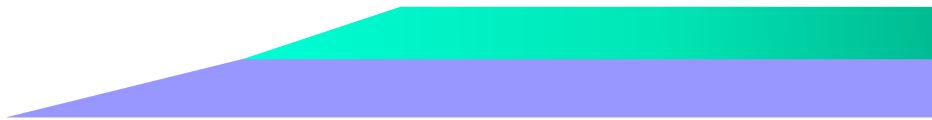


Linköping Studies in Science and Technology
Dissertation No. 2523

Travel Demand Estimation and Network Supply Calibration for Large-Scale Urban Networks

Guang Wei



Linköping Studies in Science and Technology.
Dissertation No. 2523

Travel Demand Estimation and Network Supply Calibration for Large-Scale Urban Networks

Guang Wei



Department of Science and Technology
Linköping University, SE-601 74 Norrköping, Sweden

Norrköping 2026



Except where otherwise noted, this work is licensed under a Creative Commons Attribution 4.0 International License.
<https://creativecommons.org/licenses/by/4.0/>

Travel Demand Estimation and Network Supply Calibration for Large-Scale Urban Networks

Guang Wei

ISBN 978-91-8118-562-1 (print), 978-91-8118-563-8 (pdf)

ISSN 0345-7524

DOI <https://doi.org/10.3384/9789181185638>

Copyright ©, Guang Wei, 2026

Cover illustration by Guang Wei

Generative AI tools are used for language editing and to improve readability.

Linköping University

Department of Science and Technology

SE-601 74 Norrköping

Printed by LiU Tryck, Linköping, Sweden 2026

Abstract

Estimation of origin-destination (OD) vehicle flows and link capacity calibration are fundamental processes in transportation science, especially in the context of transport modelling and simulation. They ensure that transportation models accurately reflect real-world travel behaviour and network conditions.

This thesis develops a simulation-based optimization algorithm for network-wide link capacity calibration. To address the high dimensionality of large-scale networks, the algorithm is integrated with partial least squares (PLS) regression, which reduces the number of variables and enhances computational efficiency. The algorithm is evaluated on an urban road network in Stockholm, Sweden, where it demonstrates feasibility and higher efficiency compared to the simultaneous perturbation stochastic approximation (SPSA) method.

For large-scale OD estimation, this thesis advances the field in two main directions. First, it develops a data fusion framework that integrates multiple heterogeneous data sources, including mobile network data, link count observations, and turning proportion data. Second, it proposes several methods to enhance the computational efficiency of OD estimation in large urban networks. These include: (i) implementing data-driven network assignment (DDNA) using GPS data to construct a fixed OD-to-link mapping, thereby eliminating the need for iterative assignment within a bi-level optimization structure; (ii) applying non-negative matrix factorization (NNMF) for dimensionality reduction, which simplifies the optimization by reducing the number of variables; and (iii) developing a numerical solver based on an interior-point method that exploits structural properties of the assignment matrix, such as sparsity and linearity, to enhance computational performance. The proposed OD estimation methods are evaluated on real-world networks in central Stockholm and Norrköping, Sweden, demonstrating accurate and stable OD and link flow estimates with substantial gains in computational efficiency compared to solving the OD estimation problem without dimensionality reduction techniques or numerical solver improvement.

Sammanfattning

Skattning av fordonsflöden mellan olika geografiska områden, så kallad OD-estimering från engelskans origin och destination, samt kalibrering av länkkapacitet är grundläggande processer inom transportmodellering och trafiksimulering. Dessa processer säkerställer att transportmodeller på ett korrekt sätt speglar verklig resefterfrågan och nätverksförhållanden.

Denna avhandling utvecklar en simuleringsbaserad optimeringsalgoritm för kalibrering av länkkapacitet på nätverksnivå. För att hantera den höga dimensionaliteten i storskaliga nätverk integreras algoritmen med Partial Least Squares (PLS)-regression, vilket minskar antalet variabler och förbättrar beräkningseffektiviteten. Det föreslagna ramverket utvärderas på ett vägnät för Stockholm, som visar på att algoritmen producerar resultat med hög noggrannhet och har högre effektivitet jämfört med metoden Simultaneous Perturbation Stochastic Approximation (SPSA) som är vanlig för detta problem.

När det gäller OD-estimering för storskaliga nätverk bidrar denna avhandling till området i två huvudsakliga riktningar. För det första utvecklas ett ramverk för datafusion som integrerar flera heterogena datakällor, inklusive mobilnätverksdata, observationer av länkflöden och data över fordons svängandelar i korsningar. För det andra föreslås flera metoder för att förbättra beräkningseffektiviteten vid OD-estimering i stora urbana nätverk. Dessa metoder bygger på: (i) Implementering av datadriven nätutläggning (DDNA) med hjälp av GPS-data för att skapa en fast avbildning mellan OD-par och länkar, vilket eliminerar behovet av iterativ nätverksutläggning; (ii) Tillämpning av Non-negative Matrix Factorization (NNMF) för dimensionsreduktion, vilket förenklar optimeringen genom att minska antalet variabler; samt (iii) Utveckling av en lösningsmetod baserad på en inrepunktsmetod som utnyttjar strukturella egenskaper i den matris som beskriver avbildningen mellan OD-par och länkar, såsom gleshet och linjäritet, för att förbättra beräkningsprestandan.

De föreslagna metoderna för OD-estimering utvärderas på verkliga nätverk i centrala Stockholm och Norrköping. Resultaten visar på god måluppfyllelse och stabila estimat av OD- och länkflöden, med betydande vinster i beräkningseffektivitet jämfört med att lösa OD-estimeringsproblem utan tekniker för dimensionsreduktion eller förbättrad lösningsalgoritm.

Acknowledgments

The research included in this thesis has been funded by the Swedish Transport Administration (TRV 2018/134731 and TRV 2021/22404).

First of all, I would like to thank my supervisors Clas Rydergren and David Gundlegård for their guidance and inspiration. I am grateful to them who have supported my work and put a lot of effort into helping me improve my writing skills.

I also would like to thank my previous supervisors Gunnar Flötteröd and Joakim Ekström for their guidance and instruction during my first two-year PhD career.

I thank all my colleagues in the division of Communication and Transport Systems (KTS), who have helped me greatly both in my study and in my daily life. I am also grateful to Rasmus Ringdahl, who has helped me greatly in solving technical problems and enlightened me greatly in improving my programming skill.

Finally, I would like to show my appreciation to my family and friends for your love and support.

Norrköping, March 2026
Guang Wei

Contents

Abstract	iii
Sammanfattning	v
Acknowledgments	vii
1 Introduction	1
1.1 Research motivation	1
1.2 Aim of the thesis	2
1.3 Outline	2
2 Traffic flow models and traffic assignment	5
2.1 Traffic models	5
2.2 Traffic flow models	6
2.2.1 Macroscopic models	6
2.2.2 Microscopic models	7
2.2.3 Mesoscopic models	8
2.3 Traffic assignment	8
2.3.1 Traffic assignment problem	9
2.3.2 Data-driven network assignment	10
3 OD estimation	15
3.1 Mathematical formulation	16
3.2 Methods for OD estimation	17
4 Link capacity calibration	19
4.1 Mathematical formulation	19
4.2 Simulation-based optimization algorithms	20
4.3 Framework for network-wide link capacity calibration	21
5 Dimensionality reduction	25
5.1 Principal component analysis	26
5.2 Partial least squares regression	26
5.3 Non-negative matrix factorization	27
5.4 Importance of dimensionality reduction	27
6 Research challenges, questions and methodology	29

Contents

6.1	Research challenges	29
6.1.1	Data fusion	30
6.1.2	Computational efficiency	30
6.2	Research objectives and research questions	31
6.3	Research methodology	32
7	Contributions of the thesis	35
7.1	Summary of papers	35
7.2	Scientific contributions	40
8	Conclusions and future work	41
8.1	Conclusions	41
8.2	Future work directions	41
	Bibliography	43
	Abbreviations	49
	Paper I	53
	Paper II	71
	Paper III	81
	Paper IV	91
	Paper V	111
	Paper VI	119

Chapter 1

Introduction

Since the invention of the car, vehicle traffic has been steadily increasing for over a century. Cars and road transport have become integral to modern society, shaping urban development and mobility. However, the rapid growth in vehicle numbers has caused significant congestion problems in many cities, reducing overall traffic efficiency. Traffic congestion increases travel costs and reduces accessibility by increasing travel time. Constant exposure to congested roads has harmful consequences, including noise pollution, increased driver stress, deterioration of mental well-being, and disruptions in urban development (Bigazzi and Clifton, 2015). In addition, traffic congestion has significant effects on the environment, leading to higher fuel consumption and elevated levels of air pollution.

1.1 Research motivation

Origin-destination (OD) demand¹, which is typically in the form of a matrix, describes how many trips² are intended to occur between different zones in an urban network. Each entry in an OD matrix represents the number of trips for a given origin-destination pair. An OD estimation problem is the task of determining the number of trips between origins and destinations in a given network, using indirect data such as link flow measurements. The problem arises because OD demands are rarely observed directly. It is essentially about reconstructing travel demand patterns from observations such as link flow observations, and traffic assignment provides the mathematical link that translates OD demands, how many trips start to travel between each OD pair, into the resulting link flows on each road segment. For large-scale urban networks, OD estimation is crucial because it provides the fundamental picture

¹In this thesis, the term *OD demand* is used interchangeably with *OD flow*.

²In this thesis, trips always refer to vehicle trips.

of the movement of vehicles throughout a city, supporting effective planning, traffic management, and infrastructure investment. Identifying when, where, and how travel demand emerges, OD demands enable planners to analyse congestion patterns, assess the effectiveness of existing infrastructure, and forecast future mobility requirements. OD demand also serves as a key input for simulation models, policy assessments, and operational strategies, enabling decision makers to develop measures that improve network efficiency, enhance safety, and support sustainable urban development. Cities that lack accurate OD demand data risk misallocating resources, worsening congestion, and failing to meet sustainability targets.

Congestion occurs when the number of vehicles trying to use a link exceeds its link capacity, causing vehicles to slow down, trips to take longer, and queues to form. Link capacity refers to the maximum sustainable flow of vehicles that a link can accommodate under prevailing traffic, geometric, and control conditions. It is a foundational concept used in traffic engineering, network modelling, and transportation planning. The relationship between link capacities and OD demands is constraining: OD demands determine how much traffic starts to use the network because they represent the total number of trips that travellers start to make between every OD pair, and link capacities determine how much traffic can actually be accommodated in a link level. A link capacity calibration problem is the task of adjusting the flow capacity of links in a network so that the predicted link flows match the observed link flows. Link capacity calibration is critically important in urban networks because it ensures that traffic models and simulations reflect real-world road performance. Without proper calibration, models can misrepresent congestion, leading to poor planning and ineffective traffic management.

1.2 Aim of the thesis

Transport models require both link capacity information and travel demand to function effectively. The main aim of this thesis is to develop new methods for OD estimation and link capacity calibration in large-scale urban networks, improving the accuracy and reliability of transport models.

Contributing to this aim requires addressing major challenges, including the high computational cost caused by the large dimensionality of urban networks and the need to combine multiple data sources.

1.3 Outline

In the following chapters, the knowledge framework guiding this thesis is presented. In Chapter 2, the theoretical background of traffic flow models and traffic assignment is presented, followed by mathematical formulations and state-of-the-art approaches for OD estimation (Chapter 3) and link

capacity calibration (Chapter 4) in urban networks, respectively. Chapter 5 presents typical dimensionality reduction techniques which have been used, including principal component analysis (PCA) and non-negative matrix factorization (NNMF), or could be used, such as partial least squares (PLS), in transport estimation or calibration problems. Chapter 6 and 7 outline the research contribution by identifying the research challenges the thesis aims to address, presenting the research questions, and providing summary of the included papers. Finally, Chapter 8 concludes the thesis and outlines directions for future work.

Chapter 2

Traffic flow models and traffic assignment

2.1 Traffic models

A traffic model is a mathematical representation of real-world traffic conditions. It describes how vehicles move through a transportation network by capturing the relationships among travel demand, network supply, and traveller behaviour. By abstracting complex interactions into analytical or computational frameworks, traffic models enable the study of large urban networks that cannot be fully understood through observation alone.

Fundamentally, traffic models aim to represent how many trips occur, where they begin and end, which routes travellers choose, and how these choices interact with network capacity. Depending on the required level of detail, models may operate at an aggregate scale, focusing on flows between traffic analysis zones (TAZs), which are fundamental spatial units used in transportation planning and modelling to represent where trips originate and where they end, or at a more detailed scale, simulating individual travellers or vehicles. This flexibility allows traffic models to support applications ranging from long-term planning to short-term operational analysis.

Traffic models are commonly classified as static or dynamic, depending on whether temporal variation is represented. Static models identify equilibrium conditions in traveller decisions, such as mode and route choice, under steady-state assumptions. They do not account for time-dependent changes and are therefore time-independent. In contrast, dynamic traffic models simulate how traffic evolves over time, capturing variations in demand, congestion, and traveller behaviour throughout the analysis period. By incorporating time-dependent route choices, speeds, and flows, dynamic

models provide a more realistic representation of network conditions, albeit with higher computational and data requirements.

Traffic models also play a central role in evaluating the impacts of policy measures, infrastructure investments, and emerging mobility technologies. By simulating alternative scenarios, they help planners assess how changes in demand, network design, or control strategies influence congestion, travel times, emissions, and safety. As cities grow and transportation systems become more complex, the ability of traffic models to provide reliable, data-driven insights becomes increasingly important for informed and sustainable decision making.

2.2 Traffic flow models

Traffic flow models describe the movement of vehicles at a more detailed, physical level than broader traffic models. Instead of focusing on travel demand or route choice, they aim to capture how traffic behaves on individual road segments or within small parts of a network. A wide range of traffic flow models exists, each offering different levels of detail and computational complexity (Treiber and Kesting, 2013).

Macroscopic models treat traffic flow similarly to fluid dynamics or electric currents, where vehicles are considered as a continuous stream. These models often draw from fields like fluid mechanics and emphasize statistical properties. Common variables include flow, density, and speed.

Microscopic models focus on individual vehicles, modelling their behaviour and interactions. Subtypes of microscopic models include car-following models, cellular automata, and others. Variables typically represent detailed characteristics such as a vehicle's position, velocity, and acceleration.

Mesoscopic models lie between the macroscopic and microscopic levels, combining microscopic realism (drivers accelerate, decelerate, and interact) and macroscopic property (flow–density–speed) to balance detail and computational efficiency.

2.2.1 Macroscopic models

The foundational development of macroscopic traffic models is accomplished by Lighthill and Whitham (1955a), and Lighthill and Whitham (1955b), who first propose a theoretical form of wave motion closely analogous to traffic flow behaviour. Later, Richards (1956) incorporates the concept of shock waves into this framework, resulting in the renowned Lighthill-Whitham-Richards (LWR) model, which is mathematically described using partial differential equations. In macroscopic modelling, three key variables, density (ρ), flow (Q), and local speed (V), are considered, each dependent on time (t) and one-dimensional space (x) for simplicity. These variables are related

by the following equation:

$$Q(x, t) = \rho(x, t)V(x, t). \quad (2.1)$$

Density and flow are also governed by the principle of vehicle conservation, which is expressed through the continuity equation:

$$\frac{\partial \rho(x, t)}{\partial t} + \frac{\partial Q(x, t)}{\partial x} = 0. \quad (2.2)$$

In the LWR model, speed and flow are rigidly linked to traffic density, which can result in unrealistic outcomes such as infinite accelerations or unstable traffic behaviour. To address these limitations, second-order models like Payne's Model (Payne, 1971) are later introduced, offering a more realistic representation of traffic dynamics. These models incorporate acceleration characteristics, allowing speed variations to be expressed as functions of density, local velocity, and their respective gradients. Following their introduction, these second-order models undergo extensive scrutiny and revisions, with the goal of enhancing their accuracy and effectiveness in capturing real-world traffic conditions (Daganzo, 1995a, Aw and Rascle, 2000, and Helbing and Johansson, 2009).

An important area of research in macroscopic traffic modelling involves discretizing these models to enable practical computational implementation. Daganzo (1994) and Daganzo (1995b) are the first to introduce the cell transmission model (CTM) framework, which discretizes both time and space. Although initially perceived as cumbersome, this approach proves highly suitable for massively parallel computing, allowing large regional networks to be divided into smaller subnetworks aligned with specific subareas of the study region. Later, Yperman (2007) proposes the link transmission model (LTM), which focuses on link-level dynamics. It models how traffic enters, travels through, and exits road links over time. LTM is recognized for its computational efficiency in the handling of large-scale networks (Gun et al., 2019).

2.2.2 Microscopic models

Microscopic traffic models treat each vehicle as an individual unit. The position and velocity of each vehicle are strongly influenced by those of its neighbours. These models incorporate human driving behaviours such as acceleration, deceleration, and lane changes. Although they offer greater accuracy in simulating real-world traffic dynamics, this comes at the cost of increased computational complexity and input data needs. A key challenge lies in accurately capturing driver behaviour, which varies significantly across individuals. For instance, reaction time is influenced by numerous personal and situational factors.

The car-following model is one of the most investigated microscopic models, where the movement of the i -th vehicle is governed by the actions

of its immediate predecessor, the $i - 1$ -th vehicle. In this framework, the acceleration $a_i(t)$ is expressed as a function of $v_i(t)$, $v_{i-1}(t)$ and $s_i(t)$.

$$a_i(t) = a_i(v_i(t), v_{i-1}(t), s_i(t)), \quad (2.3)$$

where $a_i(t)$ denotes the acceleration of the i -th vehicle at time t , while $v_i(t)$ and $v_{i-1}(t)$ represent the velocities of i -th and $i - 1$ -th vehicle at time t , respectively. The term $s_i(t)$ refers to the spacing or distance between these two vehicles at time t .

Nagel and Schreckenberg (1992) is the first to apply Cellular Automata (CA) models to traffic models. The mathematical foundation of CA models dates back to 1948, originally developed to explore biological systems (Neumann, 1951). In Neumann's work, the concept of the automaton is introduced, which is viewed as a black box with a defined set of internal states. Its key characteristic lies in the rules that govern how it transitions from one state to another. In CA models applied to transportation science, each cell can exist in one of two states: vacant or occupied by a single vehicle. The system evolves through four sequential steps, which include acceleration, deceleration, randomization, and vehicle movement. These four steps are executed simultaneously for all vehicles.

CA models are less frequently used today because they are computationally intensive. Each road link is segmented into multiple cells, and the number of cells directly influences the computation time, which makes CA models inefficient when modelling large-scale transportation networks. Modern microscopic traffic modelling has largely shifted toward continuous car-following and lane-changing models, which offer higher realism, smoother dynamics, and better calibration capabilities.

2.2.3 Mesoscopic models

Mesoscopic traffic models are a middle ground between macroscopic (aggregate flows) and microscopic (individual vehicle behaviour) models. In a mesoscopic model, each vehicle is tracked, but simplified rules govern speed and interactions. In addition, travel times and capacities are modelled using queueing theory (Horni et al., 2016, and Flötteröd and Rohde, 2011) or simplified speed-flow relationships. Mesoscopic traffic models balance computational efficiency with behavioural detail, making them useful for simulating large networks with realistic driver interactions without the heavy data and processing demands of microscopic models (Burghout et al., 2006).

2.3 Traffic assignment

Traffic models provide the behavioural and physical foundations for describing how vehicles interact, how congestion forms, and how traffic conditions evolve over time. However, understanding these dynamics at the link level is only

part of the picture. In real networks, travellers continuously make route choices that shape where and when congestion appears. This is where traffic assignment becomes essential. By combining the behavioural insights captured in traffic models with network-wide representations of travel demand, traffic assignment determines how flows are distributed across alternative paths and how individual decisions collectively produce the traffic patterns observed in a city. In this way, assignment serves as the bridge between local traffic dynamics and the broader spatial organisation of movement within a transport system.

2.3.1 Traffic assignment problem

Traffic assignment problems revolve around determining how travel demand is distributed across a network, and both static and dynamic formulations share a common conceptual core while differing in how they treat time, congestion, and driver behaviour. At the heart of both lies the idea that travellers choose routes in response to travel costs, and that these choices collectively shape the flows and congestion patterns that emerge on the network.

The focus of this thesis is on the dynamic traffic assignment (DTA) problem. DTA is a modelling framework that captures how traffic conditions evolve over time in response to time-varying demand, network congestion, and traveller behaviour. Unlike static assignment, which assumes steady-state conditions and time-independent flows, DTA explicitly represents the temporal dimension of traffic, allowing flows, speeds, and route choices to change throughout the analysis period. This makes DTA particularly suitable for analysing peak-hour dynamics, congestion propagation, and the effects of time-dependent control strategies.

The main elements of a dynamic traffic assignment problem are:

- 1 Time-dependent OD demand which is represented as departure rates over time.
- 2 Dynamic network loading (DNL) in which traffic flow model (macroscopic, mesoscopic, or microscopic) that simulates how vehicles move through the network, how queues form, and how travel times evolve.
- 3 Time-dependent travel costs that depend on departure time, route, and interactions with congestion waves.
- 4 Route choice behaviour, with which travellers aim to minimise their perceived travel time or generalised travel cost, leading to flow patterns that reflect individual decision-making principles.
- 5 Dynamic user equilibrium (DUE) which requires that all used routes and departure times for an OD pair have equal and minimal experienced cost, considering time-varying congestion.

In traffic assignment problems, the assignment matrix provides a fast and computationally efficient way to estimate link flows (Witheyford, 1963). It is a mathematical construct that specifies how OD demand is distributed across the network. Formally, the assignment matrix maps OD demands to link flows by indicating the proportion of each OD pair’s demand that uses each link. By linking where trips begin and end with the paths they take through the network, the assignment matrix enables computing link-level traffic volumes and identify resulting congestion patterns.

2.3.2 Data-driven network assignment

Recent studies, such as X. Yang et al. (2017), Krishnakumari et al. (2020), and Tsanakas et al. (2023), have explored the use of GPS data from probe vehicles to approximate the assignment matrix under current traffic conditions, shifting away from traditional DTA methods toward data-driven network assignment (DDNA). These approaches aim to establish an exogenous linear relationship from OD demands to link flows, enabling more efficient and accurate OD matrix estimation.

The main idea of DDNA is to introduce data from alternative sources (such as GPS data) during the same analysis period, and further to produce an empirical or externally derived assignment matrix \mathbf{A} , which is independent of the OD demands (Tsanakas et al., 2023). The assignment matrix is linear in the data-driven framework, as congestion effects are incorporated through external variables. These external variables represent traffic state information that affects travel times and route choice. It is important to note that the resulting assignment matrix is exogenous which does not offer a generic mapping from OD demands to link flows. Consequently, this assignment matrix is case-specific, and it cannot be reliably applied across different analysis periods.

In a DDNA approach, the construction of the assignment matrix \mathbf{A} is based on calculating a network loading matrix \mathbf{Q} and a route choice matrix \mathbf{R} , and $\mathbf{A} = \mathbf{QR}$. Matrix \mathbf{Q} describes the propagation pattern of the network, mapping route demands¹ to link flows. Matrix \mathbf{R} describes the route choice probability of each route for all OD pairs, mapping OD demands to route demands. Note that route demand represents the travel demand on a specific route in a specific hour, while the link flow represents the observed number of vehicles passing a link in a specific hour.

The network loading matrix \mathbf{Q} is constructed from empirical traffic measurements collected by sensors measuring the link flows and the GPS data which indirectly measure the time-dependent travel speed. A deeper conceptual interpretation of the DDNA is that vehicles are propagated across the network along the route which consists of a number of links. The link travel time is time-dependent for all links, which is generated from GPS data using a linear regression model (Tsanakas et al., 2021, Algorithm 1). When

¹In this thesis, the term *route demand* is interchangeable with *route flow*.

Discretised trajectory diagram of a vehicle along the route

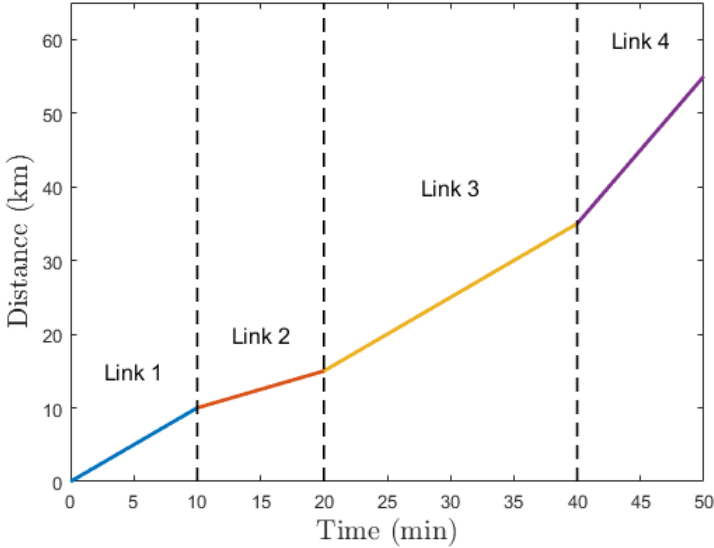


Figure 2.1: An example of discretised vehicle propagation along the route.

a vehicle joins a link, link 1, along its route, the link travel time for it on link 1 depends on the current time (t_0) from the pre-obtained regression model, which determines the future time, t_1 , when it leaves link 1 (i.e., joins the next link, link 2). When this vehicle joins link 2, the link travel time for it on link 2 depends on the new time (t_1), which determines the future time, t_2 , when it leaves link 2, joining link 3, etc. This process continues until the vehicle reaches its destination.

The vehicle propagation process, represented continuously by the trajectory diagram and discretely by the network loading matrix \mathbf{Q} , describes the same underlying physical phenomenon: how vehicles move through space and time along a route. These two representations differ in their level of abstraction. The trajectory diagram provides a continuous depiction of a vehicle's position as a function of time, whereas \mathbf{Q} offers a discrete, algebraic formulation of this movement within the DDNA framework. In effect, \mathbf{Q} corresponds to a discretised version of the trajectory diagram, capturing vehicle progression link-by-link under the assumption that travel speed changes only when the vehicle enters a new link.

A detailed explanation of the network loading procedure and the computation of \mathbf{Q} is provided in Algorithm 2 in Tsanakas et al. (2021). Figure 2.1 illustrates this discretised propagation: Each linear segment represents the vehicle's movement along a single link, and the slope of each segment reflects the travel speed at the moment the vehicle joins that link.

The route choice matrix \mathbf{R} is constructed through a model-based pro-

cedure. It encodes the proportion of OD demand that uses each feasible route. These proportions are derived from enumerated or algorithmically generated route sets, a behavioural route choice model, and external variables such as link travel times, which are obtained from GPS data. The detailed mathematical formulation for calculating \mathbf{R} can be found in Tsanakas et al. (2023).

The calculation of the assignment matrix \mathbf{A} consists following major steps, and Table 2.1 gives input and output information for each step:

- 1 Scenario loading. A scenario for the investigated region is constructed, incorporating the road network, TAZs (origins and destinations), link-flow measurements, GPS trajectory data, and the target OD demand.
- 2 Link travel time estimation. Link-level travel times are inferred from GPS trajectory data, using timestamped vehicle positions to derive travel times for each link within the specified analysis period.
- 3 Route set generation. For every OD zone pair, GPS trajectory data are processed to identify all observed routes. These routes are then added to the feasible route set for that OD pair.
- 4 Calculation of route travel times. Using the GPS trajectory data, travel times are calculated for each identified route in the route set, providing empirical route-level travel time estimates.
- 5 Calculation of loading matrix \mathbf{Q} . A matrix \mathbf{Q} is built by estimating the proportional contribution of each route demand to each link flow.
- 6 Estimation of route choice matrix \mathbf{R} . A matrix \mathbf{R} is built by estimating the proportional contribution of each OD demand to each route flow, using a Logit model.
- 7 Calculation of assignment matrix \mathbf{A} . The assignment matrix \mathbf{A} is obtained by $\mathbf{A} = \mathbf{QR}$.

For obtaining the assignment matrix, DDNA offers greater adaptability, scalability, and behavioural realism compared to DTA, making it more effective for modern transportation systems. Data-driven models learn directly from observed traffic patterns, capturing complex and stochastic driver behaviours without relying on rigid assumptions, while traditional DTA often assumes equilibrium conditions, which may not reflect real-world variability in driver decisions. Moreover, data-driven approaches can handle large-scale networks efficiently, while traditional DTA requires intensive computation and manual tuning, especially for large urban networks (Elimadi et al., 2024).

Table 2.1: Inputs and outputs of each step in the DDNA-based assignment matrix calculation.

Step	Input	Output
Scenario loading	Road network; TAZs; link-flow data; GPS trajectories; target OD demand	Complete scenario with all required data
Link travel time estimation	GPS trajectories with timestamps; network topology	Estimated link travel times
Route set generation	GPS trajectories; OD zone definitions; network topology	Feasible route set for each OD pair
Route travel time calculation	Route set; estimated link travel times	Route-level travel time estimates
Loading matrix Q	Route set; link travel times; network topology	Matrix Q (route-to-link proportions)
Route choice matrix R	OD demand; route travel times; route set; Logit model	Matrix R (OD-to-route proportions)
Assignment matrix A	Matrices Q and R	Assignment matrix A = QR

Chapter 3

OD estimation

OD estimation involves determining how many trips occur between different origin and destination pairs within a transportation network. This information is foundational for nearly all aspects of traffic analysis and urban mobility planning, including traffic flow prediction, traffic management and control, public transportation optimization, and incident and event impact assessment.

When estimating time-dependent OD demands, it is intuitive to employ a dynamic traffic assignment approach. This estimation process typically relies on two key types of input data: link flow measurements collected from stationary traffic sensors, and target OD demands, floating car data (Y. Yang et al., 2010), or survey data (Cascetta, 1984). The core objective is to iteratively adjust the target OD demands for each time interval so that the resulting assigned link flows align with observed traffic counts, while keeping the Euclidean distance between the estimated OD demands and the target OD demands relatively small.

Figure 3.1 presents an example urban network from central Norrköping, Sweden. The solid black lines denote TAZ boundaries, while the light dashed lines indicate the links. Red dots mark the locations of sensors. The goal of OD estimation is to infer the travel demand between TAZs defined by the solid boundaries over specific time periods, using the link flow measurements collected by sensors.

Estimating the OD demands is a complex task, primarily due to the nonlinear nature of the DTA process that translates OD demands into link flows. This complexity transforms the OD estimation into a bi-level optimization problem, where the upper level adjusts the OD matrix and the lower level solves the traffic assignment (Chen and Florian, 1995).



Figure 3.1: Illustration of the example Norrköping network.

3.1 Mathematical formulation

Consider an urban network with a total number of OD pairs I and a temporal analysis period H hours. For simplicity of the presentation, we use a temporal resolution of 1 hour for both the demand and the flows, and the start time and the end time for the demand and the flows are the same. B is the total number of links where link count observations, by hour, are known.

Let $x_{i,h}$ denote the number of travellers in OD pair i , whose departure time from the origin is within the h -th analysis period. The time-sliced OD demand is represented by a vector¹: $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_h, \dots, \mathbf{x}_H) \in \mathbb{R}_+^{IH \times 1}$, where $\mathbf{x}_h = (x_{1,h}, \dots, x_{i,h}, \dots, x_{I,h}) \in \mathbb{R}_+^{I \times 1}$. The observed link flow on link b within the analysed period h is denoted by $y_{b,h}$, and a flow vector $\tilde{\mathbf{y}} = (\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_h, \dots, \tilde{\mathbf{y}}_H) \in \mathbb{R}_+^{BH \times 1}$, where $\tilde{\mathbf{y}}_h = (\tilde{y}_{1,h}, \dots, \tilde{y}_{b,h}, \dots, \tilde{y}_{B,h}) \in \mathbb{R}_+^{B \times 1}$, is defined. Given an estimated assignment matrix $\mathbf{A} \in \mathbb{R}_+^{BH \times IH}$, which is constant through a DDNA approach, and the target OD demands, $\hat{\mathbf{x}}$, which has the same structure as \mathbf{x} , a non-negative least squares problem is formulated as

$$\begin{aligned} \min_{\mathbf{x}} F(\mathbf{x}) &= \gamma_1 \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \gamma_2 \|\mathbf{A}\mathbf{x} - \tilde{\mathbf{y}}\|_2^2, \\ \text{s.t. } \mathbf{x} &\geq \mathbf{0}, \end{aligned} \tag{3.1}$$

where $\|\cdot\|_2$ represents the Euclidean norm (l^2 -norm). The non-negative

¹In this thesis, bold lowercase letters denote column vectors, writing horizontally with commas (,) differentiating the entries, e.g., $\mathbf{a} = (1, 2)$. A vector without commas represents a row vector, e.g., $(1 \ 2)$. It is clear that $(1 \ 2)^T = \begin{pmatrix} 1 \\ 2 \end{pmatrix} = (1, 2)$. $\mathbb{R}_+^{m \times n}$ denotes the set of all $m \times n$ matrices whose entries are non-negative real numbers.

parameters γ_1 and γ_2 represent the weights for the corresponding terms.

Another counterpart optimization formulation as a derivation of entropy-maximising formulation is

$$\begin{aligned} \min_{\mathbf{x}} G(\mathbf{x}) = & \gamma_1 \sum_i (x_i \log(\frac{x_i}{\hat{x}_i}) - x_i + \hat{x}_i) \\ & + \gamma_2 \sum_j (\mathbf{A}_j \mathbf{x} \log(\frac{\mathbf{A}_j \mathbf{x}}{\tilde{y}_j}) - \mathbf{A}_j \mathbf{x} + \tilde{y}_j), \end{aligned} \quad (3.2)$$

where i and j represent the indices for OD demands and link flows, respectively, \hat{x}_i is the i -th component of $\hat{\mathbf{x}}$, \mathbf{A}_j is the j -th row of \mathbf{A} , and \tilde{y}_j is the j -th component of $\tilde{\mathbf{y}}$.

The quadratic (l^2) formulation in (3.1) offers a simple and computationally efficient way to estimate, with intuitive tuning and well-behaved solvers, but it penalises absolute errors, can overweight high-flow links, and enforces non-negativity only through constraints. In contrast, the entropy-based formulation in (3.2) naturally handles positivity, emphasises relative rather than absolute deviations, but it introduces logarithmic terms that complicate optimization, requires careful handling near zero, and is less intuitive to tune. In short, the quadratic model is easier to solve and interpret, while the entropy-based model provides more realistic behaviour for non-negative, flow-like quantities at the cost of increased numerical complexity.

Given that real OD estimates can legitimately contain zeros, reflecting absent or extremely rare flows, we adopt the l^2 formulation, which handles zero components gracefully and avoids the numerical and modelling issues inherent to entropy-based formulation.

It should be noted that in this thesis, OD demands (and later, link capacities in Chapter 4) are called input variables, and link flows are called output variables. These callings originate from the perspective of control theory: Input variables are the signals or actions applied to a system to influence its behaviour, while output variables are the measurable responses of the system that result from those inputs.

Since \mathbf{x} has a row dimension of IH , which is a huge number for large-scale urban networks, the computational complexity of 3.1 is very high even with the DDNA approach that \mathbf{A} can be considered as a constant matrix. It emphasizes the necessity of using dimensionality reduction techniques or improving computational algorithms to solve OD estimation problems in large-scale urban networks.

3.2 Methods for OD estimation

The literature on OD estimation typically addresses two main challenges arising from the optimization problem in (3.1). The first is determining the assignment matrix $\mathbf{A}(\mathbf{x})$, which generally depends on the travel demand

itself. The second is developing efficient methods for solving the resulting optimization problem.

Among the widely implemented methods in OD matrix estimation, one of the most popular methods is called bi-level optimization (Verbas et al., 2011). It includes an upper level, which minimizes the difference between observed and assigned link flows, and a lower level, which solves a DTA problem to determine link flows from the OD demands. In other words, the lower level performs traffic assignment to generate the assignment matrix $\mathbf{A}(\mathbf{x})$ iteratively. Bi-Level optimization has the advantage in capturing network dynamics and congestion effects, but it is computationally intensive, especially for large-scale networks (Patil et al., 2023).

Several studies have applied dimensionality reduction techniques (such as PCA and kernel PCA) to reduce the number of input variables, and further to improve computational efficiency of OD estimation in large urban networks. These methods reduce the complexity of high-dimensional OD demands while preserving accuracy, making real-time and large-scale estimation feasible (Peng et al., 2025, Djukic et al., 2014, and Qurashi et al., 2022).

Some recent studies use DDNA approaches for urban OD estimation, which often rely on large-scale mobility data such as GPS data (Tsanakas et al., 2023 and Fekih et al., 2021). This approach gives a constant assignment matrix, which eliminates the need to rerun traffic assignment models for every OD update. This drastically reduces computational cost, making OD estimation feasible in large-scale urban networks.

Chapter 4

Link capacity calibration

The network-wide link calibration problem is formulated as the problem of adjusting the initial (out-dated) link capacities, such that the assigned link flows match link flow observations, while keeping the Euclidean distance between the calibrated link capacities and the initial link capacities relatively small. It indicates that link flow measurements and initial link capacities from some data sources are necessary in a link capacity calibration problem for a given large-scale network.

4.1 Mathematical formulation

For an urban work with m links, consider a mathematical model in which there are m input variables (link capacities), x_1, \dots, x_m , and m output variables (link flows in a specific time period), y_1, \dots, y_m . The input vector and output vector are defined as

$$\mathbf{x} = (x_1, \dots, x_m) \in \mathbb{R}_+^{m \times 1}, \quad (4.1)$$

$$\mathbf{y} = (y_1, \dots, y_m) \in \mathbb{R}_+^{m \times 1}. \quad (4.2)$$

The corresponding link capacity calibration problem is formulated as

$$\begin{aligned} \min_{\mathbf{x}} F(\mathbf{x}) &= \gamma_1 \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \gamma_2 \|\tilde{\mathbf{f}}(\mathbf{x}) - \tilde{\mathbf{y}}\|_2^2, \\ \text{s.t. } \mathbf{x} &\geq \mathbf{0}, \end{aligned} \quad (4.3)$$

where the non-negative parameters γ_1 and γ_2 represent the weights for the corresponding terms. $\hat{\mathbf{x}} \in \mathbb{R}_+^{m \times 1}$ represents the initial capacities and $\tilde{\mathbf{y}} = (\tilde{y}_1, \dots, \tilde{y}_m) \in \mathbb{R}_+^{m \times 1}$ represents the observed link flows. The function $\tilde{\mathbf{f}} : \mathbb{R}_+^m \rightarrow \mathbb{R}_+^m$ represents the mapping from the link capacities to the link

flows. $\tilde{\mathbf{f}}(\mathbf{x})$ remains explicitly unknown due to the complexity between link capacities and link flows at the network level.

For the proposed calibration problem (4.3), high dimensionality (i.e., the number of links, m) not only leads to complexity in solving the optimization problem, but also greatly increases the time consumption to find the local analytical approximation of $\tilde{\mathbf{f}}(\mathbf{x})$. This demonstrates the importance of implementing dimensionality reduction techniques to reduce the variables in both input and output spaces.

4.2 Simulation-based optimization algorithms

Due to the lack of work focusing on network-wide link capacity calibration and the black-box property of the mapping from link capacities to link flows, a particular technique for OD estimation, known as the simulation-based optimization algorithms, is adopted and modified into the investigated link capacity calibration problem. This technique has been presented in Zhang et al. (2017), and Chong and Osorio (2018). This technique is designed to handle large-scale networks with a large number of links and time-varying OD demand.

This approach includes the following main steps:

- 1 A dynamic optimization problem is formulated with time-dependent input variables and analytical constraints.
- 2 A dynamic analytical network model is built to approximate traffic patterns. This metamodel is inspired by transient queuing theory and captures key dynamics such as congestion and flow propagation.
- 3 The model is used to evaluate the true performance of candidate solutions, which is known as the simulation-based evaluation. The simulation captures stochastic and nonlinear traffic phenomena that the metamodel simplifies.
- 4 Apply a gradient-based algorithm to optimize the input variables. Gradients are computed using the metamodel, which is differentiable and computationally efficient.
- 5 Steps 3 and 4 are done iteratively until the current estimate of input variables meets the convergence criteria.

The method shows the strengths of both analytical modelling (computationally fast and differentiability) and simulation (accurate and stochastic). This approach is very suitable for the cases where the mapping from the input variables to the output variables is not explicitly known, including both the investigated OD estimation problems in the paper and the network-wide link capacity calibration problems to be investigated in this thesis.

4.3 Framework for network-wide link capacity calibration

Enlightened by the work by Zhang et al. (2017), and Chong and Osorio (2018), the framework of network-wide link capacity calibration can be formulated as follows, which includes following steps. The first step is to find the local analytical approximation of the simulator, which maps link capacities to link flows. In this step, measurement equations are constructed to relate the available synthetic observations (link flows) to the simulator's calibration parameters (link capacities). A common analytical simplification is to assume a linear relationship, which requires estimating the Jacobian matrix $\frac{\partial(y_1, \dots, y_m)}{\partial(x_1, \dots, x_m)}$ in a neighbourhood of the current estimate. Once the analytical approximation is constructed, an auxiliary solution point can be found based on Equation (4.3), and the last step is solution update through iterations.

The first step aims to obtain a numerical approximation of the Jacobian matrix $\mathbf{J} \in \mathbb{R}^{m \times m}$, whose (i, j) -th entry is defined as $\mathbf{J}_{ij} = \frac{\partial y_i}{\partial x_j}$, evaluated at the estimate \mathbf{x}^{k-1} , where k denotes the current iteration index and $\mathbf{x}^0 = \hat{\mathbf{x}}$. The Jacobian characterizes how a small perturbation in the j -th input variable x_j influences the i -th output variable y_i .

To estimate \mathbf{J} at \mathbf{x}^{k-1} , a simulation-based procedure is used: The current input estimate \mathbf{x}^{k-1} is encoded in the system (network), then perturbed randomly within a narrow range, and the resulting outputs are recorded. For each input variable x_j , its perturbation in the sensitivity analysis, denoted by Δx_j , is drawn from a uniform distribution over the interval $[-\delta_k x_j^{k-1}, \delta_k x_j^{k-1}]$, where δ_k is the variation parameter and x_j^{k-1} is the estimated value of the variable x_j obtained from $k - 1$ -th iteration.

The local Jacobian matrix is estimated by using the method of regression analysis. A regression model is built to map inputs to outputs, and its resulting coefficients serve as approximate values of the Jacobian. More mathematically, suppose that in a given iteration k , we generate n trial points (equivalently, n observations). The input matrix $\mathbf{X} \in \mathbb{R}_+^{n \times m}$ is then defined as

$$\mathbf{X} = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_m^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_m^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(n)} & x_2^{(n)} & \dots & x_m^{(n)} \end{pmatrix}, \quad (4.4)$$

where $x_i^{(l)}$ represents the x_i value of the l -th trial point. The iteration index k is omitted here for readability. Similarly, the output matrix $\mathbf{Y} \in \mathbb{R}_+^{n \times m}$

can be defined as

$$\mathbf{Y} = \begin{pmatrix} y_1^{(1)} & y_2^{(1)} & \dots & y_m^{(1)} \\ y_1^{(2)} & y_2^{(2)} & \dots & y_m^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ y_1^{(n)} & y_2^{(n)} & \dots & y_m^{(n)} \end{pmatrix}, \quad (4.5)$$

where $y_i^{(l)}$ represents the value of y_i obtained from inputting the l -th trial point $\mathbf{x}^{(l)}$ into the simulator. A regression problem then arises, asking for finding the Jacobian matrix estimate $\hat{\mathbf{J}} \in \mathbb{R}^{m \times m}$ and the constant vector \mathbf{b} , such that:

$$\mathbf{Y}^T \approx \hat{\mathbf{J}} \mathbf{X}^T + \mathbf{B}, \quad (4.6)$$

where $\hat{\mathbf{J}}$ approximates the local Jacobian matrix at the current input estimate \mathbf{x}^{k-1} ,

$$\hat{\mathbf{J}} \approx \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \dots & \frac{\partial y_1}{\partial x_m} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \dots & \frac{\partial y_2}{\partial x_m} \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \dots & \frac{\partial y_m}{\partial x_m} \end{pmatrix}_{\mathbf{x}^{k-1}}, \quad (4.7)$$

and $\mathbf{B} \in \mathbb{R}^{m \times n}$ is the constant matrix:

$$\mathbf{B} = \begin{pmatrix} b_1 & b_1 & \dots & b_1 \\ b_2 & b_2 & \dots & b_2 \\ \vdots & \vdots & \vdots & \vdots \\ b_m & b_m & \dots & b_m \end{pmatrix} = (\mathbf{b} \ \mathbf{b} \ \dots \ \mathbf{b}), \quad (4.8)$$

where $\mathbf{b} = (b_1, b_2, \dots, b_m) \in \mathbb{R}^{m \times 1}$.

Substituting the explicit forms of \mathbf{X} , \mathbf{Y} , $\hat{\mathbf{J}}$, and \mathbf{B} into (4.6) yields

$$\begin{pmatrix} y_1^{(1)} & y_1^{(2)} & \dots & y_1^{(n)} \\ y_2^{(1)} & y_2^{(2)} & \dots & y_2^{(n)} \\ \vdots & \vdots & \vdots & \vdots \\ y_m^{(1)} & y_m^{(2)} & \dots & y_m^{(n)} \end{pmatrix} \approx \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \dots & \frac{\partial y_1}{\partial x_m} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \dots & \frac{\partial y_2}{\partial x_m} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \dots & \frac{\partial y_m}{\partial x_m} \end{pmatrix} \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & \dots & x_1^{(n)} \\ x_2^{(1)} & x_2^{(2)} & \dots & x_2^{(n)} \\ \vdots & \vdots & \vdots & \vdots \\ x_m^{(1)} & x_m^{(2)} & \dots & x_m^{(n)} \end{pmatrix} + \begin{pmatrix} b_1 & b_1 & \dots & b_1 \\ b_2 & b_2 & \dots & b_2 \\ \vdots & \vdots & \vdots & \vdots \\ b_m & b_m & \dots & b_m \end{pmatrix}. \quad (4.9)$$

Equation (4.9) is equivalent to the component-wise regression model

$$y_i^{(l)} \approx \frac{\partial y_i}{\partial x_1} \Big|_{\mathbf{x}^{k-1}} x_1^{(l)} + \frac{\partial y_i}{\partial x_2} \Big|_{\mathbf{x}^{k-1}} x_2^{(l)} + \dots + \frac{\partial y_i}{\partial x_n} \Big|_{\mathbf{x}^{k-1}} x_n^{(l)} + b_i, \quad (4.10)$$

$$\forall i = 1, 2, \dots, m, \quad \forall l = 1, 2, \dots, n, \quad \forall k = 1, 2, \dots$$

Thus, estimating the Jacobian matrix $\hat{\mathbf{J}}$ and the intercept vector \mathbf{b} (or equivalently, the matrix \mathbf{B}) provides the linear regression model that will be used in the subsequent steps.

Given the approximate Jacobian matrix $\hat{\mathbf{J}}$ and the constant vector \mathbf{b} , the following optimization problem, derived from (4.3), is solved to get the newest input estimate:

$$\min_{\mathbf{x}} F(\mathbf{x}) = \gamma_1 \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \gamma_2 \sum_i (\tilde{y}_i - \sum_j \hat{\mathbf{J}}_{ij} x_j - b_i)^2, \quad (4.11)$$

$$s.t. \quad \mathbf{x} \geq \mathbf{0},$$

where $\hat{\mathbf{J}}_{ij}$ represents the (i, j) -th entry of $\hat{\mathbf{J}}$.

The multivariate linear regression used above is only mathematically valid when the underlying relationship is linear. Consequently, when estimating the local Jacobian, each variable must vary only within a sufficiently small neighbourhood to preserve this linearity assumption. This also implies that the resulting optimal solution is a local optimum rather than a global one. To address this limitation, the procedure is repeated iteratively: at each iteration, new trial points are generated, the linear regression model in (4.6) is constructed, and the corresponding optimization problem (4.11) is solved. The mathematical conditions that justify iteratively repeating this procedure are presented in Spall (1992).

Chapter 5

Dimensionality reduction

Dimensionality reduction is the process of transforming data from a high-dimensional space into a lower-dimensional space while preserving the most meaningful properties of the original data. OD estimation and link capacity calibration in large-scale networks often involve recovering a large number of input variables (OD demands, link capacities) from a limited set of observed link flows. This creates an ill-posed inverse problem, where the number of unknowns far exceeds the number of equations. Dimensionality reduction helps by reducing computational complexity, improving stability and robustness, and enabling estimation with minimal prior information due to the reduction of input variables. The essence of dimensionality reduction is to identify dominant patterns in input variables by projecting them onto a lower-dimensional space. It is useful for capturing temporal or spatial correlations between input variables.

Dimensionality reduction techniques in literature in various research fields span linear, nonlinear, and feature selection based approaches. They include classical methods like PCA, independent component analysis (ICA), partial least squares regression (PLSR), and NNMF, manifold learning methods such as uniform manifold approximation and projection (UMAP), and modern approaches such as random forests. Comprehensive references on dimensionality reduction techniques include Geladi and Kowalski (1986) and Sorzano et al. (2014).

In this thesis, three different dimensionality reduction techniques are introduced and investigated for link capacity calibration and OD estimation problems: PCA, PLSR and NNMF. All of these three approaches share very similar ideas, and the differences are in how the principal directions (i.e., latent variables) are determined.

5.1 Principal component analysis

PCA is widely used as a dimensionality reduction technique in OD estimation (Qurashi et al., 2022 and Cao et al., 2024). The core idea of PCA in OD estimation is to gather historical OD demands to stack them into a matrix form (known as a data matrix), in which rows represent different data pieces and columns represent values of each input variable (OD demand), and then find the eigenvectors (i.e., principal components) and eigenvalues (amount of variance explained by each component) of the covariance matrix calculated from the standardized data matrix. The top k principal components can be used to reconstruct or estimate OD demands, where k is a much smaller number compared to the number of OD demand terms. In short, PCA captures latent structure in OD demand and it can handle large OD matrices efficiently.

5.2 Partial least squares regression

PLSR is a powerful statistical technique used to model complex relationships between input and output variables, especially when input variables are highly collinear or when the number of input variables exceeds the number of observations. Unlike PCA, which ignores the output variables, PLSR finds components (i.e., latent variables) that not only capture variance in the input variables but also maximize their covariance with the output variables. Frank and Friedman (1993) gives a statistical and geometrical interpretation of PLSR.

PLSR works by projecting both the input data matrix and the output data matrix into a new space. It identifies new in- and output variables (latent variables) that are linear combinations of the original high-dimensional in- and output variables. These latent variables are chosen to maximize the covariance between the input data matrix and the output data matrix. This process of extracting latent variables is iterative and continues until a predefined number of components is extracted or performance criteria are met. Detailed mathematical formulation and relevant algorithms can be found in Wold et al. (2001) and Geladi and Kowalski (1986). PLSR has been widely applied in the field of chemometrics, e.g., Höskuldsson (2001), Kvalheim (2010), Godoy et al. (2014), Hanrahan et al. (2005) and Martens (2001), but to the best of our knowledge, it had not been explored by researchers in transportation before the work of this thesis.

In link capacity calibration in large-scale urban networks, PLSR can be implemented to identify latent components that explain both the inputs (link capacities) and the outputs (link flows). These components are used to estimate adjusted link capacities.

5.3 Non-negative matrix factorization

NNMF is a dimensionality reduction technique similar to PCA. The only difference is that the latent components and scores obtained by decomposing the data matrix only have non-negative element values. Although NNMF and PCA share the similar idea in dimensionality reduction, their mathematical framework is very different. In NNMF, uniqueness of factorization result cannot be guaranteed, since it depends heavily on initialization and high dimensional sparse data often has multiple nearly optimal decompositions. One of the most popular algorithms for NNMF is provided by Lee and Seung (2001).

Several studies have applied NNMF to estimation and calibration problems in transportation fields, particularly for OD estimation, traffic data recovery, and spatio-temporal traffic pattern analysis. Atif et al. (2021) proposes a structured NNMF approach for estimating OD traffic flows in large-scale networks. The results show that NNMF can solve the ill-posed inverse problem of OD estimation by enforcing non-negativity and structure constraints. Yu et al. (2023) uses NNMF to recover missing or corrupted traffic data in intelligent transportation systems (ITS), which improves calibration accuracy by handling noise. Wang et al., 2022 uses NNMF in urban traffic pattern analysis to decompose large-scale traffic states into meaningful components for calibration and prediction.

5.4 Importance of dimensionality reduction

The significance of using dimensionality reduction techniques in traffic related estimation and calibration problems does not only include reducing computational cost, but also helps in another two perspectives: First, it could enable a better understanding of the problem structure, such as visualizing clusters or manifolds, detecting bottlenecks or symmetries, and identifying which variables matter most. Second, it could convert an underdetermined problem to an (over-)determined problem with limited observations by eliminating degrees of freedom.

Chapter 6

Research challenges, questions and methodology

This chapter outlines the two main research challenges in OD estimation and link capacity calibration, which together motivate the formulation of the research objectives and research questions. Building on these foundations, the chapter concludes by presenting the overall research methodology.

6.1 Research challenges

OD estimation and link capacity calibration in large-scale urban transportation networks are complex, and previous research work, such as Zhang et al. (2025) and Klein (2019), has identified two key research challenges: Data fusion and computational efficiency.

- 1 Data fusion: Various data types (e.g., GPS data, traffic sensor data, and mobile network data) are available in urban networks, but integrating these effectively into link capacity calibration and OD estimation methods remains a challenge from two perspectives: First, data from various sources usually share different temporal and spatial resolutions. The other is that it is not easy to construct the mapping from the input variables, i.e., OD demands or link capacities, to the observed data.
- 2 Computational efficiency: Traditional calibration and estimation methods struggle with large-scale urban networks due to the sheer number

of variables, for example, one per road segment for link capacity calibration. For OD estimation, the difficulty increases substantially when temporal patterns are incorporated, as this introduces an additional time-dependent dimension of variables that must be inferred alongside the spatial ones.

6.1.1 Data fusion

Data fusion in urban network estimation and calibration problems, where multiple data sources are integrated to improve the accuracy of the estimate from the traffic model, offers powerful benefits. However, it also introduces several challenges: The first challenge is temporal and spatial misalignment. Data streams may not be synchronized in time (e.g., GPS data vs. hourly link counts) or aligned spatially (e.g., different partitions of the traffic analysis zones from the traffic planning model and GPS data). This misalignment can worsen the calibration or estimation results unless corrected through a proper way of integration. Another challenge is that constructing a reliable mapping from the input variables to the observed traffic data is inherently difficult.

6.1.2 Computational efficiency

The high dimensionality of urban network calibration problems leads to the challenge of calibrating models with a vast number of parameters, often one for each road segment or OD pair, while having limited observational data and computational resources. This number is even larger when the temporal pattern is taken into account. High dimensionality leads to two main issues: The first one is sparse observations. Typically, only a small subset of traffic flows or travel times is observed, making it difficult to accurately estimate all parameters, or in other words, the corresponding optimization problem is under-determined. The other issue from high dimensionality is computational bottlenecks. Full Jacobian estimation or brute-force optimization through the calibration process become infeasible due to the sheer number of variables.

Simultaneous perturbation stochastic approximation (SPSA) is a popular method for calibrating urban traffic models, especially in high-dimensional settings. However, SPSA can struggle to converge in high-dimensional spaces, especially where input variables (OD demands and link capacities) are weakly correlated or noisy (Kostic et al., 2017 and Antoniou et al., 2015).

Dimensionality reduction approaches such as PCA have been widely used in urban network calibration, but other approaches such as PLSR and NNMF are rarely seen in urban network calibration problem research. PLS often achieves a lower value of mean square error (MSE) than principal component regression (PCR), especially when the output variables are correlated with directions in the data that have low variance. In addition, PCR might discard components that are predictive but have low variance, while PLS retains

them due to their relevance to the dependent variables. Moreover, PLS is more effective when the number of independent variables exceeds the number of observations (Liu et al., 2022). NNMF respects the natural constraints of many traffic related datasets (e.g., OD demands, link flows), which are inherently non-negative, while PCA does not enforce non-negativity, which requires that the non-negativity should be added as constraints.

Although there are various investigated approaches in improving the computational efficiency in urban network calibration and estimation problems, such as stochastic optimization (Lu et al., 2015 and Antoniou et al., 2015), simulation-based calibration (Osorio et al., 2011 and Zhang et al., 2025) and dimensionality reduction (Qurashi et al., 2022), it has been rarely seen that there is research work focusing on improving the numerical algorithm in solving a given optimization formulation. Considering the mathematical properties in urban network estimation and calibration problems, such as sparsity of the mapping from input variables to link flows, non-negativity, and much more variables than observations, a specific numerical solver can be developed to improve the computational efficiency. Generic solvers in commercial software, such as *lsnonneg* in MATLAB, are slow since the mathematical properties of large-scale urban network calibration and estimation problems, such as scarcity of the mapping, non-negativity of the variables and convexity of the optimization formulation, have not been taken into consideration.

6.2 Research objectives and research questions

Based on the two main research challenges mentioned in Section 6.1, this thesis focuses on the following two main research objectives.

The first objective is to fuse multiple sources of data. In some OD estimation problems, data of OD demands with different temporal and spatial patterns are provided, and it is necessary to build up a new target demand by fusing information from multiple data sources. Moreover, when turning proportion data is available, which indicates the share of vehicles choosing each turning movement at an intersection, the OD estimation method can be refined to produce more reliable and accurate results. The question is how to find a mapping from the OD demands to the turning proportion observations so that the corresponding term can be embedded in the optimization formulation.

The second objective is to improve the computational efficiency in large-scale urban network calibration and estimation. This can be done in different ways, such as using dimensionality reduction approaches to reduce the number of variables or developing an optimization solver in which mathematical properties of the estimation or calibration problem are taken into consideration.

The quality of the results obtained from the OD estimation and link capacity calibration process is fundamentally shaped by the ability to integrate heterogeneous data sources and to exploit structural information embedded in the transportation network. By fusing multiple datasets, the OD estimation framework can generate demand patterns that more accurately reflect real-world traffic dynamics. Incorporating turning movement observations is particularly valuable, as it constrains the solution space and enhances the identifiability of OD demands, thereby improving both the reliability and the granularity of the resulting estimates. At the same time, the quality of the estimation or calibration in large-scale urban networks depends critically on the computational tractability of the underlying optimization problem. Approaches that reduce dimensionality, simplify the functional relationships between inputs and outputs, or leverage mathematical properties of the estimation formulation contribute to more stable and efficient solution procedures. These methodological improvements not only accelerate computation but also reduce the likelihood of convergence to suboptimal or noisy solutions. In general, the combination of data fusion and computationally informed model design leads to estimation and calibration results that are more robust, interpretable, and operationally meaningful for large-scale traffic systems.

The following two research questions are formulated to achieve the two objectives:

- **RQ1: How can data from multiple sources be fused to improve the quality of transport model input data?**
- **RQ2: How to improve the computational efficiency in large-scale urban network estimation and calibration problems?**

Among the six included papers, Paper I addresses RQ2 by integrating a simulation-based optimization algorithm with PLSR. Paper II also answers RQ2 by applying DDNA to OD estimation. Papers III and IV respond to RQ1, progressively incorporating additional data types based on Paper II. Paper V returns to RQ2 by combining DDNA with NNMF, while Paper VI answers RQ2 by introducing an efficient numerical solver.

6.3 Research methodology

The methodology employed in this thesis is fundamentally quantitative and utilizes simulated or empirical traffic data. The analytical framework is built around linear formulations designed to ensure computational efficiency and scalability to large urban networks. Across all components of the work, a central objective is to maintain internal consistency between OD demands or link capacities, and observed link flows, based on evaluations conducted using simulated or empirical data.

The methods for both OD estimation and link capacity calibration problems in this thesis include two major steps: The first one is to build up

the corresponding mathematical formulations, depending on the source of data and what methods are adopted. The second step is to develop efficient computational algorithms for the two problems.

Since there are two major research questions (RQ1 and RQ2) to be answered, the research methods for them are illustrated separately.

To achieve fusion of multiple sources of data, two different approaches for different data types are proposed: 1) For the given target demands with different spatial and temporal patterns, two mathematical formulae are provided for time slicing OD matrices. In this research, mobile network data, a 24-hour OD matrix from a traffic planning model, vehicle probe data, and link count data are used (Paper III). 2) With data on the turning proportion provided, a new optimization formulation is constructed, with new terms minimizing the distance between the observed and calculated turning proportion from the route choice model and the assignment matrix from the DDNA approach (Paper VI).

To improve computational efficiency in large urban network calibration and estimation problems, three different approaches are investigated: 1) Using dimensionality reduction techniques. PLS regression is investigated and implemented in Paper I for link capacity calibration and NNMF is investigated and implemented in Paper V for OD estimation, with the results being compared to a corresponding benchmark method, respectively. In both works, not only a low dimensionality representation needs to be found from historical simulation results, but also a new optimization problem needs to be formulated accordingly in the low-dimensional space. 2) Using DDNA to achieve an efficient optimization approach in OD estimation. Papers II, III, and IV present a DDNA-based approach which provides an exogenous linear mapping between OD demand and link flows, and it can be utilised for efficient OD estimation. 3) Developing a new algorithm based on interior-point method and Newton's method to solve the high-dimensional OD estimation problem, which is presented in Paper VI. In this new algorithm, properties of OD calibration with DDNA approaches are considered, including non-negativity, convexity and sparsity of the assignment matrix. The non-negativity constraints are transformed into a penalty function added in the optimization formulation.

Figure 6.1 provides a visual overview of how each paper addresses the research questions (RQ1 and RQ2), along with the key methodologies employed. The figure organises the structure of the thesis into three aligned columns. The left column lists the two overarching research questions that guide the work. The middle column maps each of the six papers to the research question(s) they address, showing how the contributions built from methodological development to applications. The right column indicates the type of evaluation method for each paper: Paper I focuses on simulation experiments to validate methodological components, while the remaining papers rely on empirical data to demonstrate method performance and practical relevance.

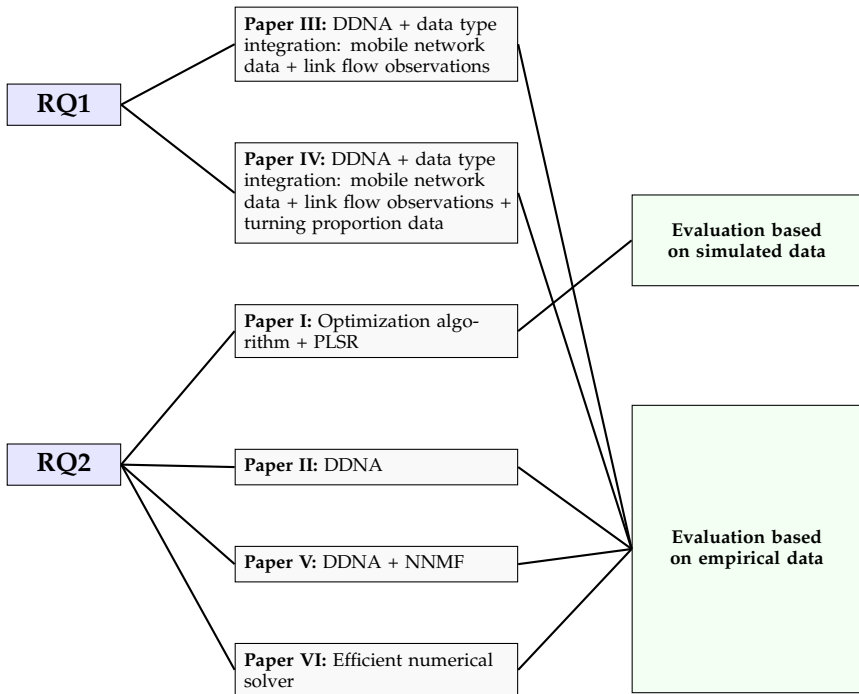


Figure 6.1: Mapping of papers to research questions and evaluation.

Chapter 7

Contributions of the thesis

7.1 Summary of papers

Six research papers are included in this thesis, and they are aligned with the research objectives. The description of each paper is given below.

Paper I: Network-wide Calibration of Link Capacities for Dynamic Traffic Assignment Models

In Paper I, a new approach is developed for calibrating link capacities across large urban networks using flow observations from only a subset of links. The paper begins by utilising a dynamic traffic assignment model in MATSim, where link capacities are treated as unknown parameters to be inferred. To relate simulated link flows to these capacities, the paper introduces a regression-based framework built on PLSR, which is well suited for handling high dimensionality in network-wide calibration, combined with simulation-based optimization algorithms. Trial simulations are generated to train the PLSR model, enabling efficient prediction of capacities for all links in the network from synthetic flow data. The method is then benchmarked against the SPSA algorithm, demonstrating substantially improved computational efficiency and strong accuracy even when only limited flow observations are available. Overall, the results show that the proposed PLSR-based approach offers a practical and scalable alternative for network-wide link capacity calibration.

Paper I is co-authored by Clas Rydbergren, David Gundlegård, Joakim Ek-

ström, and Gunnar Flötteröd. The author of this thesis contributed to the mathematical formulation and the programming of the whole proposed method, design of numerical experiments and analysis of the results. The author also took the leading role of writing.

Paper I has been published in:

Wei, G., Rydergren, C., Gundlegård, D., Ekström, J. and Flötteröd, G., 2025. Network-Wide Calibration of Link Capacities for Dynamic Traffic Assignment Models. *Journal of Advanced Transportation*, 2025(1), p.8854907.

Contents of Paper I have been presented at the *Swedish Transportation Research Conference - STRC 2020, 2021 and 2022*, and at *hEART 2022: 10th Symposium of the European Association for Research in Transportation*, June 1-3, 2022, KU Leuven, Belgium. An earlier version of this paper was presented as the main author's Licentiate thesis in Linköping University, 2022 (Wei, 2022).

Paper II: Consistent OD and link flow estimation based on data-driven network assignment

Paper II investigates a data-driven network assignment approach that replaces traditional model-based route choice with an empirically derived assignment matrix constructed from GPS probe data. This matrix captures how trips actually propagate through the network, reflecting observed route preferences and congestion patterns, and enables a linear mapping from OD demands to link flows that greatly simplifies the estimation task. Building on this representation, the paper formulates a joint optimization framework that estimates OD demands and link flows in a consistent manner using observed link counts. The method is applied to a real-world case study in Stockholm, where its explanatory power is evaluated on both training and test datasets. The results highlight the importance of carefully calibrating the weight parameters in the OD estimation step and the Logit parameter governing route choice: Too little weight on the target OD leads to overfitting and high variance, while too much weight introduces bias. Overall, the study demonstrates that the data-driven assignment approach can provide reliable and interpretable estimates when these parameters are properly tuned.

Paper II is co-authored by David Gundlegård, and Clas Rydergren. The author of this thesis contributed to mathematical optimization formulation, design of numerical experiments and analysis of the results. The author also took the leading role of writing.

Paper II has been published in:

Wei, G., Gundlegård, D. and Rydergren, C., 2025. Consistent origin-destination and link flow estimation based on data-driven network assignment. *Transportation Research Procedia*, 86, pp.668-675.

Contents of Paper II have been presented at the 26th EURO Working Group on Transportation Conference (Lund, Sweden, 2024).

Paper III: Time Slicing Origin-Destination Matrices Using Mobile Network Data, Link Counts and Vehicle Probe Data

Paper III develops a data-driven approach for generating time-sliced OD matrices by integrating mobile network data, link counts, and vehicle probe data within the data-driven network assignment framework. The method begins by combining a 24-hour OD demand from a transport planning model with hourly mobile network data to construct target hourly demands for the traffic analysis zone pairs. Vehicle probe data are then used to estimate hourly assignment matrices, capturing how traffic distributes across the network throughout the day. These components are brought together in an OD estimation step that produces consistent hourly OD matrices aligned with both the planning model and observed traffic patterns. The approach is validated using real-world data from Norrköping, Sweden, and shows strong agreement between estimated and observed link flows across training and test sets. Compared with benchmark methods, the proposed framework delivers substantially improved performance, preserves the structural characteristics of the mobile network data and transport model, and provides OD estimates suitable for detailed traffic analysis and dynamic modelling applications.

Paper III is co-authored by David Gundlegård, and Clas Rydergren. The author of this thesis played a key role in developing the mathematical framework, designing the numerical experiments, and analysing the outcomes. In addition, the author was primarily responsible for writing the manuscript.

Paper III has been published in:

Wei, G., Gundlegård, D. and Rydergren, C., 2026. Time Slicing Origin-Destination Matrices Using Mobile Network Data, Link Counts and Vehicle Probe Data. *Transportation Research Procedia*, 95, pp.928-935.

Contents of Paper III have been presented at the 27th EURO Working Group on Transportation Conference (Edinburgh, UK, 2025).

Paper IV: Leveraging Turning Proportion Data for Origin-Destination Estimation in Urban Networks

Paper IV presents an extended OD estimation approach that incorporates turning proportion data into the data-driven network assignment framework. Building on earlier studies that integrate multiple data sources, such as link counts, GPS trajectories, and mobile network data, the method in this paper adds measured turning proportions as additional penalty terms, enabling the model to match both link-level traffic volumes and intersection-level turning behaviour within a single optimization formulation. The approach is implemented on the urban road network of Norrköping, Sweden, using real traffic counts and probe-derived turning proportions extracted from GPS-equipped vehicles. The results show that including turning movement information preserves the model’s ability to accurately reproduce observed link flows while simultaneously achieving a close fit to the turning proportions. This demonstrates that the enhanced formulation can effectively utilise widely available probe-based turning data to generate OD estimates that more correctly reflect real intersection-level traffic patterns.

Paper IV is co-authored by David Gundlegård, and Clas Rydergren. The author of this thesis played a key role in formulating the mathematical framework and analysis, running the numerical experiments, and assessing the numerical results. The author also took primary responsibility for drafting the manuscript.

Paper V: Combining Data-driven Network Assignment and Non-negative Matrix Factorization for Origin-Destination Estimation in Urban Networks

Paper V proposes a computationally efficient OD estimation framework that combines link count data and vehicle probe information within the data-driven network assignment model, supported by a non-negative matrix factorization representation. Historical traffic data are used to construct a low-dimensional basis for OD demand, allowing the estimation problem to be solved more rapidly while retaining the dominant structural patterns of real-world travel behaviour. The method is applied to empirical data from central Stockholm, where it produces OD matrices at high computational speed and yields link flow estimates that closely match observed counts in both training and test evaluations.

The approach consists of two main components. First, OD demands are estimated without dimensionality reduction for a set of historical days, and these estimates are used to derive the NNMF-based low-rank structure. Second, this low-dimensional representation is employed to estimate hourly OD patterns for new days, using the set of latent variables. The results show that the low-dimensional formulation improves predictive performance on test links by mitigating overfitting through a more compact feature space. Although some loss of detail is expected compared with the full dimensional model, the substantial gains in computational efficiency make the method

well suited for large-scale OD estimation problems.

Paper V is co-authored by David Gundlegård, Rasmus Ringdahl, and Clas Rydbergren. The author of this thesis contributed to the mathematical formulation and the programming of the whole proposed method, design of numerical experiments and analysis of the results. The author also took the leading role of writing.

Paper V has been published in: Wei, G., Gundlegård, D., Ringdahl, R. and Rydbergren, C., 2025, September. Combining Data-Driven Network Assignment and Non-Negative Matrix Factorization for Origin-Destination Estimation in Urban Networks. *In 2025 9th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS) (pp. 1-6). IEEE.*

Contents of Paper V have been presented at 9th International Conference on Models and Technologies for Intelligent Transportation Systems (Luxemburg, Luxemburg, 2025).

Paper VI: A Scalable Interior-Point Method for OD Estimation based on Data-Driven Network Assignment in Urban Networks

This paper introduces a scalable algorithm for origin–destination (OD) estimation in large urban networks, built around a customised interior-point optimization framework. The key innovation lies in designing an interior-point method that exploits the structural properties of the OD estimation problem, enabling substantial gains in both computational speed and memory efficiency. Central to the approach is an efficient procedure for solving the Newton system within each interior-point iteration, which removes the scalability bottlenecks that typically limit OD estimation in medium- and large-scale networks. The algorithm is evaluated through two urban case studies, demonstrating significant reductions in computation time while maintaining high-quality solutions. By alleviating the heavy memory demands of traditional methods, the proposed approach offers a practical and robust solution for OD estimation across networks of varying size and complexity, making it well suited for modern, data-rich urban traffic systems.

Paper VI is co-authored by David Gundlegård and Marian Codreanu. The author of this thesis played a key role in formulating the mathematical framework and analysis, designing the algorithm, and assessing the numerical results. The author also took primary responsibility for drafting the manuscript.

7.2 Scientific contributions

This thesis advances the field of traffic modelling and analytics by developing a set of data-driven, computationally scalable methods that address two core challenges: integrating traffic data, and solving large-scale estimation or calibration problems efficiently.

The thesis addresses RQ1 by demonstrating that DDNA offers a systematic way to fuse heterogeneous data sources into a unified framework, thereby enhancing the quality of transport model input data. By deriving assignment relationships directly from empirical observations, GPS trajectories, mobile network data, link counts, and turning movements, the thesis shows that realistic traffic patterns can be captured without relying on behavioural models that can be difficult to calibrate. This data-driven mapping enables consistent and computationally tractable OD estimation, and the framework is extended to incorporate additional data sources such as turning proportions and mobile network data. These extensions allow the resulting OD matrices to reproduce not only link-level volumes but also intersection-level turning dynamics, demonstrating the versatility of the DDNA approach.

The thesis answers RQ2 by showing that computational efficiency in large-scale urban network estimation and calibration can be substantially improved through the use of low-dimensional representations. Two complementary methods, PLSR and NNMF, are developed to reduce the complexity of both capacity calibration and OD estimation. PLSR is used to extract the dominant relationships between link flows and link capacities, enabling efficient network-wide capacity inference from sparse observations. NNMF, in turn, identifies low-dimensional structure in historical OD patterns, allowing new OD demands to be estimated rapidly while mitigating overfitting. Together, these techniques demonstrate how low-rank representations can preserve essential traffic patterns while dramatically reducing computational burden, making large-scale estimation feasible in operational settings. Complementing these methods, the thesis also contributes to the computational scalability of OD estimation by introducing a customised interior-point algorithm tailored to the structure of the problem. By designing an efficient solver for the Newton system at the core of the interior-point iterations, the method overcomes the memory and runtime limitations that typically restrict OD estimation to small or medium-sized networks. Case studies show that the algorithm delivers substantial reductions in computation time while maintaining high solution quality, enabling fast and reliable estimation across networks of varying size and complexity.

Together, the contributions form a coherent methodological framework that integrates diverse data sources, leverages data-driven assignment principles, and exploits low-dimensional structures to achieve computational scalability. The result is a suite of computationally efficient tools capable of producing accurate and consistent OD and capacity estimates for urban traffic systems.

Chapter 8

Conclusions and future work

8.1 Conclusions

This thesis demonstrates that progress on OD estimation and link capacity calibration for large-scale urban networks requires both methodological innovation and strategic data integration. The thesis shows that fusing mobile network data, probe trajectories, traffic counts, turning proportion data, and data-driven behavioural models yields richer and more reliable OD estimates. The thesis also illustrates that computational efficiency can be achieved through gradient-based calibration, data-driven network assignment models, numerical algorithm development, and dimensionality reduction techniques. Together, these contributions help bridge the gap between emerging data sources and practical modelling needs, potentially supporting both scientific inquiry and real-world transportation planning.

8.2 Future work directions

Future research can build on the contributions of this thesis by addressing several methodological and practical limitations that remain in the integration of heterogeneous data sources and the efficient OD estimation and link capacity calibration of large-scale traffic models.

One important direction concerns the limited diversity and representativeness of data sources used in current OD estimation frameworks. While the thesis successfully combine mobile network data, probe trajectories, turning proportion data and traffic counts, they do not incorporate emerging data streams such as connected vehicle data. Future work could explore how these

additional sources, each with distinct biases, sampling rates, and spatial coverage, can be harmonized within unified estimation frameworks. This may not only improve OD demand accuracy but also enhance the robustness of models in mixed-traffic environments, where multiple types of vehicles and technologies coexist on the road at the same time.

A second avenue for advancement lies in the behavioural assumptions of data-driven assignment methods. The current approaches rely on route choice patterns extracted from data, but they do not fully capture the dynamic behavioural responses of travellers to congestion, incidents, or policy interventions. Future research could integrate machine learning-based behavioural models or hybrid simulation-learning frameworks that allow route choice to adapt dynamically to changing network conditions. Such developments would help bridge the gap between static or quasi-static data-driven assignment and fully dynamic traffic behaviour, improving the applicability of these methods in real-time traffic management and scenario analysis.

Another limitation concerns the scalability and computational demands of the proposed methods when applied to very large metropolitan networks or high temporal resolutions. Although the studies introduce several efficiency-enhancing techniques, such as gradient-based calibration and non-negative matrix factorization, there remains a need for methods that can operate at city-wide scale with minute-level temporal granularity. Future research could explore parallel computing or surrogate modelling approaches with significantly reduced computational cost. Additionally, online or incremental estimation methods could be developed to update OD demands continuously as new data streams arrive, enabling real-time applications.

One more useful direction for future work is to explore entropy-based methods for link capacity calibration. Because link capacities must always be positive, an entropy-maximising approach can naturally enforce this requirement while avoiding unrealistic or extreme values. Adding an entropy term to the calibration problem could help stabilise the estimates when flow data are sparse or noisy, and prevent the model from overfitting. Studying how such entropy-based ideas interact with simulation-based models could lead to more reliable and practical methods for calibrating link capacities in large networks.

Finally, this thesis highlights the potential of data fusion but do not fully address the uncertainty quantification associated with combining heterogeneous data sources. Mobile network data, probe data, and traffic counts each carry different types of noise, sampling biases, and temporal inconsistencies. Future research could incorporate probabilistic modeling or Bayesian inference to explicitly quantify and propagate uncertainty through the OD estimation and link capacity calibration process. Such approaches would provide practitioners with confidence intervals or reliability measures, enhancing the interpretability and operational value of the resulting models.

Bibliography

- Antoniou, C., Azevedo, C. L., Lu, L., Pereira, F., and Ben-Akiva, M. (2015). “W-SPSA in Practice: Approximation of Weight Matrices and Calibration of Traffic Simulation Models”. In: *Transportation Research Procedia* 7. 21st International Symposium on Transportation and Traffic Theory Kobe, Japan, 5-7 August, 2015, pp. 233–253. URL: <https://www.sciencedirect.com/science/article/pii/S2352146515000812>.
- Atif, S. M., Gillis, N., Qazi, S., and Naseem, I. (2021). “Structured non-negative matrix factorization for traffic flow estimation of large cloud networks”. In: *Computer Networks* 201, p. 108564.
- Aw, A. and Rascle, M. (2000). “Resurrection of ”Second Order” Models of Traffic Flow”. In: *SIAM Journal on Applied Mathematics* 60.3, pp. 916–938. eprint: <https://doi.org/10.1137/S0036139997332099>. URL: <https://doi.org/10.1137/S0036139997332099>.
- Bigazzi, A. Y. and Clifton, K. J. (2015). “Modeling the effects of congestion on fuel economy for advanced power train vehicles”. In: *Transportation Planning and Technology* 38.2, pp. 149–161. eprint: <https://doi.org/10.1080/03081060.2014.997449>. URL: <https://doi.org/10.1080/03081060.2014.997449>.
- Burghout, W., Koutsopoulos, H., and Andreasson, I. (2006). “A discrete-event mesoscopic traffic simulation model for hybrid traffic simulation”. In: *2006 IEEE Intelligent Transportation Systems Conference*, pp. 1102–1107.
- Cao, Y., Lint, H. van, Krishnakumari, P., and Bliemer, M. (2024). “Data driven origin–destination matrix estimation on large networks—A joint origin–destination–path–choice formulation”. In: *Transportation Research Part C: Emerging Technologies* 168, p. 104850.
- Cascetta, E. (1984). “Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator”. In: *Transportation Research Part B: Methodological* 18.4, pp. 289–299. URL: <http://www.sciencedirect.com/science/article/pii/0191261584900122>.

- Chen, Y. and Florian, M. (1995). “The nonlinear bilevel programming problem: Formulations, regularity and optimality conditions”. In: *Optimization* 32.3, pp. 193–209.
- Chong, L. and Osorio, C. (2018). “A Simulation-Based Optimization Algorithm for Dynamic Large-Scale Urban Transportation Problems”. In: *Transportation Science* 52.3, pp. 637–656. URL: <https://doi.org/10.1287/trsc.2016.0717>.
- Daganzo, C. F. (1994). “The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory”. In: *Transportation Research Part B: Methodological* 28.4, pp. 269–287. URL: <https://www.sciencedirect.com/science/article/pii/S0191261594900027>.
- Daganzo, C. F. (1995a). “Requiem for second-order fluid approximations of traffic flow”. In: *Transportation Research Part B: Methodological* 29.4, pp. 277–286. URL: <https://www.sciencedirect.com/science/article/pii/S019126159500007Z>.
- Daganzo, C. F. (1995b). “The cell transmission model, part II: Network traffic”. In: *Transportation Research Part B: Methodological* 29.2, pp. 79–93. URL: <https://www.sciencedirect.com/science/article/pii/S019126159400022R>.
- Djukic, T., Lint, H. van, and Hoogendoorn, S. P. (2014). *Methodology for Efficient Real-Time OD Demand Estimation on Large-Scale Networks*. Technical Report. Delft, The Netherlands: Delft University of Technology, ITS Edulab.
- Elimadi, M., Abbas-Turki, A., Koukam, A., Dridi, M., and Mualla, Y. (2024). “Review of Traffic Assignment and Future Challenges”. In: *Applied Sciences* 14.2. URL: <https://www.mdpi.com/2076-3417/14/2/683>.
- Fekih, M., Bellemans, T., Smoreda, Z., Bonnel, P., Furno, A., and Galland, S. (2021). “A data-driven approach for origin–destination matrix construction from cellular network signalling data: a case study of Lyon region (France)”. In: *Transportation* 48.4, pp. 1671–1702.
- Flötteröd, G. and Rohde, J. (2011). “Operational Macroscopic Modeling of Complex Urban Intersections”. In: *Transportation Research Part B: Methodological* 45.6, pp. 903–922.
- Frank, I. E. and Friedman, J. H. (1993). “A Statistical View of Some Chemometrics Regression Tools”. In: *Technometrics* 35.2, pp. 109–135. URL: <http://www.jstor.org/stable/1269656>.
- Geladi, P. and Kowalski, B. R. (1986). “Partial least-squares regression: a tutorial”. In: *Analytica Chimica Acta* 185, pp. 1–17. URL: <http://www.sciencedirect.com/science/article/pii/S0003267086800289>.

- Godoy, J. L., Vega, J. R., and Marchetti, J. L. (2014). “Relationships between PCA and PLS-regression”. In: *Chemometrics and Intelligent Laboratory Systems* 130, pp. 182–191. URL: <https://www.sciencedirect.com/science/article/pii/S0169743913002189>.
- Gun, J. P. T. van der, Pel, A. J., and Arem, B. van (2019). “The Link Transmission Model with Variable Fundamental Diagrams and Initial Conditions”. In: *Transportmetrica B: Transport Dynamics* 7.1, pp. 834–864.
- Hanrahan, G., Udeh, F., and Patil, D. G. (2005). “Chemometrics and Statistics: Multivariate Calibration Techniques”. In: *Encyclopedia of Analytical Science*. Ed. by P. Worsfold, A. Townshend, and C. Poole. 2nd ed. Oxford: Elsevier, pp. 27–32. ISBN: 978-0-12-369397-6. URL: <https://www.sciencedirect.com/science/article/pii/B0123693977000777>.
- Helbing, D. and Johansson, A. (2009). “On the Controversy Around Daganzo’s Requiem for and Aw-Rasclé’s Resurrection of Second-Order Traffic Flow Models”. In: *The European Physical Journal B* 69.4, pp. 567–574.
- Horni, A., Nagel, K., and Axhausen, K. W. (2016). *The Multi-Agent Transport Simulation MATSim*. London, GBR: Ubiquity Press. ISBN: 1909188751.
- Höskuldsson, A. (2001). “Variable and subset selection in PLS regression”. In: *Chemometrics and Intelligent Laboratory Systems* 55.1, pp. 23–38. URL: <http://www.sciencedirect.com/science/article/pii/S0169743900001131>.
- Klein, L. A. (2019). “Sensor and Data Fusion for Intelligent Transportation Systems”. In: *Proceedings of SPIE: Sensors and Systems for Space Applications XII*. Vol. 11155. Society of Photo-Optical Instrumentation Engineers (SPIE), p. 1115502.
- Kostic, B., Gentile, G., and Antoniou, C. (2017). “Techniques for improving the effectiveness of the SPSA algorithm in dynamic demand calibration”. In: *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*. IEEE, pp. 368–373.
- Krishnakumari, P., Van Lint, H., Djukic, T., and Cats, O. (2020). “A data driven method for OD matrix estimation”. In: *Transportation Research Part C: Emerging Technologies* 113, pp. 38–56.
- Kvalheim, O. (July 2010). “Interpretation of partial least squares regression models by means of target projection and selectivity ratio plots”. In: *Journal of Chemometrics* 24, pp. 496–504.
- Lee, D. D. and Seung, H. S. (2001). “Algorithms for Non-negative Matrix Factorization”. In: *Advances in Neural Information Processing Systems 13*. MIT Press, pp. 556–562.

- Lighthill, M. J. and Whitham, G. B. (1955a). “On kinematic waves I. Flood movement in long rivers”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 229.1178, pp. 281–316. URL: <https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1955.0088>.
- Lighthill, M. J. and Whitham, G. B. (1955b). “On kinematic waves II. A theory of traffic flow on long crowded roads”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 229.1178, pp. 317–345. URL: <https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1955.0089>.
- Liu, C., Zhang, X., Nguyen, T. T., Liu, J., Wu, T., Lee, E., and Tu, X. M. (2022). “Partial Least Squares Regression and Principal Component Analysis: Similarity and Differences Between Two Popular Variable Reduction Approaches”. In: *General Psychiatry* 35.1, e100662. URL: <https://gpsych.bmj.com/content/35/1/e100662>.
- Lu, L., Xu, Y., Antoniou, C., and Ben-Akiva, M. (2015). “An enhanced SPSA algorithm for the calibration of Dynamic Traffic Assignment models”. In: *Transportation Research Part C: Emerging Technologies* 51, pp. 149–166. URL: <http://www.sciencedirect.com/science/article/pii/S0968090X14003295>.
- Martens, H. (2001). “Reliable and relevant modelling of real world data: a personal account of the development of PLS Regression”. In: *Chemometrics and Intelligent Laboratory Systems* 58.2. PLS Methods, pp. 85–95. URL: <https://www.sciencedirect.com/science/article/pii/S0169743901001538>.
- Nagel, K. and Schreckenberg, M. (1992). “A Cellular Automaton Model for Freeway Traffic”. In: *Journal de Physique I* 2.12, pp. 2221–2229.
- Neumann, J. von (1951). “The General and Logical Theory of Automata”. In: *Cerebral Mechanisms in Behaviour: The Hixon Symposium*. Ed. by L. A. Jeffress. New York: Wiley, pp. 1–41.
- Osorio, C., Flötteröd, G., and Bierlaire, M. (2011). “Dynamic network loading: a stochastic differentiable model that derives link state distributions”. In: *Procedia - Social and Behavioral Sciences* 17. Papers selected for the 19th International Symposium on Transportation and Traffic Theory, pp. 364–381. URL: <https://www.sciencedirect.com/science/article/pii/S1877042811010810>.
- Patil, S. N., Behara, K. N., Khadhir, A., and Bhaskar, A. (2023). “Methods to enhance the quality of bi-level origin–destination matrix adjustment process”. In: *Transportation Letters* 15.2, pp. 77–86.

- Payne, H. J. (1971). “Models of Freeway Traffic and Control”. In: *Mathematical Models of Public Systems*. Vol. 1. La Jolla, California: Simulation Council, pp. 51–61.
- Peng, C., Xu, C., Wang, C., Tong, H., and Ren, W. (2025). “Efficient calibration of high-dimensional dynamic OD matrix in metropolitan networks: combining dimensionality reduction and bagging-based metamodel”. In: *Transportmetrica A: Transport Science*, pp. 1–26.
- Qurashi, M., Lu, Q.-L., Cantelmo, G., and Antoniou, C. (2022). “Dynamic demand estimation on large scale networks using Principal Component Analysis: The case of non-existent or irrelevant historical estimates”. In: *Transportation Research Part C: Emerging Technologies* 136, p. 103504. URL: <https://www.sciencedirect.com/science/article/pii/S0968090X21004903>.
- Richards, P. I. (1956). “Shock Waves on the Highway”. In: *Operations Research* 4.1, pp. 42–51. eprint: <https://doi.org/10.1287/opre.4.1.42>. URL: <https://doi.org/10.1287/opre.4.1.42>.
- Sorzano, C. O. S., Vargas, J., and Montano, A. P. (2014). “A survey of dimensionality reduction techniques”. In: *arXiv preprint arXiv:1403.2877*.
- Spall, J. (1992). “Multivariate stochastic approximation using a simultaneous perturbation gradient approximation”. In: *IEEE Transactions on Automatic Control* 37.3, pp. 332–341.
- Treiber, M. and Kesting, A. (2013). *Traffic flow dynamics*. Vol. 1. Springer.
- Tsanakas, N., Ekström, J., Gundlegård, D., Olstam, J., and Rydergren, C. (2021). “Data-driven network loading”. In: *Transportmetrica B: Transport Dynamics* 9.1, pp. 237–265.
- Tsanakas, N., Gundlegård, D., and Rydergren, C. (2023). “O–D matrix estimation based on data-driven network assignment”. In: *Transportmetrica B: Transport Dynamics* 11.1, pp. 376–407.
- Verbas, I. Ö., Mahmassani, H. S., and Zhang, K. (2011). “Time-dependent origin–destination demand estimation: challenges and methods for large-scale networks with multiple vehicle classes”. In: *Transportation research record* 2263.1, pp. 45–56.
- Wang, Y., Zhang, Y., Wang, L., Hu, Y., and Yin, B. (2022). “Urban Traffic Pattern Analysis and Applications Based on Spatio-Temporal Non-Negative Matrix Factorization”. In: *IEEE Transactions on Intelligent Transportation Systems* 23.8, pp. 12752–12765.
- Wei, G. (2022). “Calibration of Urban Network Capacities”. Licentiate Thesis. Linköping, Sweden: Linköping University. ISBN: 978-91-7929-540-8. URL: <https://liu.diva-portal.org/smash/record.jsf?pid=diva2:1690200>.

- Withford, D. K. (1963). “Traffic Assignment Analysis and Evaluation”. In: *Highway Research Record* 6.
- Wold, S., Sjöström, M., and Eriksson, L. (2001). “PLS-regression: a basic tool of chemometrics”. In: *Chemometrics and Intelligent Laboratory Systems* 58.2. PLS Methods, pp. 109–130. URL: <http://www.sciencedirect.com/science/article/pii/S0169743901001551>.
- Yang, X., Lu, Y., and Hao, W. (2017). “Origin–Destination Estimation Using Probe Vehicle Trajectory and Link Counts”. In: *Journal of Advanced Transportation* 2017.1. URL: <https://doi.org/10.1155/2017/4341532>.
- Yang, Y., Lu, H.-p., and Hu, Q. (2010). “A Bi-level programming model for origin-destination estimation based on FCD”. In: *ICCTP 2010: Integrated Transportation Systems: Green, Intelligent, Reliable*, pp. 117–124.
- Yperman, I. (2007). “The Link Transmission Model for Dynamic Network Loading”. PhD Thesis. Leuven, Belgium: Katholieke Universiteit Leuven, pp. 1–215. URL: <https://lirias.kuleuven.be/retrieve/93668>.
- Yu, L., Wang, H., He, Y., and Wen, Y. (2023). “Traffic Data Recovery and Outlier Detection Based on Non-negative Matrix Factorization and Truncated-Quadratic Loss Function”. In: *International Conference on Neural Information Processing*. Springer, pp. 41–52.
- Zhang, C., Li, Y., Arora, N., Pierce, D., and Osorio, C. (2025). *Traffic Simulations: Multi-City Calibration of Metropolitan Highway Networks*. arXiv preprint. URL: <https://arxiv.org/abs/2501.04783>.
- Zhang, C., Osorio, C., and Flötteröd, G. (2017). “Efficient calibration techniques for large-scale traffic simulators”. In: *Transportation Research Part B: Methodological* 97, pp. 214–239. URL: <https://www.sciencedirect.com/science/article/pii/S0191261516302831>.

Abbreviations

CA Cellular Automata

CTM cell transmission model

DDNA data-driven network assignment

DNL dynamic network loading

DTA dynamic traffic assignment

LTM link transmission model

MSE mean square error

NNMF non-negative matrix factorization

OD origin-destination

PCA principal component analysis

PCR principal component regression

PLS partial least squares

PLSR partial least squares regression

Bibliography

SPSA simultaneous perturbation stochastic approximation

TAZ traffic analysis zone

Erratum

In Paper II, the second and third sentences of the second paragraph in Section *Problem Formulation*:

”The time-sliced demand is represented by a vector $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_h, \dots, \mathbf{x}_H)^T$, where $\mathbf{x}_h = (x_{1,h}, \dots, x_{i,h}, \dots, x_{I,h})^T$. The observed link flow on link b within the analysed period h is denoted by $y_{b,h}$, and a flow vector $\tilde{\mathbf{y}} = (\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_h, \dots, \tilde{\mathbf{y}}_H)^T$, where $\tilde{\mathbf{y}}_h = (\tilde{y}_{1,h}, \dots, \tilde{y}_{b,h}, \dots, \tilde{y}_{B,h})^T$, is defined.” should be corrected as :

”The time-sliced demand is represented by a vector $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_h, \dots, \mathbf{x}_H)$, where $\mathbf{x}_h = (x_{1,h}, \dots, x_{i,h}, \dots, x_{I,h})$. The observed link flow on link b within the analysed period h is denoted by $y_{b,h}$, and a flow vector $\tilde{\mathbf{y}} = (\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_h, \dots, \tilde{\mathbf{y}}_H)$, where $\tilde{\mathbf{y}}_h = (\tilde{y}_{1,h}, \dots, \tilde{y}_{b,h}, \dots, \tilde{y}_{B,h})$, is defined.”

Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

<https://doi.org/10.3384/9789181185638>

FACULTY OF SCIENCE AND ENGINEERING

Linköping Studies in Science and Technology. Dissertation No. 2523, 2026
Department of Science and Technology

Linköping University
SE-581 83 Linköping, Sweden

www.liu.se