384. https://doi.org/10.1016/j.trb.2014.07.001

Bliemer, M.C.J., 2007. 'Dynamic Queuing and Spillback in Analytical Multiclass Dynamic Network Loading Model'. *Transportation Research Record* 2029(1), pp.14–21. https://doi.org/10.3141/2029-02

Bliemer, M.C.J., Bovy, P.H.L., 2003. 'Quasi-variational inequality formulation of the multiclass dynamic traffic assignment problem'. *Transportation Research Part B: Methodological* 37(6), pp.501–519. https://doi.org/10.1016/S0191-2615(02)00025-5

Bliemer, M.C.J., Brederode, L.J.N., Wismans, L.J.J., Smits, E.S., 2012. 'Quasi-dynamic network loading: adding queuing and spillback to static traffic assignment', *in: Proceedings of the 91st Transportation Research Board (TRB) Annual Meeting*. Presented at the 91st Transportation Research Board (TRB) Annual Meeting 2012, Washington, D.C.

Blum, J.R., 1954. 'Approximation Methods which Converge with Probability one'. *The Annals of Mathematical Statistics* 25(2), pp.382–386. https://doi.org/10.1214/aoms/1177728794

Bovy, P.H., 1991. 'Zusammenfassung des Schweizerischen Kreiselhandbuchs'. *Straße und Verkehr* 3, pp.129–139.

Bovy, P.H.L., 1990. *Toedeling van verkeer in congestievrije netwerken* (PhD thesis). Delft.

Boyce, D., Ralevic-Dekic, B., Bar-Gera, H., 2004. 'Convergence of Traffic Assignments: How Much is Enough?'. *Journal of Transportation Engineering* 130(1), pp.49–55. https://doi.org/10.1061/(ASCE)0733-947X(2004)130:1(49)

Branston, D., 1976. 'Link capacity functions: A review'. *Transportation Research* 10(4), pp.223–236. https://doi.org/10.1016/0041-1647(76)90055-1

Brederode, L., Bliemer, M., Wismans, L., 2010. 'STAQ: Static Traffic Assignment with Queing', *in: Proceedings of the European Transport Conference*. Presented at the European Transport Conference, Glasgow, UK.

Brederode, L., Hardt, T., Rijksen, B., 2020. 'Development of a microscopic tour based demand model without statistical noise'. Presented at the European Transport Conference, Dublin.

Brederode, L., Heynicks, M., Koopal, R., 2016a. 'Quasi Dynamic Assignment on the Large Scale Congested Network of Noord-Brabant', *in: Proceedings of the 44th European Transport Conference*. Presented at the European transport conference, AET and contributors, Barcelona, p. 17.

Brederode, L., Hofman, F., van Grol, R., 2017. 'Testing of a demand matrix estimation method incorporating observed speeds and congestion patterns on the Dutch strategic model system using an assignment model with hard capacity constraints', *in: Proceedings of the 45th*

European Transport Conference 2017. Presented at the 45th European Transport Conference, Barcelona.

Brederode, L., Pel, A., Wismans, L., de Romph, E., Hoogendoorn, S., 2018. 'Static Traffic Assignment with Queuing: model properties and applications'*. Transportmetrica A: Transport Science* pp.1–36. https://doi.org/10.1080/23249935.2018.1453561

Brederode, L., Pel, A.J., Hoogendoorn, S.P., 2014. 'Matrix estimation for static traffic assignment models with queuing'*, in: Proceedings of HEART 2014 - 3rd Symposium of the European Association for Research of Transportation*. Leeds UK.

Brederode, L., Pel, A.J., Wismans, L., de Romph, E., 2016b. 'Improving convergence of quasi dynamic assignment models'*, in: Proceedings of the 6th International Symposium on Dynamic Traffic Assignment*. Presented at the 6th International Symposium on Dynamic Traffic Assignment, Sydney, Australia.

Brederode, L., Pel, A.J., Wismans, L., de Romph, E., Hoogendoorn, S.P., 2019. 'Static Traffic Assignment with Queuing: model properties and applications'*. Transportmetrica A: Transport Science* 15(2), pp.179–214. https://doi.org/10.1080/23249935.2018.1453561

Brederode, L., Pel, A.J., Wismans, L., Rijksen, B., Hoogendoorn, S.P., 2023. 'Travel demand matrix estimation for strategic road traffic assignment models with strict capacity constraints and residual queues'*. Transportation Research Part B: Methodological* 167, pp.1–31. https://doi.org/10.1016/j.trb.2022.11.006

Brederode, L., Verlinden, K., 2019. 'Travel demand matrix estimation methods integrating the full richness of observed traffic flow data from congested networks'. *Transportation Research Procedia, Modeling and Assessing Future Mobility ScenariosSelected Proceedings of the 46th European Transport Conference 2018, ETC 2018* 42, pp.19–31. https://doi.org/10.1016/j.trpro.2019.12.003

Bui, T.T., Nakayama, S., Yamaguchi, H., Koike, K., 2019. 'Link-Based Approach for Semi-Dynamic Stochastic User Equilibrium Traffic Assignment with Sensitivity Analysis Model'*. Journal of Japan Society of Civil Engineers, Ser. D3 (Infrastructure Planning and Management)* 75(5), pp.I_615-I_625. https://doi.org/10.2208/jscejipm.75.I_615

Buisson, C., Lebacque, J.P., Lesort, J.B., 1999. 'Travel Times Computation for Dynamic Assignment Modelling'*, in: Transportation Networks: Recent Methodological Advances*. Elsevier, pp. 303–317. https://doi.org/10.1016/B978-008043052-2/50019-6

Bundschuh, M., Vortisch, P., Vuren, T.V., 2006. 'Modelling queues in static traffic assignment'*, in: Proceedings of the European Transport Conference (ETC) 2006*. Presented at the European Transport Conference, Association for European Transport (AET), Strasbourg, France.

Bureau of Public Roads, 1964. *Traffic assignment manual for application with a large, high speed computer.* U.S. Dept. of Commerce, Bureau of Public Roads, Office of Planning, Urban Planning Division; for sale by the Superintendent of Documents, U.S. Govt. Print. Off., Washington.

Caliper, 2010. *What Transcad Users Should Know About Traffic Assignment.* (Caliper Transportation Publication). Available at: https://pdfs.semanticscholar.org/b473/3cbdd5bf3d9cc216f34335c69c7e8e10baaf.pdf

Caliper Publications [WWW Document], n.d. URL https://www.caliper.com/press/transportationlibrary.htm (accessed 7.25.23).

Cantarella, G.E., Watling, D., Luca, S. de, Pace, R.D., 2019. *Dynamics and Stochasticity in Transportation Systems: Tools for Transportation Network Modelling*. Elsevier.

Cantelmo, G., Viti, F., Cipriani, E., Nigro, M., 2017. 'A Utility-based Dynamic Demand Estimation Model that Explicitly Accounts for Activity Scheduling and Duration.'*. Transportation Research Procedia* 23, pp.440–459. https://doi.org/10.1016/j.trpro.2017.05.025

Canudas-de-Wit, C., Ferrara, A., 2018. 'A variable-length Cell Transmission Model for road traffic systems'. *Transportation Research Part C: Emerging Technologies* 97, pp.428–455. https://doi.org/10.1016/j.trc.2018.07.023

Casas, J., de Villa, A.X., Breen, M., Perarnau, J., Delgado, M., Torday, A., 2015. 'Quasi-dynamic model in Aimsun compared to static and dynamic models'*, in: Australasian Transport Research Forum 2015 Proceedings*. Presented at the Australasian Transport Research Forum (ATRF), Sydney, Australia.

Cascetta, E., 2009. *Transportation Systems Analysis, Springer Optimization and Its Applications*. Springer US, Boston, MA. https://doi.org/10.1007/978-0-387-75857-2

Cascetta, E., Nuzzolo, A., Russo, F., Vitetta, A., 1996. 'A Modified Logit Route Choice Model Overcoming Path Overlapping Problems. Specification and some Calibration Results for Interurban Networks'*, in: Proceedings of the 13th International Symposium on Transportation and Traffic Theory*. Presented at the 13th international symposium on transportation and traffic theory (ISTTT), Lyon, France, pp. 697–711.

Castiglione, J., 2015. *Activity-based travel demand models: a primer*. Transportation Research Board, Washington, DC.

Castiglione, M., Cantelmo, G., Qurashi, M., Nigro, M., Antoniou, C., 2021. 'Assignment Matrix Free Algorithms for On-line Estimation of Dynamic Origin-Destination Matrices'*. Front. Future Transp.* 2. https://doi.org/10.3389/ffutr.2021.640570

Chakirov, A., Fourie, P.J., 2014. 'Enriched Sioux Falls scenario with dynamic and disaggregate demand' pp.39 p. https://doi.org/10.3929/ETHZ-B-000080996

Chan, C., Kuncheria, A., Zhao, B., Cabannes, T., Keimer, A., Wang, B., Bayen, A., Macfarlane, J., 2021. 'Quasi-Dynamic Traffic Assignment using High Performance Computing'*. arXiv:2104.12911 [cs]*.

Chiu, Y.-C., Bottom, J., Mahut, M., Paz, A., Balakrishna, R., Waller, T., Hicks, J., 2011. *Dynamic Traffic Assignment* (No. E-C153)*, Transportation Research Circular*.

Chu, C., 1989. 'A Paired Combinatorial Logit Model For Travel Demand Analysis'*, in: Selected Proceedings Of The Fifth World Conference On Transport Research*. Presented at the World Conference On Transport Research (WCTRS), Yokohama, pp. 295–309.

Cipriani, E., A. Gemma, M. Nigro, 2013. 'A bi-level gradient approximation method for dynamic traffic demand estimation: Sensitivity analysis and adaptive approach'*, in: 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*. Presented at the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), pp. 2100–2105. https://doi.org/10.1109/ITSC.2013.6728539

Cipriani, E., Florian, M., Mahut, M., Nigro, M., 2011. 'A gradient approximation approach for adjusting temporal origin–destination matrices'*. Transportation Research Part C: Emerging Technologies* 19(2), pp.270–282. https://doi.org/10.1016/j.trc.2010.05.013

Clerx, W., 2022. 'Strategische verkeersmodellen en de mobiliteitstransitie – NM Magazine'*. NM Magazine* 17(2), pp.40.

Correa, J.R., Schulz, A.S., Stier-Moses, N.E., 2004. 'Selfish Routing in Capacitated Networks'*. Mathematics of OR* 29(4), pp.961–976. https://doi.org/10.1287/moor.1040.0098

Dafermos, S., 1980. 'Traffic Equilibrium and Variational Inequalities'*. Transportation Science* 14(1), pp.42–54. https://doi.org/10.1287/trsc.14.1.42

Daganzo, C.F., 1995. 'The cell transmission model, part II: Network traffic'*. Transportation Research Part B: Methodological* 29(2), pp.79–93. https://doi.org/10.1016/0191-2615(94)00022-R

Daganzo, C.F., 1994. 'The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory'*. Transportation Research Part B: Methodological* 28(4), pp.269–287. https://doi.org/10.1016/0191-2615(94)90002-7

Daganzo, C.F., 1977. 'On the traffic assignment problem with flow dependent costs—II'*. Transportation Research* 11(6), pp.439–441. https://doi.org/10.1016/0041-1647(77)90010-7

Daganzo, C.F., Sheffi, Y., 1977. 'On Stochastic Models of Traffic Assignment'*. Transportation Science* 11(3), pp.253–274. https://doi.org/10.1287/trsc.11.3.253

Davidson, K.B., 1966. 'A flow travel time relationship for use in transportation planning'. Presented at the Australian Road Research Board (ARRB) Conference, 3rd, 1966, Sydney.

Davidson, P., Thomas, A., Teye-Ali, C., Clarke, P., Shanin, M., 2011. 'Clocktime assignment: a new mesoscopic junction delay highway assignment approach to continuously assign traffic over the whole day'. Presented at the European Transport Conference, AET, Glasgow, UK, p. 12.

de Graaf, S., 2021. *Transities in verkeersmodellen* (Whitepaper). Goudappel, Deventer. Available at: https://www.goudappel.nl/sites/default/files/2022-01/Whitepaper%20Transities%20in%20verkeersmodellen.pdf (accessed 7.27.23).

Di, X., Liu, H.X., Pang, J.-S., Ban, X. (Jeff), 2013. 'Boundedly rational user equilibria (BRUE): Mathematical formulation and solution sets'. *Transportation Research Part B: Methodological* 57, pp.300–313. https://doi.org/10.1016/j.trb.2013.06.008

Djukic, T., Hoogendoorn, S., Van Lint, H., 2013. 'Reliability Assessment of Dynamic OD Estimation Methods Based on Structural Similarity Index'. Presented at the Transportation Research Board 92nd Annual MeetingTransportation Research Board.

Djukic, T., Masip, D., Breen, M., Perarnau, J., Casas, J., 2017. 'Modified bi-level framework for dynamic OD demand estimation in the congested networks'. Presented at the the 97th Annual Meeting of the Transportation Research Board, Washington, D.C.

Engelson, L., Vyas, G., Vovsha, P., 2022. 'Systematic investigation of microsimulation variability of the Swedish national travel model with different methods for handling stochasticity'. Presented at the European Transport Conference, Milan.

Erhardt, G.D., Hoque, J., Chen, M., Souleyrette, R., Schmitt, D., Chaudhary, A., Rapolu, S., Kim, K., Weller, S., Sall, E., Wachs, M., National Cooperative Highway Research Program, Transportation Research Board, National Academies of Sciences, Engineering, and Medicine, 2020. *Traffic Forecasting Accuracy Assessment Research*. Transportation Research Board, Washington, D.C. https://doi.org/10.17226/25637

Fafieanie, M.E., 2009. *Calibrating route set generation by map matching GPS data* (Master Thesis). Twente University, Enschede.

Filippi, F., 2022. 'A Paradigm Shift for a Transition to Sustainable Urban Transport'. *Sustainability (Switzerland)* 14(5). https://doi.org/10.3390/su14052853

Fiorenzo-Catalano, M.S., 2007. *Choice set generation in multi-modal transportation networks* (PhD thesis). Delft University of Technology, Delft.

Fisk, C., 1980. 'Some developments in equilibrium traffic assignment'. *Transportation Research Part B: Methodological* 14(3), pp.243–255. https://doi.org/10.1016/0191-2615(80)90004-1

Florian, M., Spiess, H., 1982. 'The convergence of diagonalization algorithms for asymmetric network equilibrium problems'. *Transportation Research Part B: Methodological* 16(6), pp.477–483. https://doi.org/10.1016/0191-2615(82)90007-8

Flötteröd, G., Flügel, S., 2015. 'Traffic assignment for strategic urban transport model systems'. Presented at the ITEA Conference, Oslo, p. 18.

Flötteröd, G., Rohde, J., 2011. 'Operational macroscopic modeling of complex urban road intersections'. *Transportation Research Part B: Methodological* 45(6), pp.903–922. https://doi.org/10.1016/j.trb.2011.04.001

Flügel, S., Flötteröd, G., Kwong, C.K., Steinsland, C., 2014. 'Evaluation of methods for calculating traffic assignment and travel times in congested urban areas with strategic transport models'. *TØI report* 1358, pp.2014.

Frederix, R., 2012. *Dynamic origin-destination matrix estimation in large-scale congested networks.* (PhD thesis). K.U.Leuven. Faculteit Ingenieurswetenschappen, Leuven.

Frederix, R., Viti, F., Tampère, C.M.J., 2013. 'Dynamic origin–destination estimation in congested networks: theoretical findings and implications in practice'. *Transportmetrica A: Transport Science* 9(6), pp.494–513. https://doi.org/10.1080/18128602.2011.619587

Friesz, T.L., Han, K., 2019. 'The mathematical foundations of dynamic user equilibrium'. *Transportation Research Part B: Methodological* 126, pp.309–328. https://doi.org/10.1016/j.trb.2018.08.015

Friesz, T.L., Han, K., Neto, P.A., Meimand, A., Yao, T., 2013. 'Dynamic user equilibrium based on a hydrodynamic model'. *Transportation Research Part B: Methodological* 47, pp.102–126. https://doi.org/10.1016/j.trb.2012.10.001

Fujita, M., Matsui, H., Mizokami, S., 1988. 'Modelling of the Time-of-Day Traffic Assignment Over a Traffic Network'. *Doboku Gakkai Ronbunshu* 1988(389), pp.111–119. https://doi.org/10.2208/jscej.1988.111

Fujita, M., Yamamoto, K., Matsui, H., 1989. 'Modelling of the Time-of-Day Traffic Assignment Over a Congested Network'. *Doboku Gakkai Ronbunshu* 1989(407), pp.129–138. https://doi.org/10.2208/jscej.1989.407_129

Fusco, G., Colombaroni, C., Lo Sardo, S., 2013. 'A Quasi-Dynamic Traffic Assignment Model for Large Congested Urban Road Networks'. *International Journal of Mathematical Models and Methods in Applied Sciences* 7, pp.63–74.

Ge, Q., Fukuda, D., Han, K., Song, W., 2020. 'Reservoir-based surrogate modeling of dynamic user equilibrium'. *Transportation Research Part C: Emerging Technologies* 113, pp.350–369. https://doi.org/10.1016/j.trc.2019.10.010

Gentile, G., 2010. 'The General Link Transmission Model for Dynamic Network Loading and a Comparison with the DUE Algorithm'*, in: New Developments in Transport Planning: Advances in Dynamic Traffic Assignment*. Edward Elgar Publishing, pp. 153–178.

Gentile, G., Velonà, P., Cantarella, G.E., 2015. 'Uniqueness of stochastic user equilibrium with asymmetric volume-delay functions for merging and diversion'. *EURO Journal on Transportation and Logistics* 3(3), pp.309–331. https://doi.org/10.1007/s13676-013-0042-0

Gibb, J., 2011. 'Model of Traffic Flow Capacity Constraint through Nodes for Dynamic Network Loading with Queue Spillback'. *Transportation Research Record* 2263(1), pp.113–122. https://doi.org/10.3141/2263-13

Givoni, M., Perl, A., 2020. 'Rethinking Transport Infrastructure Planning to Extend Its Value over Time'. *Journal of Planning Education and Research* 40(1), pp.82–91. https://doi.org/10.1177/0739456X17741196

Gordon, A., Lalanne-Tauzia, X., 2020. 'The Broken Algorithm That Poisoned American Transportation'. *Vice*. URL https://www.vice.com/en/article/v7gxy9/the-broken-algorithm-that-poisoned-american-transportation-v27n3 (accessed 7.28.23).

Greenshields, B.D., 1935. 'A study of traffic capacity'. *Highway research board proceedings* 14(1), pp.448–477.

Hall, M.D., van Vliet, D., Willumsen, L., 1980. 'SATURN—a simulation-assignment model for the evaluation of traffic management schemes'. *Traffic Engineering and Control* 21, pp.168–176.

Han, K., Friesz, T.L., Szeto, W.Y., Liu, H., 2015. 'Elastic demand dynamic network user equilibrium: Formulation, existence and computation'. *Transportation Research Part B: Methodological* 81, pp.183–209. https://doi.org/10.1016/j.trb.2015.07.008

Heynickx, M., Koopal, R., Zantema, K., 2016. 'The approach of traffic modelling in Noord-Brabant'. Presented at the European Transport Conference, Barcelona.

Himpe, W., Corthout, R., Tampère, C., 2016. 'An efficient iterative link transmission model'. *Transportation Research Part B: Methodological, Within-day Dynamics in Transportation Networks* 92, pp.170–190. https://doi.org/10.1016/j.trb.2015.12.013

Himpe, W., Ginestou, R., Tampère, C., 2019. 'High performance computing applied to dynamic traffic assignment'*, in: Procedia Computer Science*. pp. 409–416. https://doi.org/10.1016/j.procs.2019.04.056

Hofman, F., 2018. 'Frank Hofman: “Hoe beter modelgebruik, hoe meer we de kracht ervan kunnen benutten”'.

Horni, A., Nagel, K., Axhausen, K.W., 2011. 'High-resolution destination choice in agent-based demand models'. *Arbeitsberichte Verkehrs- und Raumplanung* 682. https://doi.org/10.3929/ethz-a-006686309

Huang, W., Xu, G., Lo, H.K., 2020. 'Pareto-Optimal Sustainable Transportation Network Design under Spatial Queuing'. *Netw Spat Econ*. https://doi.org/10.1007/s11067-020-09494-6

Jabari, S.E., 2016. 'Node modeling for congested urban road networks'. *Transportation Research Part B: Methodological* 91, pp.229–249. https://doi.org/10.1016/j.trb.2016.06.001

Janson, B.N., 1991. 'Dynamic traffic assignment for urban road networks'. *Transportation Research Part B: Methodological* 25(2), pp.143–161. https://doi.org/10.1016/0191-2615(91)90020-J

Jin, W.-L., 2012a. 'A kinematic wave theory of multi-commodity network traffic flow'. *Transportation Research Part B: Methodological* 46(8), pp.1000–1022. https://doi.org/10.1016/j.trb.2012.02.009

Jin, W.-L., 2012b. 'The traffic statics problem in a road network'. *Transportation Research Part B: Methodological* 46(10), pp.1360–1373. https://doi.org/10.1016/j.trb.2012.06.003

Jin, W.L., Zhang, H.M., 2004. 'Multicommodity Kinematic Wave Simulation Model for Network Traffic Flow'. *Transportation Research Record* 1883(1), pp.59–67. https://doi.org/10.3141/1883-07

Jin, W.L., Zhang, H.M., 2003. 'On the distribution schemes for determining flows through a merge'. *Transportation Research Part B: Methodological* 37(6), pp.521–540. https://doi.org/10.1016/S0191-2615(02)00026-7

Kager, R., 2007. *Vijf mythes over verkeersmodellen*. Milieudefensie, Amsterdam. Available at: https://www.leefmilieu.nl/sites/www3.leefmilieu.nl/files/imported/pdf_s/verkeersmodellen.pdf (accessed 7.28.23).

Kikuchi, S., Akamatsu, T., 2007. 'A semi-dynamic traffic equilibrium assignment model with link arrival and departure rates'. *Journal of Infrastructure Planning and Management* 24, pp.577–585.

Kitthamkesorn, S., Chen, A., 2013. 'A path-size weibit stochastic user equilibrium model'. *Transportation Research Part B: Methodological* 57, pp.378–397. https://doi.org/10.1016/j.trb.2013.06.001

Köhler, E., Strehler, M., 2012. 'Combining static and dynamic models for traffic signal optimization inherent load-dependent travel times in a cyclically time-expanded network model'. *Procedia-Social and Behavioral Sciences* 54, pp.1125–1134.

Koike, K., Nakayama, S., Yamaguchi, H., 2022. 'A link-based semi-dynamic user equilibrium traffic assignment model considering signal effect'. *Asian Transport Studies* 8, pp.100062. https://doi.org/10.1016/j.eastsj.2022.100062

Kotz, S., Kozubowski, T., Podgorski, K., 2001. *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*. Birkhäuser Basel.

Kouwenhoven, M., de Jong, G.C., Koster, P., van den Berg, V.A.C., Verhoef, E.T., Bates, J., Warffemius, P.M.J., 2014. 'New values of time and reliability in passenger transport in The Netherlands'. *Research in Transportation Economics, Appraisal in Transport* 47, pp.37–49. https://doi.org/10.1016/j.retrec.2014.09.017

L, M., 2020. Is transport modelling junk science? [WWW Document]. *Greater Auckland*. URL https://www.greaterauckland.org.nz/2020/08/31/is-transport-modelling-junk-science/ (accessed 7.28.23).

Lam, W.H.K., Lo, H.P., Zhang, N., 1996. 'A Quasi-dynamic Traffic Assignment Model with Time-dependent Queues'. *HKIE Transactions* 3(2), pp.7–14. https://doi.org/10.1080/1023697X.1996.10667698

Lam, W.H.K., Zhang, Y., 2000. 'Capacity-constrained traffic assignment in networks with residual queues'. *Journal of Transportation Engineering* 126(2). https://doi.org/10.1061/(ASCE)0733-947X(2000)126:2(121)

Larsson, T., Patriksson, M., 1999. 'Side constrained traffic equilibrium models— analysis, computation and applications'. *Transportation Research Part B: Methodological* 33(4), pp.233–264. https://doi.org/10.1016/S0191-2615(98)00024-1

Lawphongpanich, S., Hearn, D.W., 1984. 'Simplical decomposition of the asymmetric traffic assignment problem'. *Transportation Research Part B: Methodological* 18(2), pp.123–133. https://doi.org/10.1016/0191-2615(84)90026-2

Lebacque, J.P., Khoshyaran, M.M., 2013. 'A Variational Formulation for Higher Order Macroscopic Traffic Flow Models of the GSOM Family'. *Procedia - Social and Behavioral Sciences* 80, pp.370–394. https://doi.org/10.1016/j.sbspro.2013.05.021

Lebacque, J.P., Khoshyaran, M.M., 2005. 'First-Order Macroscopic Traffic Flow Models: Intersection Modeling, Network Modeling', *in: 16th International Symposium on Transportation and Traffic Theory*. Presented at the Transportation and Traffic Theory, College Park Maryland, US, pp. 365–386.

Levin, M.W., Pool, M., Owens, T., Juri, N.R., Waller, T., 2015. 'Improving the Convergence of Simulation-based Dynamic Traffic Assignment Methodologies'. *Networks and Spatial Economics* 15(3), pp.655–676. https://doi.org/10.1007/s11067-014-9242-x

Lighthill, M.J., Whitham, G.B., 1955. 'On Kinematic Waves. II. A Theory of Traffic Flow on Long Crowded Roads'. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 229(1178), pp.317–345.

Lindveld, C., 2006. *O-D matrix estimation using the Combined Calibration method as applied to the LMS/NRM*. Imperial College London.

Liu, H.X., He, X., He, B., 2009. 'Method of Successive Weighted Averages (MSWA) and Self-Regulated Averaging Schemes for Solving Stochastic User Equilibrium Problem'. *Netw Spat Econ* 9(4), pp.485–503. https://doi.org/10.1007/s11067-007-9023-x

Lo, H.K., Szeto, W.Y., 2002. 'A cell-based variational inequality formulation of the dynamic user optimal assignment problem'. *Transportation Research Part B: Methodological* 36(5), pp.421–443. https://doi.org/10.1016/S0191-2615(01)00011-X

Lockwood, D., 2018. *Prism 5.0 - Model validation report*. Mott MacDonald, Birmingham. Available at: https://corporate.tfwm.org.uk/media/3502/prism5_reports2_modelvalidationreport_v18_20180531-final.pdf (accessed 1.31.23).

Long, J., Gao, Z., Szeto, W.Y., 2011. 'Discretised link travel time models based on cumulative flows: Formulations and properties'. *Transportation Research Part B: Methodological* 45(1), pp.232–254. https://doi.org/10.1016/j.trb.2010.05.002

Lyons, G., Davidson, C., 2016. 'Guidance for transport planning and policymaking in the face of an uncertain future'. *Transportation Research Part A: Policy and Practice* 88, pp.104–116. https://doi.org/10.1016/j.tra.2016.03.012

Ma, W., Pi, X., Qian, S., 2020. 'Estimating multi-class dynamic origin-destination demand through a forward-backward algorithm on computational graphs'. *Transportation Research Part C: Emerging Technologies* 119, pp.102747. https://doi.org/10.1016/j.trc.2020.102747

Maher, M.J., Zhang, X., Vliet, D.V., 2001. 'A bi-level programming approach for trip matrix estimation and traffic control problems with stochastic user equilibrium link flows'. *Transportation Research Part B: Methodological* 35(1), pp.23–40. https://doi.org/10.1016/S0191-2615(00)00017-5

Mahut, M., Florian, M., 2010. 'Traffic Simulation with Dynameq', *in: Barceló, J. (Ed.), Fundamentals of Traffic Simulation, International Series in Operations Research & Management Science*. Springer, New York, NY, pp. 323–361. https://doi.org/10.1007/978-1-4419-6142-6_9

Marzano, V., Papola, A., Simonelli, F., Papageorgiou, M., 2018. 'A Kalman Filter for Quasi-Dynamic o-d Flow Estimation/Updating'. *IEEE Trans. Intell. Transport. Syst.* 19(11), pp.3604–3612. https://doi.org/10.1109/TITS.2018.2865610

Miller, S.D., Payne, H.J., Thompson, W.A., 1975. 'An algorithm for traffic assignment on capacity transportation networks with queues.'. Presented at the Johns Hopkins conference on information sciences and systems, Baltimore, MD.

Miyagi, T., Makimura, K., 1991. 'A study on semi-dynamic traffic assignment method'. *Traffic Engineering* 26, pp.17–28.

Murty, K.G., 1991. *Linear programming*. Wiley.

Nagurney, A., 1993. *Network economics: a variational inequality approach.* Kluwer Academice Publishers, Boston, USA.

Nakayama, S., 2009. 'A semi-dynamic assignment model considering space-time movement of traffic congestion'. *JSCE Journal of Infrastructure Planning & Management* 64 D, pp.340–353.

Nakayama, S., Connors, R., 2014. 'A quasi-dynamic assignment model that guarantees unique network equilibrium'. *Transportmetrica A: Transport Science* 10(7), pp.669–692. https://doi.org/10.1080/18128602.2012.751685

Nakayama, S., Takayama, J., Nakai, J., Nagao, K., 2012. 'Semi-dynamic traffic assignment model with mode and route choices under stochastic travel times'. *Journal of Advanced Transportation* (46), pp.269–281. https://doi.org/10.1002/atr.208

Newell, G.F., 1993. 'A simplified theory of kinematic waves in highway traffic, part I: General theory'. *Transportation Research Part B: Methodological* 27(4), pp.281–287. https://doi.org/10.1016/0191-2615(93)90038-C

Ni, D., Leonard, J.D., 2005. 'A simplified kinematic wave model at a merge bottleneck'. *Applied Mathematical Modelling* 29(11), pp.1054–1072. https://doi.org/10.1016/j.apm.2005.02.008

Nie, Y., Zhang, H.M., Lee, D.-H., 2004. 'Models and algorithms for the traffic assignment problem with link capacity constraints'. *Transportation Research Part B: Methodological* 38(4), pp.285–312. https://doi.org/10.1016/S0191-2615(03)00010-9

NM Magazine, 2015. 'Thema: Over de zin en onzin van verkeersmodellen'. *NM Magazine* 2015(3), pp.8–9.

Nocedal, J., 1980. 'Updating quasi-Newton matrices with limited storage'. *Mathematics of computation* 35(151), pp.773–782.

Ortuzár, J.D., Willumsen, L.G., 2011. *Modelling Transport, 4th Edition*, 4th edition. ed. Wiley.

Osorio, C., 2019a. 'Dynamic origin-destination matrix calibration for large-scale network simulators'. *Transportation Research Part C: Emerging Technologies* 98, pp.186–206. https://doi.org/10.1016/j.trc.2018.09.023

Osorio, C., 2019b. 'High-dimensional offline origin-destination (OD) demand calibration for stochastic traffic simulators of large-scale road networks'. *Transportation Research Part B: Methodological* 124, pp.18–43. https://doi.org/10.1016/j.trb.2019.01.005

Patil, P.N., Ross, K.C., Boyles, S.D., 2021. 'Convergence behavior for traffic assignment characterization metrics'. *Transportmetrica A: Transport Science* 17(4), pp.1244–1271. https://doi.org/10.1080/23249935.2020.1857883

Payne, H.J., Thompson, W.A., 1975. 'Traffic assignment on transportation networks with capacity constraints and queueing.'. Presented at the 47th national ORSA meeting/TIMS, North American meeting, Chicago, IL.

Peeta, S., Ziliaskopoulos, A.K., 2001. 'Foundations of Dynamic Traffic Assignment: The Past, the Present and the Future'. *Networks and Spatial Economics* 1(3), pp.233–265. https://doi.org/10.1023/A:1012827724856

Perederieieva, O., Ehrgott, M., Raith, A., Wang, J.Y.T., 2015. 'A framework for and empirical study of algorithms for traffic assignment'. *Computers & Operations Research* 54, pp.90–107. https://doi.org/10.1016/j.cor.2014.08.024

Petprakob, W., Wijerathne, L., Iryo, T., Urata, J., Fukuda, K., Hori, M., 2018. 'On the Implementation of High Performance Computing Extension for Day-to-Day Traffic Assignment'. *Transportation Research Procedia, International Symposium of Transport Simulation (ISTS'18) and the International Workshop on Traffic Data Collection and its Standardization (IWTDCS'18)Emerging Transport Technologies for Next Generation Mobility* 34, pp.267–274. https://doi.org/10.1016/j.trpro.2018.11.041

Prato, C.G., 2009. 'Route choice modeling: past, present and future research directions'. *Journal of Choice Modelling* 2(1), pp.65–100. https://doi.org/10.1016/S1755-5345(13)70005-8

Pravinvongvuth, S., Chen, A., 2005. 'Adaptation of the paired combinatorial logit model to the route choice problem'. *Transportmetrica* 1(3), pp.223–240. https://doi.org/10.1080/18128600508685649

PTV Visum Help [WWW Document], n.d. URL https://cgi.ptvgroup.com/vision-help/VISUM_2023_ENG/Content/TitelCopyright/Index.htm (accessed 8.4.23).

Qurashi, M., Ma, T., Chaniotakis, E., Antoniou, C., 2020. 'PC–SPSA: Employing Dimensionality Reduction to Limit SPSA Search Noise in DTA Model Calibration'. *IEEE Transactions on Intelligent Transportation Systems* 21(4), pp.1635–1645. https://doi.org/10.1109/TITS.2019.2915273

Raadsen, M., Bliemer, M., 2019a. 'Continuous-time general link transmission model with simplified fanning, Part II: Event-based algorithm for networks'. *Transportation Research Part B: Methodological* 126, pp.471–501. https://doi.org/10.1016/j.trb.2018.01.003

Raadsen, M., Bliemer, M., 2019b. 'Steady-state link travel time methods: Formulation, derivation, classification, and unification'. *Transportation Research Part B: Methodological* 122, pp.167–191. https://doi.org/10.1016/j.trb.2019.01.014

Raadsen, M., Bliemer, M., Bell, M., 2016. 'An efficient and exact event-based algorithm for solving simplified first order dynamic network loading problems in continuous time'. *Transportation Research Part B: Methodological* 92, pp.191–210. https://doi.org/10.1016/j.trb.2015.08.004

Raadsen, M.P.H., Mein, H.E., Schilpzand, M.P., Brandt, F., 2010. 'Implementation of a single dynamic traffic assignment model on mixed urban and highway transport networks including junction modelling'*, in: Proceedings of the Third International Symposium on Dynamic Traffic Assignment*. Presented at the Third International Symposium on Dynamic Traffic Assignment, Takayama, Japan, p. 19.

Ran, B., Boyce, D., 1996. *Modeling Dynamic Transportation Networks*. Springer.

Richards, P.I., 1956. 'Shock Waves on the Highway'. *Operations Research* 4(1), pp.42–51.

Rijksen, B., 2018. *Matrix Estimation With STAQ* (Masters Thesis). University of Twente, Deventer.

Rijkswaterstaat, W., 2015. *Handboek Capaciteitswaarden Infrastructuur Autosnelwegen.pdf*. Available at: https://staticresources.rijkswaterstaat.nl/binaries/Handboek%20Capaciteitswaarden%20Infrastructuur%20Autosnelwegen_tcm21-76101.pdf (accessed 11.9.16).

Ros-Roca, X., Montero Mercadé, L., Barceló Bugeda, J., 2018. *Notes on the measure of the structural similarity of OD matrices* (Research Report). Available at: http://hdl.handle.net/2117/126659 (accessed 3.26.20).

SATURN 11.6 Manual [WWW Document], n.d. URL https://saturnsoftware2.co.uk/SATURNhelp/index.html#!Documents/saturn116manual.htm (accessed 7.25.23).

Shafiei, S., Saberi, M., Zockaie, A., Sarvi, M., 2017. 'Sensitivity-Based Linear Approximation Method to Estimate Time-Dependent Origin–Destination Demand in Congested Networks'. *Transportation Research Record: Journal of the Transportation Research Board* 2669, pp.72–79. https://doi.org/10.3141/2669-08

Significance, 2021. *Dcumentatie van GM4 Deel D7-6, programma QBLOK4 (versie 4.21)* (manual).

Simoni, M.D., Claudel, C.G., 2020. 'A fast lax–hopf algorithm to solve the lighthill–whitham–richards traffic flow model on networks'. *Transportation Science* 54(6), pp.1526–1534. https://doi.org/10.1287/trsc.2019.0951

Smith, M., 2012. 'Traffic control and route choice: modelling and optimisation'. Presented at the JCT Symposium, University of Warwick.

Smith, M.J., 2013. 'A link-based elastic demand equilibrium model with capacity constraints and queueing delays'. *Transportation Research Part C: Emerging Technologies* 29, pp.131–147. https://doi.org/10.1016/j.trc.2012.04.011

Smith, M.J., 1987. 'Traffic Control And Traffic Assignment In A Signal-controlled Network With Queueing'*, in: Proceedings of the 12th International Symposium On The Theory Of Traffic Flow And Transportation (ISTTT)*. Presented at the 12th International Symposium On The Theory Of Traffic Flow And Transportation (ISTTT), Berkeley, California, US, pp. 61–77.

Smits, E.-S., 2010. *Origin-Destination Matrix Estimation in OmniTRANS* (Masters Thesis). Utrecht University, Deventer.

Smits, E.-S., Bliemer, M., Pel, A.J., van Arem, B., 2015. 'A family of macroscopic node models'. *Transportation Research Part B: Methodological* 74, pp.20–39. https://doi.org/10.1016/j.trb.2015.01.002

Smits, E.-S., Pel, A.J., Bliemer, M., van Arem, B., 2018. 'Generalized Multivariate Extreme Value Models for Explicit Route Choice Sets'. https://doi.org/10.48550/arXiv.1808.04280

Soria-Lara, J.A., Banister, D., 2018. 'Collaborative backcasting for transport policy scenario building'. *Futures* 95, pp.11–21. https://doi.org/10.1016/j.futures.2017.09.003

Spiess, H., 1990. 'Technical Note—Conical Volume-Delay Functions'. *Transportation Science* 24(2), pp.153–158. http://dx.doi.org/10.1287/trsc.24.2.153

Szeto, W.Y., Lo, H.K., 2006. 'Dynamic Traffic Assignment: Properties and Extensions'. *Transportmetrica* 2(1), pp.31–52. https://doi.org/10.1080/18128600608685654

Tajaddini, A., Rose, G., Kockelman, K.M., Vu, H.L., Tajaddini, A., Rose, G., Kockelman, K.M., Vu, H.L., 2020. *Recent Progress in Activity-Based Travel Demand Modeling: Rising Data and Applicability, Models and Technologies for Smart, Sustainable and Safe Transportation Systems*. IntechOpen. https://doi.org/10.5772/intechopen.93827

Tajtehranifard, H., 2017. *Incident Duration Modelling and System Optimal Traffic Re-Routing* (PhD thesis). Queensland University of Technology. https://doi.org/10.5204/thesis.eprints.110525

Tampère, C.M.J., Corthout, R., Cattrysse, D., Immers, L.H., 2011. 'A generic class of first order node models for dynamic macroscopic simulation of traffic flows'. *Transportation Research Part B: Methodological* 45(1), pp.289–309. https://doi.org/10.1016/j.trb.2010.06.004

Taylor, M.A.P., 1984. 'A note on using Davidson's function in equilibrium assignment'. *Transportation Research Part B: Methodological* 18(3), pp.181–199. https://doi.org/10.1016/0191-2615(84)90031-6

Taylor, N., 2003. 'The CONTRAM dynamic traffic assignment model'. *Networks and Spatial Economics* 3, pp.297–322. https://doi.org/10.1023/A:1025394201651

Toledo, T., Kolechkina, T., 2013. 'Estimation of dynamic origin–destination matrices using linear assignment matrix approximations'. *IEEE Transactions on Intelligent Transportation Systems* 14(2), pp.618–626. https://doi.org/10.1109/TITS.2012.2226211

Toledo, T., Kolechkina, T., Wagner, P., Ciuffo, B., Azevedo, C., Marzano, V., Flötteröd, G., 2015. 'Network model calibration studies'. *Daamen W., Buisson C. and S. Hoogendoorn (eds). Traffic simulation and data: validation methods and applications, CRC Press, Taylor and Francis, London* pp.141–162.

Transportation Networks for Research Core Team, 2019. Transportation Networks for Research. [WWW Document]. URL https://github.com/bstabler/TransportationNetworks (accessed 8.26.19).

Transportation Research Board, 2014. *Dynamic, integrated model system: Sacramento-area application. Volume 1.* (SHRP 2 Research Reports No. S2- C10B- RW-1)*, second Strategic Highway Research Program (SHRP 2)*. Available at: https://dx.doi.org/10.17226/22381

Transpute, 2003. *Flowsimulator; beschrijving van het model*. Transpute BV, Gouda. Available at: http://www.transpute.nl/applicaties/flowsim (accessed 8.28.17).

Tsanakas, N., Ekström, J., Olstam, J., 2020. 'Estimating Emissions from Static Traffic Models: Problems and Solutions'. *Journal of Advanced Transportation* 2020, pp.1–17. https://doi.org/10.1155/2020/5401792

Tympakianaki, A., Koutsopoulos, H.N., Jenelius, E., 2015. 'c-SPSA: Cluster-wise simultaneous perturbation stochastic approximation algorithm and its application to dynamic origin–destination matrix estimation'. *Transportation Research Part C: Emerging Technologies* 55, pp.231–245. https://doi.org/10.1016/j.trc.2015.01.016

van der Gun, J.P.T., Pel, A.J., van Arem, B., 2020. 'Travel times in quasi-dynamic traffic assignment'. *Transportmetrica A: Transport Science* 16(3), pp.865–891. https://doi.org/10.1080/23249935.2020.1720862

Van der Zijpp, N.J., 1996. *Dynamic origin-destination matrix estimation on motorway networks* (PhD thesis).

Van Vliet, D., 1982. 'Saturn - A Modern Assignment Model'. *Traffic Engineering & Control* 23(12).

van Vuren, T., n.d. The future of modelling – evolution, not revolution [WWW Document]. URL https://www.mottmac.com/views/the-future-of-modelling-evolution-not-revolution (accessed 1.12.23).

Verlinden, K., van Grol, R., 2022. *Handboek kalibratie LMS/NRM basisjaar 2018 met behulp van SigKal* (No. 22046-R01- 1.1). Significance.

VID, 2017. File top 50 [WWW Document]. *VID | file top 50 over 2017*. URL https://rijkswaterstaatverkeersinformatie.nl/top50.2017.html (accessed 9.7.18).

Vovsha, P., 2019. 'Decision-Making Process Underlying Travel Behavior and Its Incorporation in Applied Travel Models'*, in: Bucciarelli, E., Chen, S.-H., Corchado, J.M. (Eds.), Decision Economics. Designs, Models, and Techniques for Boundedly Rational Decisions*. Springer International Publishing, Cham, pp. 36–48. https://doi.org/10.1007/978-3-319-99698-1_5

Waltz, R.A., Morales, J.L., Nocedal, J., Orban, D., 2006. 'An interior algorithm for nonlinear optimization that combines line search and trust region steps'. *Mathematical Programming* 107(3), pp.391–408. https://doi.org/10.1007/s10107-004-0560-5

Wardrop, J.G., 1952. 'Road paper. some theoretical aspects of road traffic research.'. *Proceedings of the Institution of Civil Engineers* 1(3), pp.325–362. https://doi.org/10.1680/ipeds.1952.11259

Wright, M.A., Gomes, G., Horowitz, R., Kurzhanskiy, A.A., 2017. 'On node models for high-dimensional road networks'. *Transportation Research Part B: Methodological* 105, pp.212–234. https://doi.org/10.1016/j.trb.2017.09.001

Wu, X., Guo, J., Xian, K., Zhou, X., 2018. 'Hierarchical travel demand estimation using multiple data sources: A forward and backward propagation algorithmic framework on a layered computational graph'. *Transportation Research Part C: Emerging Technologies* 96, pp.321–346. https://doi.org/10.1016/j.trc.2018.09.021

Yang, H., 1995. 'Heuristic algorithms for the bilevel origin-destination matrix estimation problem'. *Transportation Research Part B: Methodological* 29(4), pp.231–242. https://doi.org/10.1016/0191-2615(95)00003-V

Yang, H., Yagar, S., 1994. 'Traffic assignment and traffic control in general freeway-arterial corridor systems'. *Transportation Research Part B: Methodological* 28(6), pp.463–486. https://doi.org/10.1016/0191-2615(94)90015-9

Yperman, I., 2007. *The Link Transmission Model for Dynamic Network Loading* (PhD thesis). Katholieke Universiteit Leuven, Leuven.

Yperman, I., Logghe, S., Tampère, C., Immers, B., 2005. 'The link transmission model: an efficient implementation of kinematic wave theory in traffic networks'.

Zhou, Z., Chen, A., Bekhor, S., 2012. 'C-logit stochastic user equilibrium model: formulations and solution algorithm'. *Transportmetrica* 8(1), pp.17–41. https://doi.org/10.1080/18128600903489629

Zill, J.C., Veitch, T., 2022. 'Frozen randomness at the individual utility level'*, in: Australasian Transport Research Forum 2022 Proceedings*. Presented at the Australasian Transport Research Forum 2022, Adelaide, Australia.

Summary

Strategic transport model systems are used to support decision making by assessing the long-term impact of transport policies and land-use scenarios. This thesis focuses on traffic assignment (TA) models for road traffic within strategic transport model systems. TA models describe route choices of road traffic and the resulting traffic state on the network (i.e., traffic conditions, including congestion).
There are two use cases for TA models in strategic transport model systems, both of which are subject of this thesis. The primary use case is the application of a strategic transport model system to evaluate (policy) scenarios. The secondary use case is the estimation of travel demand from observed network data, which is only conducted for a base year (reference scenario) when constructing a new (version of a) strategic transport model system.
To facilitate fair comparison of model outcomes for different scenarios, user equilibrium (UE) conditions are imposed on strategic TA models outcomes. Because this requires an iterative approach, TA models are often the most computationally expensive component in strategic model systems. The above implies that for strategic TA models, there exists a clear trade-off between maximizing model accuracy and minimizing model complexity while ensuring model stability (i.e.: the satisfaction of UE conditions).
The figure below summarizes the contents and coherence of the different chapters in this thesis along with the relationship to the research objectives. In context of the primary use case, the first research objective is to develop a strategic TA model with better accuracy in congested conditions compared to static capacity restrained TA models which are currently mostly used in strategic transport model systems. In context of the secondary use case, the second research objective is to embed this TA model in a travel demand estimation methodology that fuses data on observed flows, congestion patterns and (route) queuing delays.

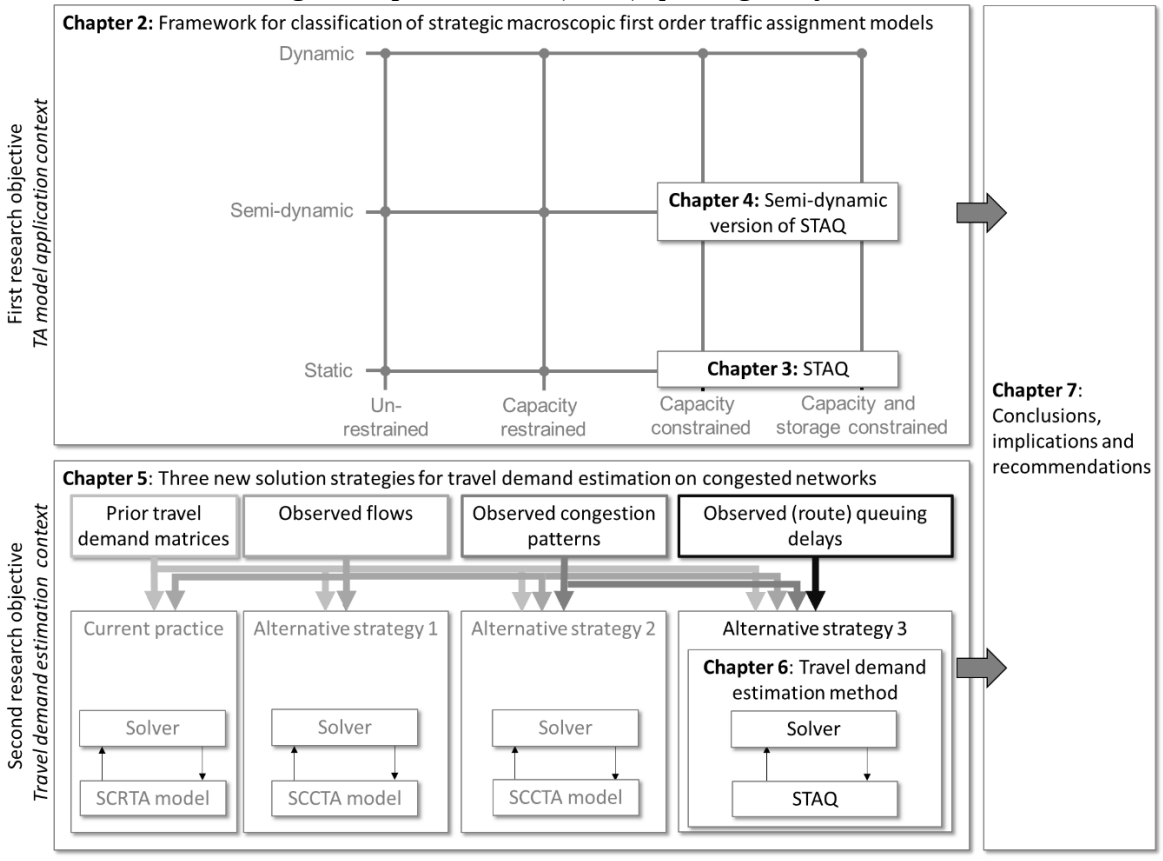


Figure: graphical thesis summary

A framework for classification of strategic macroscopic first order TA models
To provide deeper understanding of the often implicit assumptions made in traffic assignment models and to aid strategic transport model users confronted with the trade-off described above, chapter 2 presents a theoretical framework that classifies all macroscopic first order TA models using concepts analogous to genetics in biology. When only considering equilibrium models, the framework uses two genes to determine the spatial capability (unrestrained, capacity restrained, capacity constrained, capacity and storage constrained) and temporal capability (static, semi-dynamic, dynamic) of TA models. The framework allows for comparing different models in terms of functionality, and paves the way for developing novel traffic assignment models.

Static Traffic Assignment with Queuing (STAQ) and its semi-dynamic sibling
Chapters 3 and 4 describe the development, implementation, testing and large-scale applications of a new static capacity constrained TA model called STAQ (chapter 3) and a semi-dynamic version of this TA model (chapter 4). Both models solve explicitly formulated mathematical problems (a fixed point problem within a (series of) variational inequality problem(s)), from which an efficient solution algorithm is derived.
Compared to the most commonly used static capacity restrained TA models, STAQ adds strict capacity constraints and hence modelling of queues. Additionally, its semi-dynamic sibling removes the assumption that the network is empty at the start of each assignment. Instead, it initiates with a network with traffic that it has transferred from residual queues from the previous time period.
The tests and applications demonstrate that both TA models satisfy the stability and complexity requirements whilst, contrary to their capacity restrained counterparts, successfully model flow reduction and spillback effects of primary bottlenecks, albeit that spillback effects are not included in the route choice behaviour. Both TA models can be run including these effects, but at the expense of stability. Furthermore, both TA models still allow for simultaneous assignment of different vehicle classes and are suitable for application on both urban roads and motorways due to the inclusion of a junction modelling component.
Application of STAQ and its semi-dynamic sibling on the large-scale strategic transport model of Noord-Brabant demonstrates that the addition of capacity constraints causes large differences in terms of collective losses and thus societal benefits of policy measures influencing travel times. This means that shifting from a capacity restrained towards a capacity constrained TA model has substantial effects on the outcomes of a cost benefit analysis for study areas with structural congestion. The applications also show that the empty network assumption in static models causes omission of up to ~75% of collective losses in busy periods, which makes it very likely that this assumption influences (policy) decisions based upon queue size and delay related model outcomes on congested networks.
Chapters 3 and 4 demonstrate that STAQ and its semi-dynamic sibling are viable alternatives to static capacity restrained TA models, providing more accuracy on congested networks without reducing stability and without increasing input requirements, whilst keeping computational requirements to acceptable levels. This makes these models suitable for applications where both static capacity restrained and dynamic TA models may fail: strategic applications on large-scale congested networks.

Solution strategies for travel demand estimation on congested networks
Contrary to capacity restrained TA models, the capacity constraints in STAQ and its semi-dynamic sibling directly influence observed link flow values both up- as well as downstream from active bottlenecks. In strategic transport models, this means that an observed link flow value is either unaffected, metered, partially metered or in queue due to active bottlenecks.

Flow observed on unaffected or partially metered links contain information about travel demand that can be directly used for travel demand estimation, whereas flow observed on links in queue is only useful when supplemented by observed link speeds or queue lengths. Flows observed on metered links only contain information on network supply and can therefore not be used for travel demand estimation. Data and sensitivity analysis on the transport model describing the most congested region of the Netherlands indicates that it is highly unlikely that more than 3% of observed link flows of any Dutch strategic transport model is flow metered, meaning that more than 97% contains information on travel demand.
However, the capacity restrained TA model based solution strategy from current practice can only handle flow values that are unaffected by active bottlenecks (~34% of observed links). Therefore, current practice is to derive unconstrained link demand values from flows affected by active bottlenecks and then, instead of the actual observed flows, use these link demand values during matrix estimation. As such, the solution strategy from current practice exhibits poor tractability and robustness and does not integrate any information from the assignment model about the composition of routes on the observed links.
In chapter 5, a conceptual framework for travel demand matrix estimation methods based upon the four types of observed flow values is presented. The framework is used to identify three novel static capacity constrained TA model based solution strategies for travel demand estimation that can also handle observations on partially metered links and links in queue and hence increases the proportion of usable observations from ~34% to ~97%. The three novel strategies are assessed by comparison to the solution strategy from current practice. The comparison demonstrates that the capacity constrained based methods are more tractable and robust and allow for usage of observed congestion patterns and travel times from (big) data sources. Furthermore, these methods reveal inconsistencies between model link capacities and observed congestion patterns and between count values, allowing the modeler to correct the model network and other matrix estimation input.

Proposed travel demand estimation method

Chapter 6 presents an efficient solution method for the matrix estimation problem using STAQ. The solution method allows for inclusion of route queuing delays and congestion patterns besides the traditional link flows and prior demand matrix whilst the tractability of STAQ avoids the need for tedious tuning of application specific algorithmic parameters.
The proposed solution method solves a series of simplified optimization problems, thereby avoiding costly additional assignment model runs. Link state constraints are used to prevent usage of approximations outside their valid range as well as to include observed congestion patterns. The proposed solution method is relatively fast, scalable, robust, tractable and reliable because conditions under which a solution to the simplified optimization problem exist are known and because the problem is convex and has a smooth objective function.
Four test case applications on the small Sioux Falls model are presented, each consisting of 100 runs with varied input for robustness. The applications demonstrate the added value of inclusion of observed congestion patterns and route queuing delays within the solution method. In addition, application on the large scale BBMB model demonstrates that the proposed solution method is indeed scalable to large scale applications and clearly outperforms the method mostly used in current practice.

Conclusions and implications

The combination of STAQ and the proposed travel demand estimation method in their current form allows for deployment of a new generation of strategic transport model systems that, for the first time, are suited for application on large scale, structurally congested networks

consisting of both highways and urban roads and for different road user classes, whilst maintaining the (low) complexity and (high) stability levels required for strategic applications. The shift from static capacity restrained to the more advanced TA models presented in this thesis has implications for policy makers. The addition of capacity constraints in STAQ has substantial effects on the outcomes of cost benefit analysis for study areas with structural congestion. On top of that, removal of the empty network assumption in the semi-dynamic version of STAQ causes increases in collective losses of up to 76% in peak periods. This makes it very likely that the empty network assumption influences (policy) decisions based upon outcomes from static TA models on congested networks.

From a methodological standpoint, it would be straightforward to extend the proposed travel demand estimator to handle multiple time periods. A combination of this extended version with STAQs semi-dynamic sibling would enable demand estimation on observed flows, delays, and congestion patterns at any temporal aggregation level, ultimately leading to continuous (24-hour) estimation.

Besides providing building blocks for new generations of strategic transport model systems, the software and methods developed in this thesis can also be used to incrementally improve current strategic transport model systems. The travel demand estimation method presented in chapter 6 allows to directly use measured link flows, thereby removing the need to estimate link demands and use these as input (as in current practice). This means that compared to current practice, the presented method is much more transparent and less input sensitive, resulting in better tractability, comparability and transferability of the estimation process. This reduces lead times for application of the estimation method, whilst delivering better accuracy.

This thesis provides another improvement to current strategic transport model systems by including observed queuing delays in the travel demand estimation method. This reduces the under-specification of the mathematical problem that it solves, leading to more consistent results. This ensures similar quality of results when e.g. updating the base year of a transport model system, or comparing base year outcomes of two different transport model systems.

Discussion

Based on the TA model applications in chapters 3 and 4, the author concludes that semi-dynamic capacity constrained TA models are currently the most capable models that still possess the stability and complexity properties required for the strategic application context. Although TA models with storage constraints are more capable with respect to the spatial assumptions, they fail to satisfy the stability requirement as their (implicit) route-cost function can become more sensitive to changes in demand on other routes than to changes in demand on the considered route. Similarly, dynamic TA models are more capable with respect to the temporal assumptions, but fail to meet the stability requirement due to sensitivity of the route cost function as (all) network conditions are transferred between many short time periods. Being also storage constrained, the dynamic TA models used in practice suffer from both causes of instability. Based on literature and own research and application experiences, the author expects that both problems can be solved only to some extent by development of more enhanced algorithms, and at the cost of computational efficiency.

The addition of capacity constraints in TA models requires a different solution method for travel demand estimation, as the relationship between OD demand and link flows is no longer strictly monotonic. Current practice circumvents this problem by feeding a traditional matrix estimation method with unconstrained link demand values estimated from the observed flows in a preprocessing step. The author strongly advices against this approach, as preprocessing methods assume general (exogenous) relationships between link demands and link flows. This yields estimation results that fit well to the link demand estimates, but poorly to the observed link flows. Therefore, in practice, manual changes to the link demand estimates are conducted

in an iterative fashion, causing high and uncertain lead times for demand matrix estimation projects with only reasonable outcomes. Instead, the author advises to shift to the proposed travel demand estimation method which allows for direct estimation on observed link flows and supports observed congestion patterns and travel times as additional input data types.
A further shift to the semi-dynamic capacity constrained TA models would not require a different travel demand estimation methodology. Only when observed flows, delays and congestion patterns are to be included on varying temporal aggregation levels, (expected minor) extension of the proposed estimation method is required. This is left for further research.
This thesis represents only part of the research output, the software implementations of STAQ, its semi-dynamic counterpart and the demand estimation method are an equally, if not more important result. At the time of writing, STAQ is included in OmniTRANS transport planning software and is used in eight different Dutch strategic transport model systems, while its semi-dynamic counterpart and the travel demand estimation method have already successfully been applied in pilots on full scale Dutch strategic transport model systems.

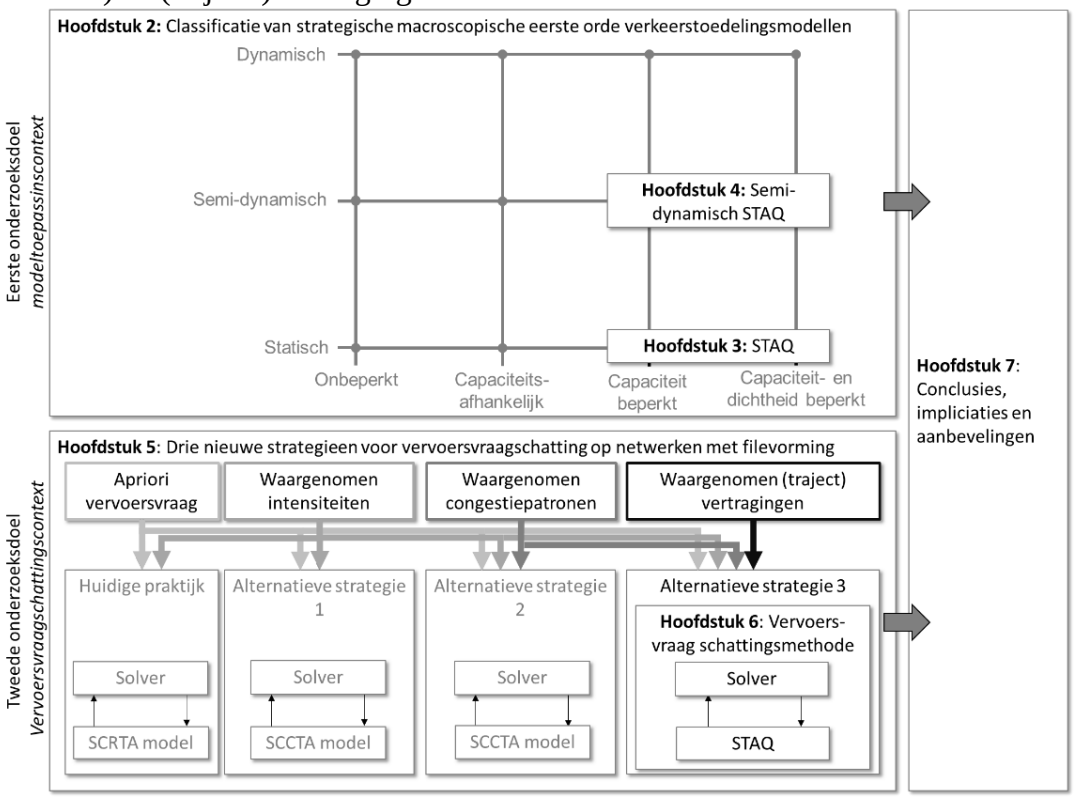
Samenvatting

Strategische verkeersmodelsystemen zijn beslissingsondersteunende instrumenten die de verwachte lange termijneffecten van mobiliteitsbeleid en ruimtelijke scenario's bepalen. Dit proefschrift richt zich op verkeerstoedelingsmodellen voor wegverkeer ('toedelingsmodellen') binnen strategische verkeersmodelsystemen. Toedelingsmodellen beschrijven de routekeuze van weggebruikers en daarmee de verkeerssituatie op het auto-netwerk.

Toedelingsmodellen in strategische verkeersmodelsystemen spelen een rol in zowel de modeltoepassingscontext (waarin het verkeersmodelsysteem wordt gebruikt om (beleids-) scenario's te evalueren) als de modelbouw- (of actualisatie-) context waarin de vervoers-vraagschatting op basis van o.a. waargenomen wegvakintensiteiten plaats vindt. Beide contexten zijn onderwerp zijn van dit proefschrift.

Om modeluitkomsten voor verschillende scenario's eerlijk te kunnen vergelijken moeten uitkomsten van strategische toedelingsmodellen voldoen aan de gebruikersevenwicht-condities van Wardrop. Omdat het rekenen onder deze condities een iteratieve aanpak vereist, zijn toedelingsmodellen vaak de meest rekenintensieve component van strategische verkeersmodel-systemen. Het voorgaande impliceert dat een strategisch toedelingsmodel een compromis is tussen (maximaal) realisme en (minimale) complexiteit onder randvoorwaarde van stabiliteit (het voldoen aan evenwichtscondities).

Onderstaande figuur vat de inhoud en samenhang van de verschillende hoofdstukken in dit proefschrift samen in relatie tot de onderzoeksdoelen. Het eerste onderzoeksdoel (modeltoepassingscontext) is het ontwikkelen van een strategisch toedelingsmodel wat effecten van filevorming beter beschrijft dan de 'capaciteitsafhankelijke' toedelingsmodellen die nu het meeste worden gebruikt in strategische verkeersmodelsystemen. Het tweede onderzoeksdoel (modelbouwcontext) is om het ontwikkelde toedelingsmodel in te bedden in een vervoersvraag-schattingsmethode waarin waargenomen wegvakintensiteiten, congestiepatronen (koplocaties van files) en (traject-) vertragingen fuseert.



Figuur: grafische samenvatting van dit proefschrift

Classificatie van strategische macroscopische eerste orde verkeerstoedelingsmodellen
Om een beter begrip te krijgen van de vaak impliciete aannames in toedelingsmodellen en om gebruikers van strategische verkeersmodellen te helpen die geconfronteerd worden met het hierboven beschreven compromis, presenteert hoofdstuk 2 een theoretisch raamwerk dat alle macroscopische eerste orde toedelingsmodellen classificeert met behulp van concepten ontleend uit de genetica in biologie. Voor evenwichtsmodellen gebruikt het raamwerk twee 'genen' om aannamen over de ruimtelijke interactie (onbeperkt, capaciteitsafhankelijk, capaciteit beperkt, capaciteit en dichtheid beperkt) en temporele interactie (statisch, semi-dynamisch, dynamisch) van weggebruikers in toedelingsmodellen te definiëren. Het raamwerk maakt het mogelijk om verschillende modellen functioneel te vergelijken en maakt de weg vrij voor de ontwikkeling van nieuwe toedelingsmodellen.

Static Traffic Assignment with Queuing (STAQ) en semi-dynamisch STAQ
Hoofdstukken 3 en 4 beschrijven de ontwikkeling, implementatie, het testen en grootschalige toepassingen van een nieuw statisch capaciteitsbeperkt toedelingsmodel STAQ (hoofdstuk 3) en een semi-dynamische versie van STAQ (hoofdstuk 4). Beide modellen lossen expliciet geformuleerde wiskundige problemen op (een fixed point probleem binnen een (reeks) variatie-ongelijkheidsproble(e)m(en)), op basis waarvan een efficiënt oplossingsalgoritme is afgeleid.
De meeste strategische verkeersmodelsystemen gebruiken een statisch capaciteitsafhankelijk toedelingsmodel. Ten opzichte van dit modeltype voegt STAQ strikte capaciteitsbeperkingen toe en daarmee de modellering van wachtrijvorming. Daarbovenop laat de semi-dynamische versie van STAQ de aanname los dat het netwerk leeg is bij de start van elke toedeling. In plaats daarvan zijn resterende wachtrijen van de vorige tijdsperiode nog aanwezig.
Toepassingen tonen aan dat beide toedelingsmodellen voldoen aan de stabiliteits- en complexiteitseisen terwijl ze, in tegenstelling tot hun capaciteitsafhankelijke tegenhangers, doorstroomreductie en terugslageffecten als gevolg van wachtrijvorming rondom primaire knelpunten modelleren. Om aan de stabiliteits-eis te voldoen worden de terugslageffecten alleen meegenomen in het routekeuzegedrag na (en niet tijdens) het bepalen van de evenwichtscondities. Beide modellen kunnen meerdere gebruikers- en/of voertuigklassen simultaan toedelen en zijn geschikt voor toepassing op zowel het stedelijk/regionale als hoofdwegennet omdat ze een kruispuntmodelleringscomponent bevatten.
Toepassing van STAQ en semi-dynamisch STAQ op het grootschalige strategische verkeersmodelsysteem van de provincie Noord-Brabant laat zien dat het toevoegen van strikte capaciteitsbeperkingen grote verschillen veroorzaakt in voertuigverliesuren en dus maatschappelijke baten van beleidsmaatregelen die reistijden van weggebruikers beïnvloeden. Dit betekent dat de verschuiving van een capaciteitsafhankelijk naar een capaciteitsbeperkt toedelingsmodel substantiële effecten heeft op de uitkomsten van een kosten-batenanalyse voor studiegebieden met structurele filevorming. De toepassingen laten ook zien dat de lege netwerk aanname in statische toedelingsmodellen zorgt voor een onderschatting tot ca ~75% van de voertuigverliesuren in drukke periodes, wat het zeer waarschijnlijk maakt dat deze aanname (beleids-)beslissingen beïnvloedt die gebaseerd zijn op wachtrijomvang- en vertragings-gerelateerde modeluitkomsten op netwerken waarin filevorming voorkomt.
Hoofdstukken 3 en 4 laten zien dat STAQ en semi-dynamisch STAQ serieuze alternatieven zijn voor capaciteitsafhankelijke toedelingsmodellen. De STAQ-modellen bieden meer realisme op overbelaste netwerken, voldoen aan de stabiliteitscondities, handhaven de lage invoervereisten, terwijl de rekenvereisten op een aanvaardbaar niveau worden gehouden. Dit maakt de modellen geschikt voor toepassingen waarbij zowel statische capaciteitsafhankelijke als dynamische toedelingsmodellen het laten afweten: strategische toepassingen op grootschalige overbelaste netwerken.

Strategieën voor vervoersvraagschatting op netwerken met filevorming

In tegenstelling tot in capaciteitsafhankelijke toedelingsmodellen, beïnvloeden infrastructurele capaciteitsbeperkingen in zowel de werkelijkheid als in (semi-dynamisch) STAQ de betekenis van waargenomen wegvakintensiteiten, zowel stroomopwaarts als stroomafwaarts van actieve knelpunten. Een waargenomen wegvakintensiteit is ofwel niet-beïnvloed, (doorstroom-) gereduceerd, gedeeltelijk gereduceerd of in de wachtrij van een actief knelpunt waargenomen. Intensiteiten die zijn waargenomen op niet-beïnvloede of gedeeltelijk gereduceerde wegvakken bevatten informatie die direct kan worden gebruikt voor de schatting van de vervoersvraag, terwijl in wachtrijen waargenomen wegvakintensiteiten alleen nuttig zijn als deze vergezeld zijn van waargenomen wegvaksnelheden of wachtrijlengtes. Waargenomen intensiteiten op volledig gereduceerde wegvakken bevatten alleen informatie over de infrastructurele capaciteit en kunnen daarom niet worden gebruikt voor het schatten van de vervoersvraag. Een data- en gevoeligheidsanalyse op het strategische verkeersmodelsysteem van Rijkswaterstaat dat de volledige Randstad beschrijft, laat zien dat het hoogst onwaarschijnlijk is dat meer dan 3% van de waargenomen wegvakintensiteiten in Nederland volledig gereduceerd is, wat betekent dat meer dan 97% informatie bevat over de vervoersvraag.

De oplossingsstrategie uit de huidige praktijk gebruikt een capaciteitsafhankelijk toedelings-model en kan daardoor alleen waargenomen intensiteiten gebruiken die niet beïnvloed zijn door actieve knelpunten (~ 34% van de waargenomen wegvakken). Daarom worden in de huidige praktijk zogenaamde ‘wensvraag’ waarden afgeleid voor alle door actieve knelpunten beïnvloede waargenomen wegvakintensiteiten om vervolgens deze ‘wensvraag’ waarden te gebruiken tijdens de vervoersvraagschatting. Deze oplossingsstrategie uit de huidige praktijk is daardoor moeilijk traceerbaar en weinig robuust. Bovendien gebruikt het geen informatie uit het toedelingsmodel over de verzameling van routes die gebruik maken van de bemeten wegvakken.

Hoofdstuk 5 presenteert een conceptueel raamwerk voor vervoersvraagschattingsmethoden op basis van de vier soorten intensiteits-waarnemingen. Het raamwerk wordt gebruikt om drie nieuwe oplossingsstrategieën voor vervoersvraagschatting te identificeren, alle gebaseerd op een statisch capaciteitsbeperkt toedelingsmodel. Alle drie de strategieën kunnen gebruik maken van waarnemingen op gedeeltelijk gereduceerde wegvakken en wegvakken met een wachtrij, waardoor het aandeel bruikbare waarnemingen wordt verhoogd van ~34% naar ~97%. Een vergelijking van de drie nieuwe strategieën met de oplossingsstrategie uit de huidige praktijk toont aan dat de nieuwe strategieën transparanter en robuuster zijn en het gebruik van waargenomen congestiepatronen en reistijden uit (big) databronnen mogelijk maken. Bovendien onthullen deze methoden zowel kruislingse als onderlinge inconsistenties tussen gemodelleerde wegvakcapaciteiten, waargenomen congestiepatronen, waargenomen wegvakintensiteiten en waargenomen traject-vertragingen, op basis waarvan de modelleur het modelnetwerk en andere invoer voor de vervoersvraagschatting kan corrigeren.

Een nieuwe vervoersvraagschattingsmethode op basis van STAQ

Hoofdstuk 6 presenteert een efficiënte oplossingsmethode voor het matrixschattingsprobleem die gebruik maakt van STAQ. De oplossingsmethode kan gebruik maken van waargenomen (traject-)vertragingen en congestiepatronen naast de traditionele databronnen (waargenomen wegvakintensiteiten en een a priori vervoersvraagschatting). Doordat de methode een expliciet geformuleerd wiskundig probleem oplost is deze traceerbaar en transparant en hoeven er geen toepassings-specifieke parameters bepaald te worden.

De nieuwe oplossingsmethode lost een reeks vereenvoudigde optimalisatieproblemen op, waardoor gebruik van het (relatief rekenintensieve) toedelingsmodel wordt geminimaliseerd tot één toepassing per iteratie. De oplossingsmethode bevat randvoorwaarden die de toestand van knelpunten in het netwerk (actief/ passief) vastzet, waardoor alleen in het domein binnen de oplossingsruimte waarin het vereenvoudigde optimalisatieprobleem geldig is wordt gezocht.

Deze randvoorwaarden worden ook gebruikt om waargenomen congestiepatronen te specificeren. De voorgestelde oplossingsmethode is relatief snel, schaalbaar, robuust, transparant en betrouwbaar omdat de omstandigheden waaronder een oplossing voor het vereenvoudigde optimalisatieprobleem bestaat bekend zijn en omdat het probleem convex is en een gladde doelfunctie heeft.
Er zijn vier test-toepassingen op het (kleine) Sioux Falls-model uitgevoerd, elk bestaande uit 100 runs met gevarieerde input voor robuustheid. De toepassingen tonen de toegevoegde waarde aan van het gebruik van waargenomen congestiepatronen en (traject-) vertragingen binnen de oplossingsmethode. Bovendien toont toepassing op het grootschalige provinciale model van de Provincie Noord-Brabant aan dat de methode inderdaad schaalbaar is voor grootschalige toepassingen en duidelijk beter presteert dan de methoden uit de huidige praktijk.

Conclusies en implicaties
De combinatie van STAQ en de nieuwe vervoersvraagschattingsmethode maakt voor het eerst de inzet van een nieuwe generatie strategische verkeersmodelsystemen mogelijk die geschikt zijn voor toepassing op grootschalige, structureel overbelaste netwerken op zowel snelwegen als stedelijke wegen en voor verschillende klassen weggebruikers, met behoud van de (lage) complexiteit en (hoge) stabiliteitsniveaus die nodig zijn voor strategische toepassingen.
De verschuiving van statische capaciteitsafhankelijke naar de meer geavanceerde toedelingsmodellen modellen die in dit proefschrift worden gepresenteerd, heeft implicaties voor beleidsmakers. De toevoeging van capaciteitsbeperkingen in STAQ heeft substantiële effecten op de uitkomsten van kosten-batenanalyse voor studiegebieden met structurele congestie. Bovendien zorgt het wegnemen van de lege netwerk-aanname in de semi-dynamische versie van STAQ voor een toename van de collectieve verliezen tot 76% in spitsperiodes in het verkeersmodelsysteem van de provincie Noord-Brabant. Dit maakt het zeer waarschijnlijk dat de lege netwerk-aanname van invloed is op (beleids-)beslissingen op basis van uitkomsten van statische toedelingsmodellen op overbelaste netwerken.
Vanuit methodologisch oogpunt is de ontwikkelde vervoersvraagschatter eenvoudig uit te breiden zodat deze met meerdere tijdsperioden kan werken. Een combinatie van deze uitgebreide versie met semi-dynamisch STAQ zou de vervoersvraagschatting op basis van waargenomen wegvakintensiteiten, (traject-)vertragingen en congestiepatronen op elk temporeel aggregatieniveau mogelijk maken, ultimo leidend tot een continue (24-uurs) schatter.
Naast het leveren van bouwstenen voor nieuwe generaties strategische verkeersmodelsystemen, kunnen de software en methodieken ontwikkeld in dit proefschrift, ook worden gebruikt om huidige strategische verkeersmodelsystemen stapsgewijs te verbeteren. De in hoofdstuk 6 gepresenteerde vervoersvraagschattingsmethode maakt het mogelijk om direct gebruik te maken van waargenomen wegvakintensiteiten, waardoor het niet meer nodig is om vooraf zogenaamde ‘wensvraag’ te bepalen en deze als invoer te gebruiken (zoals in de huidige praktijk). Dit betekent dat de gepresenteerde methode in vergelijking met de huidige praktijk transparanter en minder inputgevoelig is, wat resulteert in een betere traceerbaarheid, vergelijkbaarheid en overdraagbaarheid van het schattingsproces. Dit verkort de doorlooptijden voor het toepassen van de schattingsmethode en levert tegelijkertijd een hogere kwaliteit op.
Dit proefschrift biedt een verdere verbetering van de huidige strategische verkeersmodel-systemen door waargenomen (traject-)vertragingen mee te nemen in de vervoersvraag-schattingsmethode. Dit vermindert de onderspecificatie van het wiskundige probleem dat het oplost, wat leidt tot consistentere resultaten. Dit leidt tot een consistent kwaliteitsniveau van resultaten, waardoor tussen verschillende toepassingen beter vergelijkbaar zijn. Dit is relevant wanneer het basisjaar van een verkeersmodelsysteem wordt geactualiseerd of wanneer uitkomsten van verschillende verkeersmodelsystemen worden vergeleken.

Discussie

Op basis van de toepassingen van toedelingsmodellen in hoofdstukken 3 en 4 wordt geconcludeerd dat semi-dynamische capaciteitsbeperkte toedelingsmodellen momenteel de meest realistische modellen zijn die nog voldoen aan de stabiliteit en complexiteits-eisen die de strategische toepassingscontext stelt. Hoewel dichtheidsbeperkte toedelingsmodellen realistischer zijn met betrekking tot de aannames over ruimtelijke interactie van reizigers, voldoen ze niet aan de stabiliteitseis omdat hun (impliciete) routekostenfunctie gevoeliger kan zijn voor veranderingen in de vraag op andere routes dan voor veranderingen in de vraag op de beschouwde route. Evenzo zijn dynamische toedelingsmodellen realistischer met betrekking tot de aannames over temporele interactie van reizigers, maar voldoen ze niet aan de stabiliteitsvereiste vanwege de gevoeligheid van de routekostenfunctie, aangezien in dit type model (alle) netwerkcondities tussen vele korte tijdsperioden worden overdragen. Omdat ze ook dichtheidsbeperkt zijn, hebben de dynamische toedelingsmodellen uit de praktijk last van beide oorzaken van instabiliteit. Op basis van literatuur, eigen onderzoek en toepassingservaringen verwacht de auteur dat beide problemen slechts deels kunnen worden opgelost door meer geavanceerde algoritmen, en ten koste zal gaan van rekenefficiëntie.

De toevoeging van capaciteitsbeperkingen in toedelingsmodellen vereist een andere vervoersvraagschattingsmethode omdat de relatie tussen vervoersvraag enerzijds en wegvakintensiteiten anderzijds niet langer strikt monotoon is. De huidige praktijk omzeilt dit probleem door een traditionele vervoersvraagschattingsmethode te voeden met zogenaamde ‘wensvraag’ waarden die in een voorverwerkingsstap worden bepaald op basis van de waargenomen wegvakintensiteiten. De auteur raadt deze benadering ten zeerste af, aangezien de methoden gebruikt in de voorverwerkingsstap uit gaan van algemene (exogene) relaties tussen vervoersvraag en wegvakintensiteit. Dit levert schattingsresultaten op die goed passen bij de ‘wensvraag’ waarden, maar slecht bij de waargenomen wegvakintensiteiten. Daarom worden in de praktijk handmatige wijzigingen aan de ‘wensvraag’ waarden doorgevoerd in een iteratief proces, wat leidt tot lange en onzekere doorlooptijden voor projecten met vervoersvraagschatting resulterend in slechts redelijke resultaten. In plaats daarvan adviseert de auteur om over te stappen op de voorgestelde vervoersvraagschattingsmethode, die direct gebruik maakt van waargenomen wegvakintensiteiten en ook congestiepatronen en (traject-) vertragingen kan verwerken.

Een verdere verschuiving van statische naar de semi-dynamische capaciteitsbeperkte toedelingsmodellen vereist geen andere vervoersvraagschattingsmethode. Alleen wanneer waargenomen wegvakintensiteiten, (traject-)vertragingen en congestiepatronen op verschillende temporele aggregatieniveaus moeten worden meegenomen, is een (naar verwachting kleine) uitbreiding van de voorgestelde schattingsmethode nodig. Realisatie van deze uitbreiding is een aanbeveling voor toekomstig onderzoek.

Dit proefschrift vertegenwoordigt slechts een deel van het onderzoeksresultaat, de software-implementaties van STAQ, zijn semi-dynamische tegenhanger en de vervoersvraagschatter zijn een even belangrijk, zo niet belangrijker resultaat. Op het moment van schrijven is STAQ opgenomen in de OmniTRANS verkeersmodelleringssoftware van DAT.Mobility en wordt het gebruikt in acht verschillende Nederlandse strategische verkeersmodelsystemen. De semi-dynamische versie van STAQ en de vervoersvraagschattingsmethode zijn al succesvol toegepast in pilots op grootschalige Nederlandse strategische verkeersmodelsystemen.

About the Author

Luuk Brederode was born in 1980 in Tilburg, where he also grew up. After completing secondary school, he moved to Enschede in 1999 where he obtained his Msc degree in Civil engineering at the University of Twente in 2005. In the same year, Luuk moved to Deventer where he started working at Goudappel as a consultant in transport modelling, building and applying urban and regional strategic transport model systems for Dutch clients.

In 2008 Luuk switched to a job as transport model innovator at Goudappel where he specialized in research, prototyping and valorization of methodologies for strategic transport model systems. Luuk moved over with the founding of Goudappels IT-related sister company DAT.Mobility in 2014, widening his scope to include data fusion methods. He also supervised numerous Bachelor's and Master's thesis projects related to transport modelling and data fusion from students in mainly Mathematics, Civil engineering and Econometrics.

Parallel to his job at Goudappel and DAT.Mobility, he conducted a PhD research from 2013 to 2023 at the transport and planning department of Delft University of Technology resulting in this thesis. In 2014, he spent three months as an occupational trainee at the university of Sydney, Australia, working out the first concepts that would ultimately result in the travel demand matrix estimation methodology developed in chapter 6. At DAT.Mobility, Luuk continues to actively develop and realize the methodological roadmap of the company's transport model related products.

List of publications

Journal Articles

Brederode, L., Gerards, L., Wismans, L., Pel, A., Hoogendoorn, S., 2023. Extension of a static into a semi-dynamic traffic assignment model with strict capacity constraints. Transportmetrica A: Transport Science 0(0), pp.1–34. https://doi.org/10.1080/23249935.2023.2249118

Brederode, L., Pel, A.J., Wismans, L., Rijksen, B., Hoogendoorn, S.P., 2023. Travel demand matrix estimation for strategic road traffic assignment models with strict capacity constraints and residual queues. Transportation Research Part B: Methodological 167, 1–31. https://doi.org/10.1016/j.trb.2022.11.006

Brederode, L., Pel, A.J., Wismans, L., de Romph, E., Hoogendoorn, S.P., 2019. Static Traffic Assignment with Queuing: model properties and applications. Transportmetrica A: Transport Science 15, 179–214. https://doi.org/10.1080/23249935.2018.1453561

Bliemer, M., Raadsen, M., Brederode, L., Bell, M., Wismans, L., Smith, M., 2017. Genetics of traffic assignment models for strategic transport planning. Transport Reviews 37, 56–78. https://doi.org/10.1080/01441647.2016.1207211

Peer-reviewed conference contributions

Bliemer, M., Raadsen, M., Wismans, L., Brederode, L., 2019. Dynamic speed control and lane management in the general link transmission model. Presented at the Tenth Triennial Symposium on Transportation Analysis (TRISTAN X), Hamilton Island, Australia.

Bliemer, M.C.J., Brederode, L.J.N., Wismans, L.J.J., Smits, E.S., 2012. Quasi-dynamic network loading: adding queuing and spillback to static traffic assignment. Presented at the 91st Transportation Research Board (TRB) Annual Meeting 2012, Washington, D.C.

Brederode, L., Verlinden, K., 2019. Travel demand matrix estimation methods integrating the full richness of observed traffic flow data from congested networks. Transportation Research Procedia, Modeling and Assessing Future Mobility ScenariosSelected Proceedings of the 46th European Transport Conference 2018, ETC 2018 42, 19–31. https://doi.org/10.1016/j.trpro.2019.12.003

Brederode, L., Pots, M., Fransen, R., Brethouwer, J.-T., 2019. Big Data fusion and parametrization for strategic transport demand models, in: 2019 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). Presented at the 2019 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), pp. 1–8. https://doi.org/10.1109/MTITS.2019.8883333

Brederode, L., Pel, A.J., Wismans, L., de Romph, E., 2016. Improving convergence of quasi dynamic assignment models. Presented at the 6th International Symposium on Dynamic Traffic Assignment, Sydney, Australia.

Brederode, L., Pel, A.J., Hoogendoorn, S.P., 2014. Matrix estimation for static traffic assignment models with queuing. hEART 2014 – 3rd symposium of the European association for research of transportation, Leeds UK.

Wismans, L., Van den Brink, R., Brederode, L., Zantema, K., van Berkum, E., 2013. Comparison of Estimation of Emissions Based on Static and Dynamic Traffic Assignment Models. Presented at the Transportation Research Board 92nd Annual Meeting of the Transportation Research Board, Washington, DC.

Non-peer-reviewed conference contributions

Possel, B., Graaf, S.D., Brederode, L., 2020. The Mobility Spectrum: A data driven strategic transport model for the whole of The Netherlands. Presented at the European Transport Conference 2020, online due to COVID19, p. 12.

Brederode, L., Koopal, R., 2019. Estimating the Potential for Mobility-As-A-Service in the Netherlands Using Mobile Phone Data. Presented at the European Transport Conference, Association for European Transport (AET), Dublin, Ireland.

Wismans, L., Suijs, L., Brederode, L., Palm, H., Beek, P. van, 2018. State Estimation, Short Term Prediction and Virtual Patrolling Providing a Consistent and Common Picture for Traffic Management and Service Providers. Presented at the 25th ITS World Congress 2018.

Brederode, L., Hofman, F., van Grol, R., 2017. Testing of a demand matrix estimation method incorporating observed speeds and congestion patterns on the Dutch strategic model system using an assignment model with hard capacity constraints, Presented at the 45th European Transport Conference, Barcelona.

Suijs, L., Wismans, L., Brederode, L., 2017. Model–based short term predictor of traffic states. Presented at the 45th European Transport Conference, ETC 2017, Barcelona.

Zantema, K., Heynickx, M., Brederode, L., Koopal, R., 2016. Time of Day and Demand Equilibrium in Large Scale Models, in: European Transport Conference 2016. Presented at the European transport conference, Association for European Transport (AET), Barcelona.

Brederode, L., Heynicks, M., Koopal, R., 2016. Quasi Dynamic Assignment on the Large Scale Congested Network of Noord-Brabant. Presented at the European transport conference, AET 2016 and contributors, Barcelona, p. 17.

Brederode, L., Bliemer, M., Wismans, L., 2010. STAQ: Static Traffic Assignment with Queuing, in: Proceedings of the European Transport Conference. Presented at the ETC 2010: European Transport Conference, Glasgow, UK.

Professional magazine articles

Koopal, R., Brederode, L., Boomsma, R., 2020. MaaS-potentiescan voor heel Nederland op basis van gsm-data. Tijdschrift vervoerswetenschap 56, 49–62.

Brederode, L., Hoogendoorn, S.P., 2018. Strategische verkeersmodellen – hoe krijgen we ze realistisch én snel? NM-magazine 2018.

TRAIL Thesis Series

The following list contains the most recent dissertations in the TRAIL Thesis Series. For a complete overview of more than 275 titles see the TRAIL website: www.rsTRAIL.nl.

The TRAIL Thesis Series is a series of the Netherlands TRAIL Research School on transport, infrastructure and logistics.

Brederode, L.J.N., *Incorporating Congestion Phenomena into Large Scale Strategic Transport Model Systems*, T2023/12, October 2023, TRAIL Thesis Series, the Netherlands

Hernandez, J.I., *Data-driven Methods to study Individual Choice Behaviour: with applications to discrete choice experiments and Participatory Value Evaluation experiments*, T2023/14, October 2023, TRAIL Thesis Series, the Netherlands

Aoun, J., *Impact Assessment of Train-Centric Rail Signaling Technologies,* T2023/13, October 2023, TRAIL Thesis Series, the Netherlands

Pot, F.J., *The Extra Mile: Perceived accessibility in rural areas*, T2023/12, September 2023, TRAIL Thesis Series, the Netherlands

Nikghadam, S., *Cooperation between Vessel Service Providers for Port Call Performance Improvement*, T2023/11, July 2023, TRAIL Thesis Series, the Netherlands

Li, M., *Towards Closed-loop Maintenance Logistics for Offshore Wind Farms: Approaches for strategic and tactical decision-making*, T2023/10, July 2023, TRAIL Thesis Series, the Netherlands

Berg, T. van den, *Moral Values, Behaviour, and the Self: An empirical and conceptual analysis,* T2023/9, May 2023, TRAIL Thesis Series, the Netherlands

Shelat, S., *Route Choice Behaviour under Uncertainty in Public Transport Networks: Stated and revealed preference analyses,* T2023/8, June 2023, TRAIL Thesis Series, the Netherlands

Zhang, Y., Flexible, *Dynamic, and Collaborative Synchromodal Transport Planning Considering Preferences*, T2023/7, June 2023, TRAIL Thesis Series, the Netherlands

Kapetanović, M., *Improving Environmental Sustainability of Regional Railway Services*, T2023/6, June 2023, TRAIL Thesis Series, the Netherlands

Li, G., *Uncertainty Quantification and Predictability Analysis for Traffic Forecasting at Multiple Scales,* T2023/5, April 2023, TRAIL Thesis Series, the Netherlands

Harter, C., *Vulnerability through Vertical Collaboration in Transportation: A complex networks approach,* T2023/4, March 2023, TRAIL Thesis Series, the Netherlands

Razmi Rad, S., *Design and Evaluation of Dedicated Lanes for Connected and Automated Vehicles,* T2023/3, March 2023, TRAIL Thesis Series, the Netherlands

Eikenbroek, O., *Variations in Urban Traffic*, T2023/2, February 2023, TRAIL Thesis Series, the Netherlands

Wang, S., *Modeling Urban Automated Mobility on-Demand Systems: an Agent-Based Approach*, T2023/1, January 2023, TRAIL Thesis Series, the Netherlands

Szép, T., *Identifying Moral Antecedents of Decision-Making in Discrete Choice Models*, T2022/18, December 2022, TRAIL Thesis Series, the Netherlands
Zhou, Y., *Ship Behavior in Ports and Waterways: An empirical perspective*, T2022/17, December 2022, TRAIL Thesis Series, the Netherlands

Yan, Y., *Wear Behaviour of A Convex Pattern Surface for Bulk Handling Equipment*, T2022/16, December 2022, TRAIL Thesis Series, the Netherlands

Giudici, A., *Cooperation, Reliability, and Matching in Inland Freight Transport*, T2022/15, December 2022, TRAIL Thesis Series, the Netherlands

Nadi Najafabadi, A., *Data-Driven Modelling of Routing and Scheduling in Freight Transport*, T2022/14, October 2022, TRAIL Thesis Series, the Netherlands

Heuvel, J. van den, Mind Your Passenger! The passenger capacity of platforms at railway stations in the Netherlands, T2022/13, October 2022, TRAIL Thesis Series, the Netherlands

Haas, M. de, Longitudinal Studies in Travel Behaviour Research, T2022/12, October 2022, TRAIL Thesis Series, the Netherlands

Dixit, M., *Transit Performance Assessment and Route Choice Modelling Using Smart Card Data*, T2022/11, October 2022, TRAIL Thesis Series, the Netherlands

Du, Z., *Cooperative Control of Autonomous Multi-Vessel Systems for Floating Object Manipulation*, T2022/10, September 2022, TRAIL Thesis Series, the Netherlands

Larsen, R.B., *Real-time Co-planning in Synchromodal Transport Networks using Model Predictive Control*, T2022/9, September 2022, TRAIL Thesis Series, the Netherlands

Zeinaly, Y., Model-based Control of Large-scale Baggage Handling Systems: Leveraging the theory of linear positive systems for robust scalable control design, T2022/8, June 2022, TRAIL Thesis Series, the Netherlands

Fahim, P.B.M., *The Future of Ports in the Physical Internet*, T2022/7, May 2022, TRAIL Thesis Series, the Netherlands

Huang, B., *Assessing Reference Dependence in Travel Choice Behaviour*, T2022/6, May 2022, TRAIL Thesis Series, the Netherlands

Reggiani, G., *A Multiscale View on Bikeability of Urban Networks*, T2022/5, May 2022, TRAIL Thesis Series, the Netherlands